

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Elucidating Mechanisms Governing Enhancer Function in Development and Disease

Permalink

<https://escholarship.org/uc/item/66f0w896>

Author

Solvason, Jesse Joseph

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Elucidating Mechanisms Governing Enhancer Function in Development and Disease

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

in

Bioinformatics & Systems Biology

by

Jesse Joseph Solvason

Committee in charge:

Professor Emma Farley, Chair
Professor Hannah Carter, Co-Chair
Professor Joseph Ecker
Professor Kelly Frazer
Professor Michael Perry

2024

Copyright

Jesse Joseph Solvason, 2024

All rights reserved.

The Dissertation of Jesse Joseph Solvason is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To my parents Nanette and Brent Solvason,
And to my brother Stefan Solvason.
I would not have made this journey without their guidance and support.

To my one-of-a-kind grandmother Edna Sargeant. From my taste in music,
to my sense of humor, to my work ethic, to my personal values,
I would not be the same without her.

To my life-long friends who maintained our friendships despite great distances:
Madi J, Will S, Steven C, Ryan W, Joseph H, Arisa C, Mariah R,
Marshall W, Hiroki B, The Brothers Choi, and Fabian L.

To my previous mentors who contributed, each in their own unique way,
to my development as an independent scientist.

A scientist capable of completing a PhD, posing the next question,
strategizing the approach, executing a well-controlled experiment,
and performing careful analysis:
From Chimera Bioengineering - Melissa F, Gus Z, and Ben W.
From NASA Ames Research Center - George C and Andro R.

To my lab mates who contributed to a positive and supportive lab culture,
we persevered together and supported each other in pursuit of our own personal goals.

To the undergrad research interns Simran Jandu and Maggie Ma whom I advised,
it was a pleasure watching you learn and grow as scientists.

EPIGRAPH

I'm not listening to you. You only believe in Science. That's probably why we never win.

- *Ignacio Libre*

Truth suffers from too much analysis.

- *Ancient Fremen Saying*

TABLE OF CONTENTS

Dissertation Approval Page	iii
Dedication	iv
Epigraph	v
Table of Contents	vi
List of Figures	ix
Acknowledgements	xii
Vita.....	xiii
Abstract of the Dissertation	xiv
Introduction.....	1
0.1 Enhancers drive traits and diseases	1
0.2 Methodologies to measure transcription factor binding affinity	3
0.3 Utilizing the low-affinity regulatory principle in causal variant prediction	5
0.4 Broadening our repertoire of enhancer regulatory principles	6
0.5 Epigenomics studies of enhancer grammar	8
0.6 Sequence based deep learning approaches to design tissue-specific enhancers	11
0.7 Synthetic screens to study enhancer grammar	12
0.8 Design considerations of targeted and controlled synthetic screens to study enhancer grammar	13
Chapter 1 Single-nucleotide variants within heart enhancers increase binding affinity and disrupt heart development	15
1.1 Abstract	15
1.2 Introduction	15
1.3 Results.....	17
1.3.1 FoxF is expressed within the TVCs and triggers migration of the TVCs to the ventral midline	17
1.3.2 Low-affinity ETS sites are necessary for tissue-specific expression	17
1.3.3 Affinity-optimizing SNVs lead to loss of tissue-specific expression	21
1.3.4 Affinity-optimizing SNVs cause migration defects and disrupt heart development	24
1.3.5 Low-affinity ETS sites are prevalent in putative DHEs across chordates, and these enhancers are vulnerable to affinity-optimizing SNVs	27
1.3.6 Human and mouse Gata4 heart enhancers contain low-affinity ETS sites .	28
1.3.7 Affinity-optimizing SNVs within the human GATA4 G9 enhancer cause GOF expression in human iPSC-derived cardiomyocytes	30
1.4 Discussion	30
1.4.1 Low-affinity ETS sites are a prevalent feature of DHEs.....	30

1.4.2	Affinity-optimizing SNVs cause GOF enhancer activity, which can contribute to phenotypes.....	30
1.4.3	Affinity-optimizing SNVs may contribute to evolution of novel traits.....	31
1.4.4	Violations in regulatory principles governing enhancers can pinpoint causal enhancer variants	31
1.4.5	Limitations of the study	32
1.5	Methods	33
1.5.1	Tunicates	33
1.5.2	iPSC-derived cardiomyocytes	33
1.5.3	Mutagenesis and cloning	34
1.5.4	Clonal Electroporation.....	34
1.5.5	Electrophoretic mobility shift assay	36
1.5.6	Scoring Relative Affinities	36
1.5.7	Computational Enhancer Analysis	37
1.5.8	Enhancer Library	38
1.5.9	Data Availability	39
1.5.10	Code Availability	39
1.5.11	Acknowledgements	40
1.5.12	Author Contributions.....	40
1.5.13	Competing Interests	40
Chapter 2	Affinity-optimizing enhancer variants disrupt development	41
2.1	Abstract	41
2.2	Introduction	42
2.3	Results.....	43
2.3.1	Redundant low-affinity ETS sites regulate the ZRS	43
2.3.2	Human polydactyly SNVs subtly increase affinity	47
2.3.3	Affinity-optimizing SNVs cause polydactyly	48
2.3.4	Predicting penetrance and severity	53
2.3.5	Affinity-optimizing SNVs prevalent across the ZRS	53
2.3.6	The enhanceosome contains affinity-optimizing SNVs	56
2.3.7	Other transcription factors and disease enhancers	56
2.3.8	Regulatory principles predict causal SNVs	60
2.4	Discussion	61
2.5	Methods	64
2.5.1	Mice	64
2.5.2	Calculation of binding affinity	69
2.5.3	Analysis of MPRA Data	69
2.5.4	Analysis of eQTL data.....	70
2.5.5	Statistics	71
2.5.6	Data Availability	71
2.5.7	Code Availability	72
2.5.8	Acknowledgements	72
2.5.9	Author Contributions.....	72
2.5.10	Competing Interests	72
Chapter 3	Dependency grammar of developmental enhancers.	73

3.1	Abstract	73
3.2	Introduction	74
3.3	Results.....	75
3.3.1	ETS and GATA are necessary and sufficient for neural enhancer activity	75
3.3.2	Testing every possible organization of Otx-a enhancer in developing embryos	77
3.3.3	50,000 active elements regulated by GATA and ETS.....	78
3.3.4	Not all clusters of sites are equal: Despite having the same sequence content, binding sites and affinities, only a fraction of elements are functional. ...	79
3.3.5	Syntax Contributes to Enhancer Activity	82
3.3.6	Single syntax changes convert inert elements into active enhancers	86
3.3.7	Searching for signatures of functional grammar	86
3.3.8	Syntax features and syntax scores with active enhancers.	88
3.3.9	Syntax score defines functional and non-active enhancers.	92
3.3.10	Genomic enhancers encode activity using dependency grammar	92
3.3.11	Designing tissue specific enhancers using grammar	93
3.4	Discussion	97
3.5	Methods	99
3.5.1	Ciona Electroporations	99
3.5.2	Scoring Relative Affinities of Binding Sites	100
3.5.3	MPRA Libraries	101
3.5.4	Analysis of grammar	104
3.5.5	Design of Synthetic Enhancers.....	106
3.5.6	Data Availability	106
3.5.7	Code Availability	106
3.5.8	Acknowledgements	106
3.5.9	Author Contributions.....	106
3.5.10	Competing Interests	107
	Epilogue	108
4.1	Discussion	108
4.1.1	The prevalence of low affinity sites creates a vulnerability in genomes	108
4.1.2	Towards increasing prediction accuracy of causal mutations using grammar	110
	Appendix A Supplemental Figures for Chapter 1	114
	Appendix B Supplemental Figures for Chapter 2	120
	Appendix C Supplemental Figures for Chapter 3	132
	Bibliography	148

LIST OF FIGURES

Figure 1.1.	Low-affinity ETS sites are necessary for tissue-specific expression.	20
Figure 1.2.	Affinity-optimizing SNVs lead to loss of tissue-specific expression.	23
Figure 1.3.	Affinity-optimizing SNVs cause migration defects and disrupt heart development.	26
Figure 1.4.	Low-affinity ETS binding sites are prevalent within developmental heart enhancers, and affinity-optimizing SNVs lead to gain-of-function expression in human iPSC-CMs.	29
Figure 2.1.	An ETS-A site in the ZRS enhancer contains two human variants that are associated with polydactyly, both of which subtly increase ETS binding affinity.	46
Figure 2.2.	Synthetic changes to the ETS-A site that create a 0.25-affinity site and a LOF site cause predicted phenotypes.	49
Figure 2.3.	All mice with the approximately 0.25-affinity ETS-A are indistinguishable in terms of the penetrance, laterality and severity of polydactyly.	51
Figure 2.4.	A greater increase in affinity at the ETS-A site causes more severe and penetrant polydactyly and long-bone defects.	52
Figure 2.5.	Affinity-optimizing SNVs drive GOF expression in the ZRS and IFNb enhancosome.	55
Figure 2.6.	Affinity-optimizing SNVs drive GOF expression in a wide variety of disease-associated enhancers.	59
Figure 3.1.	Otx-a neural enhancer is regulated by a common logic: GATA and ETS.	76
Figure 3.2.	Characterizing the activity of 460,800 versions of Otx-a enhancer.	81
Figure 3.3.	Robust and moderate syntaxes drive enhancer activity.	85
Figure 3.4.	Orientation of individual sites convert inert elements into active enhancers.	87
Figure 3.5.	Syntax features and syntax scores with active enhancers.	91
Figure 3.6.	Dependency grammar encode genomic and synthetic enhancers.	96
Figure A.1.	Counting data for ETS site ablation within the FoxF enhancer.	115
Figure A.2.	Representative images of each of the FoxF constructs tested in this study.	117
Figure A.3.	Differential binding affinity with ETS1 single nucleotide variant.	118

