UNIVERSITY OF CALIFORNIA SAN DIEGO

Decoding the genomic regulatory syntax driving notochord development

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

in

Bioinformatics & Systems Biology

by

Michelle Franc Ragsac

Committee in charge:

Professor Emma K. Farley, Chair
Professor Theresa Gaasterland, Co-Chair
Professor Vineet Bafna
Professor Christopher Benner
Professor Xin Sun

2022

The Dissertation of Michelle Franc Ragsac is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

Ang pamilya ay hindi isang mahalagang bagay lamang.
Ito ay ang lahat.

. . .

This dissertation is dedicated to the family that fostered a warm,
multi-generational home that valued love, curiosity, and education above all else.

To my parents, *Francisco Gregorio Ragsac* and *Lea Reyes Ragsac*,
for paving the way for a better life in the United States.

To my younger brother, *Thomas Jonathan Ragsac*,
for spending time with me through playing video games because you had to.

To my aunt, *Charleen Rodriguez Ragsac*,
for always keeping things loud and lively.

Finally, to my grandparents, *Hermogenes Riego de Dios Reyes* and *Josefina Tardeo Reyes*,
for making sure I spent my days listening to classical music on the mandolin and eating
comforting Filipino food during my youth.

. . .

I would also like to dedicate this dissertation to *Clarence Kuang-Le Mah*,
the nerdy best friend who got me interested in bioinformatics in the first place.
I'd still walk from Camp Snoopy to the Village for you.

EPIGRAPH

It is not birth, marriage, or death,
but gastrulation which is the most
important time in your life.

- *Lewis Wolpert*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I have to thank my thesis advisor and committee chair, *Professor Emma Farley*. Emma is one of the brightest scientists I've met, and my approach to research has definitely been shaped by her guidance over the past few years. Her endless support throughout graduate school—especially during difficult moments in research and my personal life—ultimately helped me get to where I am today. I would also like to thank my long-time advisor and thesis committee co-chair, *Professor Theresa "Terry" Gaasterland*. I have known Terry since my early undergraduate years at UCSD, and her infectious enthusiasm for interesting problems in genetics and genomics is something that I will always carry with me. I would also like to show gratitude to the rest of my committee, *Professors Vineet Bafna*, *Christopher Benner*, and *Xin Sun*. Without their assistance and helpful feedback throughout the later portion of my graduate school journey, this work would have never been accomplished.

I also received invaluable feedback throughout graduate school from members of the Farley Lab. Their conversations, encouragement, and camaraderie throughout the years helped me get through some of the toughest moments of my life. Thanks to *Benjamin Song* for being the best experimental scientist and person I have ever been blessed to cross paths with—and for generating all of the data for our joint thesis projects. Our projects had highs and (many) lows, but we made it to the end! I would also like to thank *Granton Jindal*, *Genevieve Ryan*, *Hannah Finnegan*, *Jesse "Joe" Solvason*, *Alexis Bantle*, *Jessica Grudzien*, *Sophia Le*, *Joanna Encarnacion*, *Krissie Tellez*, and *Fabian Lim*. Also, a special shout out to members of the Evo/Devo Journal Club-including *Professors Jim Posakony* and *Michael Perry*, as well as their students–for teaching me how to properly synthesize research papers, even if it meant learning more about insects than I ever would have expected on a medical school campus.

Next, I would like to thank *Professor Hannah Carter*, *Michelle Dow*, *Andrea Castro*, *Kivilcim Ozturk*, *Meghana Pagadala*, *Clarence Mah*, *James Talwar*, *Adam Klie*, *Cameron Waller*, *Sunduk Hwang*, *Jeanna Sheen*, and *David Laub* for welcoming me as an honorary member of their lab. There were many days during graduate school when I looked forward to coming to the (mostly windowless) huts on the medical school campus just so I could chat with them about anything on my mind. From debating the benefits of specific neural network models to growing

my wine palette over holiday dinners to drinking copious amounts of sparkling water, you all made my graduate school experience much more enjoyable!

During my second year of graduate school, I had the opportunity to join the Genetics Training Program (GTP). Even though I had to roll out of bed at times for the early morning—8:00 AM each Wednesday—Genetics Journal Club, I am thankful to *Professor Bruce Hamilton* for teaching the other GTP students and me how to think like a better researcher over free bagels (but no coffee, sadly). I'll cherish our journal club's fruitful discussions about statistics, data visualization, experimental design, and scientific ethics for years to come.

Student advocacy and support is something I will always hold dear to my heart, and thus, I would also like to recognize the individuals I worked with on various program-specific and general UCSD campus initiatives. Despite not being overtly related to research, the student experience is important, and I worked closely with the Graduate Bioinformatics Council (GBIC) to help build a better community for my current and future peers. Thank you to *Daniela "Dana" Nachmanson*, *Jonathan Pekar*, *Clarence Mah*, *Jennifer Havens*, *Owen Chapman*, *Hannah Mummey*, *Caitlin Guccione*, and many others for spending time with me writing emails, sitting in meetings, and planning events for our students with limited budgets. I would also like to thank the multitude of people I worked with during my years in the Graduate and Professional Student Association (GPSA) as a member of the Finance Committee, as Finance Committee Chair, and finally as Vice President of Financial Affairs. Thank you to everyone in GPSA for supporting me and countless other individuals across the UCSD campus, especially *Graduate Division Dean James "Jim" Antony*, *Graduate Division Assistant Dean Judy Kim*, *John Hughes*, *Breana Clark*, *Anna Dickson*, *Krish Bhutwala*, *Andy Ryan*, *Ximena Garcia-Arceo*, *Hema Kopalle*, *Mia Rose*, *Chiaki Santiago*, *Hayden Schill*, *Linda Li*, *Angus Chapman*, *Kristin Leadbetter*, *Sushil S*, *Ross Turner*, *Ben Du*, *Giulia Corno*, *Matthew Fain*, *Joseph Rainaldi*, *Becca Rose*, *Kane Wu*, and *Gabriel Zalles-Ballivian.*

Before graduate school, I was fortunate to work with the knowledgeable people at Agena Bioscience at their San Diego headquarters. Without the encouragement of people at Agena, specifically in Assays by Agena (AbA) and in Molecular Tools, I would have never considered applying to graduate school in the first place, nor would I have gained the confidence I have

today as a bioinformatician. I would, however, like to give a special thanks to *Julie Vanhnasy* and *Huimin "Helen" Tao*. Julie and Helen were my first two managers coming out of college, and I wouldn't have had it any other way. They were the most encouraging people I've had the pleasure of working with, and under their wings, I learned how to be a confident scientist for the first time.

I have been blessed to have been surrounded by friends throughout graduate school from all stages of my life. Thank you to my friends from back home in the Bay Area for always being there for me, even though we are barely able to meet up online or in person due to the craziness of adult life, especially *Sasha Smirensky*, *Ashley Mae*, *Luna Chang*, *Rebecca Tien*, *Justin "JP" Petrola*, and members of the Sisterhood of the Traveling Yoga Pants. Thank you to my friends from my time at UCSD who call me crazy for still being here—yes, I still, unfortunately, eat at Price Center—including *Karina Kak*, *Jenny Woo*, *Tiffany Diep*, *Victoria Nguyen*, *Larry Zhang*, *Victor Wong*, *Richard Phouasalith*, *Kritin Karkare*, and the entire Camp Snoopy Building 705 crew. Thank you to my dear friends from the Llama Lair and Litmas Lair for always giving me something exciting to do while stressing out about graduate school in our old Berwick Drive home. Thank you to my friends at Lucera, including *Grant Wu*, *Scott Louie*, and *Ashley Mae*, for making the COVID-19 pandemic more bearable with weekly dinners and Korean drama viewing sessions. Thank you to the Internet crew for supporting me when `IRL` was too hard, including the Early 2000s Tumblr Crew, the Danger Rangers, the Cute Keyboard Club (CKC), the San Diego Mechanical Keyboard Community (SDMK), the few friends I kept in touch with from the doctoral program interview circuit, and many others—you know who you are, even if I don't mention you by name or online handle. Thank you to my fellow Bioinformatics & Systems Biology peers for making graduate school lively, especially *Jonathan Pekar*, *Gibraan Rahman*, *Adam Officer*, *George Armstrong*, *Jennifer Havens*, *Adam Jussila*, *Clarence Mah*, *Cameron Martino*, *Anthony Aylward*, *Owen Chapman*, *Carlos Guzman*, *Kiki Spaulding*, *Jessica Au*, *Avery Pong*, *Alex Jambor*, *Xiaomi Du*, *Emily Kobayashi*, and the entire "BISB et al., 2018" cohort-or at least whose names I didn't already mention here. I would also like to thank *Ashley Tess* from the Chemistry Doctoral Program, as well as *Natalie Deforest*, *Sara Elmsaouri*, *Jenna Kovsky*, *Danielle Schafer*, *James Yu*, and other students from the 2018 Cohort of the Biomedical Sciences program that adopted

me during late nights at Mesa Rim or during hangouts on Beeramar.

I would like to thank my best friend and loving partner, *Clarence Mah*, for being my rock and hospital escort throughout my entire time at UCSD—from 2013 to today. Clarence has unconditionally encouraged me to pursue my passions and been there to see me grow, and I will forever be grateful to have somebody willing to listen to my terrible jokes and help me take care of Yuuki when she's being an especially sassy Shiba Inu.

Finally, I would like to thank my wonderful but enormous family for supporting my studies throughout the years, even though they sometimes found the whole process confusing and convoluted. As one of the first in my family to go to a four-year research university then attain a doctoral degree, I am proud to have had their loving support despite missing many family events and milestones. I would especially like to thank my parents, *Francisco* and *Lea Ragsac*, and my younger brother, *Thomas "JonJon" Ragsac*, for the millions of phone calls and food. I am also grateful for the support of my *Lola*, *Josefina "Fina" Reyes*, and late *Lolo*, *Hermogenes "Moneng" Reyes*, to pursue anything that makes me happy. There are also the cousins, *titos*, and *titas* whose names I cannot mention here because that document would end up being the length of a thesis in and of itself!

This time of my life was one of the most difficult intellectually, physically, and mentally, and it is not without the support of an entire village that I could be where I am today. There are probably many names that I neglected to mention, but even if you're not listed, ***maraming salamat***.

Chapter 1, in full, has been submitted for publication of the material as it may appear in "Diverse logics encode notochord enhancers." Benjamin P. Song, Michelle F. Ragsac, Krissie Tellez, Granton A. Jindal, Jessica L. Grudzien, Sophia H. Le, Emma K. Farley. *Cell Reports*, 2022. The dissertation author was the primary investigator and co-first author of this paper.

VITA

2013–2017	B.S. in Bioengineering: Bioinformatics University of California San Diego
2015–2016	Writing Studio Mentor, Sixth College Writing Studio University of California San Diego
2017	Teaching Assistant Scripps Institute of Oceanography
2017	Research Associate, Applications and Technology Services–Assays by Agena Agena Bioscience
2018	Research Associate, Research and Development–Molecular Tools Agena Bioscience
2018–2022	Ph.D. in Bioinformatics & Systems Biology, University of California San Diego
2018	Teaching Assistant Scripps Institute of Oceanography
2020–2021	Teaching Assistant, School of Medicine University of California San Diego
2020–2021	Bootcamp Instructor, Bioinformatics & Systems Biology Doctoral Program University of California San Diego

PUBLICATIONS

Author names marked with † indicate shared first co-authorship.
Publications marked with △ are included in this text.

Granton A. Jindal†, Alexis T. Bantle†, Joe J. Solvason†, Jessica L. Grudzien, Agnieszka D'Antonio-Chronowska, Fabian Lim, Sophia H. Le, **Michelle F. Ragsac**, Benjamin P. Song, Reid O. Larsen, Adam Klie, Kelly A. Frazer, and Emma K. Farley. "Affinity-optimizing variants within cardiac enhancers disrupt heart development." *In Submission*, 2022.

Benjamin P. Song†, **Michelle F. Ragsac**†, Krissie Tellez, Granton A. Jindal, Jessica L. Grudzien, Sophia H. Le, and Emma K. Farley. "Diverse logics encode notochord enhancers." *Accepted in Principle at Cell Reports*, 2022. △

Sydney C. Morgan†, Stefan Aigner†, Catelyn Anderson†, Pedro Belda-Ferre†, Peter De Hoff†, Clarisse A. Marotz†, Shashank Sathe†, Mark Zeller†, Noorsher Ahmed, Xaver Audhya, Nathan A. Baer, Tom Barber, Bethany Barrick, Lakshmi Batachari, Maryann Betty, Steven M. Blue, Brent Brainard, Tyler Buckley, Jamie Case, Anelizze Castro-Martinez, Marisol Chacón, Willi Cheung, LaVonnye Chong, Nicole G. Coufal, Evelyn S. Crescini, Scott DeGrand, David P. Dimmock, J. Joelle Donofrio-Odmann, Emily R. Eisner, Mehrbod Estaki, Lizbeth Franco Vargas,

Michele Freddock, Robert M Gallant, Andrea Galmozzi, Nina J. Gao, Sheldon Gilmer, Edyta M. Grzelak, Abbas Hakim, Jonathan Hart, Charlotte Hobbs, Greg Humphrey, Nadja Ilkenhans, Marni Jacobs, Christopher A. Kahn, Bhavika K. Kapadia, Matthew Kim, Sunil Kurian, Alma L. Lastrella, Elijah S. Lawrence, Kari Lee, Qishan Liang, Hanna Liliom, Valentina Lo Sardo, Robert Logan, Michal Machnicki, Celestine G. Magallanes, Clarence K. Mah, Denise Malacki, Ryan J. Marina, Christopher Marsh, Natasha K. Martin, Nathaniel L. Matteson, Daniel J. Maunder, Kyle McBride, Bryan McDonald, Daniel McDonald, Michelle McGraw, Audra R. Meadows, Michelle Meyer, Amber L. Morey, Jasmine R. Mueller, Toan T. Ngo, Julie Nguyen, Viet Nguyen, Laura J. Nicholson, Alhakam Nouri, Victoria Nudell, Eugenio Nunez, Kyle O'Neill, R. Tyler Ostrander, Priyadarshini Pantham, Samuel S. Park, David Picone, Ashley Plascencia, Isaraphorn Pratumchai, Michael Quigley, **Michelle Franc Ragsac**, Andrew C. Richardson, Refugio Robles-Sikisaka, Christopher A. Ruiz, Justin Ryan, Lisa Sacco, Sharada Saraf, Phoebe Seaver, Leigh Sewall, Elizabeth W. Smoot, Kathleen M. Sweeney, Chandana Tekkatte, Rebecca Tsai, Holly Valentine, Shawn Walsh, August Williams, Min Yi Wu, Bing Xia, Brian Yee, Jason Z. Zhang, Kristian G. Andersen, Lauge Farnaes, Rob Knight, Gene W. Yeo, Louise C. Laurent. "Automated, miniaturized, and scalable screening of healthcare workers, first responders, and students for SARS-CoV-2 in San Diego County." *In Submission,* 2022.

ABSTRACT OF THE DISSERTATION

Decoding the genomic regulatory syntax driving notochord development

by

Michelle Franc Ragsac

Doctor of Philosophy in Bioinformatics & Systems Biology

University of California San Diego, 2022

Professor Emma K. Farley, Chair
Professor Theresa Gaasterland, Co-Chair

Embryonic development across all vertebrates begins upon the fertilization of an egg by a sperm cell to become a single-celled zygote. Embryogenesis continues with various stages of division to eventually make up an entire organism. The processes governing development are finely orchestrated and include many participants, such as genes involved in gene regulatory networks and non-coding regions of DNA, or enhancers, to regulate the expression of those genes. Defects or perturbations to this strictly regulated machinery can lead to various clinical conditions, such as congenital heart disease. Thus, deepening our understanding of embryogenesis may help us understand the mechanisms driving congenital abnormalities as well as the evolution of developmental pathways. One defining characteristic of all chordate embryos is the presence of a notochord during development. The notochord is a long, semi-rigid fibrous rod of mesodermal

origin that provides structural support and serves as a signaling center to pattern the neighboring neural tube, paraxial mesoderm, and gut. A complete understanding of notochord structure and function during early and late life stages is thus essential to better understand congenital vertebral defects. For example, failure of vertebral notochord cells to transition to the nucleus pulposus, the cushioning between intervertebral discs of the spine, is associated with chordomas, slow-growing tumors formed from notochord cell remnants within the spine or the base of the skull. The ascidian *Ciona intestinalis Type A* (*Ciona*) is a marine organism that is evolutionarily similar to vertebrates. Through electroporation, *Ciona* is readily amenable to high-throughput, high-resolution functional studies of cis-regulatory elements like enhancers in their native, whole-embryo context. To identify key notochord enhancers, I analyzed the importance of enhancer grammar–the transcription factor order, orientation, spacing, and binding affinity–in modulating notochord-specific expression. Next, I highlight the potential of single-cell RNA-sequencing to study the gene regulatory networks governing notogenesis and their relationship to congenital abnormalities. This body of work provides new insight into the regulatory processes governing notochord development, providing direction for future efforts to improve our understanding of notochord-based diseases across chordates. Finally, I highlight Open Educational Resources (OERs) I developed for Bioinformatics education, emphasizing accessibility and inclusion.

Introduction

As the defining structure of all chordates, the notochord plays a crucial role in signaling and coordinating development during embryogenesis. In most vertebrates, the notochord ossified into the vertebrae of the spine. However, the notochord persists throughout the life of some invertebrate chordates, such as amphioxus. This thesis dissertation focuses on understanding gene regulation in the notochord of the marine urochordate, *Ciona intestinalis* (*Ciona*), during embryonic development from the perspective of the genomic sequence and the perspective of active transcripts within this key structure.

0.1 Notochord development in *Ciona intestinalis*

Chordates are animals belonging to the phylum Chordata, which includes vertebrates (subphylum Vertebrata), tunicates (subphylum Tunicata), and cephalochordates (subphylum Cephalochordata)[1]. The key defining characteristic of all chordates is the presence of a notochord during embryonic development[1–6]. The notochord is a long, semi-rigid fibrous rod of mesodermal origin that provides structural support to the developing embryo along the anterior-posterior axis. The notochord also acts as a signaling center in the developing embryo, patterning structures such as the "neural tube"[3–5]. A sheath of collagen proteins encases the notochord, allowing this flexible yet rigid structure to provide the basis for controlled mechanical support of Chordate organisms and protection for the neural tube[2–5].

While some Chordates retain the notochord throughout life as their body's primary axial support, in most vertebrates, the notochord becomes the nucleus pulposus of the intervertebral disc[3–5;7]. Failure of vertebral notochord cells to transition to the nucleus pulposus is associated with chordomas. These slow-growing tumors form from notochord cell remnants within the spine or the base of the skull[4;6]. A complete understanding of notochord structure and function

during early and late life stages is thus essential to better understand congenital neural tube and vertebral defects.

As a close chordate relative to the vertebrates, the ascidian *Ciona intestinalis Type A* or *Ciona robusta* (*Ciona*) stands as a longstanding model for studying organogenesis in a simple embryo[8–13]. For example, the *Ciona* notochord consists of only 40 post-mitotic cells, and orthologs of many *Ciona* notochord genes have known notochord expression in vertebrate embryos[11;13;14]. Of the 40 notochord cells, 32 are grouped in the anterior of the body and compose the "primary" or "A-line" notochord. The remaining eight are located more posteriorly and form the "secondary" or "B-line" notochord[11;13]. A-line and B-line refer to the conventional nomenclature denoting particular cell lineages in *Ciona*. In the 4-cell *Ciona* embryo, "A-lineage" and "B-lineage" cells are defined as the two cells on the vegetal side of the embryo, whereas the "a-lineage" and "b-lineage" cells are defined as the two cells on the animal side. The notochord thus forms from the vegetal A-line and B-line cells of the 4-cell *Ciona* embryo[11].

Within *Ciona*, notochord precursor cells are defined as early as the eight-cell stage as the A4.1 and B4.1 blastomere pair in the developing anterior and posterior regions of the embryo, respectively[11;15;16]. The A4.1 cells then divide to form the A5.1 and A5.2 blastomere pair at the onset of the 16-cell stage, which are precursors to the A-line notochord and the endoderm, nerve cord, trunk lateral cells, and muscle[11]. On the other hand, the B4.1 cells divide to form the B5.1 and B5.2 blastomere pair and, through subsequent divisions from B5.1, divide into B6.1 and B6.2. Finally, the B6.1 blastomere descendant at the 32-cell stage will eventually develop into the B-line notochord and other mesenchymal and muscle cells[11;15;16]. When gastrulation initiates at the 110-cell stage, the *Ciona* embryo contains 16 primary and four secondary notochord precursor cells[11]. Gastrulation is the stage at which the structure of the embryo changes from a single-layered blastula into a multiple-layered gastrula; thus, the notochord precursor cells coordinately invaginate as a monolayer over the primary gut, or archenteron[13;17]. Following gastrulation is neurulation, the stage at which the embryonic neural plate develops and then forms the neural tube[11;17]. At this stage, the notochord precursor cells in the *Ciona* embryo divide for the last time to define the final set of notochord cells on the embryonic midline[18].

0.2 Elucidating the mechanisms regulating notogenesis

The massive developmental transitions during embryogenesis require accurate gene regulation to maintain and balance the differentiation process. One component of this machinery is the interactions between cis-acting DNA elements-such as promoters and enhancers-and regulatory transcription factors. Enhancers were discovered in the 1980s and are short regions of DNA that contain transcription factor binding sites (TFBSs) which proteins can bind to regulate gene transcription[19–21]. Additionally, enhancers are typically located distally from the gene promoter and are approximately 100 bp to 1,000 bp in length[19;21]. Interestingly, the presence of a collection of TFBSs alone is insufficient in encoding functional activity of a particular target gene. For example, only specific arrangements of binding sites can activate transcription. The overarching rules governing the functional arrangement of TFBSs within enhancers is termed "enhancer grammar." Enhancer grammar is the interplay between the syntax-the order, orientation, and spacing of TFBSs-and the binding affinity of TFBSs to confer expression of a given enhancer sequence[22;23]. Despite the importance of enhancers and their known association with developmental defects and disease, we still do not entirely understand how an enhancer's sequence encodes particular functions. In Chapter 1, I discuss the investigation into a notochord enhancer governed by Zic, ETS, FoxA, and Brachyury (Bra) transcription factor binding sites[24;25]. Zic and ETS are co-expressed in the developing notochord of *Ciona* and in other vertebrates and are important for notochord specification[26;27]. The preceding study which discovered a putative notochord grammar relying on Zic and ETS found an interplay between the syntax and affinity of the binding sites present, such that the organization could compensate for the affinity and vice versa[24]. In this chapter, I discuss an enhancer screen in which I search for evidence of the Zic and ETS notochord enhancer grammar across the *Ciona* genome and test for functionality in the *Ciona* notochord through a pilot screen of 90 genomic elements at the embryonic tailbud stage. From this screen, we were able to identify nine notochord enhancers, finding that enhancer grammar is critical within one of these elements. We also identify that some enhancers contain TFBSs for Zic, ETS, FoxA, and Bra, and translate that this set of binding sites may be an important signature for Brachyury enhancers across Chordates[25].

Beyond the universal quality of containing transcription factor motifs, enhancer sequences can vary significantly in the location, length, and type of transcription factor binding sites present. Additionally, these changes can be even more dramatic as you compare across species[28–30]. However, studies have suggested that even with low sequence conservation, the function of specific enhancers may be conserved across species and that this function may be partly due to combinatorial action of conserved transcription factors[30;31]. This may be because a single transcription factor across its homologs in multiple species may have similar binding properties and thus recognize identical DNA sequences[30;32]. In Chapter 2, I continue the discussion of the notochord enhancer grammar studied in Chapter 1 but in greater detail and at a larger scale across the *Ciona* genome. Within this study, we develop improvements over our initial search of *Ciona* genomic regions containing Zic and ETS, such as allowing for greater flexibility of the Zic binding site within a sequence window. We find 4,344 genomic regions that harbor at least one Zic binding site and two ETS binding sites and test these regions in a massively-parallel reporter assay. In Chapter 2, I describe our preliminary results which suggest that only 15.4% of the genomic elements we identified are functional enhancers. Further study of this enhancer library will likely identify novel notochord enhancers and help us better understand how Zic and ETS encode notochord development through particular grammatical constraints.

Within Chapter 1 and Chapter 2, I conduct high-throughput screens of genomic elements within developing whole embryos to better understand how enhancers encode notochord-specific expression patterns. Nonetheless, understanding the underlying processes driving development also requires understanding how genes are expressed, primarily how these gene expression profiles differ across cells[32]. For instance, all cells in a developing embryo contain the same set of genes. However, different cells express different sets of these genes, leading to differences in expression and, thus, molecular function[22;32]. Technological advances have enabled the cataloging of global gene expression profiles of single cells using single-cell RNA-sequencing (scRNA-seq), allowing scientists to define the heterogeneity within cell populations during embryonic development[33–35]. This new paradigm has allowed developmental biologists to identify precisely when and in which cell types genes controlling cell fate decisions are expressed[36]. Despite the availability of large cell-type atlases generated via scRNA-seq and other omics technologies, there is still

much to be learned about gene regulatory networks. *Ciona* is a particularly suitable model for understanding the transcriptional changes necessary for proper development due to its genomic and morphological simplicity and historical significance as a model organism for embryological studies. In Chapter 3, I discuss an initiative to develop a high-resolution, single-cell atlas of a gastrulating *Ciona* embryo to understand notogenesis and the formation of other early structures.

0.3 Training the next generation of bioinformaticians

In recent years, genomics technologies have become more high-throughput and affordable to all research groups, resulting in a boom in data available for all biomedical research areas. However, this also results in a backlog of data to analyze for those that conducted the experiments. Despite never receiving a formal education in computation, many researchers are then faced with the arduous task of learning how to run bioinformatics pipelines[37;38]. While computational courses have started being integrated into the standard curriculum for undergraduate biology majors, there remains a need to support graduate students and other scientists that did not experience this shift in training for the field. In Chapter 4, I discuss the pedagogical philosophy that drove the in-person and virtual bioinformatics courses I taught at the University of California, San Diego.

0.4 Conclusion

The massive developmental transitions during embryogenesis require accurate gene regulation to maintain and balance the differentiation process. In this dissertation, I present our approach to understanding regulation in the developing notochord by conducting high-throughput, whole embryo reporter screens to identify functional enhancers. I also present a novel, proof-of-concept package for performing flexible genomic searches of combinatorial arrangements of TFBSs. Additionally, I share our current understanding of *Ciona* gastrulation and notogenesis from studying single-cell transcriptional expression profiles. Finally, I also discuss my contributions to bioinformatics education.

Chapter 1

Diverse logics encode notochord enhancers

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.

1.1 Introduction

Enhancers are genomic elements that act as switches to ensure the precise patterns of gene expression required for development[21]. 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)[39–43]. 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[44–46], 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[22;47–49]. 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)[23].

We previously identified grammatical rules governing notochord enhancers regulated by Zic and ETS TFBSs[24]. 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[24]. 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 1.1) and in vertebrates[26;27]. The notochord is a key feature of chordates and acts as a signaling center to pattern the neighboring neural tube, paraxial mesoderm, and gut[3;50]. Specification of the notochord by Brachyury (Bra), also known as T, is highly conserved across chordates[51–54]. Other conserved TFs important for activation of notochord gene expression include Zic, ETS, a TF downstream of FGF signaling, and FoxA[16;26;27;55–68].

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[9]. Fertilized *Ciona* eggs can be electroporated with many enhancers in a single experiment which allows for testing of many enhancers in whole, developing embryos[69;70]. 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[69;71], make it an ideal organism to study the rules governing notochord enhancers during development.

Within the *Ciona* genome, we found 1,092 elements containing one Zic site and at least two ETS sites within 30 bp 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*[72]. 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 12 bp 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*[15] and vertebrates[73] 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.

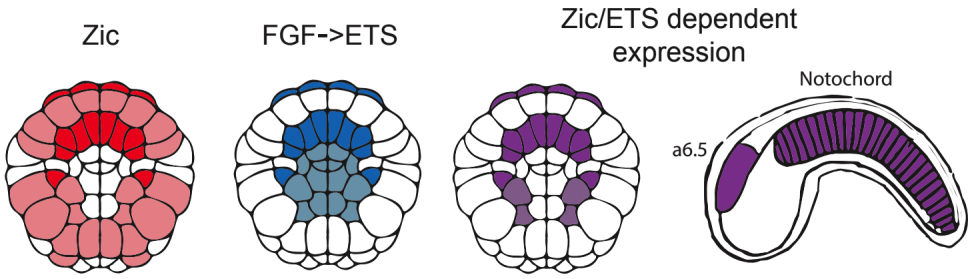

Figure 1.1. Zic and ETS expression in the 110-cell stage embryo. Co-expression of Zic and ETS is shown in purple and occurs in the notochord, a6.5 lineage, which gives rise to the anterior sensory vesicle and palps, and four mesenchyme cells shown in light purple. A schematic of the tailbud embryo shows 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.

1.2 Results

1.2.1 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)[16;27;74]. Using the Zic site as an anchor, we searched the 30 bp 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[75;76], as we have previously found that low-affinity ETS sites are required to encode notochord-specific expression[24]. This search identified 1,092 genomic regions approximately 68 bp in length. We define these regions as ZEE elements.

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

We selected 90 ZEE elements (Figure A.11 and Table S1) and synthesized these upstream of a minimal promoter (bpFog)[77;78] and a transcribable barcode to conduct an enhancer screen (experiment outlined in Figure 1.2A). 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[79] and both Zic and ETS are expressed[80;81]. 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 log_2 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 A.2).

1.2.3 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

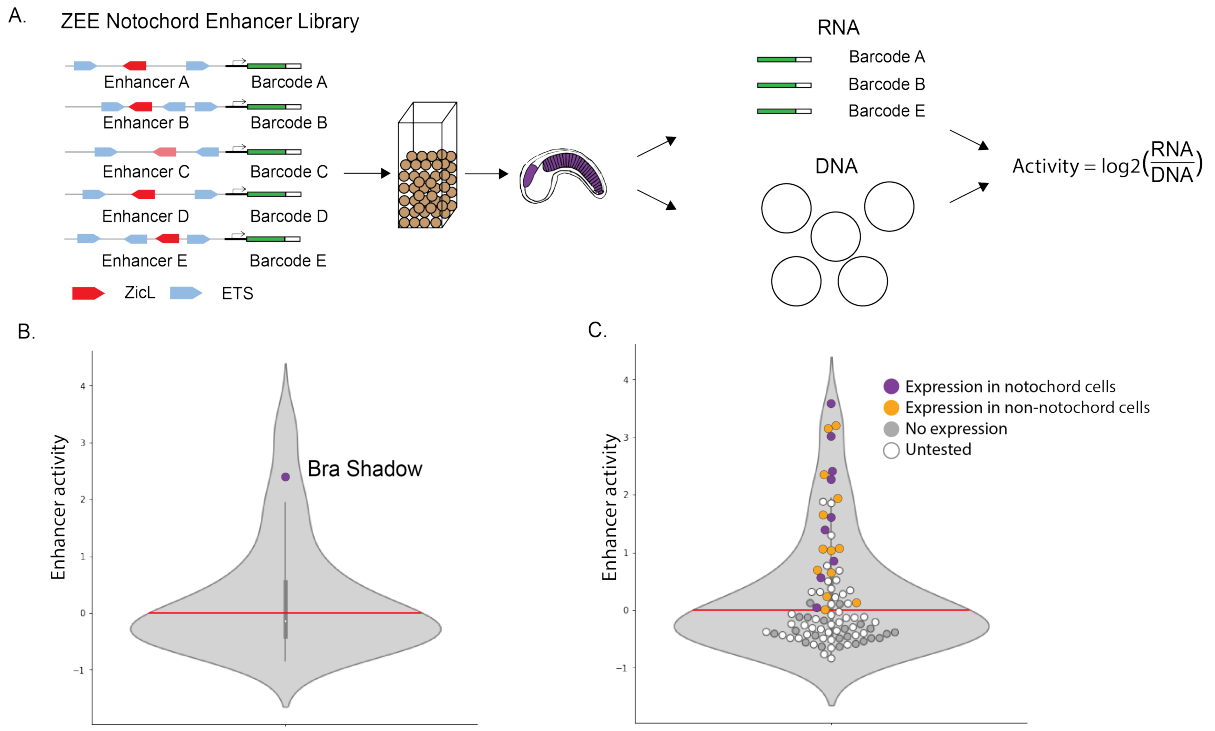

Figure 1.2. Screening Zic and ETS genomic elements in *Ciona*. **A.** Schematic of enhancer screen. 90 ZEE genomic regions, each associated with on average six unique barcodes were electroporated into fertilized *Ciona* eggs. mRNA and plasmid DNA were extracted from 5.5 hpf embryos (tailbud embryo shown 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 log_2 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.

a6.5 lineage, both locations that express Zic and ETS[24]. The BraS enhancer activity score is 2.4 (Figure 1.2B), 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 8 hpf in at least 150 embryos across three biological replicates. Collectively, we analyzed expression of these elements in over 6,600 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 (Figure 1.2C). 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 A.33 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[10;82–85] (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 1.1). Therefore, enhancers under combinatorial control of Zic and ETS are likely to

be active in the notochord and the a6.5 neural lineage[27;85;86]. Collectively these results indicate that our enhancer screen accurately detects functional enhancers, and our tissue-specific analysis provides detailed expression patterns for these enhancers.

1.2.4 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[53;55;56;60;61;64;87]. 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[61;64;88] and `TNNCAC` as the Bra motif[87;89–92].

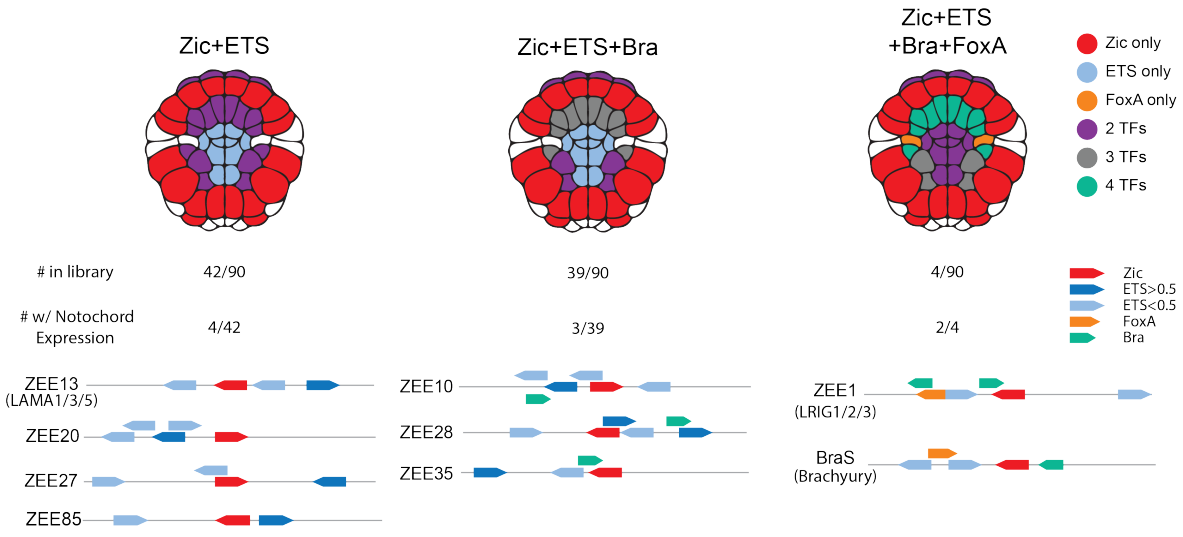

Figure 1.3. Combinations of transcription factors in ZEE enhancers that drive notochord expression. Notochord-expressing ZEE elements were grouped by the combination of transcription factor binding sites present in each element. For each combination, an embryo schematic shows the overlapping region of 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.

1.2.5 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 1.3 and Figure A.4). 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.

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

Four enhancers containing Zic and ETS sites only (ZEE13, ZEE20, ZEE27 and ZEE85) drive notochord expression. ZEE13, ZEE20 and ZEE27 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, ZEE27, and ZEE85 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[72] (Figure 1.4A). 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[76], as the binding specificity of ETS is highly conserved across bilaterians[76;93]. 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).

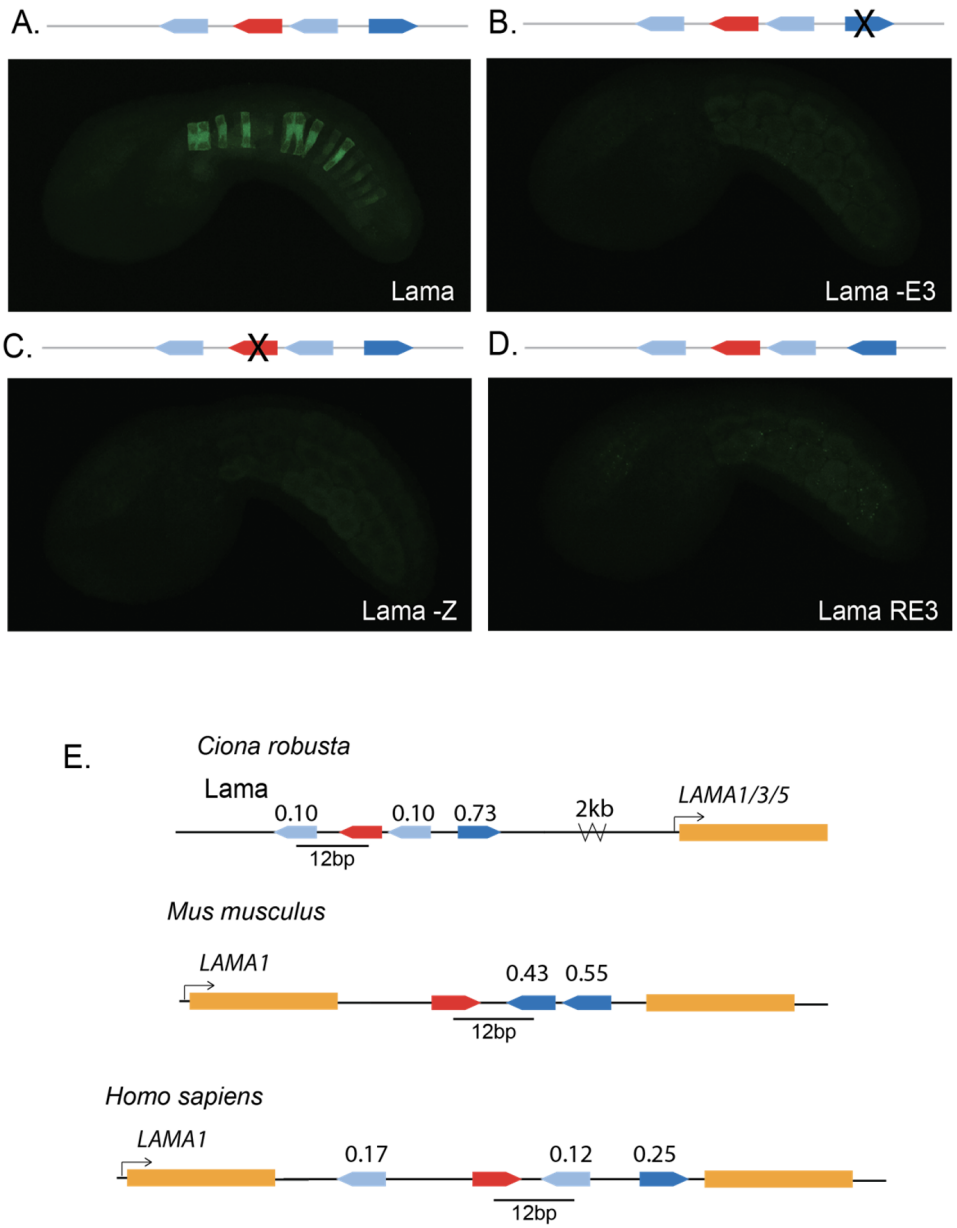

Figure 1.4. Zic and ETS grammar encodes a notochord *laminin alpha* enhancer. **A.** Embryo electroporated with the Lama enhancer (ZEE13); GFP expression can be seen in the notochord. **B.** Embryo electroporated with Lama -E3, where ETS3 was mutated to be non-functional; no GFP expression detected. **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 12 bp 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 1.4B, Figure A.5A, 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 1.4C, Figure A.5A, 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 enhancers. 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 1.4D, Figure A.5A, 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[24].

1.2.7 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[14;72;94]. 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[95–97]. 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[26;66] and ETS[98;99] 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 1.4E). We have previously seen that the spacing between Zic and adjacent ETS sites affects levels of expression, with spacings of 11 and 13 bp seen between ETS and Zic sites in the BraS enhancer and Mnx enhancer, respectively[24]. In line with this observation, the *laminin alpha-1* clusters in mouse and human and the *Ciona* Lama enhancer have a 12 bp spacing between the ETS and adjacent Zic site in all three species, suggesting that such spacings (11 to 13 bp) 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.

1.2.8 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[14]. 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[24]. 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 11 bp 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[24]. 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 1.5B, Figure A.5B, 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 1.5C, Figure A.5B, and Table S4). This experiment highlights the importance of understanding sufficiency in addition to necessity of sites.

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[60;62;64;86;100;101]. 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 1.5D and E, Figure A.4B, 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 1.5F, Figure A.5B, 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 A.4C), 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 A.4D) 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.

1.2.9 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[15;102], which drives strong expression in the notochord (Figure A.6A). 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[101;103]. 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[15;103]. Here we annotate the Bra434 enhancer using crystal structure data, enhancer mutagenesis data, EMSA and PBM data[16;27;61;64;74–76;87–92].

Our approach identifies two Zic sites, six low-affinity ETS sites, three FoxA sites, and eight Bra sites (Figure 1.5G and Figure A.6B). 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[15;101–103]. 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.

1.2.10 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 (Figure 1.5H)[73]. This mouse T enhancer

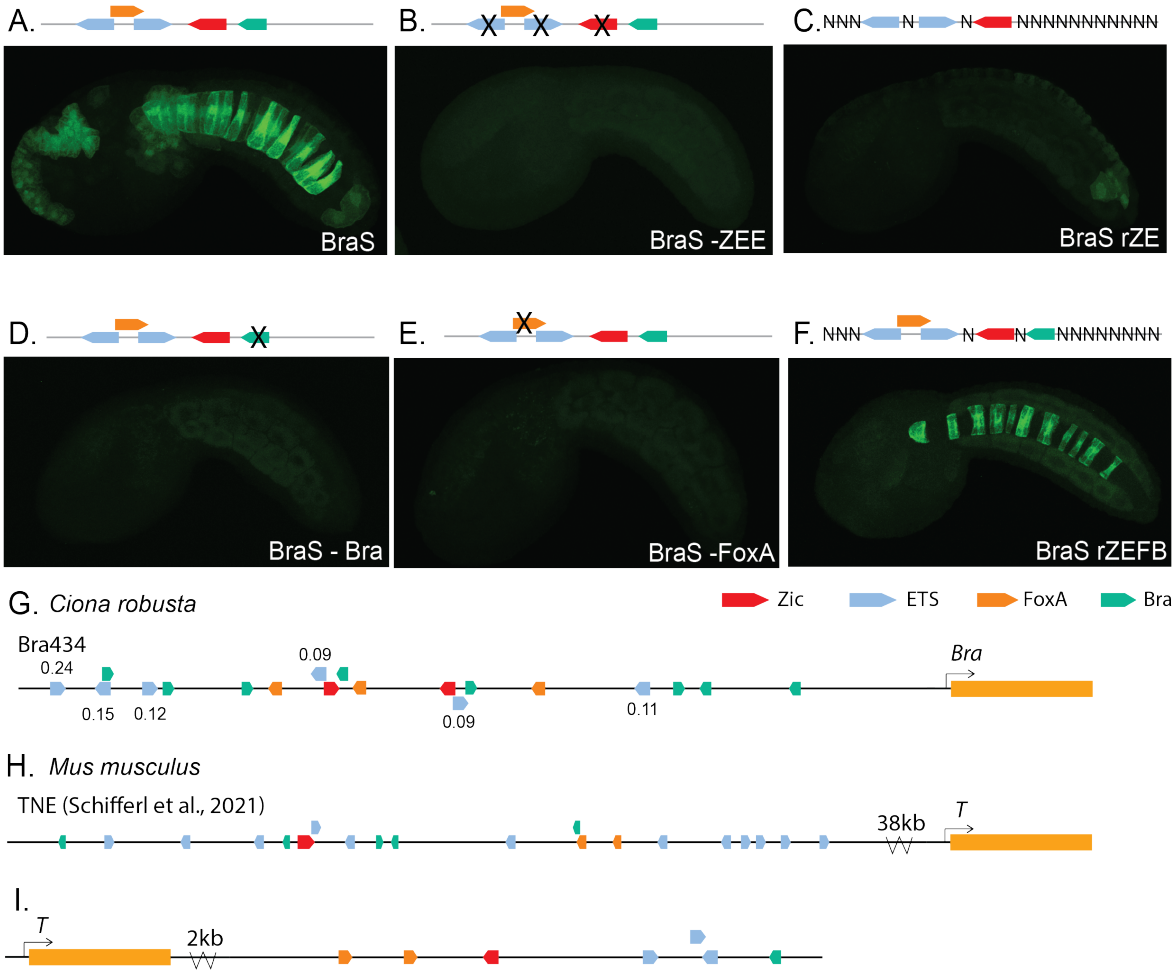

Figure 1.5. Zic, ETS, FoxA, and Bra may be a common regulatory logic for *Brachyury* enhancers. A. Embryo electroporated with the Bra Shadow (BraS) enhancer; GFP expression can be seen in the notochord. **B.** Embryo electroporated with BraS -ZEE, where the Zic and two ETS sites were mutated to be non-functional; 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.

is required for *Bra/T* expression, notochord cell specification and differentiation[73]. 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[73]. 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[104–106]. This is thought to confer the transcriptional robustness required for successful development[104;106–108]. 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 1.5I). This putative enhancer occurs within an open chromatin region in mouse E8.25 notochord cells[109], suggesting this may be another mouse T enhancer. Similarly in zebrafish, a notochord enhancer located 2.1kb upstream of the Bra ortholog *ntl*[110] 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.

1.3 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 1,092 genomic regions containing a Zic site within 30 bp 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.

1.3.1 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[111]. 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[111].

1.3.2 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[24]. *Laminin alpha* is a key gene involved in notochord development in both *Ciona* and vertebrates[72;97]. 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 12 bp 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[24]. 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 12 bp spacing within these elements.

1.3.3 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[112]. 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.

1.3.4 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[24]. 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[111]. Until then, our current knowledge of grammatical constraints may still be useful for pointing us towards putative enhancers.

1.3.5 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[73;110]. 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[30]. Our approach searching for grammatical signatures rather than sequence conservation may allow for identification of such functionally conserved enhancers.

1.3.6 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[111]. 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[23].

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[111]. Another approach is to study entirely random sequences for enhancer activity, which has been done in the context of promoters in bacteria and yeast[113;114]. 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[115]. 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.

1.3.7 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 1,092 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 12 bp 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.

1.4 STAR*Methods

1.4.1 Key resources table

Table 1.1. Key resources table

REAGENT or RESOURCE	**SOURCE**	**IDENTIFIER**
Deposited Data		
snATAC-seq mouse E8.25	Pijuan-Sala et al., 2020[109]	GEO: GSE133244
FACS-sorted notochord RNA-Seq	Reeves et al., 2017[14]	N/A
Human reference genome NCBI build 38	Genome Reference Consortium	NCBI, Human GRCh38 Reference
Mouse reference genome NCBI build 39	Genome Reference Consortium	NCBI, Mouse GRCm39 Reference
Ciona robusta genome	Satoh et al., 2005[116]	Ghost Database
mouse ETS-1 universal PBM data	Wei et al., 2010[76]	UniProbe Database
ZEE library screen	This paper	N/A
Experimental Models: Organisms/Strains		
Ciona intestinalis type A (*Ciona* robusta)	M-Rep	N/A
Oligonucleotides		
Oligonucleotides for library screen, see Table S1	This paper	N/A
Oligonucleotides for mutagenesis, see Table S4	This paper	N/A
Recombinant DNA		
Plasmid: BraS bpFog>GFP	Farley Lab	N/A
Plasmid: BraS -ZEE bpFog>GFP	This paper	N/A
Plasmid: BraS rZE bpFog>GFP	This paper	N/A

Continued on next page

Table 1.1. Key resources table, *continued from previous page*

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Plasmid: BraS -FoxA bpFog>GFP	This paper	N/A
Plasmid: BraS -Bra bpFog>GFP	This paper	N/A
Plasmid: BraS rZEFB bpFog>GFP	This paper	N/A
Plasmid: Lama1 bpFog>GFP	This paper	N/A
Plasmid: Lama1 bpFog>GFP	This paper	N/A
Plasmid: Lama1 -E3 bpFog>GFP	This paper	N/A
Plasmid: Lama1 -Z bpFog>GFP	This paper	N/A
Plasmid: Lama1 RE3 bpFog>GFP	This paper	N/A
Software and Algorithms		
Python (version 3.8.6)	Python Software Foundation	https://www.python.org
Conda (version 4.9.2)	Anaconda, Inc.	https://docs.conda.io
Bioconda	Grüning et al., 2018	https://bioconda.github.io
Biopython (version 1.78)	Cock et al., 2009	https://biopython.org
FastQC (version 0.11.9)	Babraham Institute	https://www.bioinformatics.babraham.ac.uk
MultiQC (version 1.8)	Ewels et al., 2016	https://multiqc.info
FLASH (version 1.2.11)	Magoč et al., 2011	http://www.cbcb.umd.edu/software/flash
`pandas` (version 1.2.1)	NumFOCUS	https://pandas.pydata.org

Continued on next page

Table 1.1. Key resources table, *continued from previous page*

REAGENT or RESOURCE	SOURCE	IDENTIFIER
`numpy` (version 1.20.3)	Harris et al., 2020	https://numpy.org
`matplotlib` (version 3.2.2)	Hunter, 2007	https://matplotlib.org
`scikit-learn` (version 0.24.1)	Pedregosa et al., 2011	https://scikit-learn.org
`seaborn` (version 0.11.1)	Waskom et al., 2021	https://seaborn.pydata.org
`Diverse-Logics-Notochord-Study`	Code used in this paper	Diverse-Logics-Notochord-Study GitHub

1.4.2 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 K. Farley (efarley@ucsd.edu).

Materials availability

Plasmids generated in this study are available upon request.

1.4.3 Experimental model and subject details

Tunicates

Adult *Ciona intestinalis type A*, also known as *Ciona robusta*, were obtained from M-Rep and were maintained under constant illumination in seawater (obtained from Reliant Aquariums) at 18°C. *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.

1.4.4 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 30 bp 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

Dechlorination, *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 5,000 fertilized eggs. Embryos developed until 5 hours and 30 minutes at 22°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 500 μL of Back Extraction Buffer (4 M guanidine thiocyanate, 50 mM sodium citrate, and 1 M 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 slides 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., 2007a; 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 $log_2(\frac{RNA}{DNA})$ value across a given enhancer's biological replicates.

1.4.5 Quantification and statistical analysis

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

1.5 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. The data will also be on SRA listed under the submission identifier PRJNA861319 and will be made available as of the date of publication. DOIs will be listed in the key resources table upon publication.

All original code has been deposited to GitHub (https://github.com/farleylab/Diverse-Logics-Notochord-Study) and is publicly available. DOIs will be listed in the key resources table upon publication.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

1.6 Acknowledgments

We thank the Farley Lab and Dennis Schifferl for helpful discussions. We thank Janet H.T. Song for her critical reading of the manuscript. We thank the UCSD IGM Genomics Center for their assistance with sequencing. B.P.S. was supported by NIH T32 GM133351. M.F.R. is supported by T32 GM008666. K.T. is supported by NSF 2109907 and 3DP2HG010013-01S1. G.A.J. was supported by a Hartwell Fellowship and NIH T32HL007444. E.K.F., B.P.S., M.F.R., K.T., G.A.J., J.L.G., S.H.L. were supported by NIH DP2HG010013.

Chapter 1, in full, is a reformatted reprint of the material as it appears in "Diverse logics encode notochord enhancers." Benjamin P. Song, Michelle F. Ragsac, Krissie Tellez, Granton A. Jindal, Jessica L. Grudzien, Sophia H. Le, Emma K. Farley. *In Submission*, 2022. The dissertation author was the primary investigator and co-first author of this paper.

1.6.1 Author contributions

E.K.F., B.P.S., M.F.R, K.T., G.A.J. designed experiments. B.P.S., K.T., J.L.G., S.H.L. conducted experiments. M.F.R. conducted bioinformatic analyses. E.K.F. and B.P.S wrote the manuscript. All authors were involved in editing the manuscript.

1.6.2 Declaration of interests

The authors declare no competing interests.

Chapter 2

A proof-of-concept method to identify enhancers using constraints on binding site motifs

2.1 Introduction

Enhancers are non-coding elements of the human genome that act as switches to regulate when and where genes are expressed. This feature of enhancers thus makes them a key contributor to tissue-specific gene expression during tightly controlled processes such as development and homeostasis[21;39–43]. Most disease mutations are located within enhancers. Additionally, the interplay between the syntax-the order, orientation, and spacing of transcription factor binding sites (TFBSs)-and binding affinity can finely control gene expression patterns through a mechanism known as "enhancer grammar"[23;44–46]. However, there is still a lack of understanding of how the grammar of a particular genomic sequence relates to proper or improper enhancer function[23]. With the increasing volume of genomic data collected due to next-generation sequencing (NGS), it is necessary to develop computational tools to mine genomes to pinpoint tissue-specific enhancers for further study[38;117–119].

Computational tools to identify enhancers have primarily focused on chromatin signatures[120;121]. While tissue-specific epigenomic data can sometimes pinpoint tissue-specific enhancers, this approach largely ignores the possible link between TFBS organization, or enhancer grammar, and tissue-specific activity[23;111;122–124]. Currently, four different models for enhancer-TFBS interactions have been proposed. The billboard model suggests that there are no constraints on TFBS arrangements within an enhancer—only that TFBSs be present

somewhere in the sequence[23;40;125]. In contrast, the enhanceosome model suggests that TFBSs must reside in a precise arrangement within an enhancer[23;49;126–128]. The TF-collective model suggests that in the absence of TFBS organization within an enhancer sequence, there is collective occupancy of the enhancer sequence by transcription factors (TF) through a combination of direct TF binding to TFBSs and TF-TF interactions[23;129]. There is a new model that has been proposed by our group that encompasses the previous three models as a spectrum based on the interplay of constraints. We call this "dependency grammar." Dependency grammar proposes that the interplay between TFBS syntax and affinity is shaped by biological, mechanistic, and evolutionary constraints[23]. Thus, identifying TFBSs within putative enhancers is a critical first step in determining enhancer grammar.

In a previous study, we tested 90 genomic regions containing Zic and ETS TFBSs and found additional binding sites-Brachyury (Bra) and FoxA-that may be playing a role in dictating enhancer activity in the *Ciona* notochord. From the nine sequences we found to be active, we created three groupings based on what collection of Zic, ETS, Bra, and FoxA binding sites were present: (1) Zic and ETS; (2) Zic, ETS, and Bra; and (3) Zic, ETS, Bra, and FoxA (Chapter 1)[25]. Interestingly, when testing genomic regions from the organism *Ciona intestinalis type A* (*Ciona*) containing these sites, we came to the striking conclusion that clusters of binding sites alone are not sufficient to drive expression even though all the transcription factors at play have some biological association with the nervous system or notochord (Chapter 1)[25]. Here, we focus on laying the groundwork for expanding the scope of the previous study to understand how enhancers regulated by Zic and ETS encode notochord expression within *Ciona* using an updated genomic reference sequence developed after we performed our initial screen by Satou *et al.* (2019)[12]. We then apply our methods to perform large searches for clusters of motifs within other vertebrate genomes, including chicken, mouse, zebrafish, and human. We have also developed EnGAGE (**E**ntire **G**enome se**A**rches for **G**rammars of **E**nhancers), a proof-of-concept computational framework to search for tissue-specific enhancers within genomes using one's knowledge of TFBS motif signatures. We propose that in the future, this tool can be further developed to look for enhancer grammar by allowing users to add constraints on TFBS syntax or affinity. In the following, we demonstrate the potential future synergy between EnGAGE and

massively parallel reporter assays (MPRAs) to study the *Ciona* notochord enhancers.

2.2 Results

2.2.1 Searching for clusters of Zic and ETS sites within an updated *Ciona* genome

In our previous study, we identified regions across the *Ciona* genome containing one Zic site and at least two ETS sites within 30 bp of the Zic site (Chapter 1)[24;25]. Within the study, we selected 90 of these identified regions to comprise the "ZEE Library." While we were able to identify a suitable number of ZEE elements to perform an enhancer screen, there were two key limitations in our initial search methodology that we wanted to address in an updated search. The first limitation was the genomic reference used for *Ciona*. While completing the analyses for our previous study, a new genomic reference for *Ciona* was released using more modern next-generation sequencing methods in 2019—the previous genome reference was assembled in 2008[8;12]. The second limitation of our previous search was the inherent search design itself. Because we fixed the Zic site in the center of the genomic element, we potentially missed functional ZEE elements that did not follow this constraint. To continue studying the notochord dependency grammar we previously identified at a greater scale, we developed a new search methodology that improved upon these limitations. We improved our methods by using the updated *Ciona* genomic reference and allowing for more flexibility in binding site location when searching for regions of interest containing Zic and ETS.

For the next iteration of our search, we identified 100 bp regions in the updated *Ciona* genome containing at least one Zic site and at least two non-overlapping ETS sites. Like our previous approach, we searched for ETS sites using the core motif, `GGAW` (`GGAA` or `GGAT`), to consider all ETS sites regardless of affinity[25;75;76]. We also defined Zic sites using EMSA and enhancer mutagenesis data from previous studies[16;25;27;74]. Using this approach, we identified 4,434 regions with at least one Zic and two ETS sites. Within this study, we define these regions as KYN elements to reference the new "KY" *Ciona* genome assembly, and we are looking at "N," or notochord, enhancers. In our previous study, we found that two other transcription factors expressed in the notochord, Bra and FoxA, may also contribute to the activity of some

enhancers that rely on Zic and ETS. Therefore, we also searched the updated KYN library for these other TFBSs (Chapter 1)[25]. Thus, we have three groupings of TFBSs that we were interested in: (1) Zic and ETS, (2) Zic, ETS, and Brachyury (Bra), and (3) Zic, ETS, Bra, and FoxA. In associating our KYN elements with each of these groups, we found that 65.9% belonged to the Zic/ETS group (2,863/4,344 KYN elements), 31.9% belonged to the Zic/ETS/FoxA group (1,384/4,344 KYN elements), and 4.3% belonged to the Zic/ETS/FoxA/Bra group (187 KYN elements) (Figure 2.2A).

After determining our genomic elements of interest, we wanted to evaluate how many ZEE elements (see 1) exist within the new KYN library using Magic-BLAST[130]. We created a custom BLAST database using the KYN elements as a reference, then searched for our ZEE elements within this custom database. Of the ZEE elements included in our previous study, 76.7% were present (69/90 ZEE elements) in our new KYN elements (Figure 2.1). Additionally, 77.8% of the ZEE elements expressed in the notochord were present in the KYN library (7/9 ZEE notochord-expressing elements), including the Brachyury Shadow (BraS) enhancer and the LAMA1/3/5 and LRIG1/2/3 enhancers, which our studies have previously identified (Figure 2.1). Additionally, one element, ZEE86, matched two KYN elements—KYN2713 and KYN4077 (Figure 2.1).

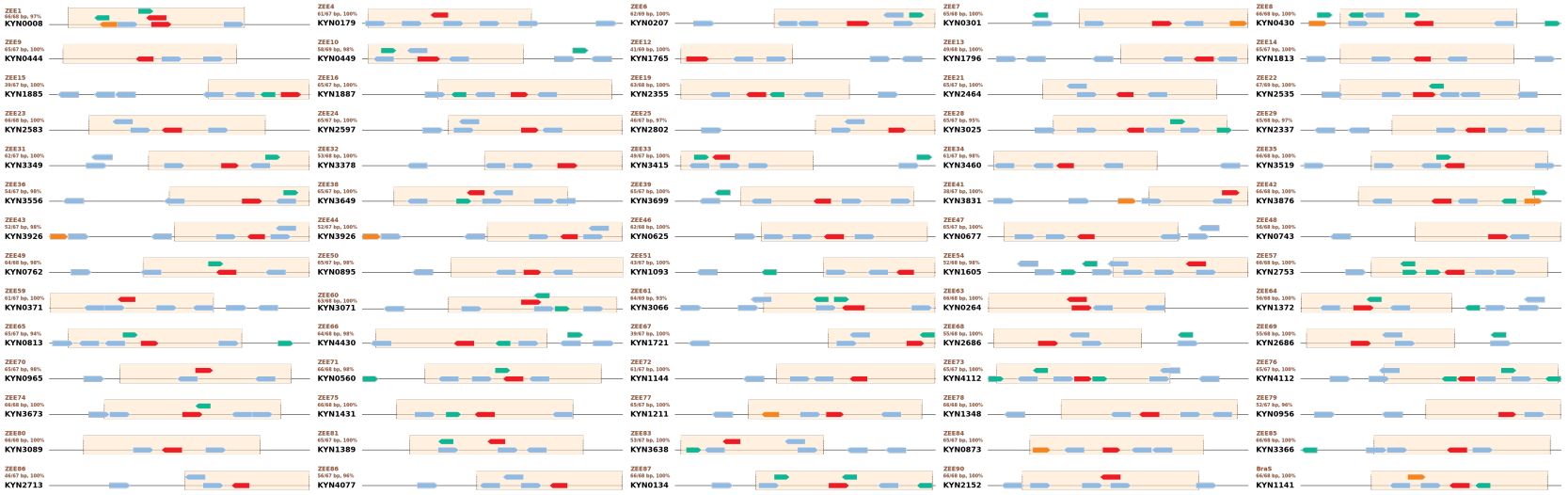

Figure 2.1. The majority of ZEE sequences can be found in the KYN library. As determined by Magic-BLAST, a total of 69/90 ZEE elements were found to overlap with members of the KYN library. Each diagram represented in the figure represents a schematic of the KYN enhancer, where the highlighted region represents the overlapping ZEE element that was detected by Magic-BLAST. The number of bases that aligned, along with the percentage alignment is also indicated. The colored binding sites represent the order, orientation, and spacing of Zic (red), ETS (blue), Bra (green), and FoxA (orange).

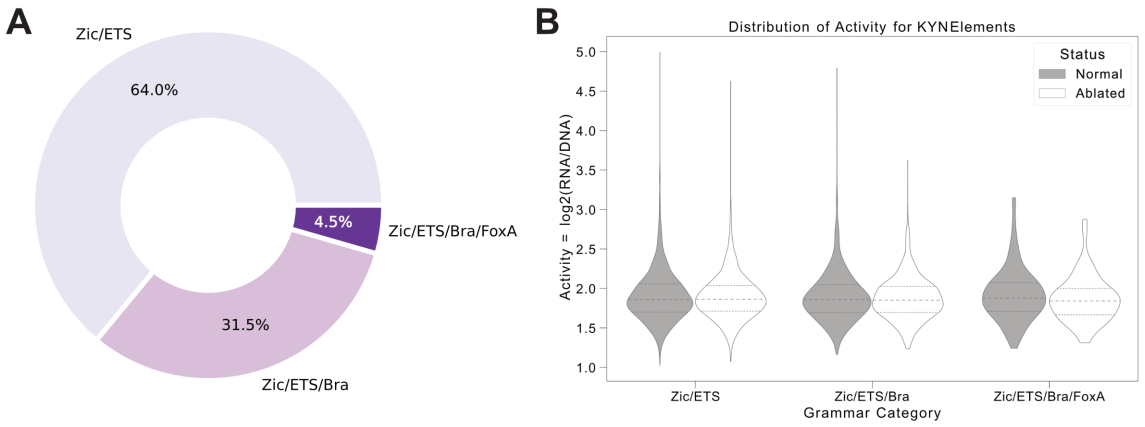

Figure 2.2. ZEE Library Contents and Expression. A. The distribution of all 4,344 KYN elements across the Zic/ETS (pale lilac), Zic/ETS/Bra (pale magenta), and Zic/ETS/Bra/FoxA (violet) grammar categories. **B.** Violin plot showing the distribution of enhancer activity for the KYN library screen split across the Zic/ETS, Zic/ETS/Bra, and Zic/ETS/Bra/FoxA grammar categories. The grammar categories are also separated by the members of the KYN library that are "normal" (grey) or "ablated" (white). The lines in the violin plot represent the 25th, 50th, and 75th percentiles.

2.2.2 Evaluating KYN genomic elements for enhancer activity in developing whole *Ciona* embryos

Next, we wanted to determine which KYN elements were functional by conducting an enhancer screen. We synthesized all 4,434 KYN elements upstream of a minimal promoter (bpFog) and a transcribable barcode. In addition to the ZEE elements that drove notochord expression in the previous study (Chapter 1)[25], we wanted to evaluate how centering the Zic site within the sequence would impact expression. Thus, we took the sequences for the ZEE element with the highest enhancer activity, ZEE1 or LRIG1/2/3, and BraS, and centered the Zic site based on the sequence present in the updated *Ciona* genome. Finally, we wanted to test the necessity of Zic and ETS for a given KYN element. To do this, we ablated the core of the Zic (`GCWG` to `GAWG`) and the core of the ETS (`GGAW` to `GCAW`) binding sites for each of the KYN elements and then included these sequences within the library for a total of 8,868 KYN elements.

Ultimately, our enhancer screen ended up including 8,872 sequences, as we had some dropouts that were not present in the final massively-parallel reporter assay (MPRA). Each enhancer sequence was associated with, on average, 88 barcodes, where each barcode represents

a specific measurement of enhancer activity. We then electroporated the enhancer library into fertilized *Ciona* eggs. We collected embryos at the late gastrula stage (5.5 hours post fertilization, hpf), as this stage is where notochord cells are developing, and both Zic and ETS are expressed[25–27]. At this time point, we isolated both the mRNA and plasmid DNA and then sequenced the mRNA and DNA barcodes present.

First, we filtered the data on if there was at least one mRNA or DNA barcode present for a given enhancer and if the Zic/ETS-ablated form of the enhancer was also present within the data. Next, we calculated an activity score for each KYN element. We first calculated the reads per million (RPM) for each mRNA and DNA barcode, then averaged the RPM across the mRNA and DNA barcodes associated with a given KYN element. To normalize the enhancer activity to differences in the amount of plasmid electroporated into each embryo, we took the log_2 of the average enhancer activity—mRNA RPM—divided by the average plasmid present—DNA RPM—for the same enhancer (Figure 2.2B). We then filtered out enhancers in each replicate that were lower than two standard deviations lower than the mean log_2 value, as well as filtered out enhancers that were higher than the 99th percentile of standard deviation of log_2 values. In total, our library had 83.0% of our expected sequences present (7,360/8,872 KYN sequences), including 85.7% of the sequences that overlapped with the ZEE library (60/70 ZEE elements with successful KYN hits). The lowest and highest activity scores calculated were 1.02 and 4.99, respectively. In the following section, we discuss our identification of active enhancer elements within the KYN library.

2.2.3 Several active KYN enhancers are proximal to genes implicated in the notochord and nervous system

In our enhancer screen, we were interested in identifying enhancers that would decrease in expression upon ablation of the Zic and ETS binding sites to ensure the importance of these sites in conferring activity in our *Ciona* embryos. We first filtered for functional enhancers by filtering for non-ablated regions with an activity score greater than or equal to the 90th percentile of the mean activity score or an activity score of 2.26. The 90th percentile was selected as an arbitrary cutoff that also acted as the most stringent. Next, we divided this group into three subsets based on the binding sites that were present: Zic/ETS, Zic/ETS/Bra, or Zic/ETS/Bra/FoxA.

To determine enhancers whose activity depended on Zic and ETS, we searched for enhancers that had ablated counterparts with an expression below the 90th percentile cutoff. In total, we identified 438 active regions containing Zic and ETS, 204 active regions containing Zic, ETS, and Bra, and 29 active regions containing Zic, ETS, Bra, and FoxA that were dependent on Zic and ETS. To identify the top candidates from each category to evaluate further, we calculated the difference in enhancer activity between the "normal" and "ablated" sequences and sorted enhancers by this value. We then determined proximal genes by expanding approximately 5 kb on either edge of the sequence and annotating genes within this expanded window as proximal to our region of interest. The top five candidates within each category can be found in Table 2.1.