Figure A.4.	Counting of migration defects in tailbud embryos with affinity-optimizing mutations.	119
Figure B.1.	Class I ETS family members have conserved DBDs.	121
Figure B.2.	PBM binding affinities correlate with the in vivo ETS-1 ChIP signal in various cell types.	121
Figure B.3.	Conservation of ZRS ETS sites between humans and mice.	122
Figure B.4.	EMSA shows the binding of human and mouse ETS-1 to the ETS-A site and ETS-A variants.	123
Figure B.5.	Ptch1 in situ hybridization in the hindlimb bud and forelimb bud of transgenic mice.	125
Figure B.6.	Syn 0.52 mice show highly penetrant polydactyly in both forelimb and hindlimb, as well as tibial hemimelia.	127
Figure B.7.	Affinity-optimizing SNVs are significantly associated with GOF enhancer activity.	128
Figure B.8.	EMSA shows stronger binding of the human HOXA13 and HOXD13 DBDs to the Dutch 2 variant relative to the WT sequence.	129
Figure B.9.	Experimental details for MPRA performed with 11 disease-associated enhancers.	130
Figure B.10.	Affinity-optimizing eQTL variants are enriched in GOF target gene expression.	131
Figure C.1.	GATA and ETS are sufficient for the neural expression pattern of Otx-a. .	133
Figure C.2.	The GATA and ETS binding sites are necessary and sufficient to drive the Otx-a neural expression pattern.	134
Figure C.3.	GATA ETS elements validated to be neural enhancers by fluorescent microscopy.	135
Figure C.4.	GATA ETS elements validated to activate neural and ectopic expression by fluorescent microscopy.	136
Figure C.5.	GATA ETS elements validated to be inert sequences by fluorescent microscopy.	137
Figure C.6.	The GATA and ETS binding sites are the drivers of neural enhancer activity in the Otx-a grammar library members.	138

Figure C.7.	GATA and ETS are required for neural expression of Otx-a grammar library members	139
Figure C.8.	Examples of Otx-a grammar library members.	140
Figure C.9.	Sequences from robust and inert organizational families reproduce MPRA enhancer activity under the microscope.	142
Figure C.10.	In a single flip comparison, there are no sequence features which distinguish inert and active enhancers.	143
Figure C.11.	Single flips cause inert sequences to convert to neural enhancers.	144
Figure C.12.	No syntax features are significantly enriched in active enhancers after permuting enhancer activity and sequence labels.	145
Figure C.13.	Optimal spacing across doublet syntax features.	146
Figure C.14.	Percent of embryos with neural expression activated by synthetic neural enhancers.....	147
Figure C.15.	Syntax and affinity alone slightly enrich for active enhancers.	147

ACKNOWLEDGEMENTS

I acknowledge my advisor Professor Emma Farley, who stands a clear example of a careful rigorous scientist. She taught me how critically important perseverance is - as much as technical ability.

I acknowledge Dr. Granton Jindal, a postdoc in the lab. He trained me as a rotation student and as a junior grad student. He contributed in great ways to my development in the lab, and passed on helpful perspectives about working in the sciences.

I acknowledge co-chair Professor Hannah Carter and my committee members Professor Kelly Frazer, Professor Michael Perry and Professor Joe Ecker. Your feedback was critical in the development of my projects.

I acknowledge Adam Klie for many helpful conversations we had across countless projects in the 5 years we overlapped in our PhDs. It was a delight to work with him. I acknowledge Simran Jandu, an undergraduate research assistant in the lab. She helped develop the code for the computational toolkit "TF Sites" which is not included in this thesis.

Chapter 1, in full, is a reprint of the material as it appears in *Developmental Cell*, 2023. Granton Jindal, Alexis Bantle, Joe J. Solvason, Jessica L Grudzien, Agnieszka D'Antonio-Chronowska, Fabian Lim, Sophia H Le, Benjamin P Song, Michelle F Ragsac, Adam Klie, Reid O Larsen, Kelly A Frazer, & Emma K Farley. Thank you to all the co-authors for their contribution to the work, especially to Dr. Granton Jindal. Chapter 2, in full, is a reprint of the material as it appears in *Nature*, 2024. Fabian Lim, Joe J. Solvason, Genevieve E. Ryan, Sophia H. Le, Granton A. Jindal, Paige Steffen, Simran K. Jandu & Emma K. Farley. Thank you to all the co-authors for their contribution to the work, especially Fabian Limb and Dr. Genevieve Ryan. Chapter 3 is a study that is in review at the time this thesis was written. I acknowledge Fabian Lim (Co-first), Jessica Grudzien, Ben Song, Sophia Le, and Emma Farley.

VITA

- 2011 High School Diploma
Palo Alto High School
- 2011–2016 B.S. in Molecular, Cellular and Developmental Biology
With a Minor in Chemistry
University of California Santa Cruz
- 2014–2016 Undergraduate Research Intern
NASA Ames Research Center
- 2016–2018 Associate Scientist
Chimera Bioengineering
- 2018–2024 Ph.D. in Bioinformatics & Systems Biology
University of California San Diego

PUBLICATIONS

Author names marked with † indicate shared co-authorship.

Publications marked with $\triangle$ are included in this text.

Joe J. Solvason[†], Fabian Lim[†], Jessica Grudzien, Ben Song, Sophia Le, & Emma Farley. “Dependency grammar of developmental enhancers.” *In Review*. $\triangle$

Fabian Lim[†], **Joe J. Solvason**[†], Genevieve E. Ryan[†], Sophia H. Le, Granton A. Jindal, Paige Steffen, Simran K. Jandu & Emma K. Farley. “Affinity-optimizing enhancer variants disrupt development.” *Nature*, 2024. $\triangle$

Granton Jindal, Alexis Bantle[†], **Joe J. Solvason**[†], Jessica L Grudzien, Agnieszka D’Antonio-Chronowska, Fabian Lim, Sophia H Le, Benjamin P Song, Michelle F Ragsac, Adam Klie, Reid O Larsen, Kelly A Frazer, & Emma K Farley. “Single-nucleotide variants within heart enhancers increase binding affinity and disrupt heart development.” *Developmental Cell*, 2023. $\triangle$

Adam Klie, David Laub, James V. Talwar, Hayden Stites, Tobias Jores, **Joe J. Solvason**, Emma K. Farley & Hannah Carter “Predictive analyses of regulatory sequences with EUGENE.” *Nature Computational Science*, 2023.

ABSTRACT OF THE DISSERTATION

Elucidating Mechanisms Governing Enhancer Function in Development and Disease

by

Jesse Joseph Solvason

Doctor of Philosophy in Bioinformatics & Systems Biology

University of California San Diego, 2024

Professor Emma Farley, Chair
Professor Hannah Carter, Co-Chair

Enhancers are genomic elements that activate time- and tissue-specific gene expression through binding of transcription factors (TFs). Enhancers are therefore critical for successful development and homeostasis. Although the majority of disease-associated genetic variants reside within predicted enhancers¹⁻³, we still have little understanding of how enhancer sequence encodes its function, or the types of changes to enhancer sequence that alter enhancer activity and contribute to phenotypes. We and others have previously shown that suboptimal TF binding site affinity within enhancers are important for maintaining transcription levels and tissue-specificity^{4;5;86}. Here, we find that single nucleotide variants (SNVs) that increase binding site affinity cause gain-of-function (GOF) gene expression and developmental defects in *Ciona*

robusta⁷ (Ciona) and mice⁸. In humans, optimizing SNVs across hundreds of enhancers for a variety of TFs are enriched for GOF gene expression^{7;8}. Thus, a mechanistic understanding of enhancers, namely the importance of suboptimal affinity sites and variants that violate this principle, can pinpoint causal enhancer variants. Beyond the affinity of binding sites, the organization of sites is another component of functional enhancers. The physical properties of TFs and the DNA to which they bind are thought to constrain the way in which binding sites within a functional enhancer can be organized. The interplay of binding site syntax (the order, orientation and spacing of sites) and the affinity of sites is known as enhancer grammar⁹. To better understand how enhancer sequences encode their function, we conducted an exhaustive screen of all possible organization of the Ciona Otx-a enhancer. The Otx-a enhancer⁵ activates expression in neural cells and is regulated by a common developmental logic, a tissue determinant (GATA) and signaling pathway effector (ETS). The majority of the organizations are inert, illustrating the importance of syntax within enhancers. We identify syntax features enriched in active enhancers and use them to predict genomic enhancers. We find that considering both syntax and affinity gives the greatest accuracy of genomic enhancer predictions, highlighting the importance of grammar. We also use grammar to design tissue specific enhancers. Collectively, these studies increase our mechanistic understanding of how enhancer sequences encode their functions, and the types of changes to enhancers that can alter their function and may contribute to novel expression patterns and phenotypes.