Upon looking over the top candidates, we see the largest differences between normal and their ablated variants in the grouping containing Zic and ETS and the grouping containing Zic, ETS, and Bra, whereas we see almost minimal difference between the normal and ablated variants in the grouping containing Zic, ETS, Bra, and FoxA. When reviewing the top five candidates from each grouping of TFBSs, we were surprised to find multiple genes implicated in nervous system disorders, especially regarding brain-associated conditions (e.g., spinocerebellar ataxia), such as *RGS8*, *RGS4*, *SYS1*, *ALDH4A1*, *WARS2*, *ITPR1*, *XRN2*, *KCNQ3*, and *OTX1*. Several genes were also implicated in skeletal and bone-related disorders, such as *HOXD3*, *DYNLT2B*, and *FLG*. While we have put these enhancers into these three groups, we do not know if the Bra and FoxA sites are contributing to their activity. However, we do know that these enhancers are dependent on Zic and ETS, two transcription factors critical in neural and notochord development. Ultimately, more exploration is needed to discern if these regions are truly functional within the *Ciona* notochord through imaging studies, but their primary association with homologous human genes is promising. Additionally, more work is needed to ascertain if known disease-associated SNPs fall within regions containing potential notochord enhancer grammars consisting of Zic, ETS, Bra, and FoxA binding sites.

Table 2.1. Top five KYN elements across grammatical categories

GRAMMAR	KYN ID	LOCATION	ACTIVITY(N)	ACTIVITY(A)	ACTIVITY(N-A)	PROXIMAL GENES
Zic/ETS	KYN4389	chr9:2616602-2616702	4.57	1.98	2.59	*POM121L2*, *RGS8*, *RGS4*
Zic/ETS	KYN4390	chr9:2616623-2616723	4.26	1.97	2.29	*POM121L2*, *RGS8*, *RGS4*
Zic/ETS	KYN1765	chr2:2253506-2253606	3.38	1.58	1.79	*ERCC5*, *SYS1*, *ESD*
Zic/ETS	KYN0616	chr10:1184720-1184820	3.75	1.99	1.76	*ALDH4A1*, *CBL*
Zic/ETS	KYN0516	chr1:14250867-14250967	3.61	1.95	1.66	*HOXD3*, *DYNLT2B*
Zic/ETS/Bra	KYN2946	chr6:381260-381360	4.79	1.71	3.08	*WARS2*
Zic/ETS/Bra	KYN3554	chr8:5563238-5563338	2.79	1.33	1.46	*ITPR1*
Zic/ETS/Bra	KYN0942	chr11:2883954-2884054	2.74	1.37	1.36	*YME1L1*
Zic/ETS/Bra	KYN2661	chr4:6611261-6611361	3.12	1.82	1.31	*SEMA6A*, *ADAMTSL1*

Continued on next page

Table 2.1. Top five KYN elements across grammatical categories, *continued from previous page*

GRAMMAR	KYN ID	LOCATION	ACTIVITY(N)	ACTIVITY(A)	ACTIVITY(N-A)	PROXIMAL GENES
Zic/ETS/Bra	KYN1322	chr12:6286408-6286508	2.87	1.58	1.29	*PTPRF*, *PTPRQ*, *DNAJA1*
Zic/ETS/Bra/FoxA	KYN4030	UAContig6:165904-166004	3.12	2.03	1.08	*FLG*, *GIN1*, *XRN2*, *HUS1B*, *KCNQ3*
Zic/ETS/Bra/FoxA	KYN3090	chr7:319156-319256	2.58	1.57	1.00	*ELAC2*
Zic/ETS/Bra/FoxA	KYN4335	chr4:4509497-4509597	2.90	1.94	0.96	*OTX1*, *CETN2*
Zic/ETS/Bra/FoxA	KYN1404	chr13:639638-639738	2.72	1.77	0.95	*PIK3AP1*, *KCNJ5*
Zic/ETS/Bra/FoxA	KYN4230	chr10:4824445-4824545	2.49	1.59	0.90	*ANGPT2*

2.2.4 Identifying genomic regions containing Zic and ETS binding sites in other species

In hopes of finding additional regions for us to evaluate in the future, we applied our methodology to gather KYN regions from *Ciona* to other vertebrate species. We also searched the *Ciona savigni* for regions containing Zic and ETS binding sites to compare against *Ciona intestinalis type A*. The number of regions obtained in this approach can be found in Table 2.2. As expected, with an increase in the genome size, we see an increase in the number of sites with Zic and ETS binding sites. More exploration is needed to evaluate where these regions fall in relation to notochord and neural-associated genes and if the grammar between our various transcription factors of interest is comparable across members of Chordata.

Table 2.2. Number of regions containing Zic and ETS found across other species

NAME	SCIENTIFIC NAME	GENOME ASSEMBLY	REGIONS
Solitary sea squirt	*Ciona savignyi*	CSAV2[1,131]	8,061
Chicken	*Gallus gallus*	galGal6[2]	121,913
Zebrafish	*Danio rerio*	danRer11[3]	170,972
Mouse	*Mus musculus*	mm10[4]	182,575
Human	*Homo sapiens*	hg38[5]	179,501

2.2.5 Developing a proof-of-concept software package for clusters of binding sites within genomes

Finding our preliminary exploration into active enhancers promising, we wanted to develop a tool that could translate our genomic searches in *Ciona* and other vertebrates to other organisms that we did not feature within this study. We then developed a method to look for clusters of binding sites within genomes in Python called Entire Genome seArches for Grammars of Enhancers (EnGAGE, `engage-tools` GitHub Repository Link).

EnGAGE is a proof-of-concept Python package to search for clusters of TFBS motifs of choice within an input reference genome using regular expression definitions of binding sites. Using EnGAGE, users can define a `Cluster` parent class object to which they can add various child class `TF` objects. These `TF` objects represent individual transcription factor binding motifs in. After the parameters have been set, the user can use the `find_motif_cluster()` method to search through any genome of interest for locations of particular clusters for further exploration.

As more information is learned about enhancers, additional constraints on TFBS syntax and affinity can be added to the `Cluster` object to search for functional grammars.

2.3 Discussion

The marine chordate *Ciona* is easily amenable to high-throughput enhancer studies, making it a valuable model system for studying functional genomics. The *Ciona* genomic reference sequence was recently reassembled based on modern sequencing techniques, providing dramatic improvements over the previous reference genome from the early 2000s[8;12]. In this work, we sought to lay the groundwork for future studies to understand the regulatory logic of notochord enhancers we discovered in a previous study of Zic, ETS, FoxA, and Brachyury binding sites (Chapter 1)[25]. While we found a total of 4,344 elements in the *Ciona* genome containing at least one Zic site and two ETS sites, testing these elements in an MPRA revealed that only 15.4% of these sites (671/4,344 regions) were active *and* dependent on Zic and ETS. Further study of this enhancer library will likely identify novel notochord enhancers and help us better understand how Zic and ETS encode notochord development through particular grammatical constraints.

2.3.1 Differences between the ZEE library and KYN library

After searching for new elements in the updated *Ciona* genome to formulate the KYN library, we wanted to evaluate if there was an overlap between these elements and our previous study of the ZEE elements (Chapter 1)[25]. While we could corroborate the majority of ZEE elements within the updated KYN library using Magic-BLAST[130], the amount of overlap varied, and some sequences had perfect alignment but less than 50 bp of overlapping sequence (e.g., ZEE12, ZEE13, ZEE15, and ZEE33) (Figure 2.1). Additionally, the sizes of the regions between the ZEE library and KYN library varied. Regions tested in the ZEE library were approximately 69 bp in length (Chapter 1)[25], whereas regions tested in the KYN library were fixed at 100 bp. The additional length of the KYN library has the potential to introduce additional sequence elements that would cause discordance between the ZEE and KYN library results. Indeed, when we evaluate the expression between ZEE elements and KYN elements, there are elements from the ZEE library with notochord expression that are only moderately active in the KYN library

and vice versa.

2.3.2 Further exploration is needed to understand active KYN elements

Without much overlap between the sequences of the ZEE library and KYN library (Figure 2.1), it was difficult to determine the threshold for determining active enhancers within our study. Thus, we set a stringent threshold and strict filtering criteria to label active enhancers within our study; all elements within our study were required to have an enhancer activity greater than the 90th percentile of overall activity. Overall, this allowed us to identify potentially interesting targets for future study (Table 2.1).

Unfortunately, while the ablation studies help us understand the necessity of Zic and ETS binding sites present within the KYN elements, they do not allow us to ascertain the importance of other binding sites, such as Bra and FoxA. Indeed, some of the elements in which Zic and ETS binding site ablation leads to similar or higher levels of enhancer activity compared to their original sequence may be dependent on the Bra or FoxA sites or other TFBSs that we have not yet identified (*not featured in this study*). Thus, more exploration is needed to image these elements, conduct follow-up experiments to dissect these sequences and determine proper thresholds for future work.

2.4 Materials and Methods

2.4.1 *Ciona intestinalis* dechorionated, *in vitro* fertilization, and electroporation

Adult *Ciona intestinalis type A*, also known as *Ciona robusta*, were obtained from M-Rep and were maintained under constant illumination in seawater (obtained from Reliant Aquariums) at 18°C. *Ciona* are hermaphroditic, therefore, there is only one possible sex for individuals. Age or developmental stage of the embryos studied is indicated in the main text. Methods for dechorionated, *in vitro* fertilization, and electroporation were performed as described previously in Farley et al., 2016[24].

2.4.2 Identification of KYN putative notochord enhancers and conducting vertebrate genome searches for elements containing Zic and ETS binding sites

We identified elements in the updated *Ciona* genome by first identifying clusters of Zic and ETS sites across each chromosome. We used the following sites and their corresponding reverse complement sequence in our search for Zic binding sites: `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., 2007a; Yagi et al., 2004)[16;27]. Methods for obtaining genomic regions to include in the KYN library and the vertebrate genomes included in Table 2.2 can be found in Supplementary Figure B.1.

2.4.3 Construction of the KYN enhancer library

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 30 bp 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.

Enhancer to barcode tag assignment & enhancer dictionary analysis

We constructed a dictionary of unique barcode tag-enhancer pairs by not allowing for any mismatches in the 100 bp enhancers in our library and by not allowing barcode tag-enhancer pairs to have a read count of fewer than 25 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, there were 40 barcode tags that were matched to multiple

enhancers and thus discarded from the final dictionary. In total, the dictionary contained 748,258 total barcode tag-enhancer associations and 8,460 total enhancers that were uniquely mapped to one or more barcode tags. The median and mean number of barcode tags associated with a single enhancer were 76 and 88, respectively.

2.4.4 Conducting the KYN MPRA screen

50 μg of the KYN library was electroporated into 5,000 fertilized eggs. Embryos developed until 5 hours and 30 minutes at 22°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 500 μL of Back Extraction Buffer (4 M guanidine thiocyanate, 50 mM sodium citrate, and 1 M 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.

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 collapsed 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. All DNA and RNA libraries were processed separately for the preliminary filtering step.

Prior to normalizing our data into RPM, we first filtered out all enhancers that did not have their ablated pair present within the sample. We then filtered our data further to only include the set of barcode tags and enhancers that appear in DNA across all replicates. We then consolidated the expression for each enhancer by taking the average RPM value across barcode tags. To determine if an enhancer was active, we calculated an "enhancer activity score." This score is calculated by averaging the $log_2(\frac{RNA}{DNA})$ value across a given enhancer's biological replicates.

2.5 Acknowledgements

I would like to thank the following individuals that made this work possible: Benjamin P. Song, Jessica L. Grudzien, Sophia H. Le, Joe J. Solvason, and Emma K. Farley. I would also like to thank the Farley Lab, Hannah Carter, and the Carter Lab–especially Adam Klie–for helpful discussions during the analysis and visualizations of the data included in this chapter. I would also like to thank the team at the UCSD IGM Genomics Center for their assistance with sequencing.

2.6 Footnotes

2.6.1 Author contributions

E.K.F., B.P.S., M.F.R, designed experiments. B.P.S., J.L.G., and S.H.L. conducted experiments. M.F.R. conducted bioinformatic analyses. M.F.R. and J.J.S. were involved in the software development of EnGAGE. M.F.R. wrote the chapter. E.K.F., M.F.R., and B.P.S. were involved in editing the chapter.

2.6.2 Funding

M.F.R. was supported by NIH T32 GM008666. B.P.S. was supported by NIH T32 GM133351. J.J.S. was supported by NIH T32GM127235. E.K.F., B.P.S., M.F.R., J.L.G., S.H.L., and J.J.S. were supported by NIH DP2HG010013.

2.6.3 Data availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Emma K. Farley (efarley@ucsd.edu), upon request. All plasmids generated in this study and KYN screen sequencing data will be deposited to Addgene and GEO, respectively, when a publication is made available. All original code has been deposited to GitHub (https://github.com/farleylab/Expanded-Notochord-Logics-Study) and is publicly available. The code for the proof-of-concept tool, EnGAGE, has also been deposited to GitHub (https://github.com/farleylab/engage-tools) and is publicly available. Any additional information or data required to recapitulate the study reported in this chapter is available upon request.

2.6.4 Declaration of interests

The authors declare no competing interests.

Chapter 3

Understanding *Ciona intestinalis* gastrulation at single-cell resolution

Since its early days as a choice model organism for critical biologists such as Laurent Chabry, Ed Conklin, and Thomas Hunt Morgan, *Ciona intestinalis* (*Ciona*) has been known for its simple embryos, rapid development, and ease of manipulation for embryological studies. Additionally, as a member of the subphylum Urochordata, *Ciona* represents the simplest and most primitive chordate body plans as our closest invertebrate relative. Several groups have already provided insight into *Ciona* embryogenesis by constructing partial gene regulatory networks or focusing on tissue-specific gene expression changes during particular developmental time points. However, there is still much we can delineate from studying cell fate determination pathways.

Technological advances have enabled the cataloging of global gene expression profiles of single cells using single-cell RNA-sequencing (scRNA-seq), allowing scientists to define the heterogeneity within cell populations during embryonic development. This new paradigm has allowed developmental biologists, including those that study *Ciona*, to identify precisely when and in which cell types genes controlling cell fate decisions are expressed. Indeed, a previous study by Cao *et al.* (2019) developed a single-cell transcriptional atlas for more than 90,000 cells spanning the onset of gastrulation through the swimming tadpole stage in *Ciona*. Their atlas spanned the 4.5 hours post fertilization (hpf) time point to the 18 hpf time point, demonstrating the feasibility of atlas-scale, whole embryo single-cell methods in *Ciona* and other marine tunicates. In this chapter, I aim to lay the groundwork for studying gastrulation in *Ciona* at a higher resolution than before, incorporating approximately 350,000 cells into a transcriptional atlas

spanning just the 4.5 hpf, 5.5 hpf, and 6.5 hpf time points. By integrating only these three key time points in development, we can identify canonical and novel cell type markers and delineate pathways contributing to notogenesis.

3.1 Introduction

Embryonic development begins upon the fertilization of an egg by a sperm cell to become a single-cell zygote, which continues through many stages of cell division to form a functional organism. Developmental processes are finely orchestrated by enhancers. Non-coding elements of the genome that control the timing, location, and levels of gene expression within cells. Expression of the correct collection of genes within cells is necessary for embryonic axis formation and body plan patterning, processes required for proper development[10;20;21;132–136]. Although genetics and experimental embryology have dissected the major transcription factors and secreted signaling molecules involved in the specification of early cell lineages, the processes governing development involve many circuits beyond the well-known factors[10;80;133–136]. Thus, there is a continued need to explore the mechanisms involved in development to understand how deficiencies in cell fate specification contribute to developmental disease. Historically, gene expression studies have been limited to analyzing pooled populations of cells to obtain sufficient RNA for analysis despite the importance of cell heterogeneity in organ development[133;137–139]. Fortunately, advances in genomic technologies have allowed developmental biologists to assess the early gene expression events associated with fate specification in single cells[34;35;140–142]. Through single-cell RNA sequencing (scRNA-seq), we can now evaluate the RNA expression of every gene at single-cell resolution. In this chapter, I used scRNA-seq to explore early organ formation in the urochordate, *Ciona intestinalis type A* (also known as *Ciona robusta* or *Ciona*), to understand the transcriptional landscape in the notochord and other major cell types present during gastrulation.

Gastrulation is an early, formative developmental process that involves the reorganization of an embryo from a one-dimensional layer of epithelial cells (blastula or blastocyst) into a multi-layered, multi-dimensional structure (gastrula). It results in the formation of the major germ layers in the developing embryo (e.g., endoderm, ectoderm, and mesoderm) that act as precursors to all embryonic tissues, as well as the establishment of the dorsal/ventral and anterior/posterior

axial orientations of the embryo. After forming the major germ layers, the embryo is primed for key organ and structure formation[13;106;143;144]. The phylum Chordata is a large division of the animal kingdom that includes vertebrates, tunicates, and cephalochordates[1;8;9;69]. All chordate embryos share, among a few other hallmarks, a defining structural feature known as the notochord that forms during gastrulation that is present during some or all of their life cycle. The notochord is a hollow tube of mesodermal origin extending from the anterior to the prechordal plate. It is a flexible, midline cartilaginous rod of tissue found in very close connection with the ventral-most region of the neural tube. Beyond its structural role, the notochord plays an indispensable role in the formation of the neural tube through the secretion of various developmental morphogens, including *sonic hedgehog* (*shh*)[1–7;14;50;55;69;71;79;86;94;98;101;136]. The intricate relationship between the notochord and the formation of other key structures, such as the neural tube, renders it necessary to understand notogenesis to treat notochord-derived disorders and defects.

Within vertebrates, the notochord is a transient anatomical structure only present in the early embryo. Notochord-derived abnormalities can be traced to stress on the pathways responsible for notochord cell maintenance in adulthood or to remnants of the notochord that fail to regress during early development. The remnants of the notochord constitute the nucleus pulposus, the innermost compartment of the intervertebral discs[6;7;51;55;144]. Within the nucleus pulposus, notochord cells secrete extracellular matrix (ECM) molecules to form a proteoglycan-rich and gelatinous matrix that acts as the cushioning infrastructure responsible for the shock-absorption properties of the intervertebral discs. These properties are necessary for general movement and flexibility of the backbone in vertebrates[6;7;144]. Degeneration of notochordal cells in the nuclei pulposi causes the onset of intervertebral disc degeneration and consequent back pain, the leading cause of disability in the adult population worldwide[145–153]. Thus, many groups have focused on dissecting the factors important for notogenesis to identify potential therapeutic agents to limit or reduce the symptom-causing pathologies of intervertebral disc degeneration by targeting pathways inducing structural disruption or inflammation[149–153]. Another notochordal defect includes chordomas, a rare type of bone sarcoma that represents about 1% to 4% of primary bone tumors. While the mechanistic knowledge of chordoma formation is limited, there is evidence that they are derived from embryonic remnants of the notochord. For example, long before it was proposed as

a diagnostic marker for chordomas, brachyury was identified as a regulator for notogenesis and as a general biomarker for notochord development as well as notochord-derived tumors[6;7;152;154–161]. Brachyury is a highly conserved T-box transcription factor that helps promote cell movement and adhesion, which are fundamental for morphogenesis and tumorigenesis; it is also a known marker for the developing notochord[15;27;51–54;60;61;64;65;87;89;92;98;100;101]. With the fundamental role of brachyury in notochord development, further research into the factors involved in notogenesis is important to better understand whether aberrant activation of notochord GRNs contributes to chordomagenesis.

The marine tunicate *Ciona* is a member of the subphylum Urochordata and is thought to represent the simplest and most primitive chordate body plans[8–12;116]. While *Ciona* has been extensively studied, there is still much we can delineate from comparing *Ciona* cell fate determination pathways to other chordate species, especially concerning notochord specification and conservation. In a previous study, Cao et al. developed a single-cell transcriptional atlas spanning the onset of gastrulation through the swimming tadpole stage in *Ciona*. Within this study, they were able to construct virtual cell-lineage maps and gene networks for 41 neural subtypes that comprise the larval nervous system[162]. Other single-cell studies performed in tunicates have also proved successful in studying lineage specification in other cell types[81;163–168]. As various groups have demonstrated the feasibility of performing atlas-scale single-cell methods in *Ciona* and other tunicates, we used scRNA-seq to generate a comprehensive single-cell gene expression atlas spanning the onset of gastrulation to study the GRNs dictating notochord fate specification.

3.2 Results

3.2.1 *Ciona intestinalis* single-cell expression atlas spanning gastrulation

Ciona embryos were allowed to develop to either the 4.5 hours post fertilization (hpf), 5.5 hpf, or 6.5 hpf time points representing the early gastrula or 110-cell stage, late gastrula, and early neurula stages of development (Figure 3.1B). After developing to our time point of interest, we rapidly disassociated embryos for a particular time point in order to conduct sample processing under the 10x Genomics Chromium system and further data analysis with scanpy[169].

We had three biological replicates for each developmental stage. In total, we were able to profile 356,671 cells, allowing us to identify rare subpopulations present within the gastrula, including germ cells and the developing heart (Table 3.1, Figure 3.1A).

After conducting data normalization and Leiden clustering of neighboring cell transcriptional profiles, we performed dimensionality reduction using the uniform manifold approximation and projection (UMAP) method to visualize our *Ciona* gastrulation atlas. As the stages of gastrulation begin body plan fate specification for all cells in the growing embryo, previous studies have suggested that all major tissues within *Ciona* are specified as early as the 110-cell stage[10;11;80;162]. Thus we were pleased to see corroboration of this within our single-cell atlas (Figure 3.1A). Within our single-cell atlas, we were able to identify all major tissues–including the epidermis, endoderm, notochord, mesenchyme, nervous system, heart, muscle, and germ cells–using canonical cell type markers defined in the Aniseed and Ghost databases for ascidian research (see Methods; Figure 3.1A, Figure 3.1C-J, Table 3.1). To our surprise, we could also specify the particular starting cell lineage for subcellular clusters present within these tissues, such as the A- and B-lineages of both the notochord and mesenchyme (Figure 3.1A). By providing a higher resolution single-cell atlas just spanning gastrulation, we anticipate that this map can be used to identify conserved canonical and novel cell differentiation markers, primarily when used in conjunction with single-cell integration methods to explore conserved markers across species.

Table 3.1. Distribution of cells across annotated cell types

CELL TYPE	**NUMBER OF CELLS**	**% EMBRYO, DATA**	**% EMBRYO, LITERATURE**
Epidermis	174,113	48.82	-
Endoderm	44,816	12.57	-
Nervous System	42,338	11.87	12.68
Mesenchyme	31,695	8.89	-
Notochord	31,263	8.77	6.67
Muscle	24,661	6.91	5.33
Germ Cells	4,730	1.33	0.67
Heart	3,055	0.86	1.33

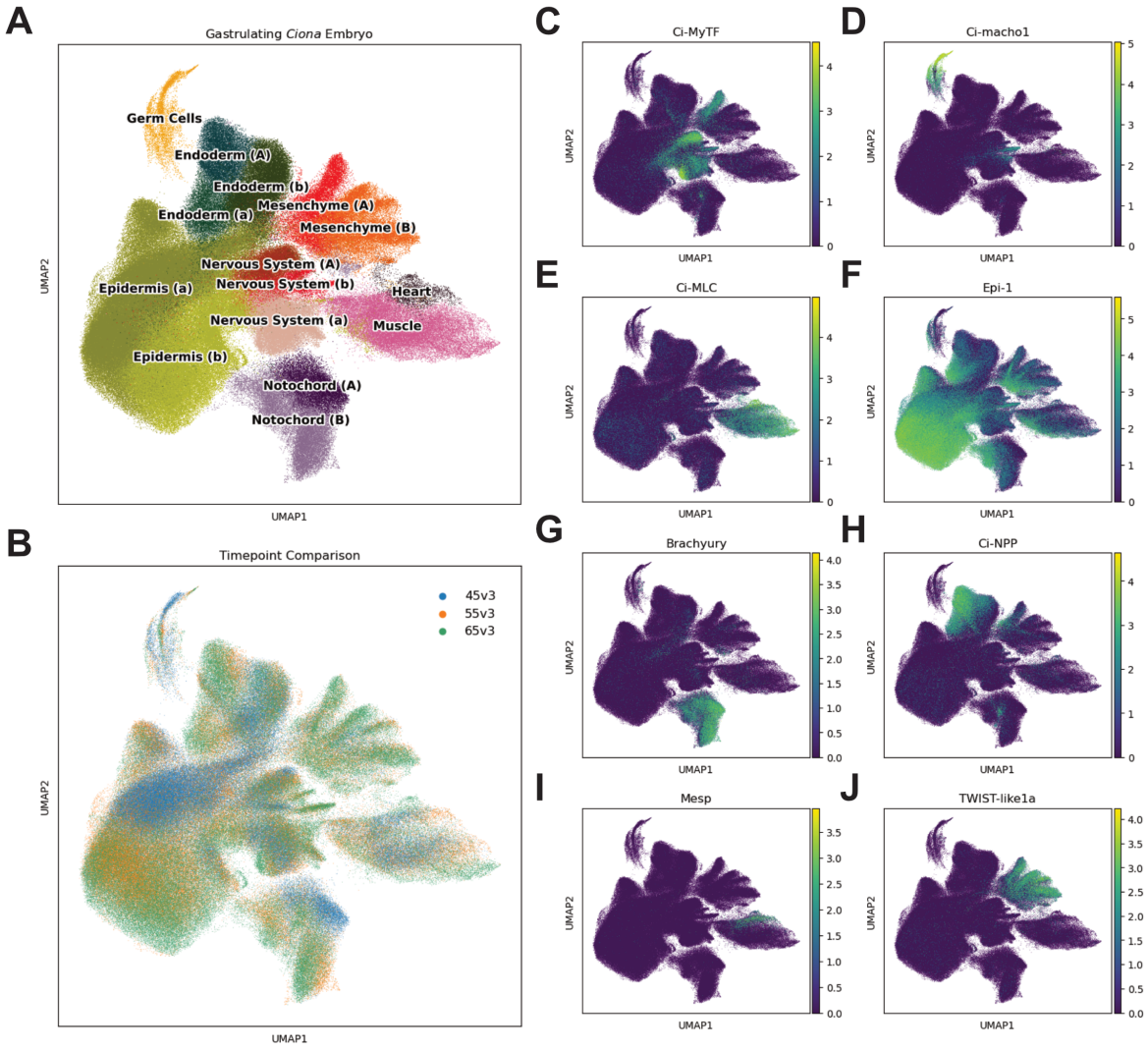

Figure 3.1. A. UMAP plot of all cell types present during *Ciona* gastrulation (consolidated across the 4.5 hpf, 5.5 hpf, and 6.5 hpf time points) resolved into the particular A, B, a, or b-lineages present in the *Ciona* embryo at the 4-cell stage. The distribution of cells across major cell types can be found in Table 3.1. **B.** UMAP plot of cells in the *Ciona* gastrula separated by time point, where `45v3` represents the 4.5 hpf time point, `55v3` the 5.5 hpf time point, and `65v3` the 6.5 hpf time point. **C-J.** UMAP visualizations of various canonical cell type marker genes used in the determination of cell type cluster identification.

3.2.2 Validating single-cell RNA-sequencing results with *in situ* hybridization studies

For some of the *in situs* we found on the Aniseed database alongside canonical cell type markers, they were either dated, poor resolution, or had ambiguous expression patterns despite being marked as exclusively expressed within a certain cell type according to the Aniseed API. Additionally, there were some genes that did not have an annotated expression pattern at all[170]. Thus, we sought to verify markers identified through our clustering methods using fluorescent *in situ* hybridization (FISH) imaging performed in our lab (see Methods). As an example, we used *Brachyury* to identify the notochord cluster in our UMAP visualization as it is known to be affiliated with notochord development in *Ciona* (Figure 3.2)[15;27;51–54;60;61;64;65;87;89;92;98;100;101]. After identifying the notochord cluster with *Brachyury*, we found another marker, *Orphan bHLH1* (Figure 3.2D), where it was unclear if it was also notochord-specific based on *in situ* images on Aniseed. We then tested the two markers in tandem using FISH (Figure 3.2E-G), ultimately confirming their co-expression in the notochord and showing the validity of our single-cell clustering results.

As expected from the high resolution of our study owing to the large number of cells encompassing each time point, we were also able to identify novel markers for particular cell types. One such marker, a *Ciona* gene that had sequence homology to the vertebrate gene *Arx*, was identified (Figure 3.3). *Ci-Arx* was found to be specifically expressed to the A-lineage nervous system of *Ciona* corresponding to Row III and Row IV of the developing neural plate during gastrulation (Figure 3.3A, Figure 3.3D-G). These rows form the anterior sensory vesicle in *Ciona* in the adult organism, correlating to the brain[11]. Our data also corroborates the finding of *Ci-Arx* and its implications in the anterior sensory vesicle from the Cao *et al.* (2019) single-cell study in *Ciona*[162]. Previous literature has shown *Arx* expression in the embryonic forebrain of both mouse and zebrafish and has implicated *Arx* in X-linked lissencephaly, a human disease marked by the absence of folds in the cerebral cortex and an abnormally small head[171;171–174]. Additionally, we found another marker within *Ciona* that had sequence homology to vertebrate gene *SWT1* and that was specifically expressed in the germ cells of the embryo (Figure 3.4). While it is not characterized in *Ciona*, the vertebrate *Swt1* has been found in human testicular tissue[175].

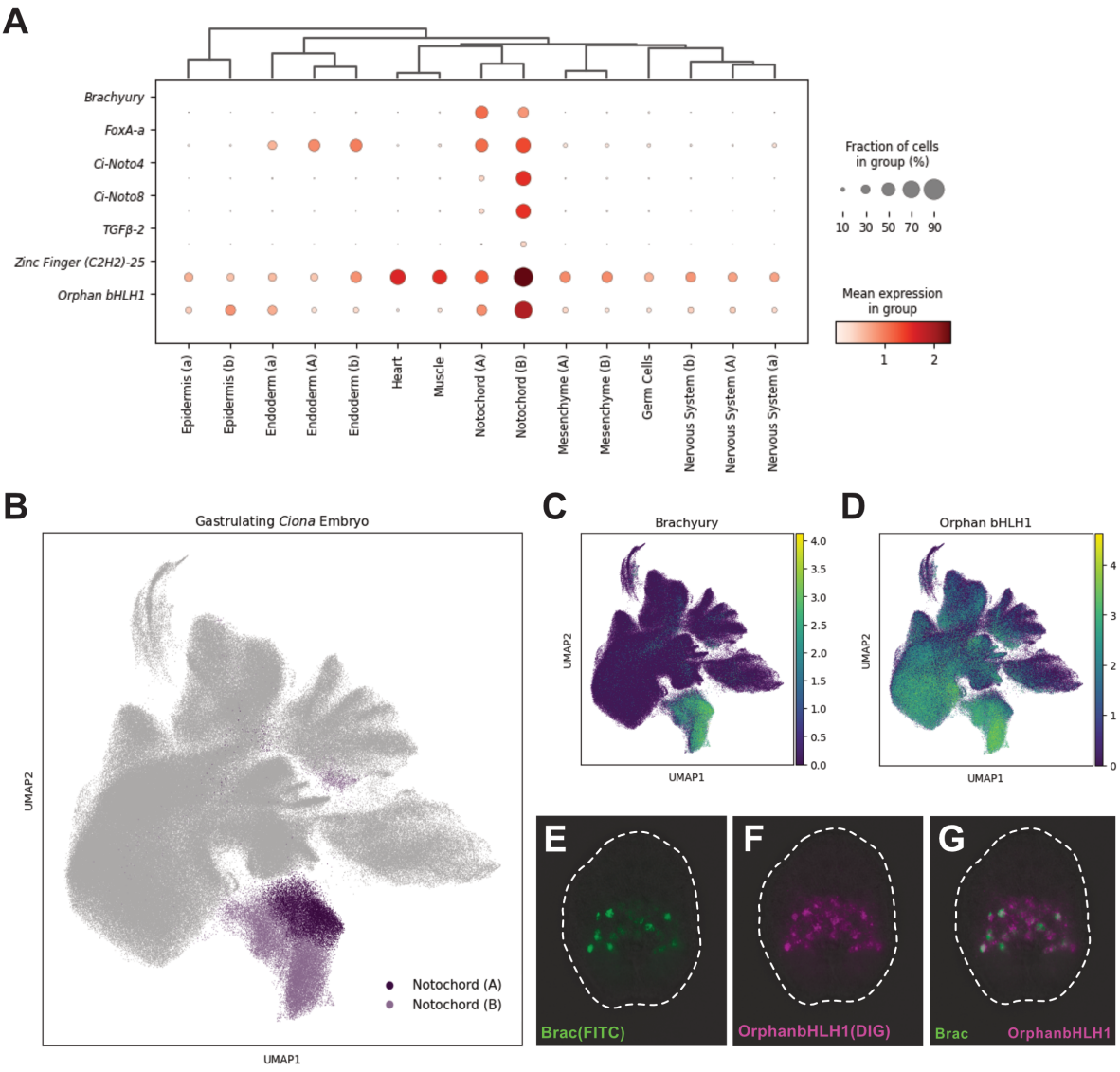

Figure 3.2. A. Dot plot of key notochord markers, including *Brachyury*, *FoxA-a*, *Ci-Noto4*, *Ci-Noto8*, *TGFβ-2*, and *Zinc Finger (C2H2)-25*, in comparison to a less-studied notochord marker, *Orphan bHLH1*. **B.** UMAP plot of cells in the *Ciona* gastrula with the A-line and B-line notochord lineages highlighted in dark and light purple, respectively. **C-D.** UMAP visualizations of *Brachyury* (C) and *Orphan bHLH1* (D) in the single-cell atlas. **E-G.** *FISH* images of *Brachyury* (E), *Orphan bHLH1* (F), and the overlay of the two (G) in 5.5 hpf *Ciona* embryos (late gastrula stage).

Additionally, vertebrate *SWT1* is also relatively understudied within germ cells, providing a potential avenue for future research on its conservation and function. The discovery of *Ci-SWT1* and *Ci-Arx* from our single-cell map of the gastrulating *Ciona* embryo represents a potential model for interrogation of novel gene expression patterns and hints at the potential usage of our atlas to uncover conserved genes delineating cell fate.

3.3 Discussion

Through scRNA-sequencing at whole embryos at a relatively small but expansive time frame of development, we were able to uncover the transcriptional signatures of major cell types in the *Ciona* gastrula. While there are already a multitude of single-cell studies that have been performed, this study constitutes one of the highest resolution atlases to date, confirming results from the previous atlas-scale effort, but with a restricted time window. With the findings of well-studied, canonical markers within *Ciona* to be specifically expressed within particular cell types alongside understudied genes, we hope that this dataset will provide suitable groundwork for future explorations into cell fate specification and its potential conservation across species[81;162–168]. We anticipate that future imaging studies into the novel marker genes we have found for each of the major cell types present during *Ciona* gastrulation will provide much needed annotations into the conservation of transcriptional pathways governing organ formation across Chordates.