Introduction

0.1 Enhancers drive traits and diseases

Development and homeostasis require enhancers to regulate when and where genes are expressed. Enhancers are genomic elements made up of transcription factor binding sites (TFBS) that acts as a scaffold to which TFs bind and activate gene expression in a time- and tissue-specific manner. Small changes to enhancer sequence can have large effects on expression levels^{7;10;11} and tissue-specificity⁹. For example, a single T>G mutation in the limb enhancer ZRS causes ectopic gene expression of the gene sonic hedgehog and results in the formation of extra digits^{12;13}. Integration of genotype-phenotype studies with epigenomic studies¹⁴ reveal that trait-associated loci are commonly found in putative enhancers¹⁻³. However, our ability is limited to predict which trait-associated genetic variation changes enhancer tissue-specificity and downstream phenotypes.

Genotype-phenotype studies vary in their approach depending on what kind of genetic variation is being studied. Most studies attempt to link phenotypes to common genetic variants, de novo variants, or somatic mutations. No matter the type of genotype-phenotype studied, a major challenge is differentiating between causal and inert variation. We call variants which contribute to the trait “causal”, while “inert” variants do not contribute. The reason inert variants significantly associate with the trait depends on the type of genetic variation studied.

Genome wide association studies (GWAS) reveal which common genetic variants are associated with human traits or disease¹⁵ (e.g. height, heart disease). GWAS typically use microarrays to determine the genotypes for thousands of participants across a known set of common genetic variants (usually >5% in minor allele frequency). The genetic variants that significantly associate with the trait are enriched for causal alleles. However, inert genetic

variation significantly associates with the trait because they are linked to the causal allele through linkage disequilibrium (LD). Fine mapping approaches enrich for causal variants through dense genotyping of previously identified GWAS loci or imputing LD patterns from large scale studies¹⁶. Integrating epigenomic studies can further prioritize which fine-mapped genomic variation are within predicted enhancers^{2;17}. Additionally, integrating expression quantitative trait loci (eQTL) can enrich for significant variants that are reside in genomic loci associated with changes in gene expression^{2;17;18}. Despite the successes of fine mapping approaches in further enriching for causal variants in enhancers, these variants are often left unvalidated by orthogonal biological assays. Thus, it is still an open question what percentage of fine-mapped genetic variation is inert versus causal. Predicting which alleles are causal is thus currently a challenge in understanding the importance of common genetic variants and disease. It also limits our accuracy in predict one's predisposition to disease based on their genotype.

De novo variants (DNV) also contribute to the emergence of traits and diseases^{11;19-21}. The contribution of DNVs to phenotypes are studied by whole-genome sequencing trios (the participant and the two parents). Genetic variants in the child but not the parents are considered de novo. In these studies, the DNVs discovered in the trait group are compared to the DNVs discovered in healthy group. Sequencing thousands of trios along with existing epigenetic data can be used to generate models which predict which DNVs are likely causal, revealing the contributions of DNVs to disease within promoters¹⁹ and enhancers¹¹. However, 45-70 germline DNVs exist per individual^{11;19;20}. It is therefore a challenge to determine which of these DNVs are driving the disease in a particular patient. One study took 31 DNVs predicted to be causal in putative enhancers and characterized how the DNV effected enhancer activity in human heart cells. 20% of the DNVs drove significant differences in enhancer activity using a reporter assay¹¹. These results suggest the importance of orthogonal validation to provide evidence for a predicted causal variant and estimate the accuracy of the model.

Somatic mutations are studied almost exclusively studied in their relevance to cancer, though their contribution to other diseases is currently being investigated²². When studying somatic mutations in cancer, tumor and healthy samples from the same patient are sequenced to determine the somatic mutations unique to the tumor. Since the same somatic mutation

between patients is extremely rare, “mutational hot spots” are instead searched for to assess which genomic loci are associated with tumor-specific somatic mutations²³. These results can be integrated with epigenomics and eQTL to understand in which of these hot spot loci are in putative enhancers. One study sequenced 930 tumors with matched transcriptomes to discover 193 noncoding loci in which tumor mutations disrupt target gene expression²⁴. This study revealed hot spots of cancer-associated somatic mutations, and the genes that are dysregulated by the mutation. However, one aspect left unaddressed was the power of these findings to enable predictions of dysregulated genes in the tumor for a given new patient. That is, even though a patient may contain a somatic mutation in one of these 193 loci, that mutation may be inert.

Currently, disease associated genetic variation in enhancers is studied through integrating epigenomics, eQTL and reporter assays with genotype-phenotype association studies. Epigenomics and eQTL can prioritize genetic variation in putative enhancers. Reporter assays can provide evidence for which precise genetic variation can cause changes in enhancer activity. Despite the important role enhancers play in the emergence of traits and diseases, we still lack an understanding for how DNA sequence encodes enhancer function. This gap in our knowledge contributes to our inability to distinguish between causal and inert genetic variation. We propose that applying mechanistic rules governing enhancer function revealed through 40 years of detailed studies^{9;25;26} can contribute an additional layer to unravelling how genetic variation in enhancers causes human disease.

0.2 Methodologies to measure transcription factor binding affinity

One important enhancer feature is the affinity with which a transcription factor binds. The first predominant method to encode binding preferences was the position weight matrix (PWM). Even today, PWMs are commonly used to predict the location and relative strength of binding sites^{27;28}. PWMs represent binding preference for a particular nucleotide (A/T/G/C) at a particular base in the binding site. Additionally, a PWM can be “scanned” across a query DNA sequence to determine predicted binding sites. The most common tools for PWM scanning are JASPAR²⁹, MEME suite³⁰ and HOMER³¹. The major issue with predicting binding sites using

PWM scanning is that it assumes that the contribution of each individual base is not affected by the context of the neighboring bases, often called the mono-nucleotide model. PWM-based predictions of transcription factor binding from the mono-nucleotide model were later shown not correlate well with empirical measurements of binding site affinity³²⁻³⁴, especially for low-affinity sites²⁷. Although not ideal for scoring affinity of binding sites, PWM represent an important first step in characterizing a TF's binding preferences.

G of a given TF to many DNA sequences in parallel, including MITOMI³² (1500 unique sequences) and BET-seq³³ (1M unique sequences). Although MITOMI and BET-seq represents an ideal methodology for empirical ΔG measurements, these methodologies have not been widely adopted. Additionally, no centralized database exists to host standardized data nor to search if a TF has been assayed using these approaches to measure ΔG .

Although not currently widely used in the field, protein binding microarray (PBM) represents a methodology to measure relative binding affinity and has already been performed for hundreds of TFs from a variety of organisms^{35;36}. For a given TF, PBM measures relative binding affinity for all possible 8-mer DNA sequences using a fluorescence readout. Although PBM does not report ΔG , methods exist to calculate a predicted ΔG from PBM data³⁷. PBM predicted ΔG s correlate well with ΔG s measured empirically by MITOMI and BET-seq for two transcription factors³⁸. In addition to ΔG , Kd can also be used to measure absolute binding affinity³⁷. Kd predicted from PBM data was shown to correlate well for two transcription factors as measured by surface plasmon resonance (SPR) and MITOMI³⁷. Additionally, relative affinities measured by PBM correlate with endogenous TF binding as measured by ChIP-seq⁸.

In summary, PBM is the ideal method to score binding affinities for 4 reasons. (1) PBM provides an affinity measurement for all possible 8-bp DNA sequences, (2) PBM affinity correlates with ΔG and Kd absolute affinities as measured by SPR, MITOMI and BET-seq, (3) PBM affinity correlates with in vivo binding as measured by ChIP-seq, (4) there exists a quality-controlled, standardized, easily searchable database called UniPROBE³⁶ which hosts PBM data for hundreds of transcription factors from yeast, flies, mouse and humans. Using PBM data, we can now predict which genetic variants cause affinity optimizations of binding sites and the downstream consequences on enhancer tissue specificity and organismal phenotypes.

0.3 Utilizing the low-affinity regulatory principle in causal variant prediction

Mechanistic studies of enhancers can lead to “enhancer regulatory principles” which can then be applied genome wide to predict the genetic variants likely to effect enhancer activity or tissue-specificity. One emerging enhancer regulatory principle is the importance of suboptimal enhancer features, such as the affinity of binding sites^{5;39}. Tissue-specificity of enhancers is achieved through combinatorial binding of transcription factors. We hypothesize that suboptimal affinities of transcription factors ensure that the enhancer will only activate in cells which contain sufficient amounts of all transcription factors⁵. If the affinity of a binding site is optimized for a single factor, that factor can unilaterally activate expression in tissues that singular factor is expressed in. Loss of enhancer tissue-specificity through affinity optimization has been previously demonstrated in the *Ciona intestinalis* nervous system⁵ and the developing fly eye³⁹. Though these two cases suggest the existence of the low-affinity enhancer regulatory principle, there are several important unanswered questions: how subtle of an affinity increase can affect enhancer tissue-specificity? What are the effects of affinity optimizations in vertebrates? Can human genetic variation cause affinity optimization of binding sites sufficient to drive loss of enhancer tissue-specificity and downstream organismal phenotypes?

In Chapter 1, we show the requirement of low-affinity sites for maintaining tissue specificity of a *Ciona robusta* (*Ciona*) heart enhancer⁷. We find that single nucleotide variants (SNVs) can optimize binding site affinity. These optimizing SNVs cause loss of heart-specific expression and drive heart malformations as severe as an extra chambered heart. We find that a conserved human heart enhancer also contains low-affinity sites and that optimizing SNVs increase enhancer activity levels in human heart cells. In Chapter 2, we find that affinity optimizing SNVs in the ZRS enhancer cause ectopic expression of the sonic hedgehog gene, driving polydactyly in mice and humans⁸. These findings demonstrate that the mechanism behind two human families with polydactyly are violations of the low-affinity enhancer regulatory principle. We then show that this enhancer regulatory principle enriches for causal variants driving increases in reporter activity and increases in endogenous gene expression. We find that this principle generalizes to a total of 3 transcription factors across hundreds of enhancers, highlighting this mechanistic

principle’s potential to reveal causal mutations across diverse contexts.

To our knowledge, this is the first time using an enhancer regulatory principle genome-wide to enrich for causal variants. We expect using this mechanistic understanding will be most successful in conjunction with population-based genotype-phenotype association studies, dense fine-mapping, eQTL fine-mapping and epigenomics to prioritize the most likely causal variants. Subsequently, it will be important to test both alleles in orthogonal biological assays to gather evidence for if the ref/alt alleles drive differential enhancer activity. Ultimately, the most convincing evidence of causality is demonstrating that introduction of these variants into the endogenous genomic locus drives changes in endogenous gene expression and downstream cellular phenotypes associated with the trait (e.g. cellular migration for metastases, or cellular conduction for some heart traits). Alternatively, causal variants can also be validated by assessing their power to predict who will exhibit a particular trait given their genotype in a new cohort.

0.4 Broadening our repertoire of enhancer regulatory principles

In the Ciona FoxF enhancer, the magnitude of affinity-optimization does not correlate with the severity of ectopic expression⁷. The most severe ectopic expression was observed for variants in sites nearby co-factor sites, suggesting our ability to predict the severity of causal variants must consider neighboring sites. Enhancer grammar could achieve this, as it describes how the dependency between enhancer features relates to gene expression. The features encoding grammar are binding site type, number, affinity and syntax (organization). Previous studies of grammar have only found the “smoking gun” of grammar, or traces of grammatical features which are enriched in predicted functional regions^{48–50;52;177}. We sought to study enhancer grammar with a more explicit and interpretable approach to provide a foundation for future studies to investigate grammar and use grammatical rules to make genomic enhancer predictions.

The grounds for grammatical rules is based on the fact that DNA and transcription factors have defined physical properties. It then follows that particular organizations of binding sites allow the formation of a transcription factors, promoting the cooperation and recruitment of co-factors that activate transcription. Enhancer grammar is defined by the binding site

type, number, affinity, and syntax⁹. Binding site syntax is the arrangement of sites and can be described by their relative order, orientation and spacing. We hypothesize that grammatical rules revealed by targeted high-throughput studies can be used to predict genomic enhancers. Additionally, genetic variation that violate these rules will change enhancer expression patterns, driving downstream changes in phenotypic traits and disease.

Enhancer grammar was first suggested almost 30 years ago^{41;43}. Since then, a body of work⁹ has emerged supporting the importance of grammar from flies to humans, and in a broad range of developmental contexts including the eye, heart, brain, and body patterning. For example, the order of binding sites in the sparkling enhancer redirects activity from the cone cells to the photo-receptor cells in the developing fly eye³⁹. In the developing mouse pituitary gland, decreasing the spacing between Pit-1 sites causes a loss of somatotrope cell expression and a gain of lactotrope expression⁴⁴. In the developing Ciona embryo, reversal of a FoxA site in a functional notochord enhancer results in complete loss of expression⁴⁵. In the human immune system, altering spacing between binding sites in the IFNb enhancer causes changes in enhancer activity levels⁴³. Collectively, these and other publications⁹ suggest that enhancer grammar is important for gene expression across diverse organisms and developmental contexts.

These and other studies led to three conceptual models for enhancer function is governed by grammar⁹. The enhancesome model is the most rigid where the particular organization of binding sites is thought to be a strict requirement. The IFNb stands the most popular example of this, however even in this enhancer we find that changes to organization do not necessarily break this enhancer, but change the activity levels. In fact, in the original paper describing the IFNb enhancesome, some spacing changes increased activity levels in their assay⁴³. The polar opposite of this model is the billboard model, which proposes that a cluster of binding sites is all that is required to be an active enhancer. A clear example of an enhancer which falls under the billboard model is the Autoregulatory Socket Enhancer 5 (ASE5 enhancer) which activates expression in the socket cells in flies⁴⁶. ASE5 contains known sites for suppressor of hairless, Vvl and an A-motif. Variants of ASE5 with perturbed organization of these sites were still active, however not all clusters of these sites were active, suggesting that the organizations of these sites are important for a cluster of these sites to be a functional enhancer. A third model, called the

TF collective model, proposes that TF-TF interactions are the most important. This model comes from a study that found that a threshold of ChIP-seq signal from a collective of TFs was important for enhancer activity using reporter assays in fly heart cells⁴⁷. Since the authors could not find any grammatical patterns, they asserted that none existed, and rather a critical mass of the TFs was required for a cluster of sites to be functional.

Overall, evidence for the existence and importance of enhancer grammar in these studies is strong. However, the case-by-case nature of these studies does not grant us a utility to use grammar to predict genomic enhancers, nor in predicting which genetic variation causes changes to enhancer function. With the advent of genome-wide and other high-throughput assays, a more recent collection of studies emerged searching for enhancer grammar at larger scales.

0.5 Epigenomics studies of enhancer grammar

Epigenomics approaches represent genome-wide assays which reveal regions contain epigenetic marks associated with enhancers¹⁴. The main types of epigenetic studies assay for histone marks (e.g. H3K27ac), open chromatin (e.g. ATAC-seq), and TF binding (e.g. ChIP-seq). Most of these studies focus on finding syntax (organizational preferences) of TF-pairs in the regions they predict to be functional. To date, several studies using genome-wide approaches were successful in finding the “smoking gun” of syntax^{48–52}. However, when no grammar is found or only weak signals are detected in their data, authors typically claim this as evidence for relaxed need for grammar^{48;52}, or that grammar isn’t important for those factors^{47;52}. However, it likely that technical or biological limitations of the experiment and analysis are contributing to insufficient power to detect the importance of syntax.

In order to find syntax signals enriched in predicted enhancers regions, one must be able to identify both binding sites with relatively high accuracy. The accuracy of predicting binding sites in the genome is dependent on how much is known about that transcription factor. ChIP-seq cannot identify TF binding sites because the signal spans 100s of bases while the binding site only spans 6-12 bases. Usually, there are many binding sites under a single ChIP-seq peak. Although advancements on increasing the resolution of ChIP-seq exist such as CUT&RUN⁵³, the resolution is still in the 50s to 100s of bases. ATAC-seq and ChIP-exo can be used to find

precise loci in which TFs bind^{48;53;54}. Although these methods appear to have good signal at some loci, it is unclear how sensitive these methods are in detecting sites genome wide that are not extremely high affinity.

To predict binding sites in ChIP-signal, PWMs are typically generated through de novo motif finding in ChIP-seq peaks using HOMER³¹ and are uploaded to JASPAR²⁹ for the community to use. The resulting PWM logos describe motif(s) in ChIP-seq regions, but their accuracy in predicting functional binding sites usually remain unvalidated. Therefore, studying the syntax of TF-pairs for which little is known or validated about the binding preferences leads to inaccurate binding sites prediction and thus an inability to determine the relative syntax of the two sites enriched in bound regions. One method to increase accuracy of PWM-based binding site predictions leverages the fact that ATAC-seq signal dips where TFs cover DNA⁵⁴. Indeed, the study which used ATAC-seq to prioritize real binding sites found that 34% of TF-pairs have evidence for preferred syntax⁵⁰, while the study which used PWM alone found 6% of TF-pairs with evidence for preferred syntax⁵². However, even the TF-pairs which are classified as having "relaxed" syntax have evidence for syntax playing a major role in binding⁵². In this experiment, the authors created many mice with indels either between or overlapping PU.1 and C/EBP β binding sites across 6 genomic loci. The authors highlight that across all 6 genomic loci, indels that overlap the motifs have larger effects than indels which cause changes to spacing. Although true, the indels which changed spacing caused dramatic decreases and increases in ChIP binding in 3 of 6 of the loci. Unfortunately, the effect of indels changing spacing on ChIP binding at these 3 loci were never compared to random background indels outside the sites, which would establish if changing spacing between the sites had a significant effect⁵². Thus, even for TF-pairs claimed to have "relaxed" syntax, it seems that there are still important, but yet unquantified, contributions spacing makes to the ability of TFs to bind DNA.