3.4 Materials and Methods

3.4.1 *Ciona* handling, collection, dissociation, and imaging of embryos

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

Ciona embryos were dechorionated as described in Christiaen *et al.* (2009)[176]. Embryos were allowed to develop to either 4.5 hours post fertilization (hpf), 5.5 hpf, or 6.5 hpf in seawater. Embryos were dissociated by resuspension 1:3 Accumax:Artificial Seawater (ASW)-Mg-Ca, followed by light vortexing and gentle pipetting with Pasteur pipettes. Dissociated cells were

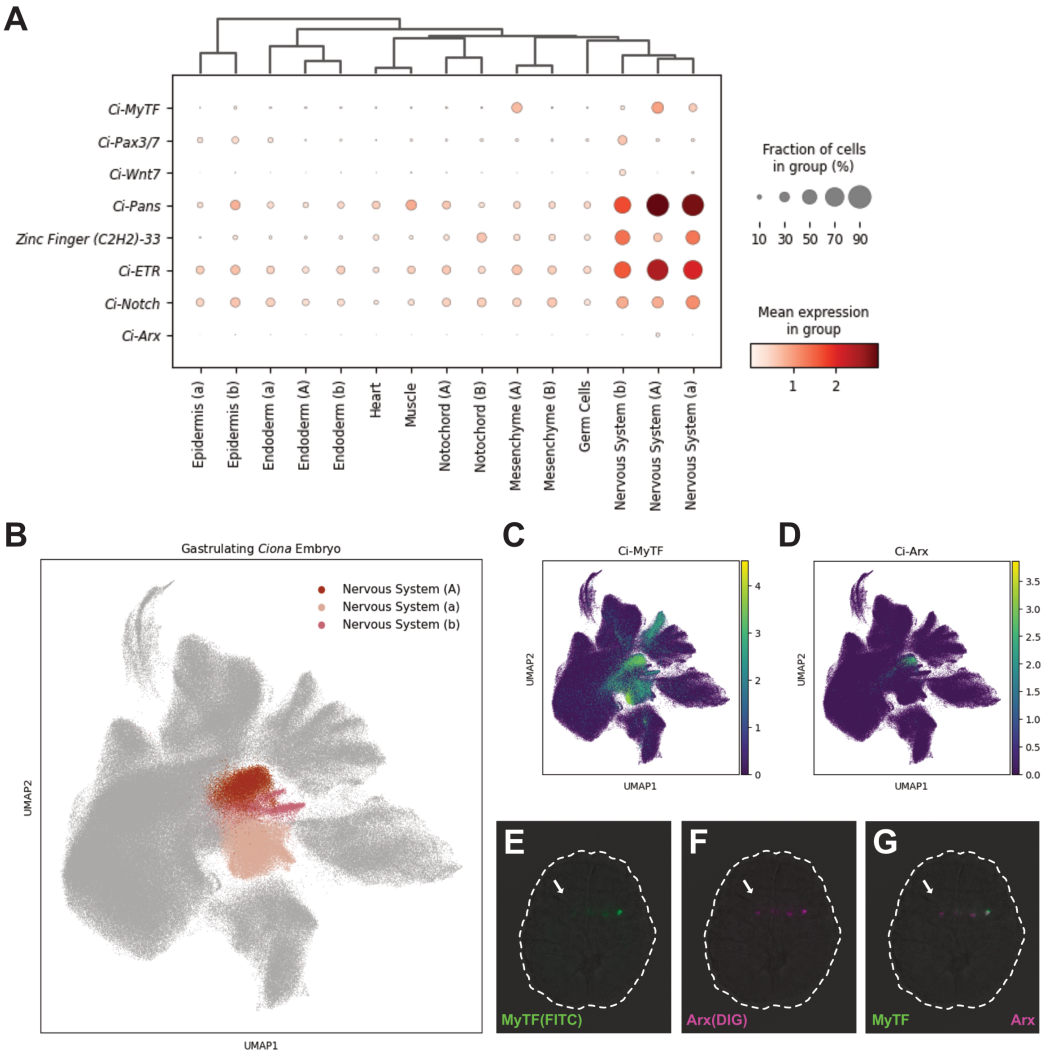

Figure 3.3. **A.** Dot plot of key nervous system markers, including *Ci-MyTF*, *Ci-Pax3/7*, *Ci-Wnt7*, *Ci-Pans*, *Zinc Finger (C2H2)-33*, *Ci-ETR*, and *Ci-Notch*, in comparison to a putative *Ciona* neural marker, *Ci-Arx*, as named via sequence homology. **B.** UMAP plot of cells in the *Ciona* nervous system with the A-line, a-line, and b-line neural lineages highlighted in burnt orange, beige, and salmon respectively. **C-D.** UMAP visualizations of *Ci-MyTF* (C) and *Ci-Arx* (D) in the single-cell atlas. **E-G.** *FISH* images of *Ci-MyTF* (E), *Ci-Arx* (F), and the overlay of the two (G) in 5.5 hpf *Ciona* embryos (late gastrula stage). The arrow indicates the developing *Ciona* neural plate Row III region of the embryo.

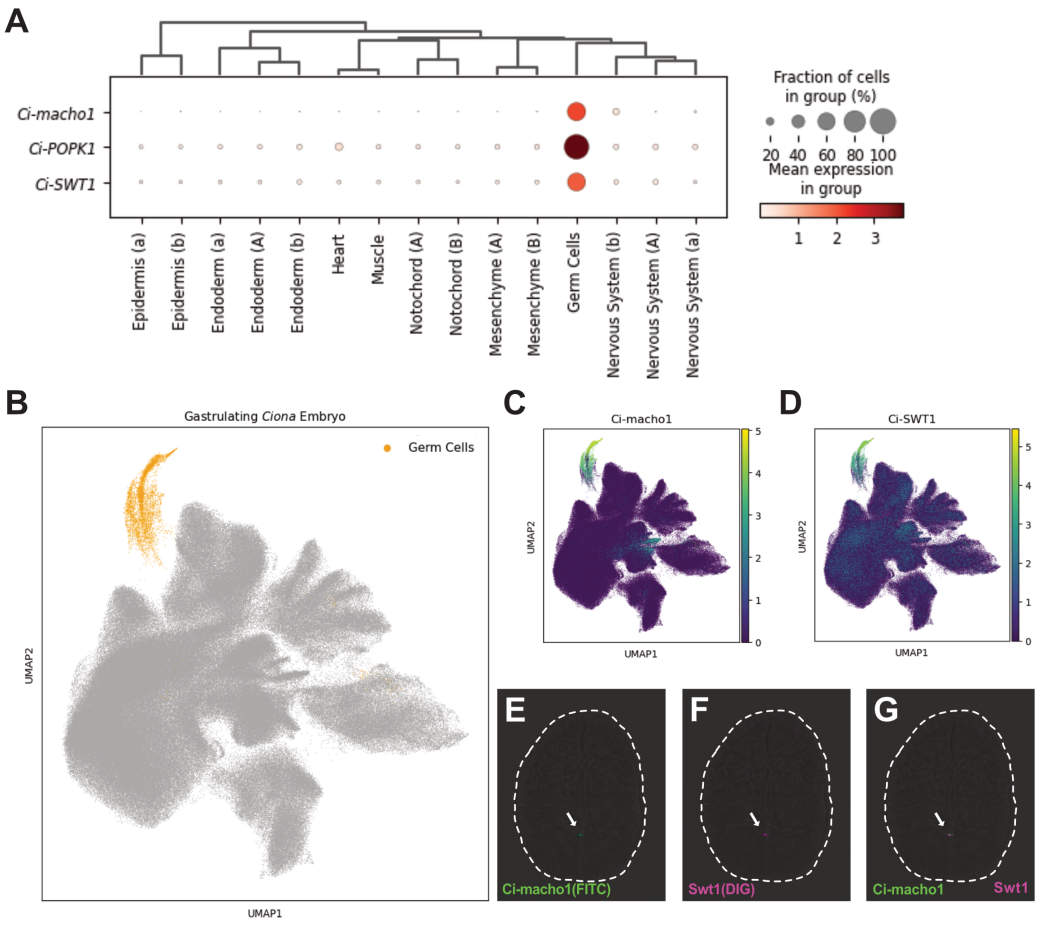

Figure 3.4. **A.** Dot plot of key germ cell markers, including *Ci-macho1* and *Ci-POPK1*, in comparison to a putative *Ciona* germ cell marker, *Ci-SWT1*, as named via sequence homology. **B.** UMAP plot of cells in the highly distinct *Ciona* germ cell cluster in orange. **C-D.** UMAP visualizations of *Ci-macho1* (C) and *Ci-SWT1* (D) in the single-cell atlas. **E-G.** *FISH* images of *Ci-macho1* (E), *Ci-SWT1* (F), and the overlay of the two (G) in 5.5 hpf *Ciona* embryos (late gastrula stage). The arrow indicates the location of the *Ciona* germ cells.

washed twice with ASW + 0.1% BSA and resuspended in 1 mL in ASW + 0.1% BSA. Cells were strained through a 50 μm cell strainer, and cell concentration was counted on a hemacytometer. Fluorescent *in situ* hybridization (FISH) assays were performed as previously described[177–180]. Embryos were counter-stained with DAPI (LifeTechnologies/Thermo Fisher Scientific, Waltham, MA). Images were taken using Leica Microsystems (Wetzlar, Germany) SP8 microscope.

3.4.2 Single-cell RNA sequencing library construction, sequencing, data preprocessing, and preliminary clustering

scRNA-seq was performed immediately after cell dissociation with the 10X Chromium 3' v2 kit (10X Genomics, Pleasanton, CA) following the manufacturer's protocol. The target number of captured cells was 10,000 for each replicate of each time point. Sequencing libraries were prepared per the manufacturer's protocol. Libraries were sequenced on the Illumina HiSeq 4000. Sequence alignment, filtering, barcode counting, and unique molecular identifier (UMI) counting were then performed using the cellranger (version 7.0.0) `count` pipeline[1] (10x Genomics, Pleasanton, CA) on each sample separately. We used the *Ciona* HT genome assembly and KY gene models[2] produced in 2019 and hosted on the Ghost database for this analysis[12]. The cellranger `count` pipeline produced an RNA count matrix for each sample included in the study—three biological replicates across each of the 4.5 hpf, 5.5 hpf, and 6.5 hpf time points of *Ciona* development. Using the Python software package scanpy (version 1.9.1)[169], all samples were combined into a single AnnData object for preprocessing. Across the three biological replicates of the 4.5 hpf, 5.5 hpf, and 6.5 hpf *Ciona* embryos, there were 147,235 cells, 121,401 cells, and 157,661 cells, respectively. Before filtering, the RNA count matrix contained 426,297 cells x 18,788 genes.

During preprocessing, doublet detection was performed using scanpy's external integration of the scrublet (version 0.2.3) tool to remove 10 cells from our RNA count matrix[181]. Next, we performed the following steps in sequence to quality filter the data: we filtered out 56,539 cells that had less than 500 counts per cell, 41 cells that had more than 10,000 counts per cell, and finally, 13,216 cells that had less than 500 genes expressed. We then filtered out 2,307 genes

[1] https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/count
[2] http://ghost.zool.kyoto-u.ac.jp/download_ht.html

detected in less than 20 cells. The resultant RNA count matrix contained 366,671 cells x 16,481 genes. Using the scanpy `normalize_total()` method, we normalized all cells to represent 10,000 reads per cell, then logarithmized the data matrix using the scanpy `log2p()` method. We then identified highly-variable genes using the dispersion-based method defined in scanpy, setting the minimum mean dispersion (`min_mean parameter`) to 0.0125, the maximum mean dispersion (`max_mean parameter`) to 3, and the minimum dispersion (`min_disp parameter`) to 0.5. This approach identified 1,561 highly-variable genes to filter the data for downstream analysis. After regressing out the total number of counts per cell with the scanpy `regress_out()` method, the data was scaled to unit variance with the scanpy `scale()` method. We denoised the data using PCA as a dimensionality reduction method, then performed batch correction with scanpy's external integration of the harmonypy (version 0.0.5) tool[182].

As a first step towards cell type clustering, we computed the neighborhood graph of cells using the PCA representation of the RNA count matrix using the scanpy `neighbors()` method with a local neighborhood size of 10 and with 10 principal components. We embedded the graph into two-dimensional space using the uniform manifold approximation and projection (UMAP) dimension reduction technique for general non-linear dimensional reduction with the scanpy `umap()` method. We performed UMAP as it is suggested by scanpy to be more faithful to the global connectivity of the manifold and, thus, better at preserving cellular trajectories. Finally, we directly clustered the neighborhood graph of cells in our data using the Leiden graph-clustering method implemented in the scanpy `leiden()` method. In total, 36 clusters were found within our data.

3.4.3 Cell type cluster identification in the *Ciona intestinalis* gastrulation atlas

After performing Leiden clustering on our single-cell *Ciona* gastrulation atlas, we annotated the clusters to correspond to tissue types present in the embryo. To expedite the clustering process, we leveraged the Aniseed API to access timepoint-specific gene location information extracted from user-submitted and published *in situ* images[170]. Currently, two gene models in circulation for *Ciona* are hosted on the Ghost database: the KH model[3] and the updated KY

[3]http://ghost.zool.kyoto-u.ac.jp/cgi-bin/gb2/gbrowse/kh/

model[4 8;12;80]. While Aniseed uses the KH models in their API, we generated our RNA count matrix with the KY gene model. As a first step, we translated the 1,561 KY gene identifiers in our RNA count matrix to KH identifiers using a chromosomal distance-based Python script (https://github.com/katarzynampiekarz/ciona_gene_model_converter). This allowed us to integrate Aniseed's gene location information at our time points of interest with our RNA count matrix to expedite the identification of cell-type clusters during gastrulation.

For the clustering applied to UMAP coordinates of the whole dataset, we refined annotation results by first comparing the expression pattern of top marker genes and known *Ciona* regulatory genes between the Leiden clusters. Clusters with similar expression patterns to key regulatory genes and known markers were considered the same cell type. We also compared our annotation results with the *in situ* records accessed via that Aniseed API or by viewing the *in situ* images recorded in the Ghost and Aniseed databases. We carefully checked the gene expression pattern for putative newly discovered cell types in clusters with poorly annotated marker genes to ensure no ambiguous expression of known markers. We identified 15 clusters representing various lineages of the following cell types: endoderm, epidermis, germ cells, heart, mesenchyme, muscle, nervous system, and notochord.

3.5 Acknowledgements

This work would not have been possible without the help of the following individuals that were fundamental to the execution of this project: Benjamin P. Song, Hannah Finnegan, and Emma K. Farley. I would also like to thank the Farley Lab for helpful discussions during the analysis of the data included in this work, especially with regards to cell cluster identification and imaging of cell type markers. I would also like to thank the UCSD IGM Genomics Center for their assistance with sequencing. Finally, I would also like to thank Alberto Stolfi and Katarzyna Piekarz from the Georgia Institute of Technology for their generous support in providing a script to translate between the *Ciona intestinalis* KH gene identifiers and the updated KY gene identifiers. Their assistance was fundamental to the identification of cell-type clusters in this work.

[4] http://ghost.zool.kyoto-u.ac.jp/default_ht.html

3.6 Footnotes

3.6.1 Author contributions

E.K.F., B.P.S., M.F.R, designed experiments. B.P.S. and H.F. conducted experiments. M.F.R. conducted bioinformatic analyses. M.F.R. wrote the chapter. E.K.F., M.F.R., and B.P.S. were involved in editing the chapter.

3.6.2 Funding

M.F.R. was supported by NIH T32 GM008666. B.P.S. was supported by NIH T32 GM133351. E.K.F., M.F.R., B.P.S., and H.F. were supported by NIH DP2HG010013.

3.6.3 Data availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Emma K. Farley (efarley@ucsd.edu), upon request. All FASTQ files and annotated scanpy `AnnData` objects will be deposited to GEO when a publication is made available. All original code for this chapter has been deposited to GitHub (https://github.com/farleylab/Ciona-Single-Cell-Gastrulation-Study) and is publicly available. Any additional information or data required to recapitulate the study reported in this chapter is available upon request.

3.6.4 Declaration of interests

The authors declare no competing interests.

Chapter 4

Generating open educational resources for university-level bioinformatics courses

Rapid advances in next-generation sequencing (NGS) technologies have improved accessibility for experimentalists to generate genomic data at scale, but the barrier to entry to learning the computational skills necessary to analyze these datasets remains high. Despite computational courses being slowly integrated into the classical undergraduate Biology curricula, the breadth of scientific and technical knowledge needed to succeed in bioinformatics courses renders them inaccessible to individuals with incomplete foundations.

For many bioinformatics graduate programs, there can be an expectation for trainees to already have a baseline knowledge of programming and bioinformatics pipeline development. Inevitably, there is usually a proportion of admitted students that are non-computational. Not addressing this knowledge gap amongst non-computational scientists contributes to issues with student retention and morale within the program, especially for students of minoritized backgrounds. To directly address this need, I made it my mission in graduate school to develop inclusive teaching strategies in academically diverse classrooms to provide students with the skills necessary to confidently perform and understand bioinformatics analysis. Additionally, I advocated for and succeeded in making expectations of incoming bioinformatics graduate students clearer to improve the retention of trainees. As a consequence of the quarantine in response to the global SARS-CoV-2 pandemic, I taught in-person and fully online modalities.

4.1 Introduction

4.1.1 Bioinformatics as a specialized data science discipline

Massively parallel or next-generation sequencing (NGS) provides researchers with an exceedingly flexible set of molecular techniques to study various types of biological sequence data at a large scale. Currently, most sequencing is performed in research laboratories that need sustainable strategies for handling computational processing and data storage[37;38;117–119]. Ergo, it has become necessary for biological and biomedical scientists at all educational levels to have some basic computational education for successful research[183–186].

The computational field of data science is an interdisciplinary discipline that utilizes algorithms, statistics, and scientific methods to extrapolate knowledge from various data types[38;183;187;188]. Thus as an amalgamation of disciplines, bioinformatics can be considered a subset of data science as it requires the ability to integrate concepts across biology, mathematics, computer science, and statistics. Additionally, bioinformatics requires substantial subfield-specific knowledge about particular computational tools and the biological context in which data was generated to generate accurate interpretations of data[117–119;183;185;186]. Therefore, universities should consider this necessary breadth of knowledge in designing new undergraduate Biology curricula for their students to apply computational skills appropriately.

While universities have started integrating computational modules into undergraduate and graduate biology student training, these modifications have not happened consistently across programs. Additionally, integrating computational coursework into current programs does not address the learning gap for scientists that wish to learn bioinformatics later in their careers when they do not have access to a classroom[37;183;189]. These learners often seek out opportunities to take computational courses in their own time, including in the university setting where many work. Unfortunately, the rising interest in computer science has imposed course enrollment caps in introductory programming and algorithms undergraduate courses due to the unmatched supply of available classes and instructors[190–194]. These enrollment caps then severely limit the opportunities for non-undergraduate individuals to supplement their professional experience with basic computational skills in an academic setting[190;191;193–195]. An added issue is that for the

few fortunate enough to enroll in a pure computer science course and learn the computational problem-solving mindset, these courses can then be difficult to translate directly into running bioinformatics pipelines. Thus, there is a need to develop accessible university-level bioinformatics course material.

4.1.2 Placing bioinformatics in the context of discipline-based education research

Educational researchers focus on scientific investigation of topics within the field of education to improve teaching and learning practices[196;197]. While educational researchers can focus on general teaching topics, such as effective teaching methods for learners of various ages, discipline-based education research (DBER) evaluates learning and teaching in a particular discipline, such as biology or computer science[198].

While DBER looks at different disciplines separately, there are concrete similarities in their multidisciplinary nature and overall goal of improving the learning experience for students at the primary, secondary, and higher education levels within their respective fields. Computer science education or computing education research addresses learning and teaching in computer science[199–204]. For this DBER field, the Association for Computing Machinery runs a special interest group (SIG) on computer science education (CSE) research known as SIGCSE [1], whose affiliated conferences are some of the top venues for educational scholars to discuss topics related to computing and teaching methods. Computing education research covers an array of questions, including studying the retention of students, the difficulties of novice programmers, and the effectiveness of learning tools employed in the classroom[199–204]. Similarly, biology education research concerns the promotion and accessibility of biology education within the classroom and teaching laboratory settings[205–211]. Many biology education research programs also evaluate the effectiveness of course-based undergraduate research experiences (CUREs) in increasing interest in science and providing a proper intervention to encourage higher representation of historically marginalized students within academia[208–211].

As a highly multidisciplinary field, bioinformatics presents a rare opportunity to understand how students learn and synthesize information spanning disparate fields and how

[1] https://www.sigcse.org/

teaching pedagogies unique to particular disciplines can be effective or ineffective. Students wishing to learn bioinformatics come from different backgrounds. Additionally, the suggested core competencies for bioinformatics also differ depending on the professional level of the individual and desired skill set for the role they are in[183;185;186;189;207;212]. Teaching methods should then differ in how they approach students with a limited programming background, students with limited molecular biology knowledge, or students with experience in both fields separately but not integrated.

4.2 Methods

4.2.1 Graduate bioinformatics training at the University of California, San Diego

The University of California, San Diego (UCSD) offers bioinformatics training at the undergraduate [2] and graduate degree [3] levels and the professional certification level at the university's extension learning center [4]. Within this chapter, I will focus on the introductory bioinformatics training that I provided to masters, doctoral, and professional students across the courses and modalities provided in Table 4.1.

[2]Students are able to get undergraduate degrees in bioinformatics from one of three departments: the Department of Bioengineering, the Department of Biology, or the Department of Computer Science and Engineering. The course requirements vary slightly depending on the department.

[3]`https://bioinformatics.ucsd.edu/`

[4]`https://extendedstudies.ucsd.edu/courses-and-programs/applied-bioinformatics`

Table 4.1. Bioinformatics courses taught at the University of California, San Diego

QUARTER	DEPARTMENT	COURSE NAME	MODALITY	WEBSITE
Spring Quarter 2019 (Apr-Jun)	Scripps Institute of Oceanography (SIO)	SIOB 242C: Marine Biotechnology III, Introduction to Bioinformatics	In-Person	*N/A*
Winter Quarter 2020 (Jan-Mar)	School of Medicine, Department of Cellular and Molecular Medicine	CMM 262/BIOM 262: Quantitative Methods in Genetics	In-Person	`cmm262-2020` GitHub Repository
September 2020 (Before Fall Quarter 2020)	Jacobs School of Engineering	Bioinformatics & Systems Biology Program Bootcamp	Online	`BISB-Bootcamp-2020` GitHub Repository
Winter Quarter 2021 (Jan-Mar)	School of Medicine, Department of Cellular and Molecular Medicine	CMM 262/BIOM 262: Quantitative Methods in Genetics	Online	`cmm262-2021` GitHub Repository
September Quarter 2021 (Before Fall Quarter 2021)	Jacobs School of Engineering	Bioinformatics & Systems Biology Program Bootcamp	Online	`BISB-Bootcamp-2021` GitHub Repository

SIOB 242C: Marine Biotechnology III, Introduction to Bioinformatics

Conceptualized and taught by Theresa (Terry) Gaasterland, Ph.D. from the Scripps Institute of Oceanography (SIO), SIOB 242C is designed to give students an introduction to using high-performance computing systems to analyze real, primary RNA-sequencing data using command-line tools. In this class, there is a lecture once a week involving file manipulation and genomic data regular expressions in Unix, along with an accompanying take-home homework assignment. For this course, I acted as the only teaching assistant and hosted a weekly problem-solving session and office hours on an as-needed basis. Due to the small size of the graduate program at SIO, there were only ten students formally enrolled in the class. Additionally, the majority of students enrolled in the course had a background in marine biology without much computational experience.

CMM 262/BIOM 262: Quantitative Methods in Genetics

CMM 262 (also cross-listed as BIOM 262) is a required course for the UCSD Genetics Training Program and is designed to teach experimental and analytical approaches in modern genetics and genomics in several topic areas. I taught CMM 262 in Winter Quarter 2020 and Winter Quarter 2021 alongside Alon Goren, Ph.D. from the UCSD School of Medicine, and three other graduate students from the BISB Program. In this class, a guest instructor specializing in a particular subtopic of genetics presents two lectures to a class of approximately fifty biomedical sciences students. The teaching assistants for CMM 262 were responsible for coordinating guest faulty speakers, managing the distribution of course materials, grading course assignments and exams, and holding office hours for students. In the 2021 iteration of the class, I served as one of the lead teaching assistants. Additionally, due to the SARS-CoV-2 pandemic, this course was taught in-person and hybrid for 2020 and entirely online for 2021. Across both years, the majority of students enrolled in CMM 262 had a background in biomedical sciences without much exposure to computer programming.

Bioinformatics & Systems Biology Program Bootcamp

Held every year during the week before the start of the academic year, the BISB Bootcamp is a student-run training course for incoming students to the BISB Doctoral Program. Through the BISB Bootcamp, incoming students are exposed to faculty research within the program and given a primer on topics in molecular biology, genetics, statistics, machine learning, computer science, and professional development meant to prepare them for their time in graduate school. As one of the course instructors, I was responsible for disseminating course materials to students before they arrived at UCSD, designing academic instructional modules, and logistical planning of the course. Due to the SARS-CoV-2 pandemic, the BISB Bootcamp was taught entirely online for 2020 and 2021. The students admitted to the BISB program are academically diverse, thus, students had varying degrees of exposure to computer programming and molecular biology.

4.2.2 Publication of locally delivered bioinformatics course materials as open educational resources

Open education is an educational movement founded on accessibility, transparency, and collaboration. Open education aims to provide broader access to the learning and training provided through formal educational systems, such as the university environment[213–217]. To provide greater access to educational materials to individuals in various time zones worldwide, open education programs typically take advantage of online platforms to distribute content, such as open educational resources (OERs). OERs are educational resources (e.g., course materials, textbooks, multimedia applications) in the public domain that are openly available for instructors or students to retain, reuse, revise, remix, or redistribute without an accompanying need to pay royalties or licensing fees[213–220].

Most course materials I developed for the bioinformatics courses I taught locally at UCSD were distributed as OERs through the GitHub platform (Table 4.1) to support the open education paradigm. By distributing the materials through GitHub, I sought to increase the reach of the high-quality bioinformatics educational materials I created for UCSD while allowing people to revise, add, or remove course content as desired while using GitHub's version-control feature for transparency of modifications. One of the fundamental guiding principles of open

education is that everyone worldwide should have access to high-quality educational experiences and resources. By publicizing the course content for CMM262 and the BISB Bootcamp, I aimed to eliminate barriers to this goal by reducing the high monetary costs of bioinformatics training and encouraging collaboration between scholars and educators in the field.

4.3 Results

Generally, bioinformatics courses often range in the course's duration and the scope of the material covered (i.e., lecturing on single versus multiple topics). One of the most common formats includes short courses that cover a particular topic or analysis pipeline (e.g., evaluating single-cell RNA-sequencing analysis, genome-wide association studies, etc.) [5]. During graduate school, I taught a total of five comprehensive Python, R, and UNIX-based bioinformatics courses that covered multiple analysis pipelines related to transcriptomics, epigenetics, and population genetics (Table 4.1). Within this section, I will discuss my strategies as a member of the teaching team for these courses to cater to the needs of students.

4.3.1 Incorporating practical computational modules into course design

There are many free bioinformatics online tutorials in the form of blog posts, GitHub-stored Jupyter Notebooks, and RMarkdown Books. Unfortunately, biological and biomedical scientists sometimes find it difficult to directly apply these generic pipelines to their data, especially when they lack programming knowledge or the computational resources needed to run a particular analysis. With any programming language, students require baseline skills in learning how to decode runtime errors and how to resolve these errors. The added complexity of data analytics requires that students analyzing biological data understand how the parameters for the tools they use impact their overall analysis and how these parameters balance with the system their study is conducted in. Thus, it can be difficult for students lacking the programming skills or theoretical biology background to apply off-the-shelf bioinformatics tools appropriately without guidance.

[5]These are common at certain institutions and bioinformatics core facilities such as Cold Spring Harbor (https://www.cshl.edu/meetings-courses-program/), the University of California, Davis campus (https://bioinformatics.ucdavis.edu/training), the Jackson Laboratory (https://www.jax.org/education-and-learning/course-and-conferences/bioinformatics-training-program), and many others.

Showcasing practical examples was important in ensuring students understood how to apply bioinformatics pipelines appropriately and in tempering expectations for bioinformatics as a whole. For example, when surveyed about a highly-interactive lecture on genome-wide association studies (GWAS) for CMM 262 taught in Winter 2021, students had high praise for the guest instructor: one student remarked in the free response section of the survey, "*I really liked the coding exercises and doing them in real-time, it made me think through what was going on in the data...*" and another student mentioned, "*The best part of the* [lecture] *was the fact that the lines were not already filled so the class was a little bit more active...*" To foster students' feelings of being active participants in lectures, we encouraged lecturers for CMM 262 to incorporate live programming in their lectures. Additionally, to ensure that students from SIOB 242C, CMM 242, and the BISB Bootcamp could apply knowledge from the courses to data produced from their present and future research labs, we specifically showcased well-known, existing community tools. Examples include `samtools`[221], `STAR`[222], `seurat`[223], `scanpy`[169], `MACS2`[224], and others. This ensured that after finishing our class, students would have access to a wealth of community resources and online forums with potential answers to their questions or answers to particular error prompts.

4.3.2 Comparison of delivery methods for deploying bioinformatics assignments

Many academic laboratories use high-performance computing (HPC) or cloud-based systems to analyze biological and biomedical datasets that cannot easily be processed on a laptop or desktop computer[38;117–119]. One example is the UCSD Triton Shared Compute Cluster (TSCC) [6] housed at the San Diego Supercomputer Center (SDSC) [7]. TSCC is a condo cluster program that researchers can buy into through hardware purchases of computing nodes or by purchasing computing hours as account credits. Despite the commonality of using Jupyter Notebooks for data exploration and visualization, academic labs can differ in how to access HPC or cloud-based computing systems based on ease of access, monetary constraints, or firewall requirements (i.e., medical data files protected by HIPAA have particular security requirements),

[6] https://sdsc.edu/services/hpc/tscc/index.html
[7] https://sdsc.edu/

monetary constraints, and ease of access. Two common methods to access Jupyter Notebooks include command line-based and on-demand-based methods. However, both methods provide pros and cons for first-time bioinformatics learners.

As part of SIOB 242C, CMM 242 taught in Winter Quarter 2020, and the BISB Bootcamp taught in September 2020, we worked with the San Diego Supercomputer Center (SDSC) to provide training accounts with enough credits for the entire quarter. Additionally, for the ChIP-sequencing analysis module taught in CMM 262 in Winter Quarter 2021, students were encouraged to complete bioinformatics pipelines similar to how you would on TSCC. With TSCC, students could learn additional skills in navigating the UNIX command line and using job scheduling systems to submit computational tasks. Students also learned how to customize software environments (e.g., conda environments) to cater to particular analysis pipelines. However, this additional layer between the student and course assignments introduced a larger learning curve toward the beginning of the course, especially for those that lacked prior programming experience. When students in CMM 262 taught in Winter 2021 were surveyed regarding their experiences learning to analyze ChIP-sequencing data through hands-on UNIX commands, many students felt the module was presented clearly. Upon being asked, "*Did the lecturer present material clearly and understandably?*", 33.3% of students indicated Strongly Agree (9/27), 37% indicated Agree (10/27), 18.5% indicated Neither Agree nor Disagree (5/27), 11.1% indicated Disagreed (3/27), and indicated 0.0% Strongly Disagreed (0/27). But when reviewing the free response section of the survey, some students felt "*. . . it was easy to fall behind. . .*" or "*. . . the speed was too fast. . .*" whereas others felt that the pace "*. . . could have been faster.*" The dichotomy in feedback in the free response section reflected the vast differences in technical background students had and their ability to follow along in the module.

In contrast to 2020, for CMM 262 taught in Winter Quarter 2021 and the BISB Bootcamp taught in September 2021, we primarily used the JupyterHub platform[8] to easily deploy data science notebooks to students that shared markdown text of lesson material alongside code blocks using bioinformatics tools. With UCSD's JupyterHub platform, DataHub[9], students could immediately jump into a particular course exercise without worrying about package installations

[8]https://jupyter.org/hub
[9]https://datahub.ucsd.edu/

or data transfers-these were all aspects handled by the teaching staff and UCSD Educational Teaching Services. Unfortunately, upon coming out of the class, these students can face a larger barrier to pursuing bioinformatics in their research as they may have an idea of how to apply analysis platforms from their coursework but not how to set up or access the bioinformatics infrastructure they need. For example, when surveyed regarding the utility of Jupyter Notebooks in CMM 262 during Winter 2021, one student commented, "*... since the code-along is being done on Jupyter Hub, I would like some additional resources or links on how to set up my machine (PC) for coding outside of the Hub.*" Additionally, several students remarked during office hours that a Jupyter Notebook-only approach without much practical, hands-on learning was less engaging. In a final course survey, there were responses such as, "*... learning with just the* [Jupyter] *notebook feels passive...*" and "*...* [if] *we had just executed* [ChIP-sequencing commands] *in the notebook, I wouldn't have understood it as well, although I'm glad to have the notebook as a reference.*"

4.3.3 Unifying students across diverse academic backgrounds in the classroom

Designing courses for students from different backgrounds can be extremely challenging regardless of the subject taught. Comprehensive introductory bioinformatics courses are no exception: the variation in course topics and the wide array of student academic backgrounds from typically non-intermixing fields make it difficult to design a course that can unify rather than alienate students in the classroom. Students entering bioinformatics courses cover various specialties, from biological and biomedical sciences to the physical and computational sciences. One of the largest challenges in designing a comprehensive bioinformatics course is to develop in-class exercises and lectures that can unify the classroom rather than unintentionally isolate groups of students based on their knowledge gaps. Typical knowledge gaps include programming, molecular biology, lab experience, and statistics.

To cater to the diverse needs of students, my main goal was to enforce a culture of inclusivity of all academic and socioeconomic backgrounds to foster a less intimidating and safer classroom environment. Between SIOB 242C, CMM 262, and the BISB Bootcamp, there are distinct differences between the backgrounds of students. For example, students in SIOB

242C and CMM 262 have a primarily experimental biology background and often have limited programming exposure. On the other hand, students in the BISB Bootcamp have a highly diverse population of students that are often more computational without having extensive experience in genetics, genomics, and molecular biology techniques. For these two groups, the teaching style varies to accommodate their unique backgrounds.

4.3.4 Teaching students with biological backgrounds to adopt a growth mindset in learning bioinformatics

In teaching students that have limited computational experience, my initial goal is to make computer science and computing more accessible and less intimidating, especially for groups of students historically excluded from these subjects. Due to inequitable access to computer science education before college, many students can feel unprepared for or unsuitable for introductory computer science coursework[225–233]. Psychological roadblocks-such as stereotype threat and imposter syndrome-can also contribute to students' perceived potential success in computer science[226;234–242]. When I teach introductory bioinformatics courses, I address these concerns to boost students' confidence and to foster a classroom environment where these concerns can be discussed openly with other students and the teaching staff. Additionally, I explicitly state that prior programming experience is not required to succeed within introductory bioinformatics courses to eliminate preconceived notions about required background knowledge before instruction takes place.