Another study took the approach of using ChIP-exo to detect the precise binding locations of TFs to search for the importance of syntax of 4 TFs important for pluripotency: Oct4, Sox2, Nanog and Klf4. ChIP-exo can find the precise genomic loci to which TF binds DNA by taking advantage of 5' exonucleases. These enzymes will chew the exposed DNA at the 5' end and stop where the TF is bound. A deep learning model was trained to predict ChIP-exo signal for an

arbitrary DNA input sequence. The authors then used the model as an “oracle” to see how alterations to binding site syntax effects the predicted ChIP-exo signal. Spacing between sites was found to be an important feature for predicted binding, suggesting the importance of syntax for these factors. Although this is another smoking gun of the importance of syntax, the authors do not validate the effects of syntax on binding, nor do they assess the functional importance of syntax for enhancer activity.

Genome-wide enhancer screens have shifted from relying on epigenetic marks associated with enhancers to directly testing a DNA sequence’s ability to drive enhancer activity using massively parallel reporter assays (MPRAs)⁵⁵. MPRAs are typically episomal, with up to hundreds of thousands of enhancers cloned into plasmids to be tested in parallel. The plasmid library is then introduced into cells and RNA sequenced to assess enhancer activity. There are two key features of MPRAs. First, the episomally measured enhancer activity of a DNA sequence has been shown to correlate well with activity measured through genomic integration⁵⁵. Second, the effect of genetic variants on endogenous gene expression as measured by eQTL correlates well with the effect on reporter expression tested in MPRA¹⁰. Together, these data suggest MPRA is an ideal model to understand how enhancer sequence activates gene expression, and how genetic variants in enhancers contribute to changes in endogenous gene expression.

One such MPRA is STARR-seq^{56;57}, which begins with shearing genomic DNA and cloning sheared fragments in the 3’UTR of a reporter gene. Genomic regions activating reporter expression are considered active enhancers. A recent STARR-seq study discovered 18,000 fly genomic enhancers active in S2 cells⁴⁹. Machine learning models trained on these fly enhancers can predict enhancer activity in S2 cells of any DNA sequence. Using the model as an oracle, the authors again find a smoking gun of syntaxes representing spacings between TF-pairs which had higher predicted activities. However, the trend of the signal was coarse, spanning 100s of bases. Recall that previous studies have found that changing spacing between sites by 1-5 bases can result in large changes in gene expression in fly neural ectoderm⁵⁸, Ciona neural system^{5;173}, Ciona notochord^{6;45}, mouse pituitary gland⁴⁴, mouse macrophage cells⁵², human immune cells⁴³ and other systems^{59;60}. Although the findings of this study support importance of grammar using a genome-wide assay, they suggest the models cannot yet pick up the effect that small

changes in spacing can have on enhancer activity.

0.6 Sequence based deep learning approaches to design tissue-specific enhancers

Recent advances in increasing the throughput of enhancer screens and more advance machine learning models has led to some studies successfully designing synthetic tissue-specific enhancers. Two main studies stand out with respect to this goal. The Aerts lab trained a deep learning model to predict accessible chromatin in cell types of brain⁶¹. They then used this model and in silico mutagenesis of DNA sequences to create enhancers that specifically activate in the kenyon cells or perineurial glia cells of the fly brain⁵¹. Again, a smoking gun is found for syntax as there are particular organizations of the TF-pairs Mef2+Ey and Onecut+Ey that optimize predicted enhancer activity. However, the authors do not validate the effects of syntax on enhancer activity, nor do they make comparisons between enhancers containing poor or optimal syntax. Additionally, the authors do not show that this model can predict how changes to DNA sequence will effect enhancer specificity or activity, limiting the utility for understanding genotype-phenotype relationships.

In a similar effort, the Tewhey lab trained a convolutional neural network on MPRA to predict the enhancer activity of a DNA sequence in K562, HepG2 and SK-N-SH cell lines⁶². They then performed in silico mutagenesis of DNA sequences to create enhancers specific to each cell line. The enhancers with the highest tissue specificity were tested in zebrafish embryos. Impressively, the authors find a large fraction of these synthetic enhancers had the predicted expression patterns in zebrafish embryos. The authors did not search for syntax in their model, nor did they assess the accuracy in the models ability to predict how changes to sequence effect enhancer activity.

Although high-throughput genomic approaches coupled with machine learning has advanced the field in designing tissue-specific enhancers, we still lack an understanding for how these enhancers function. Understanding the grammar of enhancers could in fact improve our ability to design these synthetic enhancers. For example, tuning the number, affinity and syntax of the sites to achieve different levels of activity of a previously designed synthetic enhancer.

0.7 Synthetic screens to study enhancer grammar

We propose that the genome may not have enough examples of binding site clusters to get a sufficient sample size to reveal grammatical rules that regulate them. This is not surprising given the 100s of regulatory logics at play. Attempting to find grammatical rules for all TF logics in a one-pot assay using genomic sequence alone may not be feasible. Thus, synthetic enhancer screens are becoming more popular which focus the study on a single combination of transcription factors to uncover grammatical principles.

The most in depth study of enhancer syntax to date⁶³ uses a synthetic screen to study the importance of syntax in encoding enhancer activity. The authors investigated a well-studied transcription factor logic for embryonic stem cells important for pluripotency, named OSKE, which is regulated by Oct4 (O), Sox2 (S), Klf4 (K), and Esrrb (E). A synthetic library was designed in which each synthetic enhancer receives one of each site with a randomized order and orientation. The authors detected some preferences for order of OSKE clusters. Specifically, they found that SOKE-ordered clusters drove significantly higher expression than OKES-ordered clusters. Furthermore, they detected a slight preference for SOX2 in the reverse orientation over the forward orientation for the SOKE-ordered cluster. These findings represent another smoking gun of binding site syntax. However, the findings could not be used to predict genomic enhancers.

There are several reasons that can explain why the syntax patterns they found in the syntax library design couldn't translate to making enhancer predictions in the genome. First, the OSKE library was designed using the consensus sites for each transcription factor. Given that the consensus sequence is almost always the highest affinity site, the authors thus searched for syntax rules in a context which contains the highest affinity sites. However, we know that in the *Ciona* nervous system⁵ and notochord⁶ that high-affinity sites reduce the necessity for optimal syntax. Therefore, the authors were searching for syntax rules in a screen in which we expect to see little contribution of syntax to enhancer activity. This may have limited the power to detect syntax patterns which contribute to enhancer activity. Also, there exist very few binding site clusters in the genome for which all binding sites are high-affinity. Finally, the synthetic

enhancers were designed using fixed 20bp linkers. Given that the genome contains a wide variety of spacing between OSKE clusters, it is not surprising that a screen containing fixed-spaced clusters could not translate to the genome with variable-spaced clusters.

0.8 Design considerations of targeted and controlled synthetic screens to study enhancer grammar

Given what we have learned from previous genomic and synthetic approaches, we focus on five considerations when designing synthetic libraries to ensure maximal generalizability to the genome. First, the screen should be focused on a singular transcription factor logic to maximize one's ability to detect grammatical rules. Second, the screen should generate synthetic sequences from known functional binding sites and inert linker sequences. This step is critical, as it minimizes contributions from binding sites that are not a part of the logic being studied. Third, each sequence in the library should have a paired negative control to which a mutation is applied to each binding site that ablates transcription factor binding. If this negative control is active, you know that the enhancer is not activated by the logic you are studying and should thus be omitted when searching for grammar rules. Fourth, the functional binding sites should have a range of functional affinities, maximizing generalizability to the genome which contain TFBS clusters with a large variety of affinity of binding site clusters exist. Finally, the screen should contain inert linker sequences that vary in lengths. This will ensure that the binding sites have a diverse range of spacing, maximizing generalizability to the genome in which clusters of binding sites exist with highly variable spacing.

Ciona represent an ideal model to study enhancer grammar. Ciona are a member of the urochordates, the sister group to the vertebrates⁶⁴. Ciona share many tissues with the vertebrates. For example, Ciona contain a heart, dorsal nerve cord and notochord^{65–67}. Many of the transcriptional programs which regulate the development of its tissues are also conserved, such as ETS and MesP specifying the heart⁸³. Finally, we can test the activity of hundreds of thousands of enhancers in every cell of thousands of developing embryos⁶⁸. This high-throughput approach provides a unique opportunity to reveal grammatical rules of enhancers and its use in making genomic predictions. In chapter 4, we discover grammatical rules of a conserved

transcription factor logic which uses GATA and ETS binding sites. We exhaustively test all possible syntaxes of functional GATA and ETS binding sites and inert linkers derived from the Otx-a Ciona neural enhancer⁵. The Farley lab previously identified the functional GATA and ETS binding sites, and also demonstrated that all the linker sequences are inert⁵. We generate a synthetic library containing randomized organizations of binding sites and linker sequences. We identify preferred orders, orientations and spacings of GATA/ETS in active synthetic enhancers. We then find that these activator syntaxes can be used to increase the accuracy of genomic enhancers predictions. To our knowledge, this work is the most in depth studies of enhancer grammar to date and the first to show that grammar can be used to improve sequence-based enhancer predictions at the genome scale. Additionally, we use these grammatical rules to design new synthetic tissue-specific enhancers not present in our original synthetic screen or in the genome. The importance of grammatical rules in the genome suggests that future studies may be able to predict changes in gene expression based on genetic variation disrupting functional syntaxes of sites.

Chapter 1

Single-nucleotide variants within heart enhancers increase binding affinity and disrupt heart development

1.1 Abstract

Transcriptional enhancers direct precise gene expression patterns during development and harbor the majority of variants associated with phenotypic diversity, evolutionary adaptations, and disease. Pinpointing which enhancer variants contribute to changes in gene expression and phenotypes is a major challenge. Here, we find that suboptimal or low-affinity binding sites are necessary for precise gene expression during heart development. Single-nucleotide variants (SNVs) can optimize the affinity of ETS binding sites, causing gain-of-function (GOF) gene expression, cell migration defects, and phenotypes as severe as extra beating hearts in the marine chordate *Ciona robusta*. In human induced pluripotent stem cell (iPSC)-derived cardiomyocytes, a SNV within a human GATA4 enhancer increases ETS binding affinity and causes GOF enhancer activity. The prevalence of suboptimal-affinity sites within enhancers creates a vulnerability whereby affinity-optimizing SNVs can lead to GOF gene expression, changes in cellular identity, and organismal-level phenotypes that could contribute to the evolution of novel traits or diseases.

1.2 Introduction

Enhancers are genomic elements that act as switches to control the location, level, and timing of gene expression to ensure the successful development and integrity of an organism^{14;69}. Sequence changes within enhancers can lead to changes in gene expression, phenotypic diversity,

novel traits, and disease^{70–72}. Indeed, most disease-associated variants within the genome lie within enhancers^{1–3}. Yet, pinpointing causal base-pair changes within enhancers is challenging because they are typically associated with many inert variants due to linkage disequilibrium. Furthermore, it is hard to functionally validate variants at scale in the relevant biological context⁷³. Our inability to predict causal enhancer variants is stalling efforts to harness the full potential of genomic data to understand the genetic basis of health and disease. Here, we use a mechanistic understanding of the regulatory principles governing enhancers to pinpoint causal enhancer variants that alter gene expression and heart development.

FGF (fibroblast growth factor) signaling plays a critical and conserved role in heart development across bilaterians^{74;75}. FGF signaling mediates changes in gene expression through binding of activated ETS transcription factors (TFs) to enhancers⁷⁶. Loss and gain of FGF signaling pathway components cause heart defects in organisms as diverse as *Drosophila* and humans^{77–79}. For example, gain of FGF is implicated in cardiac hypertrophy⁸⁰, whereas manipulation of ETS-1 in mice recapitulates some of the most common congenital heart defects in Jacobsen syndrome⁸¹. In *Ciona*, expression of a constitutively active form of ETS (ETS-VP16) in non-heart muscle cells leads to development of a multi-chambered heart by recruitment of more cardiac progenitor cells to the ventral midline⁸².

Ciona robusta (*Ciona*) is a marine chordate and member of the urochordates, the sister group to vertebrates⁶⁴. Similar to vertebrates, *Ciona* have a pumping heart within a pericardium, and early heart specification involves the TFs ETS and MesP⁸³. Across bilaterians, migration of the heart cells to the ventral midline is dependent on FGF signaling. In *Ciona*, the trunk ventral cells (TVCs) give rise to the heart and pharyngeal cells^{84;85}. FGF signaling and MesP activate expression of FoxF in the TVCs, this triggers migration of the TVCs to the ventral midline. *Ciona* is experimentally tractable, allowing us to investigate enhancer activity in hundreds of developing embryos and easily visualize the migration of the TVCs to the ventral midline⁸³. Thus, *Ciona* heart development, and the FoxF enhancer, provide an ideal system to investigate the relationship between enhancer sequence, tissue-specific gene expression and organismal-level phenotypes.

Developmental enhancers contain suboptimal-affinity sites (also referred to as submaximal

and low-affinity sites), which ensure tissue-specific gene expression^{5;6;86-89}. Although the use of suboptimal-affinity binding sites to encode precise gene expression has been seen in a variety of contexts, it has not been explored within heart development. Additionally, in the previous studies, many nucleotide changes are made to the sequences to turn low-affinity sites into high-affinity sites, but these large numbers of changes do not reflect the sequence changes naturally arising within genomic enhancers. Here, we seek to determine if low-affinity binding sites are important in heart development, and if single-nucleotide variants (SNVs) that increase binding affinity (affinity-optimizing SNVs) could cause gain-of-function (GOF) gene expression that contributes to phenotypic changes in heart development.

1.3 Results

1.3.1 FoxF is expressed within the TVCs and triggers migration of the TVCs to the ventral midline

A 295 bp FoxF enhancer upstream of a minimal promoter (bpFog) driving expression of nuclear mCherry recapitulates this TVC-specific expression pattern (Figures 1A, 1E, and S1A^{90;91}). An E-box site, ATTA sites and three ETS sites are important for activity of the FoxF enhancer (Figure S1B)^{84;92}. To identify additional ETS sites within the FoxF enhancer, we search for ETS cores (GGAW) within the enhancer and identify three additional putative ETS sites. We assign affinities of all six ETS sites from protein binding microarray (PBM) data, which measures the interaction of ETS1 with all possible 8-mers^{35;93}. We use PBM data for mouse ETS1 as a proxy for Ciona ETS1 binding because the DNA-binding domain and binding specificity of ETS1 is highly conserved^{5;94}. We calculate the relative affinity as the ratio of each 8-mer's fluorescence to the fluorescence of CCGGAAGT, the 8-mer with the highest affinity for ETS1 in *Drosophila*, *Ciona*, mice, and humans^{5;93;94}. The putative ETS sites within the FoxF enhancer bind ETS with low relative affinities ranging from 0.09 to 0.24 (Figure S1C).

1.3.2 Low-affinity ETS sites are necessary for tissue-specific expression

To determine if these six low-affinity ETS sites are necessary for expression within the TVCs, we ablate these ETS sites using point mutations that disrupt hydrogen bonding between the ETS TF and the binding site (GGAW to GCAW).³⁵ Ablating all of the ETS sites in

parallel leads to complete loss of enhancer activity (Figures 1B, 1E, and S1C). To investigate the contribution of each ETS site to FoxF enhancer activity, we ablate individual ETS sites with point mutations (GGAW to GCAW) and test the effect of these changes in reporter assays by scoring mCherry expression levels in embryos. To the best of our knowledge, these ablations do not create or ablate motifs for TFs present within the TVCs (Table S1). Ablating individual sites in the enhancer leads to a significant reduction in enhancer activity for five ETS sites (Figures S1C and S2). The only site without a reduction in expression upon ablation is ETS6, which has the lowest affinity at 0.09. Although loss of the ETS6 site does not alter expression, we cannot rule out that this site is redundantly involved in enhancer activity^{95–97}. These results demonstrate that five low-affinity ETS sites are necessary for FoxF enhancer activity in the TVCs and illustrate the functional role of low-affinity ETS sites within a developmental heart enhancer (DHE).

To determine if the ETS sites within the FoxF enhancer need to be low affinity for tissue-specific expression, we increase the affinity of the five ETS sites necessary for activity, keeping the GGAW core constant. Optimizing the affinity of all five ETS sites within the enhancer requires 16 nucleotide (nt) changes and leads to ectopic expression in tissues receiving FGF signaling, including the anterior tail muscle cells (ATMs), mesenchyme, nervous system, and notochord (Figures 1C and 1E). To demonstrate that it is the increase of ETS affinity driving the expression change, we made a FoxF enhancer with the 16 nt that optimize ETS affinity in combination with mutations that ablate ETS binding (GGAW to GCAW) (FoxF all opt ablated). This enhancer has no activity (Figures 1D and 1E). Thus, low-affinity sites are important for maintaining TVC-specific expression of the FoxF enhancer. Similar results have been seen upon optimizing the affinity of sites within neural and notochord enhancers, indicating that the role of low-affinity sites to encode tissue-specific expression is a general regulatory principle governing developmental enhancers^{5,6}.

Figure 1.1 Low-affinity ETS sites are necessary for tissue-specific expression. (A) The FoxF enhancer drives expression in the TVCs, marked by asterisks. (B) An enhancer with all ETS sites ablated (by changing each site from GGAW to GCAW) drives no enhancer activity. (C) Optimizing the affinity of all ETS sites leads to ectopic expression in ATMs, mesenchyme, nervous system, and notochord. (D) Ablating the ETS sites within the context of the FoxF all opt (FoxF all opt ablated) drives no enhancer activity. (E) Embryo scoring for above constructs (3 replicates, $n \geq 32$ embryos per replicate). p values are obtained using chi-squared test comparing each condition to WT with Bonferroni correction: **p < 0.0017; *p < 0.017, N.S., not significant. Schematics show electroporated enhancers, values are affinities, and dots show nucleotide changes. ETS6 is not shown in the schematic as ablation of this site did not impact expression (see Figure S1). Scale bars, 50 μ m. Data in stacked bar chart are represented as the mean. See also Figures S1 and S2 and Tables S1 and S2.