Stereotype threat is when an individual feels at risk of confirming negative stereotypes about the group of which they are a member[234–238]. Situational factors that contribute to stereotype threat include the task's difficulty at hand, the belief that the task measures their abilities, and the relevance of the stereotype to the task. Stereotype threat is believed to be a psychological barrier to students' engagement in computer science due to its ability to contribute to diminished confidence, poor performance, and loss of interest in the field, especially for minoritized students[234;236–241]. While computer science courses tend to attract more men and more white and South Asian or East Asian students, biological science courses comparatively attract more women and more Latine and Black students[238;240;241;243]. Thus, as an instructor, I try to be welcoming and compassionate towards the women and non-binary students and students

from minoritized groups that enter the classroom. In teaching bioinformatics, I address students' concerns as they arise to ensure retention and foster interest in the field.

One of the main points of concern I observed from students in SIOB 242C and CMM 262 was their inability to learn how to program late in their academic careers. Thus, the primary strategy I employed to help these students was to encourage the adoption of a growth mindset. One theory of intelligence holds that people can be categorized into two groups based on their implicit beliefs about their ability to learn. People with a fixed mindset believe that learning ability is innate, whereas people with a growth mindset believe knowledge can be acquired through effort and studying[226;244–246]. Computer science is a difficult subject for first-time learners due to (i) the steep initial learning curve in learning a new language, (ii) the detail-oriented nature required to meet syntaxial requirements, and (iii) the constructive nature of computer science as a discipline[247–249]. It is important to address each of these difficulties during instruction to encourage the development of a growth mindset in the classroom.

Most of my teaching success was derived from live programming to solve bioinformatics problems during course instruction. Because of the steep initial learning curve involved in computer science, I feel that concepts should be introduced slowly and explained explicitly. Within SIOB 242C and CMM 262, I incorporated live programming in my teaching to naturally explain new computer science concepts (e.g., variable declaration, for and while loops, conditional expressions) as they pertained to solving a bioinformatics problem in real-time. In live programming, I aimed to demystify the black box that bioinformatics can often feel like and provide a practical example of how computational concepts can be easily applied to students' work outside the classroom to encourage engagement with the material. For example, when teaching a learning module on basic statistics for CMM 262 during Winter 2021, a common point of feedback was that students "*. . . found the practical examples in notebooks extremely helpful.*" Several students also felt motivated to program on their own, and one comment indicated that "*. . . as someone who is brand new to programming it might be nice if there were a few brief practice problems we could try out on our own and see posted answer keys later. . .*" Computer programming requires that people be meticulous about noticing syntaxial errors in particular programming languages[247;250]. Through live programming, I was also able to touch on the importance of

being detail-oriented when it comes to computer programming. During live programming demos, students often pointed out errors in my programming and suggested modifications to my code to make things run successfully. Finally, we could trace through errors in programming logic together as a class, enforcing a unified community spirit in the classroom.

4.3.5 Reducing information overload in teaching bioinformatics to computational students

Introductory biology courses typically cover a multitude of topics, and it is well known that students at the secondary and undergraduate levels face difficulties in learning biological concepts[205;251–253]. For instance, biology classes have overloaded curricula and cover abstract topics[205;251;253]. These two factors combined often lead to preconceived notions of rote memorization being the defining feature of biology as a whole[205;254]. Within the BISB Bootcamp, there was a larger proportion of computational students compared to SIOB 242C and CMM 262. These students had engineering, physical, or computer science backgrounds but no extensive experience with molecular biology or genetics. My primary goal in teaching these students was to teach core biology concepts that present themselves in commonly-discussed bioinformatics problems to reduce cognitive overload.

Cognitive overload or information overload occurs when you are exposed to more details than you can process at any given time. Additionally, cognitive overload can manifest as mental fatigue, reduced attention span, and behavioral changes[255–258]. Similarly to teaching experimental biology students, I employed one strategy to reduce cognitive overload: slowly introducing biology terminology and concepts as they become relevant to the bioinformatics problem. This teaching method can also be seen through a widely-taken bioinformatics Coursera course series developed by Pavel Pevzner and Philip Compeau, as they introduce questions in biology that can use computational methods for answer generation[10].

Within the BISB Bootcamp, this teaching method was used when students were presented with the problem of looking for transcription factor binding sites within a sequence. In this example, students were introduced to several ideas in the following order:

[10] https://stepik.org/course/55789/

1. **Providing Background Information on the Biological Problem**

   (a) Consider a sentence as a string of words made up of individual letters or characters.

   (b) Also, consider that genetic information is contained within all the cells of the body as DNA.

      i. DNA consists of the nucleic acids adenine (`A`), thymine (`T`), cytosine (`C`), and guanine (`G`).

   (c) DNA can be represented as a string of different characters representing the nucleic acids.

   (d) There are regions of DNA that transcription factor proteins can bind to through the recognition of short DNA sequences (e.g., `GATA`) to activate particular genes.

      i. These regions are otherwise known as transcription factor binding sites.

2. **Defining the Biological Problem and the Associated Computational Problem**

   (a) If we consider DNA as a sentence, these binding regions can be the words in our sentence that we're trying to understand the meaning of (e.g., how come certain transcription factors activate certain genes, and are there any patterns?).

   (b) To further understand gene activation, we need to be able to recognize where transcription factor binding sites are within a DNA sequence!

   (c) Looking for transcription factor binding sites within a DNA sequence can be considered a computational "search" problem of looking for a substring within a string!

3. **Developing a Bioinformatics Solution**

   (a) We can define variables that represent the DNA sequence and the binding site.

   (b) Next, loop through each position in the DNA sequence to see if the transcription factor binding site matches the start of the sequence.

      i. If the site is found, we've successfully identified the location of a binding site!

         A. We can save the position with a match and then continue to the next position in the sequence to look for more binding site matches.

ii. If we have looked at all positions in the entire DNA sequence and haven't found a binding site match, then the site does not exist.

. . . . . .

When prompted with an optional survey to score this module from a range of one for "*Uninformative*" to five for "*Transcendent*," 37.5% of students (6/16) rated the module with a score of five, 6.3% of students (1/16) rated the module with a score of four, 37.5% of students (6/16) rated the module with a score of three, and 18.8% of students (3/16) left the field blank. In the free response section to provide feedback for this module, students were relatively satisfied with the teaching format, commenting, "[the interactive module] *was definitely needed for the rest of the week,*" "*I think it was useful and I got more out of the session I attended than I would have from the other* [lecture without programming in biology]," and "*It was very useful.*" Additionally, one student with a larger background in computer science commented, "*. . . I was able to do the CS project/presentation* [with] *no problem* [while interacting] *with the biology live.*"

While this particular example relies on prior molecular biology knowledge of DNA and proteins, we reduced the amount of background information required to understand the overall goal of this example and the applicable bioinformatics problem. In particular, we abstracted the concept of gene activation for the audience by not mentioning other parts of the system, such as transcriptional cofactors, enhancers, promoters, or genome methylation. Students were then able to easily recognize the value of computational methods when integrated with molecular biology, and our teaching methodology helped spark interest in students' interest in theoretical molecular biology across various topics. For example, one student provided the following comment, "*As someone with no biology background, it did go a bit over my head. However, I still found it useful to hear about different techniques even if I didn't fully understand them . . . I had a great discussion with* [the Bootcamp instructors] *at the end of the lecture about possibly working in a wet lab.*"

4.3.6 Using interactive teaching pedagogies to encourage student participation

The oldest teaching pedagogy is known to be "teacher-centric," where the instructor lectures students who enter the classroom as a *tabula rasa*, expected to passively receive the knowledge being disseminated. Under this paradigm, the instructor is the core regulator of knowledge in the classroom: they do most of the talking, set the rules and learning goals, and drive the direction of follow-up discussions[259]. Recently, classrooms-especially juvenile classrooms–have started adopting a "student-centric" pedagogy. In this environment, students control the direction of learning through collaborative discussions with their peers after being given the required conditions and tools by the instructor[260–262]. While teacher-centric and student-centric methods fall at opposite ends of the spectrum, instructors use varying proportions of each methodology, known as "interactive teaching"[263;264].

Bioinformatics is based on technological advancements in biology and, thus, relies heavily on access to a computer, especially for data analytics. For bioinformatics courses focused on data analysis rather than algorithmic design, we can easily incorporate interactive teaching into course lectures. Course lectures were modified in real-time based on student feedback in SIOB 242C, CMM 262, and the BISB Bootcamp. Each concept was taught as a "block" consisting of four components: (i) a molecular biology concept (e.g., genome sequences), (ii) an open question concerning the concept presented (e.g., comparing genomes), (iii) a parallel computer science concept (e.g., string comparisons), and (iv) an example computational solution to the question (e.g., genome/string alignment with dynamic programming). By being upfront with the interdisciplinary nature of bioinformatics problems, students of all backgrounds were engaged in asking questions during course instruction and providing solutions to questions provided during live programming demonstrations.

4.3.7 The impact of COVID-19 on teaching university-level bioinformatics courses in 2020 and 2021

In recent years, universities have adopted online educational tools into regular instruction to provide greater accessibility to course materials, external resources, and grading information

online. For example, many universities use the CANVAS [11] web-based learning management system (LMS)[265–267]. Many engineering and computational courses use standardized platforms for student communication and course assessments, such as Piazza [12] and GradeScope [13]. In addition to LMS platforms, some universities have started exploring "flipped classroom" formats in which students encounter lecture material independently before dedicating all in-person instructional time to discussion-like sessions[268–270]. However, towards the end of 2020, this gradual process of virtualizing traditional in-person courses was greatly accelerated by the high aerosol transmissibility of the SARS-CoV-2 virus[271–273].

The emergency of the Coronavirus Disease 2019 (COVID-19) crisis forced instructors worldwide to translate their in-person courses into virtual environments, introducing difficulties in promoting interaction between students and instructors, especially in a medium unfamiliar to many. Despite bioinformatics' reliance on virtual resources and computers as a field, many still faced challenges translating successful in-person courses to online-only mediums. Due to COVID-19, I was forced to adapt the coursework for CMM 262 and the BISB Bootcamp on the fly (Table 4.1). In addition to the commonly observed issue of student engagement, one of the largest challenges in CMM 262 and the BISB Bootcamp was losing important in-person interactions in teaching and learning programming for the first time. For example, assisting students in live programming or in-class pair-programming sessions was more difficult when they ran into individual errors with the coding module. It was also challenging to facilitate small group discussions. While it is easy to walk up to students to help them with technical difficulties during in-person instruction, we were forced to take advantage of Zoom's "breakout room" feature to assist these students. One student from CMM 262 taught in Winter 2021 commented on the interactive ChIP-sequencing analysis pipeline module: "*This module would be one that would benefit from in-person instruction, because it was easy to fall behind during the coding segment. I didn't want to interrupt the class to slow down and I would have been more comfortable asking a TA or a neighbor in a physical classroom.*" Ultimately, it was difficult to assist students with conceptual or technical difficulties during lecture time. Often, these students

[11] https://www.instructure.com/canvas
[12] https://piazza.com/
[13] https://www.gradescope.com/

would approach the teaching team during office hours to resolve any issues.

Fortunately, there were some benefits to moving the course entirely online. Students, for example, could review recorded content during their own time. Two students from the Winter 2021 iteration of CMM 262 commented, "*I think the recorded Zoom lectures help though, since I definitely needed to rewatch some parts*" and "*. . . since the lectures are recorded, I am able to go back and go through it at my own pace, which is really helpful and appreciated!*" Another student commented on the same course, "*For an online format, the course worked well when it came to being able to access the lecture recordings with captions since it can be hard to sit through an online lecture without them. I felt that the course was not adapted for longer lectures since I was experiencing Zoom fatigue and could not hold my attention for more than an hour (maybe note-taking-friendly formats or shorter, more frequent lectures may help).*" Properly deploying synchronous, practical bioinformatics classes requires instructors to consider how online mediums such as Zoom will impact students' learning experience. While we encountered logistical difficulties in interacting with students through Zoom breakout rooms or combating Zoom fatigue, the transition to online education, accelerated by the SARS-CoV-2 virus, underscored the possibility of making bioinformatics education more accessible to a broader audience.

4.4 Conclusion

We are currently in a transition period in how we approach undergraduate Biology education from one that takes a surface-level approach in introducing bioinformatics analyses in one-off modules to one that integrates traditional computational courses into the canonical curriculum. While these changes will ultimately benefit the next generation of scientists in analyzing the large-scale biological datasets of the future, there is a need to address the knowledge gap for graduate students and other professional scientists of the present. While it is important to consider incorporating practical course modules into bioinformatics and balance the amount of material to include within bioinformatics classes, one of the largest considerations is the background of the students being taught.

Students wishing to learn bioinformatics later in their careers often come from various specialties spanning biological and biomedical sciences to the physical and computational sciences.

Thus, one of the largest challenges in designing a comprehensive bioinformatics course is balancing these diverse backgrounds with designing course material that does not isolate students based on their knowledge gaps in theoretical biology, programming, and statistics. To teach bioinformatics to an academically diverse classroom, incorporating course materials that incorporate aspects of everybody's background help create a common ground for people to grow. Within SIOB 242C, CMM 262, and the BISB Bootcamp, showcasing computer science concepts of data types and looping in the context of analyzing genomic sequences proved successful in teaching biological and biomedical sciences students while cementing core instructional concepts and reducing the psychological barrier of stereotype threat. Slowly introducing theoretical biology in the context of interesting computational problems also successfully taught computational students without inflicting information overload. Balancing course content with students' learning abilities makes it possible to unify the classroom without leaving people behind. Additionally, introducing practical bioinformatics examples through student-paced live programming helps make bioinformatics accessible to new audiences and encourages an inclusive environment for all academic backgrounds.

4.5 Acknowledgments

I would like to thank Niema Moshiri, Clarence Mah, and Emma Farley for their thoughts and helpful feedback in writing this chapter. Additionally, in teaching introductory bioinformatics courses at UCSD, I learned from the students and from the other instructors I worked with in developing the course materials, lectures, and assignments. SIOB 242C, CMM 262, and the BISB Bootcamp courses would not have been successful without their dedicated support.

Firstly, I am grateful to Alon Goren, Daniela "Dana" Nachmanson, Clarence Mah, Eric Kofman, Pratibha Jagannatha, and all of our guest instructors for being wonderful and flexible members of the teaching team for CMM 262 taught during the Winter Quarters of 2020 and 2021. Through teaching this class, I fine-tuned my knowledge of diverse bioinformatics pipelines and met many of the wonderful students in the Biomedical Sciences (BMS) graduate program.

Next, I would like to thank Owen Chapman, Cameron Martino, Mike Cuoco, and Lauryn Bruce for their help in co-teaching the BISB-Biomedical Sciences (BMS) Joint Program

Bootcamp in September 2020 and the BISB Bootcamp in September 2021. Planning a student-run Bootcamp in the limbo of the early COVID-19 pandemic on top of doing my thesis research was especially stressful, and I'm thankful to have had Owen and Cameron by my side to navigate the uncertainties of whether or not we would be able to teach a 50-person, primarily experimental class on how to code on the command line in a week. I would also like to thank the numerous guest student instructors from both Bootcamp sessions who took the time outside of research obligations to teach their peers various skills needed to survive both the personal and professional aspects of graduate school. This included Alexander Wenzel, Gibraan Rahman, Clarence Mah, Adam Officer, and George Armstrong from the BISB program, as well as Alex Tankka, Sara Elmsaouri, Danielle Schafer, Maya Gosztyla, Noorsher Ahmed, and Margaret Burns from the BMS program. Without these dedicated individuals, we would have never been able to cover as much material as we did, and thus you have my thanks!

Finally, I would like to express my deepest gratitude to Terry Gaasterland for being an encouraging bioinformatics education mentor since I first took a variant of what is now SIOB 242C as an undergraduate student during Spring Quarter 2015. That class, SIO 190, sparked my interest in bioinformatics as a discipline and set the foundation for how I approach analyses today. I would also like to thank her for trusting me to assist in teaching SIOB 242C for the Spring Quarter 2017 course as a senior undergraduate student and then again during Spring Quarter 2019 as a first-year Ph.D. student. Without those experiences, I would not have been able to develop my passion for developing educational materials for students or experiment with incorporating new teaching methods into the classroom. It was also Terry's idea that I include this chapter within my thesis; for that, I will always be forever thankful that these reflections about teaching were not lost to time.

Epilogue

5.1 Conclusion

I started this work with a quote from Lewis Wolpert, "It is not birth, marriage, or death, but gastrulation that is the most important time of our lives." Indeed, gastrulation is a critical step in embryonic development. Gastrulation is when the primordial germ layers are specified, the embryonic axes manifest, and the embryo alters its morphology for the first time. Within chordates, the formation of the primary germinal layers-the ectoderm, mesoderm, and endoderm-requires meticulous control of individual cell and collective tissue behaviors with regard to space and time[5;13;143;144]. Thus, it is crucial to study the contents of a cell and the active regulatory factors at this stage to understand how defects in this machinery lead to congenital disabilities and disease.

One structure that emerges during gastrulation is the notochord, a rod-like, cartilaginous skeleton of mesodermal origin that defines chordates. The notochord serves as a signaling center for the embryonic midline and becomes an integral part of the vertebrate backbone as the nucleus pulposus of intervertebral discs[2;3;5–7;13;143;144;147;161]. Understanding notochord structure and function during gastrulation is essential to elucidate how perturbations to this machinery may lead to congenital vertebral defects. In this thesis dissertation, I demonstrated the importance of understanding the regulatory mechanisms driving gastrulation by studying the activity of enhancers and the contents of a cell during notogenesis. The ability to perform high-throughput experiments in the marine chordate *Ciona intestinalis type A* or *Ciona robusta* (*Ciona*) contributed to the ability to screen the activity of thousands of enhancers to understand the contributions of transcription factor binding sites to function (Chapter 1, Chapter 2). Likewise, *Ciona* also granted us the unique ability to profile the transcriptomes of thousands of single cells

in whole, gastrulating embryos to understand the contents of a cell across major tissues during cell type specification (Chapter 3).

In Chapter 1, we sought to understand the regulatory logic of notochord enhancers by taking advantage of the ability to perform high-throughput, massively-parallel reporter assays (MPRA) within *Ciona*. Within the *Ciona* genome, we identified 1,092 genomic regions, dubbed the ZEE library, containing a Zic binding site within 30 bp of an ETS binding site[25]. Of the 90 ZEE elements, surprisingly, only nine drove notochord expression. One of the nine we identified, the *Ciona laminin alpha* enhancer, relied on grammatical constraints on Zic and ETS for functional activity. We also find similar clusters of Zic and ETS binding sites proximal to the mouse and human *laminin alpha-1* gene with syntax similar to the *Ciona laminin* enhancer, suggesting that grammatical signatures are conserved across organisms[25]. Within this chapter, we also highlight the importance of testing the sufficiency of TFBSs to investigate if we fully understand the regulatory logic of an enhancer. Through a randomization study, we reveal within the previously identified BraS enhancer that Zic and ETS binding sites are insufficient for notochord activity. Furthermore, we also find that FoxA and Bra sites are also necessary for notochord expression with BraS and that the combination of Zic, ETS, FoxA, and Bra binding sites may be a common logic regulating Bra expression[25]. Our findings in Chapter 1 illustrate a common problem in mining genomic data for patterns, especially in mining genomes for functional enhancers based on the presence of TFBSs. We demonstrate that the presence of binding sites alone does not correlate to enhancer activity. To understand how enhancers regulate gene expression, we need to understand the number and types of TFBSs within an enhancer and the dependency between these sites, such as TFBS syntax and affinity[23]. Overall, our findings illustrate the importance of enhancer grammar within developmental enhancers and hint at the conserved role of grammar and logic across chordates.

In Chapter 2, we continue upon the framework of Chapter 1 to understand the regulatory logic of notochord enhancers consisting of Zic, ETS, Bra, and FoxA binding sites. By virtue of a new *Ciona* genome reference sequence release in 2019[12], we found a total of 4,344 genomic elements containing Zic and ETS binding sites with flexible constraints of the position of these sites within a 100 bp window. This library is otherwise known as the KYN library. After testing

these KYN elements in an MPRA in whole *Ciona* embryos, we found that only 15.4% of these sites were active and dependent on Zic and ETS binding sites. Reviewing several candidate enhancers, we find they are proximal to multiple genes implicated in nervous system disorders and skeletal and bone-related disorders. Ultimately, further study of this enhancer library through imaging studies and TFBS ablation experiments is needed to ascertain the dependency of binding sites in determining functionality. To finish this chapter, we introduced a proof-of-concept Python package, **E**ntire **G**enome se**A**rches for **G**rammars of **E**nhancers (EnGAGE), that was developed to aid in efforts to understand the connection between genomic sequence and regulatory activity. This work represents the beginnings of a new paradigm to understand enhancers through elucidating how the organization of collections of TFBSs contributes to functional activity.

Finally, we move beyond genomic sequence to the contents of a cell in Chapter 3, where we develop a high-throughput, dense transcriptional atlas of *Ciona* gastrulation. Just as the specific activity of enhancers is essential for successful development, the contents of a cell also dictate the formation of key cell types, such as the epidermis, endoderm, mesenchyme, heart, muscle, germ cells, notochord, and nervous system. In this study, we develop *Ciona* embryos to the time points dictating gastrulation-the 4.5 hours post fertilization (hpf), 5.5 hpf, and 6.5 hpf stages representing the early gastrula or 110-cell stage, late gastrula, and early neurula stages of development[11]. Once developed, the embryos are rapidly disassociated and processed for single-cell RNA sequencing. In total, we were able to profile 356,671 cells, allowing us to identify major tissues undergoing organogenesis and rare cell-type populations, such as the developing heart and germ cells. We also validate our map with fluorescent *in situ* hybridization (FISH) imaging studies, visualizing canonical marker genes and novel marker genes within late gastrula *Ciona* embryos. By providing a higher resolution single-cell atlas just spanning gastrulation, we anticipate that other groups can use the map generated in this study to identify conserved canonical and novel cell differentiation markers. Additionally, this resource will provide insight into the cell fate mechanisms governing organ formation during *Ciona* gastrulation.

In the first three chapters of this work, I interrogate the regulatory players driving notogenesis from a genomic perspective through understanding the impact of binding site dependencies within enhancers (Chapter 1, Chapter 2) and from a cellular perspective through

cataloging the contents of major cell types by creating a transcriptional atlas of the developing *Ciona* gastrula (Chapter 3). I also demonstrate that we cannot rely on bioinformatic identification of putative enhancers based on TFBS presence alone. In addition to elucidating the mechanisms driving notochord enhancer activity and the transcriptional landscape driving organogenesis, I also make the argument for studying enhancer grammar. To identify developmental enhancers accurately from genomic sequences, we need to understand the number and types of TFBSs present within a sequence and the dependency between these sites regarding syntax and binding affinity. Additionally, we need to understand the cellular context and transcriptional landscape in which these sequences are active. By studying these two elements in tandem, we can further understand how we transform from a single cell to a multicellular organism.

Finally, in the last chapter of this work, I demonstrate the importance of teaching bioinformatics and the strategies for managing an academically diverse classroom. The work presented in the first three chapters of this thesis dissertation required an understanding of programming, data visualization, molecular and developmental biology, and statistics to comment on enhancer grammar and cell type specification (Chapter 1, Chapter 2, Chapter 3). As molecular biology steps into the world of big data to understand regulatory genomics, scientists must pick up bioinformatics skills. In Chapter 4, I share my experiences teaching bioinformatics curricula at the university level through the SIOB 242C, CMM 262, and BISB Bootcamp courses offered at the University of California, San Diego. One of the most considerable challenges I encountered in developing a comprehensive bioinformatics course was balancing the diverse academic backgrounds of students with creating experiences that do not isolate students based on their knowledge gaps in molecular biology and computer programming. To create a community environment in the classroom, I discuss my strategies in slowly introducing computational and biological concepts to reduce information overload and combat stereotype threat. Additionally, I discuss the benefits of introducing practical bioinformatics examples through student-paced, classroom-wide live programming sessions. Through this work, I want to enforce that learning bioinformatics can be made accessible through proper course design and empathetic instruction.

5.2 Limitations and Future Directions

Though this work represents essential steps forward in understanding the mechanisms behind enhancer grammar and cell type specification during gastrulation, there are still many open avenues of study and important limitations to keep in mind.

In Chapter 1 and Chapter 2, there are limitations regarding identifying functional enhancers and the ability to translate grammatical principles across species. For example, within Chapter 1, we screened 90 ZEE elements for functionality; however, only 10% were active in the notochord. Additionally, in Chapter 2, we screen for 4,344 KYN elements for functionality; however, only 15.4% are active and reliant on Zic and ETS. While we anticipate that finding more notochord enhancers regulated by Zic, ETS, and possibly Bra and FoxA could better inform our understanding of the notochord enhancer grammar, finding these regions is highly limited. Combining assays of genomic regions with synthetic and random enhancer screens is thus needed to gain enough data to determine grammatical rules. With regards to our findings of possible conserved enhancer logic and grammar across chordates, we did not test the mouse *laminin alpha-1* enhancer for activity in mouse for the study presented in Chapter 1. We also did not functionally interrogate the importance of the 12 bp spacing within this enhancer in the context of *Ciona* or mouse. Conducting these additional studies would deepen our understanding of the conservation of grammar across chordates. On the other hand, for the Zic, ETS, Bra, and FoxA logic found within *Brachyury* enhancers in Chapter 1, further manipulations of these TFBSs in the context of mouse and zebrafish $Brachyury/T/TBXT$ enhancers are required to determine if the conservation of logic is essential for the regulation of *Brachyury*. Similar interrogations into the importance of binding sites in the active enhancers identified in Chapter 2 are necessary to evaluate the components of the enhancer required for activity.

Chapter 3 presents a high-resolution transcriptional atlas encompassing *Ciona* gastrulation. Despite our success in identifying key cell types and sub-clusters representing their original cell-type lineages in *Ciona* (e.g., A-line, B-line, a-line, and b-line), we did not evaluate the cell lineage specification pathways or pseudotime trajectories in the formation of these cell types. There is increasing interest in understanding the transitionary states involved in cell type

specification. Within our dataset, one avenue for future study could be the initial formation of the notochord from mesenchymal tissue or the formation of initial neural subtypes from the A-line, a-line, and b-line cell lineages of the *Ciona* embryo. Uncovering the markers delineating particular states in these trajectories, especially in a high-resolution single-cell atlas, may uncover additional essential marker genes involved in initial organ formation. In addition to studying cell type specification patterns through constructing pseudotime trajectories, another avenue for future work includes annotation of genes present in the *Ciona* gastrula. Within our high-resolution single-cell atlas, we were able to identify not only canonical markers within cell types but also many novel markers lacking clear definitions on Aniseed besides sequence homology to vertebrate homologs. A straightforward avenue for future work is visualizing these novel markers through imaging experiments to validate their expression in the cell types identified in our single-cell atlas and define these genes for the larger *Ciona* community. These studies would also confirm or deny the value of sequence homology in determining the true activity of novel markers.

5.3 Closing thoughts

The work presented in this thesis dissertation has provided novel insight into the gene regulatory programs governing a critical step of development-gastrulation. During my thesis, I developed analysis pipelines to study the genomic elements of developing embryos. These studies identified novel notochord enhancers, and we subsequently found rules governing some of these enhancers. Excitingly, we found evidence of conserved grammatical signatures across chordates, providing the promise of a universal enhancer grammar code that has yet to be fully uncovered. Additionally, my work in single-cell genomics provided profiling of over 350,000 cells collected from whole embryos, supplying one of the most comprehensive maps of the genes expressed during gastrulation. These forays into the mechanisms driving genome regulation from the sequence level to the cellular level uncovered that functional studies are essential to understanding how spatiotemporal control of gene expression occurs, leading to successful development.

Furthermore, my aside into the considerations made in bioinformatics course instruction provides viable suggestions to improve inclusivity in academically diverse classrooms that are a melange of computational scientists and molecular biologists alike. In teaching university-level

bioinformatics courses in academically diverse classrooms, I cater to the needs of computational scientists and molecular biologists through practical exercises and in-class programming demos. As advances in technology increase the amount of biological data we can extract from experiments, it becomes ever so important for individuals to learn how to parse this information with bioinformatics techniques and elucidate meaningful insights into the world around us.

Appendix A

Supplemental Material for Chapter 1

A.1 Expression patterns of ZEE elements driving notochord expression

A.1.1 Levels of expression for notochord-specific enhancers

There are four notochord-specific enhancers (ZEE10, ZEE13, ZEE20, and ZEE27). 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[14]. ZEE13, ZEE20, and ZEE27 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.

A.1.2 BraS and ZEE1 drive 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[27]; 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 potentially 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 Zic, ETS, FoxA, and Bra sites contribute to the a6.5 expression.

A.1.3 ZEE35 and ZEE85 drive weak notochord expression with stronger ectopic expression

ZEE35 and ZEE85 both drive 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 unidentified 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. 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.

A.2 Supplementary Table Captions

The Supplemental Table can be found on the GitHub repository for this study labeled as `SupplementaryTable.xlsx` at the following location:
`https://github.com/farleylab/Diverse-Logics-Notochord-Study/`.

.....

Supplementary Table S1: All ZEE elements screened

This table provides information about all ZEE elements: whether they were tested individually, their enhancer activity score, their genomic location, and their sequence.

Supplementary Table S2: Scoring of ZEE elements individually tested

This table provides scoring data for all three replicates of all ZEE elements chosen to be screened individually. Embryos were scored for no expression, expression and a6.5, b6.5, notochord, mesenchyme, and endoderm expression.

Supplementary Table S3: Sequences of notochord elements

This table provides the genomic location and sequence of notochord expressing ZEE elements.

Supplementary Table S4: Scoring of manipulations on Lama and BraS enhancers

This table provides scoring data for the manipulations of the Lama and BraS enhancers. Embryos were scored expression, no expression, notochord and a6.5 lineage expression.

Supplementary Table S5: Oligonucleotides for Lama and BraS manipulations

This table provides sequences for oligonucleotides used to mutagenize the Lama and BraS enhancers.

Supplementary Table S6: Vertebrate enhancers referenced in this study

This table provides genomic locations of vertebrate enhancers referenced in this study.

A.3 Supplementary Figures

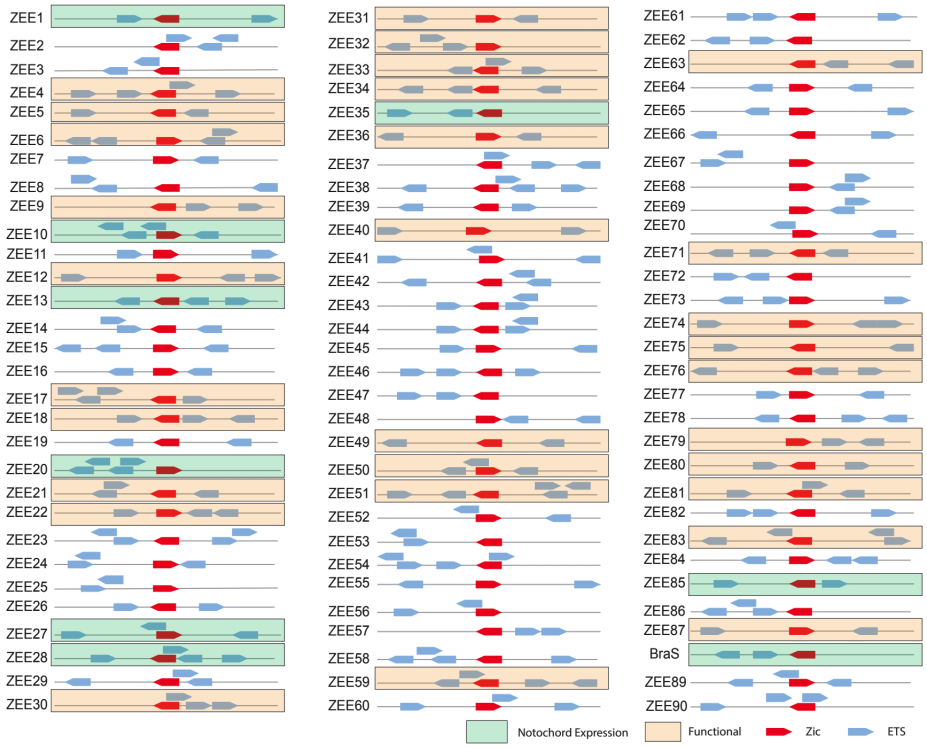

Figure A.1. ZEE elements screened. Schematic of each ZEE element tested within our MPRA assay. Zic sites are colored red and ETS sites are colored blue. ZEE elements that were functional are boxed in orange. ZEE elements that drove notochord expression are boxed in green.

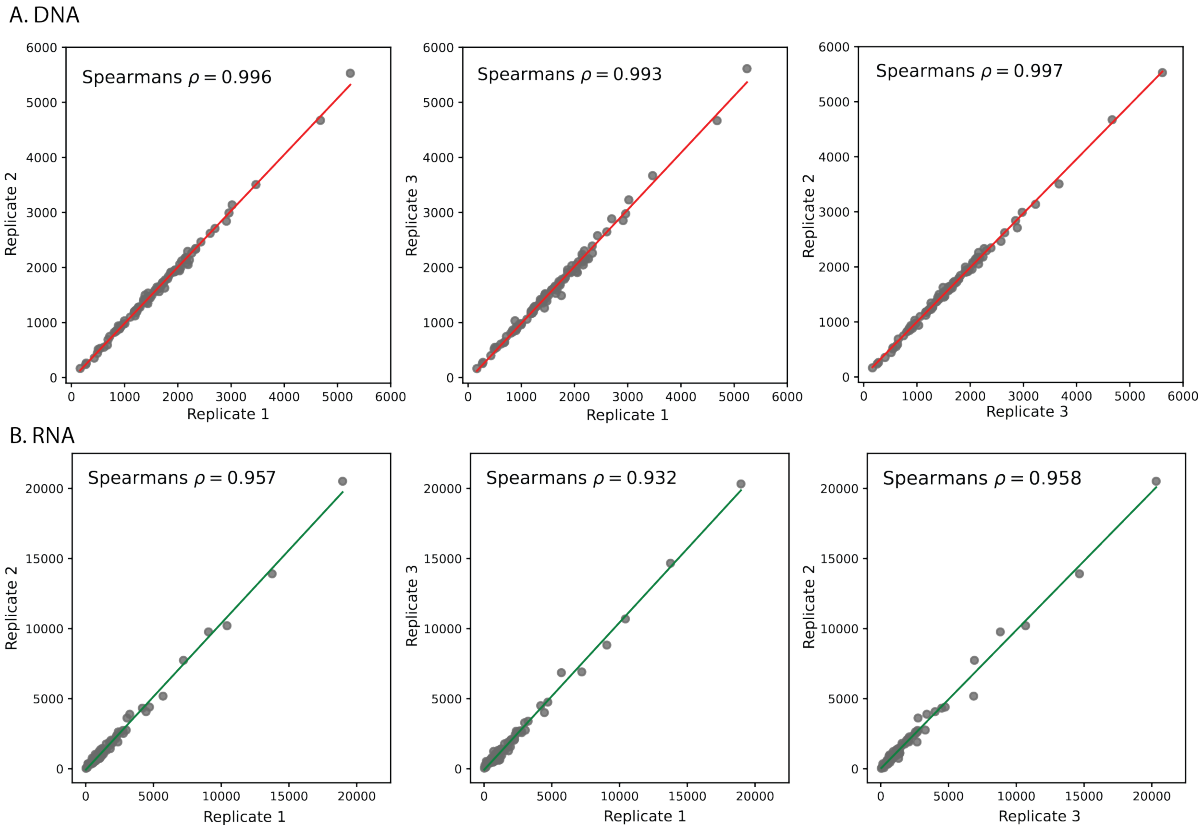

Figure A.2. Data quality metrics illustrate high robustness of ZEE genomic screen. A. Correlation of DNA plasmids detected between replicates was plotted. All Spearman correlations between replicates were >0.99. **B.** Correlation of mRNA barcodes detected between replicates was plotted. All Spearman correlations between replicates were >0.9.