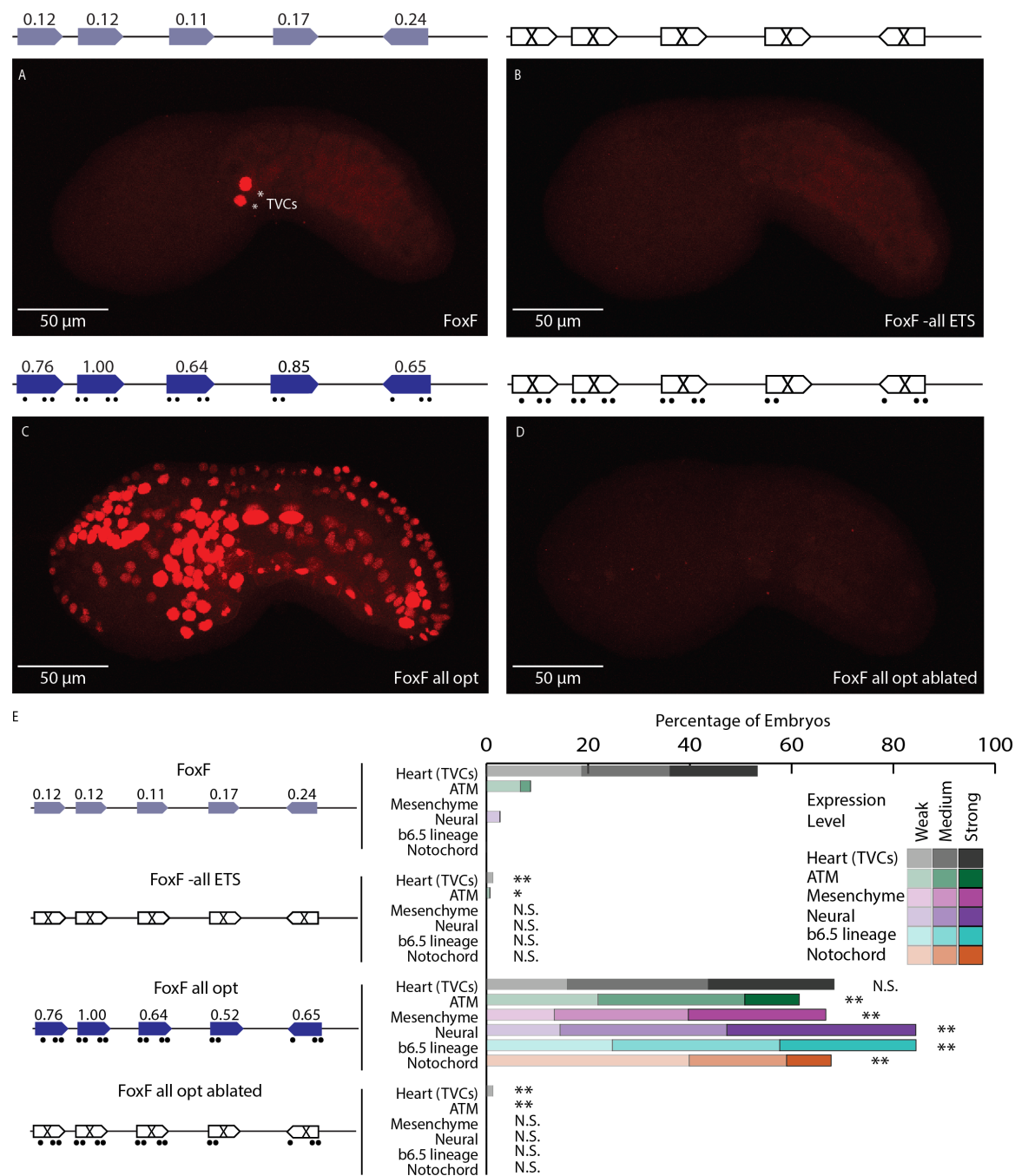


Figure 1.1. Low-affinity ETS sites are necessary for tissue-specific expression.

1.3.3 Affinity-optimizing SNVs lead to loss of tissue-specific expression

To determine if any SNVs could optimize ETS binding affinity and alter gene expression, we make every possible SNV in the flanking region of the ETS sites of the FoxF enhancer in silico and calculate the resulting affinity change. Within this manuscript, we use SNVs to denote single base-pair changes, regardless of whether they are synthetic or naturally found within Ciona or human populations. We find that SNVs in three of the five ETS sites cause a ≥ 3 -fold increase in ETS affinity (Figure 2A).

The greatest affinity change occurs within the ETS1 site, which has a starting affinity of 0.12. A SNV increases affinity 8-fold, creating an almost consensus site, with 0.97 affinity. We demonstrate the affinity change using electrophoretic mobility shift assay (EMSA) to detect differential binding of Ciona ETS1 to the wild-type (WT) and affinity-optimized ETS1 sequences (Figure S3). When we make this SNV within the FoxF enhancer (FoxF-ETS1-T-to-C), the enhancer drives ectopic mCherry expression in the ATMs and mesenchyme, both cell types that receive FGF signaling (Figures 2C and 2H).

We also find affinity-optimizing SNVs within the ETS3 and ETS4 sites. A SNV within the ETS3 site causes a 3-fold increase in affinity, from 0.11 to 0.34. This affinity-optimizing SNV drives ectopic expression in ATMs, mesenchyme, and neural tissues (Figures 2D and 2H). The affinity-optimizing SNV within the ETS4 site increases affinity from 0.17 to 0.52; this 3-fold affinity increase causes ectopic expression in the ATMs and mesenchyme (Figures 2E and 2H). These results suggest that affinity-optimizing SNVs are sufficient to cause loss of tissue-specific gene expression.

Figure 1.2 Affinity-optimizing SNVs lead to loss of tissue-specific expression. (A) Affinity fold change for SNVs within functional ETS sites of the FoxF enhancer. (B) The WT FoxF enhancer drives expression in the TVCs. (C) The FoxF-ETS1-T-to-C enhancer drives expression in the TVCs, ATMs, and mesenchyme. (D) The FoxF-ETS3-T-to-G enhancer drives expression in the TVCs, ATMs, mesenchyme, and neural tissues. (E) The FoxF-ETS4-T-to-C enhancer drives expression in the TVCs and ATMs. (F) Optimizing the affinity of three ETS sites with SNVs leads to ectopic expression in ATMs, endoderm, mesenchyme, nervous system, and notochord. (G) Ablating the ETS sites (GGAW to GCAW) within the context of the 3 SNVs no longer drives ectopic expression, only very weak TVC enhancer activity occurs in <10% of embryos. Image shows representative embryo with no expression. (H) Embryo scoring for above constructs (3 replicates, $n \geq 17$ embryos per replicate). p values are obtained by using chi-squared test comparing each condition to WT with Bonferroni correction: $**p < 0.001$; $*p < 0.01$; N.S., not significant. Asterisks mark TVCs, and arrows mark ATMs. Schematics show electroporated enhancers, values are affinities, and dots mark SNVs. Scale bars, 50 μm . Data in stacked bar chart are represented as the mean. See also Figures S2 and S3 and Tables S1 and S2.

A FoxF enhancer containing all three affinity-optimizing SNVs (FoxF 3 SNVs) drives strong activity in tissues responding to FGF signaling, similar to the FoxF all opt enhancer and the effect of these three SNVs appears to be synergistic (Figures 2F and 2H). A FoxF enhancer with all three affinity-optimizing SNVs and ablated ETS cores (GGAW to GCAW) drives almost no expression, demonstrating that the effect of these SNVs is ETS dependent (Figures 2G and 2H).

1.3.4 Affinity-optimizing SNVs cause migration defects and disrupt heart development

All three enhancers containing a single affinity-optimizing SNV show ectopic reporter expression in the ATMs (Figure 2). The ATMs and TVCs are related cell types that both receive FGF9 ligand; however, the TVCs contain higher levels of activated ETS due to FGF receptor enrichment^{98;99}. The high levels of activated ETS within the TVCs causes TVC-specific expression of FoxF and migration of the TVCs to the ventral midline⁸². Because the ATMs have low levels of activated ETS, they do not express FoxF and do not migrate (Figure 3A; Video S1)⁸². Overexpression of constitutively active ETS or FoxF in the ATMs causes the ATMs to migrate with the TVCs^{82;84}. We therefore hypothesize that the ectopic ATM expression driven by FoxF enhancers containing affinity-optimizing SNVs could cause migration of the ATMs to the ventral midline.

Figure 1.3 Affinity-optimizing SNVs cause migration defects and disrupt heart development. (A) Schematic showing a tailbud wild-type embryo, the concentration of activated ETS in the TVCs and ATMs and the role of FoxF in migration of the TVCs to the ventral midline. The TVCs have high levels of activated ETS, whereas the ATMs have lower levels of activated ETS. FoxF enhancer drives expression of FoxF in the TVCs, causing them to migrate. (B) Schematic showing a tailbud embryo containing the FoxF-ETS1-T-to-C enhancer driving FoxF. TVC cells have high levels of activated ETS, whereas ATMs have low levels of activated ETS. Embryos with the FoxF-ETS1-T-to-C enhancer drive expression of FoxF in the TVCs and ATMs, causing the ATMs to express FoxF and migrate with the TVCs. (C) Embryo with WT FoxF enhancer driving FoxF, co-electroporated with a MesP enhancer driving GFP marking the ATMs and TVCs at 16 h post fertilization (hpf). (D) Embryo with FoxF-ETS1-T-to-C enhancer driving FoxF, co-electroporated with a MesP enhancer driving GFP in the TVCs and ATMs at 16 hpf; all cells migrate to the ventral midline. Black cells in (C) and (D) are the otolith and ocellus. (E) Juvenile with WT FoxF enhancer driving FoxF with heart colored in blue; normal heart morphology is seen. (F) Juvenile with FoxF-ETS1-T-to-C enhancer driving FoxF has two hearts with main heart colored in pink, auxiliary heart colored in orange; abnormal heart morphology is seen in 79% (41/52) of embryos with migration defects. Hearts were false-colored digitally. (G) Schematic of juvenile *Ciona* with normal heart morphology. (H) Schematic of juvenile *Ciona* with auxiliary heart. Scale bars, 50 μ m. See also Figure S4 and Videos S1 and S2.