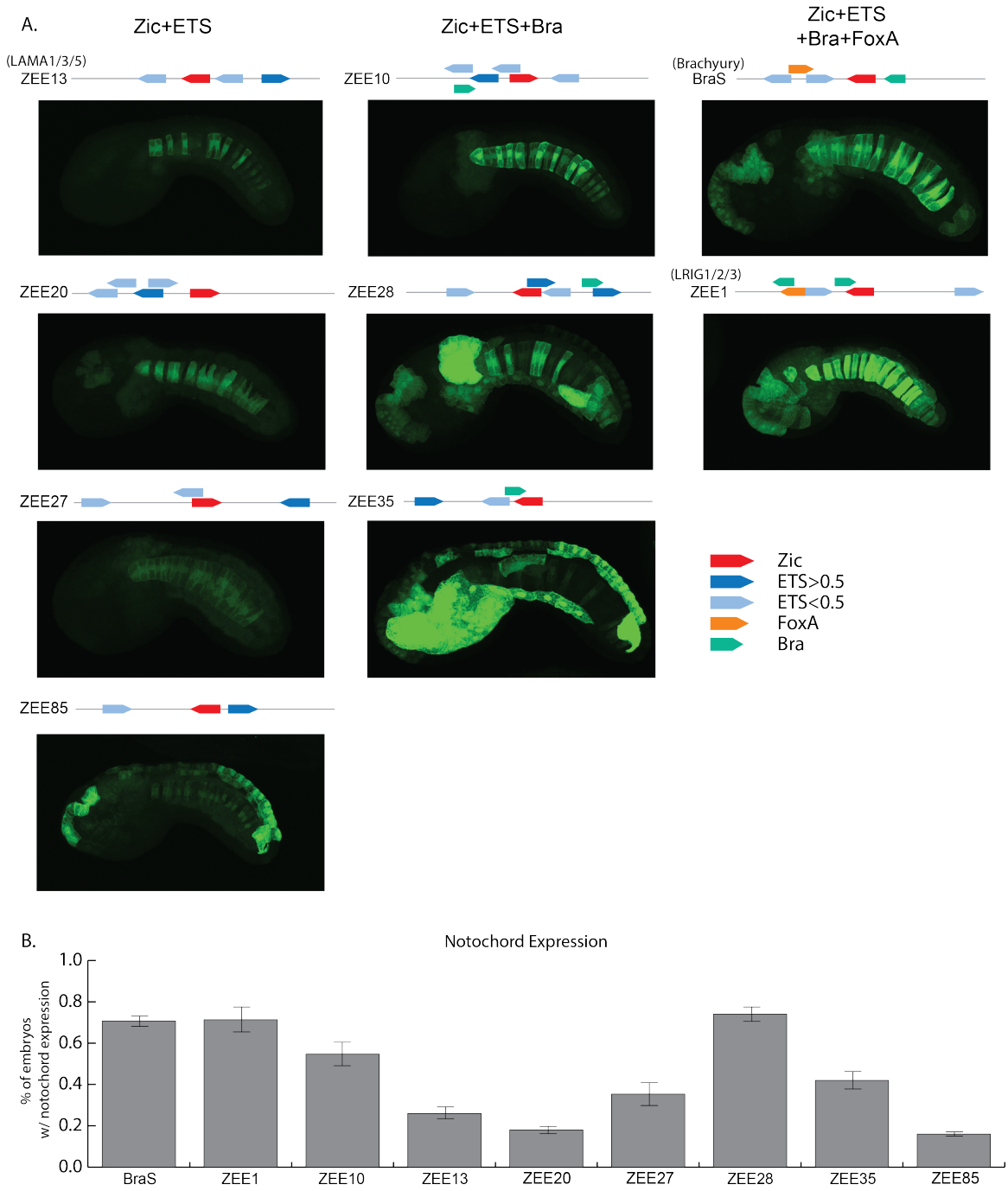

Figure A.3. Nine ZEE elements drive notochord expression. A. Images and schematics of the nine notochord enhancers in the ZEE library. Zic (red), ETS (blue), FoxA (orange), and Bra sites (green) are annotated. Dark blue ETS sites have an affinity of greater than 0.5, light blue sites have an affinity of less than 0.5. **B.** Counting data for nine ZEE elements showing the percentage of embryos with notochord expression. Three biological replicates were performed with 50 embryos per replicate analyzed.

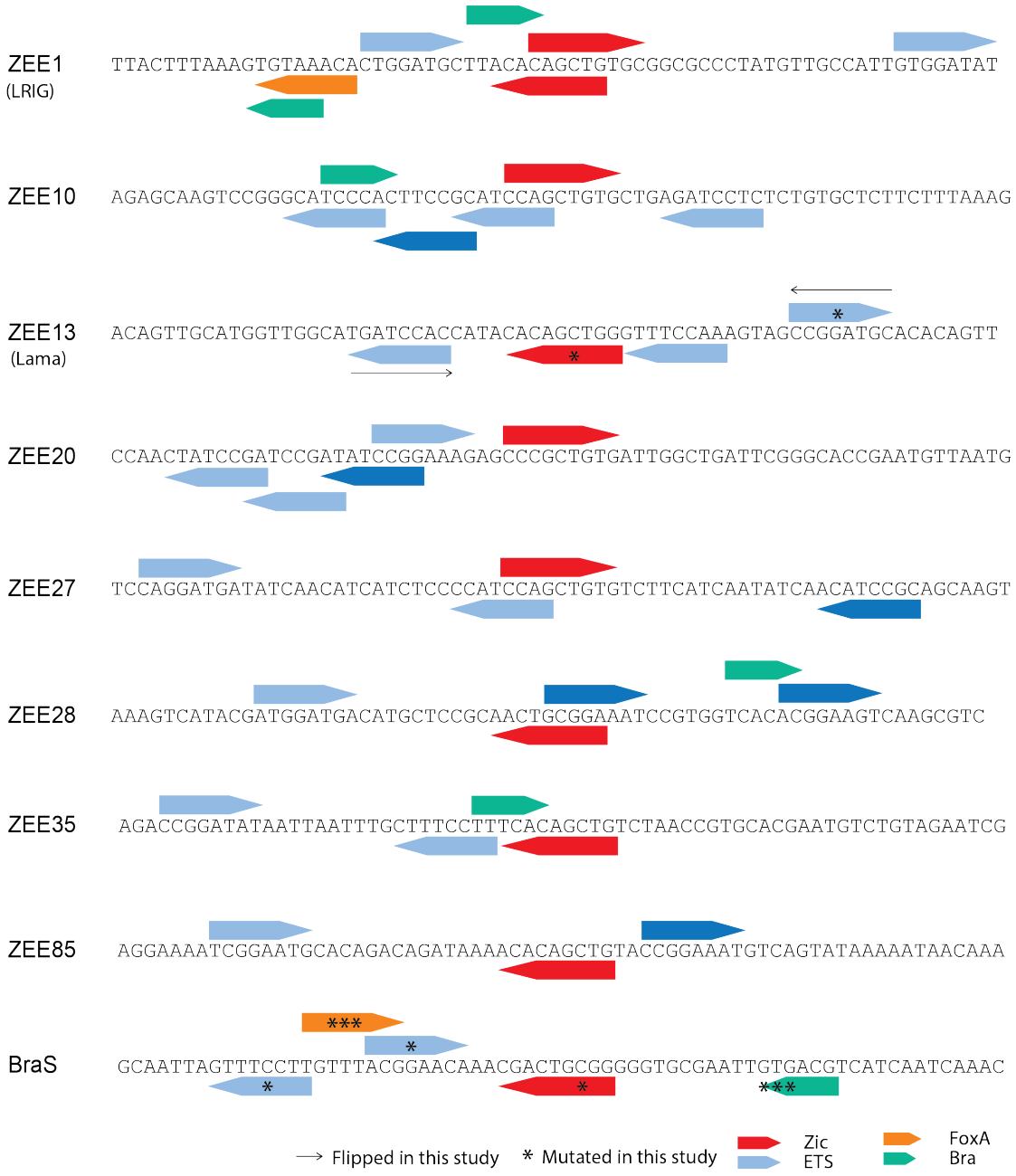

Figure A.4. Annotated sequences of the nine ZEE elements that drive notochord expression. Zic (red), ETS (blue), FoxA (orange), and Bra sites (green) are annotated. Asterisk denotes nucleotide that was mutated in this study, arrow denotes a binding site that was flipped. Dark blue ETS sites have an affinity of greater than 0.5, light blue sites have an affinity of less than 0.5.

Figure A.5. 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, * represents $P<0.05$, ** represents $P<0.01$. 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 Panel 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. (*Continued on next page.*)

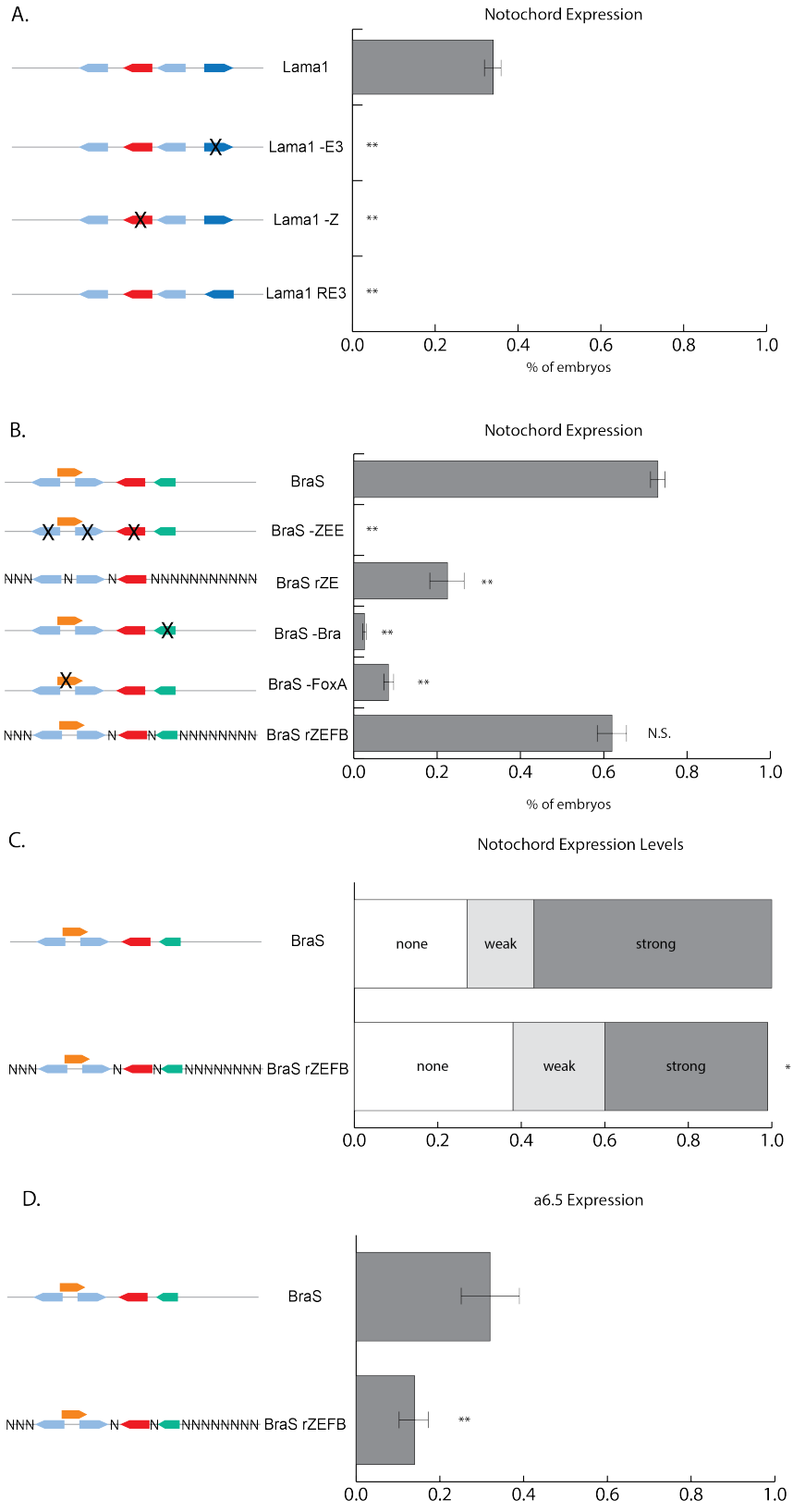

Figure A.5. (*Continued from previous page.*)

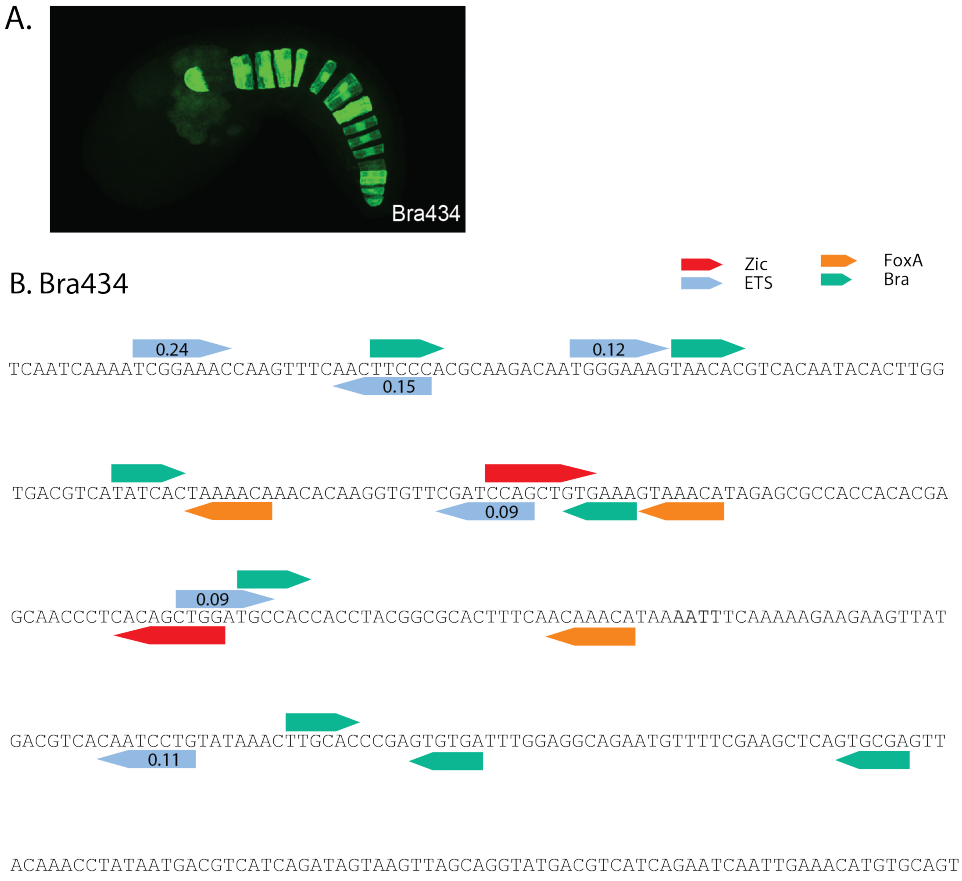

Figure A.6. Updated annotation of Bra434. A. Image of Bra434 electroporated into *Ciona* embryo. **B.** Annotation of the Bra434 using PBM, EMSA, and crystal structure data . Zic sites in red, ETS sites in light blue, FoxA sites in orange, and Bra sites in green. Affinities of ETS calculated from PBM data (Wei et al., 2010) are labeled.

Appendix B

Supplemental Material for Chapter 2

B.1 Supplementary Figures

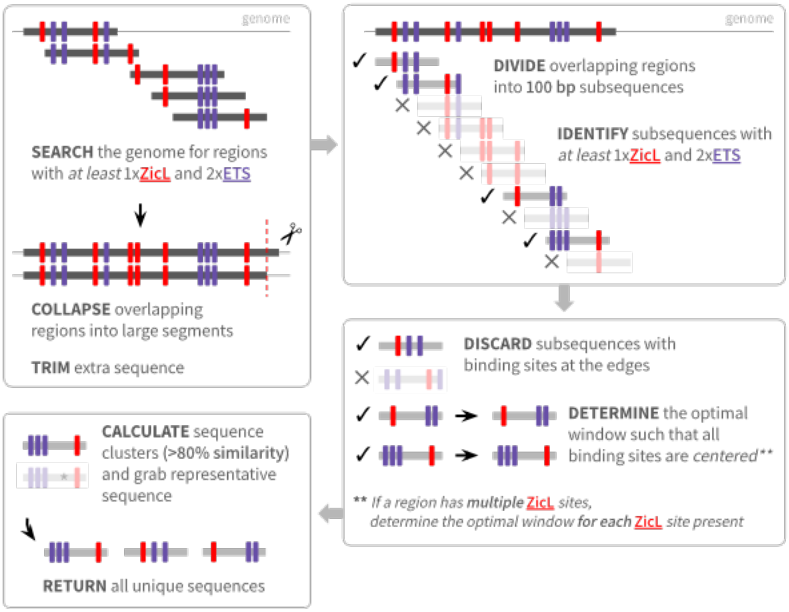

Figure B.1. KYN library search methodology. Schematic of the search methodology for identifying elements for the KYN library from the *Ciona intestinalis type A* genome. First, the genome is scanned on a chromosome by chromosome basis for large genomic blocks containing at least one Zic site and two ETS binding sites. These genomic blocks are then collapsed upon each other based on overlapping coordinates and trimmed. Then, 100 bp subsequences are extracted from the large genomic blocks and screened for if they have minimum at least one Zic site and two ETS binding sites. The subsequences are discarded if they have binding sites on the edge of the sequence, and all subsequences are modified such that all binding sites are centered within the 100 bp window. Finally, the sequence similarity is calculated between all 100 bp windows and all windows with less than 80% similarity to each other as calculated by the hamming distance are included within the KYN library.

Figure B.2. Correlation between plasmid DNA of the KYN library. All Spearman correlations between replicates were >0.70.

Appendix C

Supplemental Material for Chapter 3

C.1 Supplementary Figures

Figure C.1. A. Distribution of number of gene counts per cell across all time points (4.5 hours post fertilization (hpf), 5.5 hpf, and 6.5 hpf) and all biological replicates. **B.** Scatterplot of the number of gene counts per cell versus the number of genes per cell across the entire gastrulation atlas. **C-D.** Histograms of the number of gene counts per cell (C) and the number of genes per cell (D) across the entire gastrulation atlas. **E-F.** UMAP visualizations of the single-cell gastrulation atlas colored by time point (E) and by Leiden clustering (F). (*Continued on next page.*)

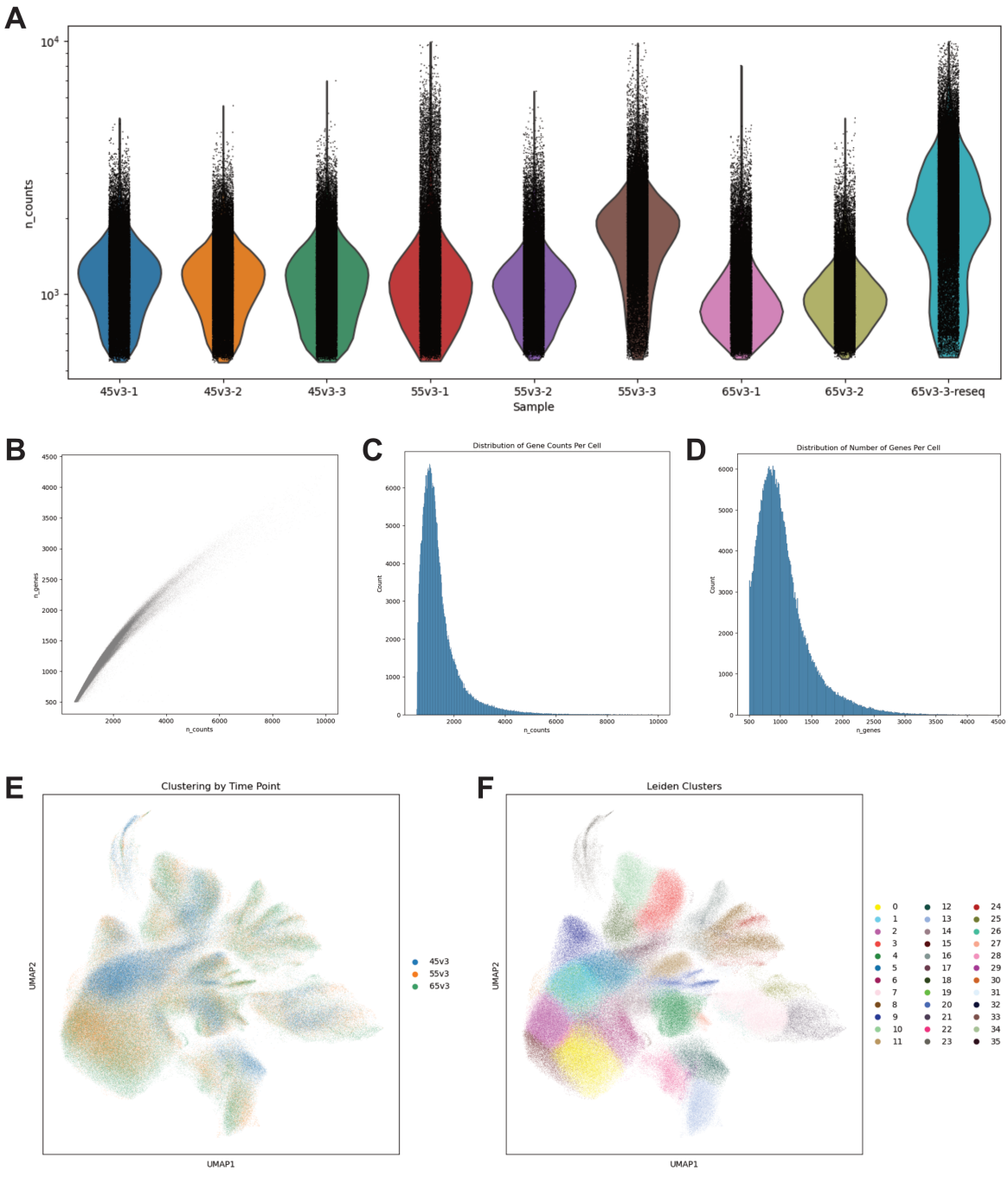

Figure C.1. (*Continued from previous page.*)

Bibliography

[1] Nicholas D. Holland. Chordates. *Current Biology*, 15(22):R911–R914, November 2005. ISSN 0960-9822. doi: 10.1016/j.cub.2005.11.008.

[2] Derek L. Stemple. The notochord. *Current Biology*, 14(20):R873–R874, October 2004. ISSN 0960-9822. doi: 10.1016/j.cub.2004.09.065.

[3] Derek L. Stemple. Structure and function of the notochord: An essential organ for chordate development. *Development*, 132(11):2503–2512, June 2005. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.01812.

[4] Diana Corallo, Valeria Trapani, and Paolo Bonaldo. The notochord: Structure and functions. *Cellular and Molecular Life Sciences*, 72(16):2989–3008, August 2015. ISSN 1420-682X, 1420-9071. doi: 10.1007/s00018-015-1897-z.

[5] Sophie Balmer, Sonja Nowotschin, and Anna-Katerina Hadjantonakis. Notochord morphogenesis in mice: Current understanding & open questions. *Developmental Dynamics*, 245 (5):547–557, 2016. ISSN 1097-0177. doi: 10.1002/dvdy.24392.

[6] Karel de Bree, Bernadette S. de Bakker, and Roelof-Jan Oostra. The development of the human notochord. *PLoS ONE*, 13(10):e0205752, October 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0205752.

[7] Lisa Lawson and Brian D. Harfe. Notochord to Nucleus Pulposus Transition. *Current Osteoporosis Reports*, 13(5):336–341, October 2015. ISSN 1544-2241. doi: 10.1007/s11914-015-0284-x.

[8] Paramvir Dehal, Yutaka Satou, Robert K. Campbell, Jarrod Chapman, Bernard Degnan, Anthony De Tomaso, Brad Davidson, Anna Di Gregorio, Maarten Gelpke, David M. Goodstein, Naoe Harafuji, Kenneth E. M. Hastings, Isaac Ho, Kohji Hotta, Wayne Huang, Takeshi Kawashima, Patrick Lemaire, Diego Martinez, Ian A. Meinertzhagen, Simona Necula, Masaru Nonaka, Nik Putnam, Sam Rash, Hidetoshi Saiga, Masanobu Satake, Astrid Terry, Lixy Yamada, Hong-Gang Wang, Satoko Awazu, Kaoru Azumi, Jeffrey Boore, Margherita Branno, Stephen Chin-Bow, Rosaria DeSantis, Sharon Doyle, Pilar Francino, David N. Keys, Shinobu Haga, Hiroko Hayashi, Kyosuke Hino, Kaoru S. Imai, Kazuo Inaba, Shungo Kano, Kenji Kobayashi, Mari Kobayashi, Byung-In Lee, Kazuhiro W. Makabe, Chitra Manohar, Giorgio Matassi, Monica Medina, Yasuaki Mochizuki, Steve Mount, Tomomi Morishita, Sachiko Miura, Akie Nakayama, Satoko Nishizaka, Hisayo

Nomoto, Fumiko Ohta, Kazuko Oishi, Isidore Rigoutsos, Masako Sano, Akane Sasaki, Yasunori Sasakura, Eiichi Shoguchi, Tadasu Shin-i, Antoinetta Spagnuolo, Didier Stainier, Miho M. Suzuki, Olivier Tassy, Naohito Takatori, Miki Tokuoka, Kasumi Yagi, Fumiko Yoshizaki, Shuichi Wada, Cindy Zhang, P. Douglas Hyatt, Frank Larimer, Chris Detter, Norman Doggett, Tijana Glavina, Trevor Hawkins, Paul Richardson, Susan Lucas, Yuji Kohara, Michael Levine, Nori Satoh, and Daniel S. Rokhsar. The draft genome of Ciona intestinalis: Insights into chordate and vertebrate origins. *Science (New York, N.Y.)*, 298 (5601):2157–2167, December 2002. ISSN 1095-9203. doi: 10.1126/science.1080049.

[9] Frédéric Delsuc, Henner Brinkmann, Daniel Chourrout, and Hervé Philippe. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, 439(7079): 965–968, February 2006. ISSN 1476-4687. doi: 10.1038/nature04336.

[10] Kaoru S. Imai, Michael Levine, Nori Satoh, and Yutaka Satou. Regulatory Blueprint for a Chordate Embryo. *Science*, 312(5777):1183–1187, May 2006. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1123404.

[11] Noriyuki Satoh. *Developmental Genomics of Ascidians*. John Wiley & Sons, Inc, Hoboken, New Jersey, 2014. ISBN 978-1-118-65604-4 978-1-118-65624-2.

[12] Yutaka Satou, Ryohei Nakamura, Deli Yu, Reiko Yoshida, Mayuko Hamada, Manabu Fujie, Kanako Hisata, Hiroyuki Takeda, and Noriyuki Satoh. A Nearly Complete Genome of Ciona intestinalis Type A (C. robusta) Reveals the Contribution of Inversion to Chromosomal Evolution in the Genus Ciona. *Genome Biology and Evolution*, 11(11):3144–3157, November 2019. doi: 10.1093/gbe/evz228.

[13] Konner M. Winkley, Matthew J. Kourakis, Anthony W. DeTomaso, Michael T. Veeman, and William C. Smith. Tunicate Gastrulation. *Current topics in developmental biology*, 136:219–242, 2020. ISSN 0070-2153. doi: 10.1016/bs.ctdb.2019.09.001.

[14] Wendy M. Reeves, Yuye Wu, Matthew J. Harder, and Michael T. Veeman. Functional and evolutionary insights from the Ciona notochord transcriptome. *Development*, 144(18): 3375–3387, September 2017. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.156174.

[15] J. C. Corbo, M. Levine, and R. W. Zeller. Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. *Development*, 124 (3):589–602, February 1997. ISSN 0950-1991, 1477-9129.

[16] Kasumi Yagi, Yutaka Satou, and Nori Satoh. A zinc finger transcription factor, ZicL, is a direct activator of Brachyury in the notochord specification of Ciona intestinalis. *Development*, 131(6):1279–1288, March 2004. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.01011.

[17] Jerry M. Rhee, Izumi Oda-Ishii, Yale J. Passamaneck, Anna-Katerina Hadjantonakis, and Anna Di Gregorio. Live imaging and morphometric analysis of embryonic development in the ascidianCiona intestinalis. *genesis*, 43(3):136–147, November 2005. ISSN 1526-954X,

1526-968X. doi: 10.1002/gene.20164.

[18] Mitsuru J. Nakamura, Jun Terai, Reiko Okubo, Kohji Hotta, and Kotaro Oka. Three-dimensional anatomy of the Ciona intestinalis tailbud embryo at single-cell resolution. *Developmental Biology*, 372(2):274–284, December 2012. ISSN 0012-1606. doi: 10.1016/j.ydbio.2012.09.007.

[19] George Khoury and Peter Gruss. Enhancer elements. *Cell*, 33(2):313–314, June 1983. ISSN 00928674. doi: 10.1016/0092-8674(83)90410-5.

[20] Evgeny Z. Kvon, Rachel Waymack, Mario Gad, and Zeba Wunderlich. Enhancer redundancy in development and disease. *Nature Reviews Genetics*, 22(5):324–336, May 2021. ISSN 1471-0064. doi: 10.1038/s41576-020-00311-x.

[21] Mike Levine. Transcriptional enhancers in animal development and evolution. *Current biology: CB*, 20(17):R754–763, September 2010. ISSN 1879-0445. doi: 10.1016/j.cub.2010.06.070.

[22] Maria I Arnone and Eric H Davidson. The hardwiring of development: Organization and function of genomic regulatory systems. *Development*, page 14, 1997.

[23] Granton A. Jindal and Emma K. Farley. Enhancer grammar in development, evolution, and disease: Dependencies and interplay. *Developmental Cell*, 56(5):575–587, March 2021. ISSN 15345807. doi: 10.1016/j.devcel.2021.02.016.

[24] Emma K. Farley, Katrina M. Olson, Wei Zhang, Daniel S. Rokhsar, and Michael S. Levine. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proceedings of the National Academy of Sciences*, 113(23):6508–6513, June 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1605085113.

[25] Benjamin P Song, Michelle F Ragsac, Krissie Tellez, Granton A Jindal, Jessica L Grudzien, Sophia H Le, and Emma K Farley. Diverse logics and grammar encode notochord enhancers, July 2022.

[26] Iain M. Dykes, Dorota Szumska, Linta Kuncheria, Rathi Puliyadi, Chiann-mun Chen, Costis Papanayotou, Helen Lockstone, Christèle Dubourg, Véronique David, Jurgen E. Schneider, Thomas M. Keane, David J. Adams, Steve D. M. Brown, Sandra Mercier, Sylvie Odent, Jérôme Collignon, and Shoumo Bhattacharya. A Requirement for Zic2 in the Regulation of Nodal Expression Underlies the Establishment of Left-Sided Identity. *Scientific Reports*, 8(1):10439, December 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-28714-1.

[27] Jun Matsumoto, Gaku Kumano, and Hiroki Nishida. Direct activation by Ets and Zic is required for initial expression of the Brachyury gene in the ascidian notochord. *Developmental Biology*, 306(2):870–882, June 2007. ISSN 0012-1606. doi: 10.1016/j.ydbio.2007.03.034.

[28] Diego Villar, Camille Berthelot, Sarah Aldridge, Tim F. Rayner, Margus Lukk, Miguel

Pignatelli, Thomas J. Park, Robert Deaville, Jonathan T. Erichsen, Anna J. Jasinska, James M.A. Turner, Mads F. Bertelsen, Elizabeth P. Murchison, Paul Flicek, and Duncan T. Odom. Enhancer Evolution across 20 Mammalian Species. *Cell*, 160(3):554–566, January 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.01.006.

[29] Lucas D. Ward and Manolis Kellis. Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. *Science*, 337(6102):1675–1678, September 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1225057.

[30] Emily S. Wong, Dawei Zheng, Siew Z. Tan, Neil I. Bower, Victoria Garside, Gilles Vanwalleghem, Federico Gaiti, Ethan Scott, Benjamin M. Hogan, Kazu Kikuchi, Edwina McGlinn, Mathias Francois, and Bernard M. Degnan. Deep conservation of the enhancer regulatory code in animals. *Science*, 370(6517), November 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aax8137.

[31] Melina Claussnitzer, Simon N. Dankel, Bernward Klocke, Harald Grallert, Viktoria Glunk, Tea Berulava, Heekyoung Lee, Nikolay Oskolkov, Joao Fadista, Kerstin Ehlers, Simone Wahl, Christoph Hoffmann, Kun Qian, Tina Rönn, Helene Riess, Martina Müller-Nurasyid, Nancy Bretschneider, Timm Schroeder, Thomas Skurk, Bernhard Horsthemke, Derek Spieler, Martin Klingenspor, Martin Seifert, Michael J. Kern, Niklas Mejhert, Ingrid Dahlman, Ola Hansson, Stefanie M. Hauck, Matthias Blüher, Peter Arner, Leif Groop, Thomas Illig, Karsten Suhre, Yi-Hsiang Hsu, Gunnar Mellgren, Hans Hauner, and Helmut Laumen. Leveraging Cross-Species Transcription Factor Binding Site Patterns: From Diabetes Risk Loci to Disease Mechanisms. *Cell*, 156(1-2):343–58, January 2014. ISSN 0092-8674. doi: 10.1016/j.cell.2013.10.058.

[32] Isabelle S. Peter and Eric H. Davidson. Evolution of Gene Regulatory Networks Controlling Body Plan Development. *Cell*, 144(6):970–985, March 2011. ISSN 00928674. doi: 10.1016/j.cell.2011.02.017.

[33] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.04.044.

[34] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.05.002.

[35] Thale Kristin Olsen and Ninib Baryawno. Introduction to Single-Cell RNA Sequencing. *Current Protocols in Molecular Biology*, 122(1):e57, 2018. ISSN 1934-3647. doi: 10.1002/cpmb.57.

[36] Allon M. Klein and Barbara Treutlein. Single cell analyses of development in the modern era. *Development*, 146(12):dev181396, June 2019. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.181396.

[37] Lindsay Barone, Jason Williams, and David Micklos. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLOS Computational Biology*, 13 (10):e1005755, October 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005755.

[38] Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. Big Data: Astronomical or Genomical? *PLOS Biology*, 13(7):e1002195, July 2015. ISSN 1545-7885. doi: 10.1371/journal.pbio.1002195.

[39] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589, May 2010. ISSN 10972765. doi: 10.1016/j.molcel.2010.05.004.

[40] Feng Liu and James W. Posakony. Role of Architecture in the Function and Specificity of Two Notch-Regulated Transcriptional Enhancer Modules. *PLOS Genetics*, 8(7):e1002796, July 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002796.

[41] S. Small, A. Blair, and M. Levine. Regulation of even-skipped stripe 2 in the Drosophila embryo. *The EMBO Journal*, 11(11):4047–4057, November 1992. ISSN 02614189. doi: 10.1002/j.1460-2075.1992.tb05498.x.

[42] François Spitz and Eileen E. M. Furlong. Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, September 2012. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3207.

[43] Christina I. Swanson, Nicole C. Evans, and Scott Barolo. Structural Rules and Complex Regulatory Circuitry Constrain Expression of a Notch- and EGFR-Regulated Eye Enhancer. *Developmental Cell*, 18(3):359–370, March 2010. ISSN 1534-5807. doi: 10.1016/j.devcel.2009.12.026.

[44] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph, Peter J Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337:7, 2012.

[45] Yu Gyoung Tak and Peggy J. Farnham. 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(1):57, December 2015. ISSN 1756-8935. doi: 10.1186/s13072-015-0050-4.

[46] Axel Visel, Edward M. Rubin, and Len A. Pennacchio. Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205, September 2009. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature08451.

[47] Scott Barolo. How to tune an enhancer. *Proceedings of the National Academy of Sciences*, 113(23):6330–6331, June 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1606109113.

[48] Michal Levo and Eran Segal. In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics*, 15(7):453–468, July 2014. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3684.

[49] Dimitris Thanos and Tom Maniatis. Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. *Cell*, 83(7):1091–1100, December 1995. ISSN 00928674. doi: 10.1016/0092-8674(95)90136-1.

[50] Bernhard G. Herrmann and Andreas Kispert. The T genes in embryogenesis. *Trends in Genetics*, 10(8):280–286, August 1994. ISSN 01689525. doi: 10.1016/0168-9525(90)90011-T.

[51] Paul Chesley. Development of the short-tailed mutant in the house mouse. *Journal of Experimental Zoology*, 70(3):429–459, May 1935. ISSN 0022-104X, 1097-010X. doi: 10.1002/jez.1400700306.

[52] Shota Chiba, Di Jiang, Noriyuki Satoh, and William C. Smith. *Brachyury* null mutant-induced defects in juvenile ascidian endodermal organs. *Development*, 136(1):35–39, January 2009. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.030981.

[53] David G. Wilkinson, Sangita Bhatt, and Bernhard G. Herrmann. Expression pattern of the mouse T gene and its role in mesoderm formation. *Nature*, 343(6259):657–659, February 1990. ISSN 0028-0836, 1476-4687. doi: 10.1038/343657a0.

[54] Hitoyoshi Yasuo and Noriyuki Satoh. Function of vertebrate T gene. *Nature*, 364(6438): 582–583, August 1993. ISSN 0028-0836, 1476-4687. doi: 10.1038/364582b0.

[55] Siew-Lan Ang and Janet Rossant. HNF-3β is essential for node and notochord formation in mouse development. *Cell*, 78(4):561–574, August 1994. ISSN 00928674. doi: 10.1016/0092-8674(94)90522-3.

[56] Sophie Dal-Pra, Christine Thisse, and Bernard Thisse. FoxA transcription factors are essential for the development of dorsal axial structures. *Developmental Biology*, 350(2): 484–495, February 2011. ISSN 00121606. doi: 10.1016/j.ydbio.2010.12.018.

[57] Paul Elms, Andrew Scurry, Jennifer Davies, Catherine Willoughby, Terry Hacker, Debora

Bogani, and Ruth Arkell. Overlapping and distinct expression domains of Zic2 and Zic3 during mouse gastrulation. *Gene Expression Patterns*, 4(5):505–511, September 2004. ISSN 1567133X. doi: 10.1016/j.modgep.2004.03.003.

[58] Kaoru S. Imai, Yutaka Satou, and Nori Satoh. Multiple functions of a Zic-like gene in the differentiation of notochord, central nervous system and muscle in *Ciona savignyi* embryos. *Development*, 129(11):2723–2732, June 2002. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.129.11.2723.

[59] Kaoru S. Imai, Nori Satoh, and Yutaka Satou. Early embryonic expression of *FGF4/6/9* gene and its role in the induction of mesenchyme and notochord in *Ciona savignyi* embryos. *Development*, 129(7):1729–1738, April 2002. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.129.7.1729.

[60] Diana S. José-Edwards, Izumi Oda-Ishii, Jamie E. Kugler, Yale J. Passamaneck, Lavanya Katikala, Yutaka Nibu, and Anna Di Gregorio. Brachyury, Foxa2 and the cis-Regulatory Origins of the Notochord. *PLOS Genetics*, 11(12):e1005730, December 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005730.

[61] Lavanya Katikala, Hitoshi Aihara, Yale J. Passamaneck, Stefan Gazdoiu, Diana S. José-Edwards, Jamie E. Kugler, Izumi Oda-Ishii, Janice H. Imai, Yutaka Nibu, and Anna Di Gregorio. Functional Brachyury Binding Sites Establish a Temporal Read-out of Gene Expression in the Ciona Notochord. *PLoS Biology*, 11(10):e1001697, October 2013. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001697.

[62] Gaku Kumano, Satoshi Yamaguchi, and Hiroki Nishida. Overlapping expression of FoxA and Zic confers responsiveness to FGF signaling to specify notochord in ascidian embryos. *Developmental Biology*, 300(2):770–784, December 2006. ISSN 00121606. doi: 10.1016/j.ydbio.2006.07.033.

[63] Takahito Miya and Hiroki Nishida. An Ets transcription factor, HrEts, is target of FGF signaling and involved in induction of notochord, mesenchyme, and brain in ascidian embryos. *Developmental Biology*, 261(1):25–38, September 2003. ISSN 00121606. doi: 10.1016/S0012-1606(03)00246-X.

[64] Yale J. Passamaneck, Lavanya Katikala, Lorena Perrone, Matthew P. Dunn, Izumi Oda-Ishii, and Anna Di Gregorio. Direct activation of a notochord cis-regulatory module by Brachyury and FoxA in the ascidian Ciona intestinalis. *Development*, 136(21):3679–3689, November 2009. ISSN 0950-1991. doi: 10.1242/dev.038141.

[65] S. Schulte-Merker and J.C. Smith. Mesoderm formation in response to Brachyury requires FGF signalling. *Current Biology*, 5(1):62–67, January 1995. ISSN 09609822. doi: 10.1016/S0960-9822(95)00017-0.

[66] Nicholas Warr, Nicola Powles-Glover, Anna Chappell, Joan Robson, Dominic Norris, and Ruth M. Arkell. Zic2 -associated holoprosencephaly is caused by a transient defect in

the organizer region during gastrulation. *Human Molecular Genetics*, 17(19):2986–2996, October 2008. ISSN 1460-2083, 0964-6906. doi: 10.1093/hmg/ddn197.

[67] Daniel C. Weinstein, Ariel Ruiz i Altaba, William S. Chen, Pamela Hoodless, Vincent R. Prezioso, Thomas M. Jessell, and James E. Darnell. The winged-helix transcription factor HNF-3β is required for notochord development in the mouse embryo. *Cell*, 78(4):575–588, August 1994. ISSN 00928674. doi: 10.1016/0092-8674(94)90523-1.

[68] Hitoyoshi Yasuo and Clare Hudson. FGF8/17/18 functions together with FGF9/16/20 during formation of the notochord in Ciona embryos. *Developmental Biology*, 302(1):92–103, February 2007. ISSN 00121606. doi: 10.1016/j.ydbio.2006.08.075.

[69] Brad Davidson and Lionel Christiaen. Linking Chordate Gene Networks to Cellular Behavior in Ascidians. *Cell*, 124(2):247–250, January 2006. ISSN 00928674. doi: 10.1016/j.cell.2006.01.013.

[70] Emma K. Farley, Katrina M. Olson, Wei Zhang, Alexander J. Brandt, Daniel S. Rokhsar, and Michael S. Levine. Suboptimization of developmental enhancers. *Science*, 350(6258):325–328, October 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aac6948.

[71] Anna Di Gregorio. The notochord gene regulatory network in chordate evolution: Conservation and divergence from Ciona to vertebrates. In *Current Topics in Developmental Biology*, volume 139, pages 325–374. Elsevier, 2020. ISBN 978-0-12-813180-0. doi: 10.1016/bs.ctdb.2020.01.002.

[72] Michael T. Veeman, Yuki Nakatani, Carolyn Hendrickson, Vivian Ericson, Clarissa Lin, and William C. Smith. *Chongmague* reveals an essential role for laminin-mediated boundary formation in chordate convergence and extension movements. *Development*, 135(1):33–41, January 2008. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.010892.

[73] Dennis Schifferl, Manuela Scholze-Wittler, Lars Wittler, Jesse V. Veenvliet, Frederic Koch, and Bernhard G. Herrmann. A 37 kb region upstream of *brachyury* comprising a notochord enhancer is essential for notochord and tail development. *Development*, 148(23):dev200059, December 2021. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.200059.

[74] H. Takahashi, Y. Mitani, G. Satoh, and N. Satoh. Evolutionary alterations of the minimal promoter for notochord-specific Brachyury expression in ascidian embryos. *Development*, 126 (17):3725–3734, September 1999. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.126.17.3725.

[75] Ekaterina P Lamber, Laurent Vanhille, Larissa C Textor, Galina S Kachalova, Michael H Sieweke, and Matthias Wilmanns. Regulation of the transcription factor Ets-1 by DNA-mediated homo-dimerization. *The EMBO Journal*, 27(14):2006–2017, July 2008. ISSN 0261-4189, 1460-2075. doi: 10.1038/emboj.2008.117.

[76] Gong-Hong Wei, Gwenael Badis, Michael F Berger, Teemu Kivioja, Kimmo Palin, Martin Enge, Martin Bonke, Arttu Jolma, Markku Varjosalo, Andrew R Gehrke, Jian Yan,

Shaheynoor Talukder, Mikko Turunen, Mikko Taipale, Hendrik G Stunnenberg, Esko Ukkonen, Timothy R Hughes, Martha L Bulyk, and Jussi Taipale. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *The EMBO Journal*, 29(13):2147–2160, July 2010. ISSN 0261-4189, 1460-2075. doi: 10.1038/emboj.2010.106.

[77] Ute Rothbächer, Vincent Bertrand, Clement Lamy, and Patrick Lemaire. A combinatorial code of maternal GATA, Ets and β-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. *Development*, 134(22):4023–4032, November 2007. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.010850.

[78] Alberto Stolfi, Kerrianne Ryan, Ian A. Meinertzhagen, and Lionel Christiaen. Migratory neuronal progenitors arise from the neural plate borders in tunicates. *Nature*, 527(7578): 371–374, November 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature15758.

[79] Di Jiang and William C. Smith. Ascidian notochord morphogenesis. *Developmental Dynamics*, 236(7):1748–1757, July 2007. ISSN 10588388, 10970177. doi: 10.1002/dvdy.21184.

[80] Kaoru S. Imai, Kyosuke Hino, Kasumi Yagi, Nori Satoh, and Yutaka Satou. Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: Towards a comprehensive understanding of gene networks. *Development*, 131(16):4047–4058, August 2004. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.01270.

[81] Konner M. Winkley, Wendy M. Reeves, and Michael T. Veeman. Single-cell analysis of cell fate bifurcation in the chordate Ciona. *BMC Biology*, 19(1):180, December 2021. ISSN 1741-7007. doi: 10.1186/s12915-021-01122-0.

[82] Clare Hudson, Sonia Lotito, and Hitoyoshi Yasuo. 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. *Development*, 134(19): 3527–3537, October 2007. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.002352.

[83] Clare Hudson, Cathy Sirour, and Hitoyoshi Yasuo. Co-expression of Foxa.a, Foxd and Fgf9/16/20 defines a transient mesendoderm regulatory state in ascidian embryos. *eLife*, 5: e14692, June 2016. ISSN 2050-084X. doi: 10.7554/eLife.14692.

[84] Vincent Picco, Clare Hudson, and Hitoyoshi Yasuo. Ephrin-Eph signalling drives the asymmetric division of notochord/neural precursors in *Ciona* embryos. *Development*, 134 (8):1491–1497, April 2007. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.003939.

[85] Eileen Wagner and Michael Levine. FGF signaling establishes the anterior border of the *Ciona* neural tube. *Development*, 139(13):2351–2359, July 2012. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.078485.

[86] Tatsuro Ikeda and Yutaka Satou. Differential temporal control of *Foxa.a* and *Zic-r.b* specifies brain versus notochord fate in the ascidian embryo. *Development*, page dev.142174,

January 2016. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.142174.

[87] E.S. Casey, M.A. O'Reilly, F.L. Conlon, and J.C. Smith. The T-box transcription factor Brachyury regulates expression of eFGF through binding to a non-palindromic response element. *Development*, 125(19):3887–3894, October 1998. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.125.19.3887.

[88] Jun Li, Ana Carolina Dantas Machado, Ming Guo, Jared M. Sagendorf, Zhan Zhou, Longying Jiang, Xiaojuan Chen, Daichao Wu, Lingzhi Qu, Zhuchu Chen, Lin Chen, Remo Rohs, and Yongheng Chen. Structure of the Forkhead Domain of FOXA2 Bound to a Complete DNA Consensus Site. *Biochemistry*, 56(29):3745–3753, July 2017. ISSN 0006-2960, 1520-4995. doi: 10.1021/acs.biochem.7b00211.

[89] Frank L. Conlon, Lynne Fairclough, Brenda M. J. Price, Elena S. Casey, and J. C. Smith. Determinants of T box protein specificity. *Development*, 128(19):3749–3758, October 2001. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.128.19.3749.

[90] A. Di Gregorio and M. Levine. Regulation of Ci-tropomyosin-like, a Brachyury target gene in the ascidian, Ciona intestinalis. *Development*, 126(24):5599–5609, December 1999. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.126.24.5599.

[91] Matthew P. Dunn and Anna Di Gregorio. The evolutionarily conserved leprecan gene: Its regulation by Brachyury and its role in the developing Ciona notochord. *Developmental Biology*, 328(2):561–574, April 2009. ISSN 00121606. doi: 10.1016/j.ydbio.2009.02.007.

[92] Christoph W. Müller and Bernhard G. Herrmann. Crystallographic structure of the T domain–DNA complex of the Brachyury transcription factor. *Nature*, 389(6653):884–888, October 1997. ISSN 0028-0836, 1476-4687. doi: 10.1038/39929.

[93] Kazuhiro R Nitta, Arttu Jolma, Yimeng Yin, Ekaterina Morgunova, Teemu Kivioja, Junaid Akhtar, Korneel Hens, Jarkko Toivonen, Bart Deplancke, Eileen E M Furlong, and Jussi Taipale. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife*, 4:e04837, March 2015. ISSN 2050-084X. doi: 10.7554/eLife.04837.

[94] Annabelle Scott and Derek L. Stemple. Zebrafish Notochordal Basement Membrane: Signaling and Structure. In *Current Topics in Developmental Biology*, volume 65, pages 229–253. Elsevier, 2004. ISBN 978-0-12-153165-2. doi: 10.1016/S0070-2153(04)65009-5.

[95] Quentin J. Machingo, Andreas Fritz, and Barry D. Shur. A $B1$,4-galactosyltransferase is required for convergent extension movements in zebrafish. *Developmental Biology*, 297(2): 471–482, September 2006. ISSN 00121606. doi: 10.1016/j.ydbio.2006.05.024.

[96] Michael J. Parsons, Isabel Campos, Elizabeth M. A. Hirst, and Derek L. Stemple. Removal of dystroglycan causes severe muscular dystrophy in zebrafish embryos. *Development*, 129 (14):3505–3512, July 2002. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.129.14.3505.

[97] Steven M. Pollard, Michael J. Parsons, Makoto Kamei, Ross N.W. Kettleborough, Kevin A. Thomas, Van N. Pham, Moon-Kyoung Bae, Annabelle Scott, Brant M. Weinstein, and Derek L. Stemple. Essential and overlapping roles for laminin α chains in notochord and blood vessel formation. *Developmental Biology*, 289(1):64–76, January 2006. ISSN 00121606. doi: 10.1016/j.ydbio.2005.10.006.

[98] Mark W. Barnett, Robert W. Old, and Elizabeth A. Jones. Neural induction and patterning by fibroblast growth factor, notochord and somite tissue in Xenopus. *Development, Growth and Differentiation*, 40(1):47–57, February 1998. ISSN 0012-1592, 1440-169X. doi: 10.1046/j.1440-169X.1998.t01-5-00006.x.

[99] Isabel Olivera-Martinez, Hidekiyo Harada, Pamela A. Halley, and Kate G. Storey. Loss of FGF-Dependent Mesoderm Identity and Rise of Endogenous Retinoid Signalling Determine Cessation of Body Axis Elongation. *PLoS Biology*, 10(10):e1001415, October 2012. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001415.

[100] Macarena Lolas, Pablo D. T. Valenzuela, Robert Tjian, and Zhe Liu. Charting Brachyury-mediated developmental pathways during early mouse embryogenesis. *Proceedings of the National Academy of Sciences*, 111(12):4478–4483, March 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1402612111.

[101] Wendy M. Reeves, Kotaro Shimai, Konner M. Winkley, and Michael T. Veeman. Brachyury controls *Ciona* notochord fate as part of a feed-forward network. *Development*, 148(3): dev195230, February 2021. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.195230.

[102] S. Fujiwara, J.C. Corbo, and M. Levine. The snail repressor establishes a muscle/notochord boundary in the Ciona embryo. *Development*, 125(13):2511–2520, July 1998. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.125.13.2511.

[103] Kotaro Shimai and Michael Veeman. Quantitative Dissection of the Proximal Ciona brachyury Enhancer. *Frontiers in Cell and Developmental Biology*, 9:804032, January 2022. ISSN 2296-634X. doi: 10.3389/fcell.2021.804032.

[104] Nicolás Frankel, Gregory K. Davis, Diego Vargas, Shu Wang, François Payre, and David L. Stern. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, 466(7305):490–493, July 2010. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature09158.

[105] Joung-Woo Hong, David A. Hendrix, and Michael S. Levine. Shadow Enhancers as a Source of Evolutionary Novelty. *Science*, 321(5894):1314–1314, September 2008. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1160631.

[106] Michael W. Perry, Alistair N. Boettiger, Jacques P. Bothma, and Michael Levine. Shadow Enhancers Foster Robustness of Drosophila Gastrulation. *Current Biology*, 20(17):1562–1567, September 2010. ISSN 09609822. doi: 10.1016/j.cub.2010.07.043.

[107] Barbora Antosova, Jana Smolikova, Lucie Klimova, Jitka Lachova, Michaela Bendova, Iryna Kozmikova, Ondrej Machon, and Zbynek Kozmik. The Gene Regulatory Network of Lens Induction Is Wired through Meis-Dependent Shadow Enhancers of Pax6. *PLOS Genetics*, 12(12):e1006441, December 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1006441.

[108] Marco Osterwalder, Iros Barozzi, Virginie Tissières, Yoko Fukuda-Yuzawa, Brandon J. Mannion, Sarah Y. Afzal, Elizabeth A. Lee, Yiwen Zhu, Ingrid Plajzer-Frick, Catherine S. Pickle, Momoe Kato, Tyler H. Garvin, Quan T. Pham, Anne N. Harrington, Jennifer A. Akiyama, Veena Afzal, Javier Lopez-Rios, Diane E. Dickel, Axel Visel, and Len A. Pennacchio. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, 554(7691):239–243, February 2018. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature25461.

[109] Blanca Pijuan-Sala, Nicola K. Wilson, Jun Xia, Xiaomeng Hou, Rebecca L. Hannah, Sarah Kinston, Fernando J. Calero-Nieto, Olivier Poirion, Sebastian Preissl, Feng Liu, and Berthold Göttgens. Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. *Nature Cell Biology*, 22(4):487–497, April 2020. ISSN 1465-7392, 1476-4679. doi: 10.1038/s41556-020-0489-9.

[110] Steven A. Harvey, Stefan Tümpel, Julien Dubrulle, Alexander F. Schier, and James C. Smith. *No Tail* integrates two modes of mesoderm induction. *Development*, 137(7): 1127–1135, April 2010. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.046318.

[111] Dana M King, Clarice Kit Yee Hong, James L Shepherdson, David M Granas, Brett B Maricque, and Barak A Cohen. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. *eLife*, 9:e41279, February 2020. ISSN 2050-084X. doi: 10.7554/eLife.41279.

[112] Timothy Fuqua, Jeff Jordan, Maria Elize van Breugel, Aliaksandr Halavatyi, Christian Tischer, Peter Polidoro, Namiko Abe, Albert Tsai, Richard S. Mann, David L. Stern, and Justin Crocker. Dense and pleiotropic regulatory information in a developmental enhancer. *Nature*, 587(7833):235–239, November 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2816-5.

[113] Avihu H. Yona, Eric J. Alm, and Jeff Gore. Random sequences rapidly evolve into de novo promoters. *Nature Communications*, 9(1):1530, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04026-w.

[114] Carl G. de Boer, Eeshit Dhaval Vaishnav, Ronen Sadeh, Esteban Luis Abeyta, Nir Friedman, and Aviv Regev. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature Biotechnology*, 38(1):56–65, January 2020. ISSN 1546-1696. doi: 10.1038/s41587-019-0315-8.

[115] Rafael Galupa, Gilberto Alvarez-Canales, Noa Ottilie Borst, Timothy Fuqua, Natalia Misunou, Kerstin Richter, Mariana R P Alves, Esther Karumbi, Melinda Liu Perkins, Tin Kocijan, Christine A Rushlow, and Justin Crocker. Enhancer architecture and chromatin accessibility constrain phenotypic space during development. *bioRxiv*, page 25, 2022.

[116] Yutaka Satou, Takeshi Kawashima, Eiichi Shoguchi, Akie Nakayama, and Nori Satoh. An Integrated Database of the Ascidian, Ciona intestinalis: Towards Functional Genomics. *Zoological Science*, 22(8):837–843, August 2005. ISSN 0289-0003. doi: 10.2108/zsj.22.837.

[117] Sabina Leonelli. The challenges of big data biology. *eLife*, 8:e47381, April 2019. ISSN 2050-084X. doi: 10.7554/eLife.47381.

[118] Vivien Marx. The big challenges of big data. *Nature*, 498(7453):255–260, June 2013. ISSN 1476-4687. doi: 10.1038/498255a.

[119] Subhajit Pal, Sudip Mondal, Gourab Das, Sunirmal Khatua, and Zhumur Ghosh. Big data in biology: The hope and present-day challenges in it. *Gene Reports*, 21:100869, December 2020. ISSN 2452-0144. doi: 10.1016/j.genrep.2020.100869.

[120] Wouter Meuleman, Alexander Muratov, Eric Rynes, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglas Dunn, Fidencio Neri, Athanasios Teodosiadis, Alex Reynolds, Eric Haugen, Jemma Nelson, Audra Johnson, Mark Frerker, Michael Buckley, Richard Sandstrom, Jeff Vierstra, Rajinder Kaul, and John Stamatoyannopoulos. Index and biological spectrum of human DNase I hypersensitive sites. *Nature*, July 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2559-3.

[121] Robert P. Zinzen, Charles Girardot, Julien Gagneur, Martina Braun, and Eileen E. M. Furlong. Combinatorial binding predicts spatio-temporal cis -regulatory activity. *Nature*, 462(7269):65–70, November 2009. ISSN 1476-4687. doi: 10.1038/nature08531.

[122] Sharon R. Grossman, Xiaolan Zhang, Li Wang, Jesse Engreitz, Alexandre Melnikov, Peter Rogov, Ryan Tewhey, Alina Isakova, Bart Deplancke, Bradley E. Bernstein, Tarjei S. Mikkelsen, and Eric S. Lander. Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proceedings of the National Academy of Sciences*, 114(7):E1291–E1300, February 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1621150114.

[123] Genevieve E. Ryan and Emma K. Farley. Functional genomic approaches to elucidate the role of enhancers during development. *WIREs Systems Biology and Medicine*, 12(2):e1467, 2020. ISSN 1939-005X. doi: 10.1002/wsbm.1467.

[124] Marc S. Halfon. Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. *Trends in Genetics*, 35(2):93–103, February 2019. ISSN 0168-9525. doi: 10.1016/j.tig.2018.11.004.

[125] Meghana M. Kulkarni and David N. Arnosti. Information display by transcriptional enhancers. *Development*, 130(26):6569–6575, 2003.

[126] D. Bazett-Jones, B Leblanc, M Herfort, and T Moss. Short-range DNA looping by the Xenopus HMG-box transcription factor, xUBF. *Science*, 264(5162):1134–1137, May 1994. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.8178172.

[127] Daniel Panne, Tom Maniatis, and Stephen C. Harrison. An Atomic Model of the Interferon-β Enhanceosome. *Cell*, 129(6):1111–1123, June 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.05.019.

[128] Alexandre Melnikov, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, Andreas Gnirke, Curtis G. Callan, Justin B. Kinney, Manolis Kellis, Eric S. Lander, and Tarjei S. Mikkelsen. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3):271–277, March 2012. ISSN 1546-1696. doi: 10.1038/nbt.2137.

[129] Guillaume Junion, Mikhail Spivakov, Charles Girardot, Martina Braun, E. Hilary Gustafson, Ewan Birney, and Eileen E. M. Furlong. A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. *Cell*, 148(3):473–486, February 2012. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2012.01.030.

[130] Grzegorz M. Boratyn, Jean Thierry-Mieg, Danielle Thierry-Mieg, Ben Busby, and Thomas L. Madden. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinformatics*, 20(1):405, July 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2996-x.

[131] Matthew M. Hill, Karl W. Broman, Elia Stupka, William C. Smith, Di Jiang, and Arend Sidow. The *C. savignyi* genetic map and its integration with the reference sequence facilitates insights into chordate genome evolution. *Genome Research*, 18(8):1369–1379, August 2008. ISSN 1088-9051. doi: 10.1101/gr.078576.108.

[132] Wei Xie and Bing Ren. Enhancing Pluripotency and Lineage Specification. *Science*, 341 (6143):245–247, July 2013. doi: 10.1126/science.1236254.

[133] Casey E. Romanoski, Christopher K. Glass, Hendrik G. Stunnenberg, Laurence Wilson, and Genevieve Almouzni. Roadmap for regulation. *Nature*, 518(7539):314–316, February 2015. ISSN 1476-4687. doi: 10.1038/518314a.

[134] Eric N. Olson. Gene Regulatory Networks in the Evolution and Development of the Heart. *Science*, 313(5795):1922–1927, September 2006. doi: 10.1126/science.1132292.

[135] Eric H. Davidson and Michael S. Levine. Properties of developmental gene regulatory networks. *Proceedings of the National Academy of Sciences*, 105(51):20063–20066, December 2008. doi: 10.1073/pnas.0806007105.

[136] Michael Levine and Eric H. Davidson. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, 102(14):4936–4942, April 2005. doi: 10.1073/pnas.0408031102.

[137] Ye Wang, Michael Mashock, Zhuang Tong, Xiaofeng Mu, Hong Chen, Xin Zhou, Hong Zhang, Gexin Zhao, Bin Liu, and Xinmin Li. Changing Technologies of RNA Sequencing and Their Applications in Clinical Oncology. *Frontiers in Oncology*, 10, 2020. ISSN 2234-943X.

[138] Mingye Hong, Shuang Tao, Ling Zhang, Li-Ting Diao, Xuanmei Huang, Shaohui Huang, Shu-Juan Xie, Zhen-Dong Xiao, and Hua Zhang. RNA sequencing: New technologies and applications in cancer research. *Journal of Hematology & Oncology*, 13(1):166, December 2020. ISSN 1756-8722. doi: 10.1186/s13045-020-01005-x.

[139] Xinmin Li and Cun-Yu Wang. From bulk, single-cell to spatial RNA sequencing. *International Journal of Oral Science*, 13(1):1–6, November 2021. ISSN 2049-3169. doi: 10.1038/s41368-021-00146-0.

[140] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B. Tuch, Asim Siddiqui, Kaiqin Lao, and M. Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5): 377–382, May 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1315.

[141] Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, 12(3):e694, March 2022. ISSN 2001-1326. doi: 10.1002/ctm2.694.

[142] Efthymia Papalexi and Rahul Satija. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1):35–45, January 2018. ISSN 1474-1741. doi: 10.1038/nri.2017.76.

[143] Lila Solnica-Krezel and Diane S. Sepich. Gastrulation: Making and Shaping Germ Layers. *Annual Review of Cell and Developmental Biology*, 28(1):687–717, 2012. doi: 10.1146/annurev-cellbio-092910-154043.

[144] Sabitri Ghimire, Veronika Mantziou, Naomi Moris, and Alfonso Martinez Arias. Human gastrulation: The embryo and its models. *Developmental Biology*, 474:100–108, June 2021. ISSN 0012-1606. doi: 10.1016/j.ydbio.2021.01.006.

[145] Sarthak Mohanty and Chitra L. Dahia. Defects in intervertebral disc and spine during development, degeneration, and pain: New research directions for disc regeneration and therapy. *WIREs Developmental Biology*, 8(4):e343, 2019. ISSN 1759-7692. doi: 10.1002/wdev.343.

[146] Jason W. Ashley, Motomi Enomoto-Iwamoto, Lachlan J. Smith, Robert L. Mauck, Danny Chan, Joseph Lee, Martin F. Heyworth, Howard An, and Yejia Zhang. Intervertebral disc development and disease-related genetic polymorphisms. *Genes & Diseases*, 3(3):171–177, September 2016. ISSN 2352-3042. doi: 10.1016/j.gendis.2016.04.006.

[147] P. Prithvi Raj. Intervertebral Disc: Anatomy-Physiology-Pathophysiology-Treatment. *Pain Practice*, 8(1):18–44, 2008. ISSN 1533-2500. doi: 10.1111/j.1533-2500.2007.00171.x.

[148] Ekta A. Patel and Michael D. Perloff. Radicular Pain Syndromes: Cervical, Lumbar, and Spinal Stenosis. *Seminars in Neurology*, 38(6):634–639, December 2018. ISSN 0271-8235, 1098-9021. doi: 10.1055/s-0038-1673680.

[149] Brian D. Harfe. Intervertebral disc repair and regeneration: Insights from the notochord. *Seminars in Cell & Developmental Biology*, 127:3–9, July 2022. ISSN 1084-9521. doi: 10.1016/j.semcdb.2021.11.012.

[150] Ajay Matta and William Mark Erwin. Current Status of the Instructional Cues Provided by Notochordal Cells in Novel Disc Repair Strategies. *International Journal of Molecular Sciences*, 23(1):427, December 2021. ISSN 1422-0067. doi: 10.3390/ijms23010427.

[151] Takashi Kamatani, Hiroki Hagizawa, Seido Yarimitsu, Miho Morioka, Saeko Koyamatsu, Michihiko Sugimoto, Joe Kodama, Junko Yamane, Hiroyuki Ishiguro, Shigeyuki Shichino, Kuniya Abe, Wataru Fujibuchi, Hiromichi Fujie, Takashi Kaito, and Noriyuki Tsumaki. Human iPS cell-derived cartilaginous tissue spatially and functionally replaces nucleus pulposus. *Biomaterials*, 284:121491, May 2022. ISSN 0142-9612. doi: 10.1016/j.biomaterials.2022.121491.

[152] Frances C. Bach, Deepani W. Poramba-Liyanage, Frank M. Riemers, Jerome Guicheux, Anne Camus, James C. Iatridis, Danny Chan, Keita Ito, Christine L. Le Maitre, and Marianna A. Tryfonidou. Notochordal Cell-Based Treatment Strategies and Their Potential in Intervertebral Disc Regeneration. *Frontiers in Cell and Developmental Biology*, 9:780749, 2021. ISSN 2296-634X. doi: 10.3389/fcell.2021.780749.

[153] D. Purmessur, M. C. Cornejo, S. K. Cho, A. C. Hecht, and J. C. Iatridis. Notochordal cell-derived therapeutic strategies for discogenic back pain. *Global Spine Journal*, 3(3):201–218, June 2013. ISSN 2192-5682. doi: 10.1055/s-0033-1350053.

[154] Jason K. Wasserman, Denis Gravel, and Bibianna Purgina. Chordoma of the Head and Neck: A Review. *Head and Neck Pathology*, 12(2):261–268, June 2018. ISSN 1936-0568. doi: 10.1007/s12105-017-0860-8.

[155] H. Gokce Yeter, Kemal Kosemehmetoglu, and Figen Soylemezoglu. Poorly differentiated chordoma: Review of 53 cases. *APMIS*, 127(9):607–615, 2019. ISSN 1600-0463. doi: 10.1111/apm.12978.

[156] Nobuyuki Fujita, Satoshi Suzuki, Kota Watanabe, Ken Ishii, Ryuichi Watanabe, Masayuki Shimoda, Keiyo Takubo, Takashi Tsuji, Yoshiaki Toyama, Takeshi Miyamoto, Keisuke Horiuchi, Masaya Nakamura, and Morio Matsumoto. Chordoma-derived cell line U-CH1-N recapitulates the biological properties of notochordal nucleus pulposus cells. *Journal of Orthopaedic Research*, 34(8):1341–1350, 2016. ISSN 1554-527X. doi: 10.1002/jor.23320.

[157] Veronica Ulici and Jesse Hart. Chordoma: A Review and Differential Diagnosis. *Archives of Pathology & Laboratory Medicine*, 146(3):386–395, July 2021. ISSN 0003-9985. doi: 10.5858/arpa.2020-0258-RA.

[158] Brian P Walcott, Brian V Nahed, Ahmed Mohyeldin, Jean-Valery Coumans, Kristopher T Kahle, and Manuel J Ferreira. Chordoma: Current concepts, management, and future directions. *The Lancet Oncology*, 13(2):e69–e76, February 2012. ISSN 1470-2045. doi:

10.1016/S1470-2045(11)70337-0.

[159] Yutaka Nibu, Diana S. José-Edwards, and Anna Di Gregorio. From Notochord Formation to Hereditary Chordoma: The Many Roles of *Brachyury*. *BioMed Research International*, 2013:e826435, March 2013. ISSN 2314-6133. doi: 10.1155/2013/826435.

[160] Youssef Yakkioui, Jacobus J. van Overbeeke, Remco Santegoeds, Manon van Engeland, and Yasin Temel. Chordoma: The entity. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1846(2):655–669, December 2014. ISSN 0304-419X. doi: 10.1016/j.bbcan.2014.07.012.

[161] Kyung-Suk Choi, Martin J. Cohn, and Brian D. Harfe. Identification of nucleus pulposus precursor cells and notochordal remnants in the mouse: Implications for disk degeneration and chordoma formation. *Developmental Dynamics*, 237(12):3953–3958, 2008. ISSN 1097-0177. doi: 10.1002/dvdy.21805.

[162] Chen Cao, Laurence A. Lemaire, Wei Wang, Peter H. Yoon, Yoolim A. Choi, Lance R. Parsons, John C. Matese, Wei Wang, Michael Levine, and Kai Chen. Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature*, 571(7765):349–354, July 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1385-y.

[163] Hanna L. Sladitschek, Ulla-Maj Fiuza, Dinko Pavlinic, Vladimir Benes, Lars Hufnagel, and Pierre A. Neveu. MorphoSeq: Full Single-Cell Transcriptome Dynamics Up to Gastrulation in a Chordate. *Cell*, 181(4):922–935.e21, May 2020. ISSN 00928674. doi: 10.1016/j.cell.2020.03.055.

[164] Tengjiao Zhang, Yichi Xu, Kaoru Imai, Teng Fei, Guilin Wang, Bo Dong, Tianwei Yu, Yutaka Satou, Weiyang Shi, and Zhirong Bao. A single-cell analysis of the molecular lineage of chordate embryogenesis. *Science Advances*, 6(45):eabc4773, November 2020. doi: 10.1126/sciadv.abc4773.

[165] Garth R. Ilsley, Ritsuko Suyama, Takeshi Noda, Nori Satoh, and Nicholas M. Luscombe. Finding cell-specific expression patterns in the early Ciona embryo with single-cell RNA-seq. *Scientific Reports*, 10(1):4961, March 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-61591-1.

[166] Wei Wang, Xiang Niu, Tim Stuart, Estelle Jullian, William M. Mauck, Robert G. Kelly, Rahul Satija, and Lionel Christiaen. A single-cell transcriptional roadmap for cardiopharyngeal fate diversification. *Nature Cell Biology*, 21(6):674–686, June 2019. ISSN 1476-4679. doi: 10.1038/s41556-019-0336-z.

[167] Takeo Horie, Ryoko Horie, Kai Chen, Chen Cao, Masashi Nakagawa, Takehiro G. Kusakabe, Noriyuki Satoh, Yasunori Sasakura, and Michael Levine. Regulatory cocktail for dopaminergic neurons in a protovertebrate identified by whole-embryo single-cell transcriptomics. *Genes & Development*, 32(19-20):1297–1302, October 2018. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.317669.118.

[168] Jingjing Wang, Huiyu Sun, Mengmeng Jiang, Jiaqi Li, Peijing Zhang, Haide Chen, Yuqing Mei, Lijiang Fei, Shujing Lai, Xiaoping Han, Xinhui Song, Suhong Xu, Ming Chen, Hongwei Ouyang, Dan Zhang, Guo-Cheng Yuan, and Guoji Guo. Tracing cell-type evolution by cross-species comparison of cell atlases. *Cell Reports*, 34(9):108803, March 2021. ISSN 2211-1247. doi: 10.1016/j.celrep.2021.108803.

[169] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, February 2018. ISSN 1474-760X. doi: 10.1186/s13059-017-1382-0.

[170] Justine Dardaillon, Delphine Dauga, Paul Simion, Emmanuel Faure, Takeshi A Onuma, Melissa B DeBiasse, Alexandra Louis, Kazuhiro R Nitta, Magali Naville, Lydia Besnardeau, Wendy Reeves, Kai Wang, Marie Fagotto, Marion Guéroult-Bellone, Shigeki Fujiwara, Rémi Dumollard, Michael Veeman, Jean-Nicolas Volff, Hugues Roest Crollius, Emmanuel Douzery, Joseph F Ryan, Bradley Davidson, Hiroki Nishida, Christelle Dantec, and Patrick Lemaire. ANISEED 2019: 4D exploration of genetic data for an extended range of tunicates. *Nucleic Acids Research*, 48(D1):D668–D675, January 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz955.

[171] Hirohito Miura, Masako Yanazawa, Kentaro Kato, and Kunio Kitamura. Expression of a novel aristaless related homeobox gene 'Arx' in the vertebrate telencephalon, diencephalon and floor plate. *Mechanisms of Development*, 65(1):99–109, July 1997. ISSN 0925-4773. doi: 10.1016/S0925-4773(97)00062-2.

[172] Kunio Kitamura, Masako Yanazawa, Noriyuki Sugiyama, Hirohito Miura, Akiko Iizuka-Kogo, Masatomo Kusaka, Kayo Omichi, Rika Suzuki, Yuko Kato-Fukui, Kyoko Kamiirisa, Mina Matsuo, Shin-ichi Kamijo, Megumi Kasahara, Hidefumi Yoshioka, Tsutomu Ogata, Takayuki Fukuda, Ikuko Kondo, Mitsuhiro Kato, William B. Dobyns, Minesuke Yokoyama, and Ken-ichirou Morohashi. Mutation of ARX causes abnormal development of forebrain and testes in mice and X-linked lissencephaly with abnormal genitalia in humans. *Nature Genetics*, 32(3):359–369, November 2002. ISSN 1546-1718. doi: 10.1038/ng1009.

[173] Youngshin Lim, Il-Taeg Cho, Xiuyu Shi, Judith B. Grinspan, Ginam Cho, and Jeffrey A. Golden. Arx Expression Suppresses Ventralization of the Developing Dorsal Forebrain. *Scientific Reports*, 9(1):226, January 2019. ISSN 2045-2322. doi: 10.1038/s41598-018-36194-6.

[174] Carl T. Fulp, Ginam Cho, Eric D. Marsh, Ilya M. Nasrallah, Patricia A. Labosky, and Jeffrey A. Golden. Identification of Arx transcriptional targets in the developing basal forebrain. *Human Molecular Genetics*, 17(23):3740–3760, December 2008. ISSN 0964-6906. doi: 10.1093/hmg/ddn271.

[175] Linn Fagerberg, Björn M. Hallström, Per Oksvold, Caroline Kampf, Dijana Djureinovic, Jacob Odeberg, Masato Habuka, Simin Tahmasebpoor, Angelika Danielsson, Karolina Edlund, Anna Asplund, Evelina Sjöstedt, Emma Lundberg, Cristina Al-Khalili Szigyarto, Marie Skogs, Jenny Ottosson Takanen, Holger Berling, Hanna Tegel, Jan Mulder, Peter Nilsson, Jochen M. Schwenk, Cecilia Lindskog, Frida Danielsson, Adil Mardinoglu, Åsa

Sivertsson, Kalle von Feilitzen, Mattias Forsberg, Martin Zwahlen, IngMarie Olsson, Sanjay Navani, Mikael Huss, Jens Nielsen, Fredrik Ponten, and Mathias Uhlén. Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Molecular & Cellular Proteomics : MCP*, 13(2):397–406, February 2014. ISSN 1535-9476. doi: 10.1074/mcp.M113.035600.

[176] Lionel Christiaen, Eileen Wagner, Weiyang Shi, and Michael Levine. Isolation of sea squirt (Ciona) gametes, fertilization, dechorionation, and development. *Cold Spring Harbor Protocols*, 2009(12):pdb.prot5344, December 2009. ISSN 1559-6095. doi: 10.1101/pdb.prot5344.

[177] Jeni Beh, Weiyang Shi, Mike Levine, Brad Davidson, and Lionel Christiaen. FoxF is essential for FGF-induced migration of heart progenitor cells in the ascidian Ciona intestinalis. *Development (Cambridge, England)*, 134(18):3297–3305, September 2007. ISSN 0950-1991. doi: 10.1242/dev.010140.

[178] Tetsuro Ikuta and Hidetoshi Saiga. Dynamic change in the expression of developmental genes in the ascidian central nervous system: Revisit to the tripartite model and the origin of the midbrain-hindbrain boundary region. *Developmental Biology*, 312(2):631–643, December 2007. ISSN 1095-564X. doi: 10.1016/j.ydbio.2007.10.005.

[179] Lionel Christiaen, Eileen Wagner, Weiyang Shi, and Michael Levine. Whole-mount in situ hybridization on sea squirt (Ciona intestinalis) embryos. *Cold Spring Harbor Protocols*, 2009(12):pdb.prot5348, December 2009. ISSN 1559-6095. doi: 10.1101/pdb.prot5348.

[180] Alberto Stolfi, Elijah K Lowe, Claudia Racioppi, Filomena Ristoratore, C Titus Brown, Billie J Swalla, and Lionel Christiaen. Divergent mechanisms regulate conserved cardiopharyngeal development and gene expression in distantly related ascidians. *eLife*, 3:e03728, 2014. ISSN 2050-084X. doi: 10.7554/eLife.03728.

[181] Samuel L. Wolock, Romain Lopez, and Allon M. Klein. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*, 8(4):281–291.e9, April 2019. ISSN 2405-4720. doi: 10.1016/j.cels.2018.11.005.

[182] Kamil Slowikowski, John Arevalo, and Jonathan Manning. Slowkow/harmonypy: Harmonypy version 0.0.9. Zenodo, November 2022.

[183] Teresa K Attwood, Sarah Blackford, Michelle D Brazas, Angela Davies, and Maria Victoria Schneider. A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, 20(2):398–404, March 2019. ISSN 1477-4054. doi: 10.1093/bib/bbx100.

[184] Pavel Pevzner and Ron Shamir. Computing Has Changed Biology—Biology Education Must Catch Up. *Science*, 325(5940):541–542, July 2009. doi: 10.1126/science.1173876.

[185] Amir Rubinstein and Benny Chor. Computational Thinking in Life Science Education.

PLOS Computational Biology, 10(11):e1003897, November 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003897.

[186] Tin Wee Tan, Shen Jean Lim, Asif M. Khan, and Shoba Ranganathan. A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the "-omics" era. *BMC Genomics*, 10(3):S36, December 2009. ISSN 1471-2164. doi: 10.1186/1471-2164-10-S3-S36.

[187] Longbing Cao. Data Science: A Comprehensive Overview. *ACM Computing Surveys*, 50 (3):43:1–43:42, June 2017. ISSN 0360-0300. doi: 10.1145/3076253.

[188] Lise Getoor. Responsible Data Science. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 1, New York, NY, USA, June 2019. Association for Computing Machinery. ISBN 978-1-4503-5643-5. doi: 10.1145/3299869.3314117.

[189] Yingqian Ada Zhan, Charles Gregory Wray, Sandeep Namburi, Spencer T. Glantz, Reinhard Laubenbacher, and Jeffrey H. Chuang. Fostering bioinformatics education through skill development of professors: Big Genomic Data Skills Training for Professors. *PLOS Computational Biology*, 15(6):e1007026, June 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007026.

[190] Carla E. Brodley. Why universities must resist GPA-based enrollment caps in the face of surging enrollments. *Communications of the ACM*, 65(8):20–22, August 2022. ISSN 0001-0782, 1557-7317. doi: 10.1145/3544547.

[191] Adams Nager and Robert D. Atkinson. The Case for Improving U.S. Computer Science Education. *SSRN Electronic Journal*, 2016. ISSN 1556-5068. doi: 10.2139/ssrn.3066335.

[192] Esther Shein. The CS teacher shortage. *Communications of the ACM*, 62(10):17–18, September 2019. ISSN 0001-0782, 1557-7317. doi: 10.1145/3355375.

[193] Tracy Camp, Stu Zweben, Ellen Walker, and Lecia Barker. Booming Enrollments: Good Times? In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, SIGCSE '15, pages 80–81, New York, NY, USA, February 2015. Association for Computing Machinery. ISBN 978-1-4503-2966-8. doi: 10.1145/2676723.2677333.

[194] Shanna Jaggars, John Fink, and Jeffrey Fletcher. A Longitudinal Analysis of Community College Pathways to Computer Science Bachelor's Degrees. Technical report, 2016.

[195] Rolf Backofen and David Gilbert. Bioinformatics and Constraints. In *Foundations of Artificial Intelligence*, volume 2, pages 905–944. Elsevier, 2006. ISBN 978-0-444-52726-4. doi: 10.1016/S1574-6526(06)80030-1.

[196] C. M. Charles. *Introduction to Educational Research. Third Edition.* Addison Wesley Longman, Inc, 1998. ISBN 978-0-8013-1872-6.

[197] National Research Council (U.S.), Lisa Towne, Lauress L. Wise, and Tina M. Winters, editors. *Advancing Scientific Research in Education*. National Academies Press, Washington, DC, 2005. ISBN 978-0-309-09321-7 978-0-309-54598-3.

[198] Stephanie J. Slater, Timothy F. Slater, Inge Heyer, and Janelle M. Bailey. *Discipline-Based Education Research: A Guide for Scientists*. Pono Publishing, 2015. ISBN 978-1-5150-2456-9.

[199] Justus J. Randolph, George Julnes, Erkki Sutinen, and Steve Lehman. A Methodological Review of Computer Science Education Research. *Journal of Information Technology Education: Research*, 7(1):135–162, January 2008. ISSN 1539-3585.

[200] Vicki L Almstrum and Orit Hazzan. Challenges to computer science education research. In *SIGCSE'05*, page 2, St. Louis, Missouri, USA., February 2005. ACM.

[201] Steve Cooper, Shuchi Grover, Mark Guzdial, and Beth Simon. A future for computing education research. *Communications of the ACM*, 57(11):34–36, October 2014. ISSN 0001-0782, 1557-7317. doi: 10.1145/2668899.

[202] Arnold Pears, Stephen Seidman, Crystal Eney, Päivi Kinnunen, and Lauri Malmi. Constructing a core literature for computing education research. *ACM SIGCSE Bulletin*, 37 (4):152–161, December 2005. ISSN 0097-8418. doi: 10.1145/1113847.1113893.

[203] Lauri Malmi, Judy Sheard, Simon, Roman Bednarik, Juha Helminen, Ari Korhonen, Niko Myller, Juha Sorva, and Ahmad Taherkhani. Characterizing research in computing education: A preliminary analysis of the literature. In *Proceedings of the Sixth International Workshop on Computing Education Research*, ICER '10, pages 3–12, New York, NY, USA, August 2010. Association for Computing Machinery. ISBN 978-1-4503-0257-9. doi: 10.1145/1839594.1839597.

[204] Mark Guzdial. Learner-Centered Design of Computing Education: Research on Computing for Everyone. *Synthesis Lectures on Human-Centered Informatics*, 8(6):1–165, November 2015. ISSN 1946-7680. doi: 10.2200/S00684ED1V01Y201511HCI033.

[205] M. Bahar, A. H. Johnstone, and M. H. Hansell. Revisiting learning difficulties in biology. *Journal of Biological Education*, 33(2):84–86, March 1999. ISSN 0021-9266. doi: 10.1080/00219266.1999.9655648.

[206] Joel J. Mintzes, James H. Wandersee, and Joseph D. Novak. Assessing understanding in biology. *Journal of Biological Education*, 35(3):118–124, June 2001. ISSN 0021-9266. doi: 10.1080/00219266.2001.9655759.

[207] Jay B. Labov, Ann H. Reid, and Keith R. Yamamoto. Integrated Biology and Undergraduate Science Education: A New Biology Education for the Twenty-First Century? *CBE—Life Sciences Education*, 9(1):10–16, March 2010. doi: 10.1187/cbe.09-12-0092.

[208] Md Zahidul I Pranjol, Paolo Oprandi, and Sarah Watson. Project-based learning in biomedical sciences: Using the collaborative creation of revision resources to consolidate knowledge, promote cohort identity and develop transferable skills. *Journal of Biological Education*, 0(0):1–17, November 2022. ISSN 0021-9266. doi: 10.1080/00219266.2022.2147576.

[209] Sara E. Brownell and Matthew J. Kloser. Toward a conceptual framework for measuring the effectiveness of course-based undergraduate research experiences in undergraduate biology. *Studies in Higher Education*, 40(3):525–544, March 2015. ISSN 0307-5079. doi: 10.1080/03075079.2015.1004234.

[210] Ashley B. Heim and Emily A. Holt. Benefits and Challenges of Instructing Introductory Biology Course-Based Undergraduate Research Experiences (CUREs) as Perceived by Graduate Teaching Assistants. *CBE—Life Sciences Education*, 18(3):ar43, September 2019. doi: 10.1187/cbe.18-09-0193.

[211] Arundhati Bakshi, Lorelei E. Patrick, and E. William Wischusen. A Framework for Implementing Course-Based Undergraduate Research Experiences (CUREs) in Freshman Biology Labs. *The American Biology Teacher*, 78(6):448–455, August 2016. ISSN 0002-7685. doi: 10.1525/abt.2016.78.6.448.

[212] Nicola Mulder, Russell Schwartz, Michelle D. Brazas, Cath Brooksbank, Bruno Gaeta, Sarah L. Morgan, Mark A. Pauley, Anne Rosenwald, Gabriella Rustici, Michael Sierk, Tandy Warnow, and Lonnie Welch. The development and application of bioinformatics core competencies to improve bioinformatics training and education. *PLOS Computational Biology*, 14(2):e1005772, February 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005772.

[213] Martin Weller. *The Battle for Open*. Ubiquity Press, November 2014. ISBN 978-1-909188-34-1 978-1-909188-36-5 978-1-909188-33-4 978-1-909188-35-8. doi: 10.5334/bam.

[214] Sanjaya Mishra. Open educational resources: Removing barriers from within. *Distance Education*, 38(3):369–380, September 2017. ISSN 0158-7919. doi: 10.1080/01587919.2017.1369350.

[215] Dr Jan Hylén. Open Educational Resources: Opportunities and Challenges. page 10.

[216] Maimoona Al Abri and Nada Dabbagh. Open Educational Resources: A Literature Review. *Journal of Mason Graduate Research*, 6(1):83–104, 2018. ISSN 2327-0764. doi: 10.13021/G8jmgr.v6i1.2386.

[217] Nicholas B. Colvard, C. Edward Watson, and Hyojin Park. The Impact of Open Educational Resources on Various Student Success Metrics. *International Journal of Teaching and Learning in Higher Education*, 30(2):262–276, 2019.

[218] Omer Faruk Islim, Nergis A. Gurel Koybasi, and Kursat Cagiltay. Use of Open Educational Resources: How, Why and Why Not? *International Journal of Teaching and Learning in*

Higher Education, 28(2):230–240, 2016.

[219] Christine Geith and Karen Vignare. Access to Education with Online Learning and Open Educational Resources: Can They Close the Gap? *Journal of Asynchronous Learning Networks*, 12(1):105–126, February 2008. ISSN 1939-5256.

[220] Robert L. Moore and Stephanie J. Blackmon. From the learner's perspective: A systematic review of MOOC learner experiences (2008–2021). *Computers & Education*, 190:104596, December 2022. ISSN 0360-1315. doi: 10.1016/j.compedu.2022.104596.

[221] Petr Danecek, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, Thomas Keane, Shane A. McCarthy, Robert M. Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, February 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab008.

[222] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, January 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts635.

[223] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021. ISSN 0092-8674. doi: 10.1016/j.cell.2021.04.048.

[224] Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, September 2008. ISSN 1474-760X. doi: 10.1186/gb-2008-9-9-r137.

[225] Jennifer Wang, Hai Hong, Jason Ravitz, and Sepehr Hejazi Moghadam. Landscape of K-12 Computer Science Education in the U.S.: Perceptions, Access, and Barriers. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE '16, pages 645–650, New York, NY, USA, February 2016. Association for Computing Machinery. ISBN 978-1-4503-3685-7. doi: 10.1145/2839509.2844628.

[226] Jane G. Stout and Jennifer M. Blaney. "But it doesn't come naturally": How effort expenditure shapes the benefit of growth mindset on women's sense of intellectual belonging in computing. *Computer Science Education*, 27(3-4):215–228, October 2017. ISSN 0899-3408. doi: 10.1080/08993408.2018.1437115.

[227] Enora R. Brown. FREEDOM FOR SOME, DISCIPLINE FOR "OTHERS": The Structure of Inequity in Education. In *Education as Enforcement*. Routledge, second edition, 2010.

ISBN 978-0-203-84322-2.

[228] Peggy C. Kirby, Jeffrey Oescher, Dave Wilson, and Karen Smith-Gratto. Computers in schools: A new source of inequity. *Computers & Education*, 14(6):537–541, January 1990. ISSN 0360-1315. doi: 10.1016/0360-1315(90)90112-K.

[229] Aleata Hubbard Cheuoua. Confronting Inequities in Computer Science Education: A Case for Critical Theory. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, SIGCSE '21, pages 425–430, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8062-1. doi: 10.1145/3408877.3432453.

[230] Colleen M. Lewis and Niral Shah. How Equity and Inequity Can Emerge in Pair Programming. In *Proceedings of the Eleventh Annual International Conference on International Computing Education Research*, ICER '15, pages 41–50, New York, NY, USA, August 2015. Association for Computing Machinery. ISBN 978-1-4503-3630-7. doi: 10.1145/2787622.2787716.

[231] Sylvia E. Rogers. Bridging the 21st Century Digital Divide. *TechTrends*, 60(3):197–199, May 2016. ISSN 1559-7075. doi: 10.1007/s11528-016-0057-0.

[232] Niral Shah, Colleen Lewis, and Roxane Caires. Analyzing Equity in Collaborative Learning Situations: A Comparative Case Study in Elementary Computer Science. In *ICLS 2014*. Boulder, CO: International Society of the Learning Sciences, June 2014.

[233] Kristen Shinohara, Michael McQuaid, and Nayeri Jacobo. Access Differential and Inequitable Access: Inaccessibility for Doctoral Students in Computing. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, pages 1–12, New York, NY, USA, October 2020. Association for Computing Machinery. ISBN 978-1-4503-7103-2. doi: 10.1145/3373625.3416989.

[234] Claude M. Steele, Steven J. Spencer, and Joshua Aronson. Contending with group image: The psychology of stereotype and social identity threat. In *Advances in Experimental Social Psychology*, volume 34, pages 379–440. Academic Press, January 2002. doi: 10.1016/S0065-2601(02)80009-0.

[235] Dustin B. Thoman, Jessi L. Smith, Elizabeth R. Brown, Justin Chase, and Joo Young K. Lee. Beyond Performance: A Motivational Experiences Model of Stereotype Threat. *Educational Psychology Review*, 25(2):211–243, June 2013. ISSN 1573-336X. doi: 10.1007/s10648-013-9219-1.

[236] Cary Stacy Smith and Li-Ching Hung. Stereotype threat: Effects on education. *Social Psychology of Education*, 11(3):243–257, August 2008. ISSN 1573-1928. doi: 10.1007/s11218-008-9053-3.

[237] Elizabeth A. Eschenbach, Mary Virnoche, Eileen M. Cashman, Susan M. Lord, and Michelle Madsen Camacho. Proven practices that can reduce stereotype threat in engineering

education: A literature review. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, pages 1–9, October 2014. doi: 10.1109/FIE.2014.7044011.

[238] Amy E. Bell, Steven J. Spencer, Emma Iserman, and Christine E.r. Logel. Stereotype Threat and Women's Performance in Engineering. *Journal of Engineering Education*, 92 (4):307–312, 2003. ISSN 2168-9830. doi: 10.1002/j.2168-9830.2003.tb00774.x.

[239] Amruth N. Kumar. A study of stereotype threat in computer science. In *Proceedings of the 17th ACM Annual Conference on Innovation and Technology in Computer Science Education*, ITiCSE '12, pages 273–278, New York, NY, USA, July 2012. Association for Computing Machinery. ISBN 978-1-4503-1246-2. doi: 10.1145/2325296.2325361.

[240] Phillip Hall Jr. and Kinnis Gosha. The Effects of Anxiety and Preparation on Performance in Technical Interviews for HBCU Computer Science Majors. In *Proceedings of the 2018 ACM SIGMIS Conference on Computers and People Research*, SIGMIS-CPR'18, pages 64–69, New York, NY, USA, June 2018. Association for Computing Machinery. ISBN 978-1-4503-5768-5. doi: 10.1145/3209626.3209707.

[241] Katrina Falkner, Claudia Szabo, Dee Michell, Anna Szorenyi, and Shantel Thyer. Gender Gap in Academia: Perceptions of Female Computer Science Academics. In *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*, ITiCSE '15, pages 111–116, New York, NY, USA, June 2015. Association for Computing Machinery. ISBN 978-1-4503-3440-2. doi: 10.1145/2729094.2742595.

[242] Adam Rosenstein, Aishma Raghu, and Leo Porter. Identifying the Prevalence of the Impostor Phenomenon Among Computer Science Students. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, SIGCSE '20, pages 30–36, New York, NY, USA, February 2020. Association for Computing Machinery. ISBN 978-1-4503-6793-6. doi: 10.1145/3328778.3366815.

[243] Women, Minorities, and Persons with Disabilities in Science and Engineering. Technical report, National Center for Science and Engineering Statistics (NCSES), Alexandria, VA, 2019.

[244] Aaron Hochanadel and Dora Finamore. Fixed And Growth Mindset In Education And How Grit Helps Students Persist In The Face Of Adversity. *Journal of International Education Research (JIER)*, 11(1):47–50, January 2015. ISSN 2158-0987. doi: 10.19030/jier.v11i1.9099.

[245] Emily Rhew, Jody S. Piro, Pauline Goolkasian, and Patricia Cosentino. The effects of a growth mindset on self-efficacy and motivation. *Cogent Education*, 5(1):1492337, January 2018. ISSN null. doi: 10.1080/2331186X.2018.1492337.

[246] Patricia Morreale, J. Jenny Li, Ching-Yu Huang, Daehan Kwak, Jean Chu, Yulia Kumar, and Paolien Wang. Framework for a Growth Mindset Classroom. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, SIGCSE '21, page 1269, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN

978-1-4503-8062-1. doi: 10.1145/3408877.3439631.

[247] Kelly Rivers, Erik Harpstead, and Ken Koedinger. Learning Curve Analysis for Programming: Which Concepts do Students Struggle With? In *Proceedings of the 2016 ACM Conference on International Computing Education Research*, pages 143–151, Melbourne VIC Australia, August 2016. ACM. ISBN 978-1-4503-4449-4. doi: 10.1145/2960310.2960333.

[248] Miranda Parker, Colleen Lewis, Harvey Mudd College, and Platt Blvd. WHAT MAKES BIG-O ANALYSIS DIFFICULT: UNDERSTANDING HOW STUDENTS UNDERSTAND RUNTIME ANALYSIS. In *Consortium for Computing Sciences in Colleges, Southwestern Conference*, page 11, Los Angeles, California, 2014.

[249] Yizhou Qian and James Lehman. Students' Misconceptions and Other Difficulties in Introductory Programming: A Literature Review. *ACM Transactions on Computing Education*, 18(1):1:1–1:24, October 2017. doi: 10.1145/3077618.

[250] Sue Sentance and Andrew Csizmadia. Teachers' perspectives on successful strategies for teaching Computing in school. *IFIP TC3 Working Conference 2015: A New Culture of Learning: Computing and Next Generations*, July 2015.

[251] Reuven Lazarowitz and Sofia Penso. High school students' difficulties in learning biology concepts. *Journal of Biological Education*, 26(3):215–223, September 1992. ISSN 0021-9266. doi: 10.1080/00219266.1992.9655276.

[252] Roisin F. Kelly-Laubscher and Kathy Luckett. Differences in Curriculum Structure between High School and University Biology: The Implications for Epistemological Access. *Journal of Biological Education*, 50(4):425–441, October 2016. ISSN 0021-9266. doi: 10.1080/00219266.2016.1138991.

[253] Trevor R. Anderson and Nancy J. Pelaez. Implementing Innovations in Undergraduate Biology Experimentation Education. In Nancy J. Pelaez, Stephanie M. Gardner, and Trevor R. Anderson, editors, *Trends in Teaching Experimentation in the Life Sciences: Putting Research into Practice to Drive Institutional Change*, Contributions from Biology Education Research, pages 547–555. Springer International Publishing, Cham, 2022. ISBN 978-3-030-98592-9. doi: 10.1007/978-3-030-98592-9_25.

[254] Mark A. McDaniel, Michael J. Cahill, Regina F. Frey, Lisa B. Limeri, and Paula P. Lemons. Learning Introductory Biology: Students' Concept-Building Approaches Predict Transfer on Biology Exams. *CBE—Life Sciences Education*, 21(4):ar65, December 2022. doi: 10.1187/cbe.21-12-0335.

[255] Marta M. Koć-Januchta, Konrad J. Schönborn, Casey Roehrig, Vinay K. Chaudhri, Lena A. E. Tibell, and H. Craig Heller. "Connecting concepts helps put main ideas together": Cognitive load and usability in learning biology with an AI-enriched textbook. *International Journal of Educational Technology in Higher Education*, 19(1):11, March 2022. ISSN 2365-9440. doi: 10.1186/s41239-021-00317-3.

[256] Daniel E. Ehrmann, Sara N. Gallant, Sujay Nagaraj, Sebastian D. Goodfellow, Danny Eytan, Anna Goldenberg, and Mjaye L. Mazwi. Evaluating and reducing cognitive load should be a priority for machine learning in healthcare. *Nature Medicine*, 28(7):1331–1333, July 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01833-z.

[257] Julia R. Fox, Byungho Park, and Annie Lang. When Available Resources Become Negative Resources: The Effects of Cognitive Overload on Memory Sensitivity and Criterion Bias. *Communication Research*, 34(3):277–296, June 2007. ISSN 0093-6502. doi: 10.1177/0093650207300429.

[258] David Kirsh. A Few Thoughts on Cognitive Overload. *Intellectica*, 1(30):19–51, 2000.

[259] Paulo Freire. *Pedagogy of the Oppressed*. Continuum, New York, 30th anniversary edition edition, 1970. ISBN 978-0-8264-1276-8.

[260] Cheryl A. Estes. Promoting Student-Centered Learning in Experiential Education. *Journal of Experiential Education*, 27(2):141–160, September 2004. ISSN 1053-8259. doi: 10.1177/105382590402700203.

[261] Gloria Brown Wright. Student-Centered Learning in Higher Education. *International Journal of Teaching and Learning in Higher Education*, 23(1):92–97, 2011. ISSN 1812-9129.

[262] Thomas Brush and John Saye. Implementation and evaluation of a student-centered learning unit: A case study. *Educational Technology Research and Development*, 48(3): 79–100, September 2000. ISSN 1556-6501. doi: 10.1007/BF02319859.

[263] S. Senthamarai. Interactive teaching strategies. *Journal of Applied and Advanced Research*, pages S36–S38, May 2018. ISSN 2519-9412. doi: 10.21839/jaar.2018.v3iS1.166.

[264] S. Kennewell, H. Tanner, S. Jones, and G. Beauchamp. Analysing the use of interactive technology to implement interactive teaching. *Journal of Computer Assisted Learning*, 24 (1):61–73, 2008. ISSN 1365-2729. doi: 10.1111/j.1365-2729.2007.00244.x.

[265] Roxana Marachi and Lawrence Quill. The case of Canvas: Longitudinal datafication through learning management systems. *Teaching in Higher Education*, 25(4):418–434, May 2020. ISSN 1356-2517. doi: 10.1080/13562517.2020.1739641.

[266] Yoany Beldarrain. Distance Education Trends: Integrating new technologies to foster student interaction and collaboration. *Distance Education*, 27(2):139–153, August 2006. ISSN 0158-7919. doi: 10.1080/01587910600789498.

[267] Anatalia N. Endozo, Solomon Oluyinka, and Richard G. Daenos. Teachers' Experiences towards Usage of Learning Management System: CANVAS. In *Proceedings of the 2019 11th International Conference on Education Technology and Computers*, ICETC 2019, pages 91–95, New York, NY, USA, January 2020. Association for Computing Machinery. ISBN 978-1-4503-7254-1. doi: 10.1145/3369255.3369257.

[268] Hakan Polat and Songül Karabatak. Effect of flipped classroom model on academic achievement, academic satisfaction and general belongingness. *Learning Environments Research*, 25(1):159–182, April 2022. ISSN 1573-1855. doi: 10.1007/s10984-021-09355-0.

[269] Taotao Long, John Cummins, and Michael Waugh. Use of the flipped classroom instructional model in higher education: Instructors' perspectives. *Journal of Computing in Higher Education*, 29(2):179–200, August 2017. ISSN 1867-1233. doi: 10.1007/s12528-016-9119-8.

[270] Keengwe Jared. *Promoting Active Learning through the Flipped Classroom Model*. IGI Global, January 2014. ISBN 978-1-4666-4988-0.

[271] David S. Jones. History in a Crisis — Lessons for Covid-19. *New England Journal of Medicine*, 382(18):1681–1683, April 2020. ISSN 0028-4793. doi: 10.1056/NEJMp2004361.

[272] Calliope Holingue, Luther G. Kalb, Kira E. Riehm, Daniel Bennett, Arie Kapteyn, Cindy B. Veldhuis, Renee M. Johnson, M. Daniele Fallin, Frauke Kreuter, Elizabeth A. Stuart, and Johannes Thrul. Mental Distress in the United States at the Beginning of the COVID-19 Pandemic. *American Journal of Public Health*, 110(11):1628–1634, November 2020. ISSN 0090-0036. doi: 10.2105/AJPH.2020.305857.

[273] Yi-Chi Wu, Ching-Sung Chen, and Yu-Jiun Chan. The outbreak of COVID-19: An overview. *Journal of the Chinese Medical Association*, 83(3):217–220, March 2020. ISSN 1726-4901. doi: 10.1097/JCMA.0000000000000270.