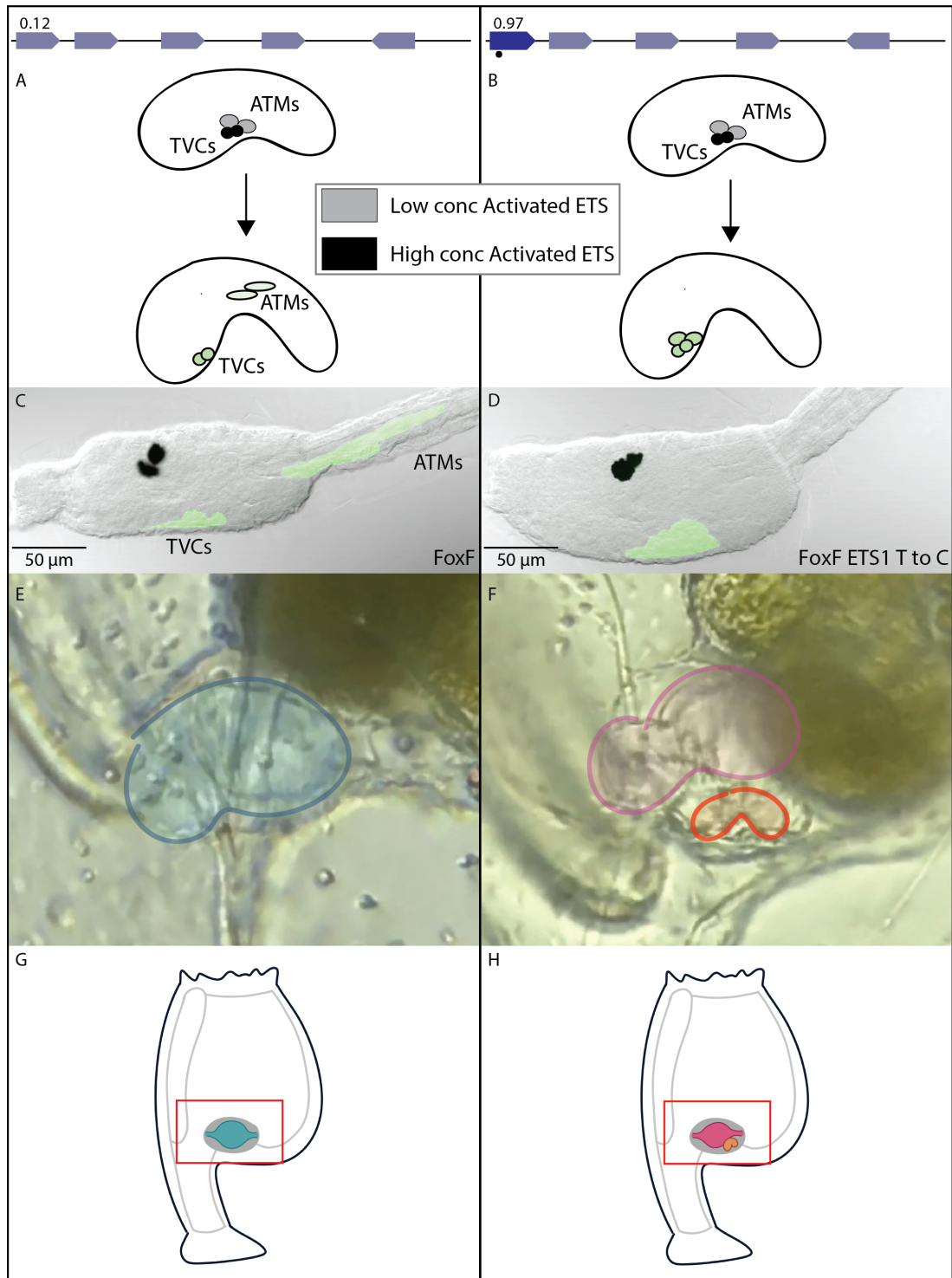


Figure 1.3. Affinity-optimizing SNVs cause migration defects and disrupt heart development.

To test this hypothesis, we co-electroporate a *Mesp-2kb>GFP* reporter to mark both the TVCs and ATMs and a construct containing the FoxF 3 kb regulatory element, including the endogenous promoter, driving FoxF mRNA (Figure S4A). We use the 3 kb regulatory element rather than a minimal enhancer to better recapitulate the regulatory region. When the WT FoxF enhancer drives FoxF, we observe no migration of ATMs to the ventral midline (Figures 3A, 3C, and S4; Video S1) and juvenile *Ciona* have normal hearts (Figures 3E and 3G; Video S2). When we electroporate the FoxF 3SNVs driving FoxF, embryos are so deformed we could not track TVC migration. When we electroporate the FoxF-ETS1-T-to-C enhancer driving FoxF, we see migration defects in 9.3% of embryos (Figures 3B, 3D, and S4; Video S1). The FoxF-ETS3-T-to-G enhancer driving FoxF causes migration defects in 9.9% of embryos, despite the fact this SNV only increases the affinity of the site to 0.34 (Figure S4). The FoxF-ETS4-T-to-C driving FoxF causes migration defects in 8.0% of embryos (Figure S4). Thus, SNVs that lead to ≥ 3 -fold increase in affinity, even those that result in relatively low-affinity sites, have the potential to contribute to migration defects during heart development. The strong impact of the optimized ETS3 site on migration could be due to nearby E-box and ATTA sites^{84;92}, suggesting that enhancer grammar⁹ may influence the functional consequences of affinity-optimizing SNVs within genomes.

To determine the impact of migration defects on heart development, we followed 52 FoxF-ETS1-T-to-C embryos with migration defects from the larval to juvenile stages. Strikingly, 79% of embryos with migration defects have abnormally developed hearts (Figures 3F and 3H). Phenotypes range from enlarged hearts to two distinct beating hearts within the same pericardium (Video S2). As the *Ciona* age, hearts from animals with the FoxF-ETS1-T-to-C SNV could not pump blood, and these animals did not survive beyond 12 days post metamorphosis.

1.3.5 Low-affinity ETS sites are prevalent in putative DHEs across chordates, and these enhancers are vulnerable to affinity-optimizing SNVs

The role of FGF signaling in specification and migration of heart cells is conserved from flies to vertebrates^{74;75}. To determine if low-affinity sites are common within DHEs, we used epigenomic datasets to define putative DHEs in both *Ciona* and humans. Using assay for

transposase-accessible chromatin with sequencing (ATAC-seq) data¹⁰⁰ for Ciona heart cells at 6.5 h post fertilization, a time at which TVCs receive FGF signaling, we identify 15,174 putative Ciona DHEs. Using ATAC-seq and chromatin immunoprecipitation sequencing (ChIP-seq) for p300 and H3K27Ac from primary fetal human hearts and induced pluripotent stem cell (iPSC)-derived embryonic-like cardiomyocytes at time points when FGF signaling is important for heart development, we find 252,931 putative human DHEs^{11;101–103}. In both Ciona and human DHEs, the median affinity of putative ETS sites is 0.12, suggesting that low-affinity sites are prevalent within chordate putative DHEs. Within 6,618 (44%) of Ciona DHEs and 154,685 (61%) of human putative DHEs, a SNV can increase the affinity of ETS sites by ≥ 3 -fold, with the average affinity after optimization being 0.52 (Figure 4A). Thus, DHEs are vulnerable to affinity-optimizing SNVs, some of which may lead to GOF gene activity and changes in cellular identity.

1.3.6 Human and mouse Gata4 heart enhancers contain low-affinity ETS sites

GATA4 is a key TF within the developing heart. One of the human DHEs is the ortholog of the mouse enhancer for Gata4 (Figure 4B). The mouse Gata4 G9 enhancer is ETS dependent and drives expression in the endocardial and myocardial heart layers¹⁰⁴. Ablating four ETS sites within the mouse Gata4 G9 enhancer leads to loss of expression, suggesting that these sites are functional and required for enhancer activity¹⁰⁴. However, the affinity of these ETS sites has not been studied. We assign the affinity of all ETS sites within the Gata4 G9 enhancer using mouse ETS1 PBM data. All of the ETS sites within the mouse Gata4 G9 enhancer have affinities ranging from 0.09 and 0.12. The four sites previously shown to contribute to enhancer activity and bind ETS-1 have affinities of 0.11, 0.12, 0.10, and 0.12, respectively. These findings demonstrate that low-affinity ETS sites are necessary for heart-specific expression. The human GATA4 G9 enhancer contains low-affinity binding sites that are almost identical to those within the mouse Gata4 G9 enhancer. A reporter construct containing the human GATA4 G9 enhancer, basal supercore promoter, and reporter (GFP and barcode) drives expression in human iPSC-derived fetal-like cardiomyocytes, which are the primary cell type found within the myocardial layers of the heart (Figure 4C).

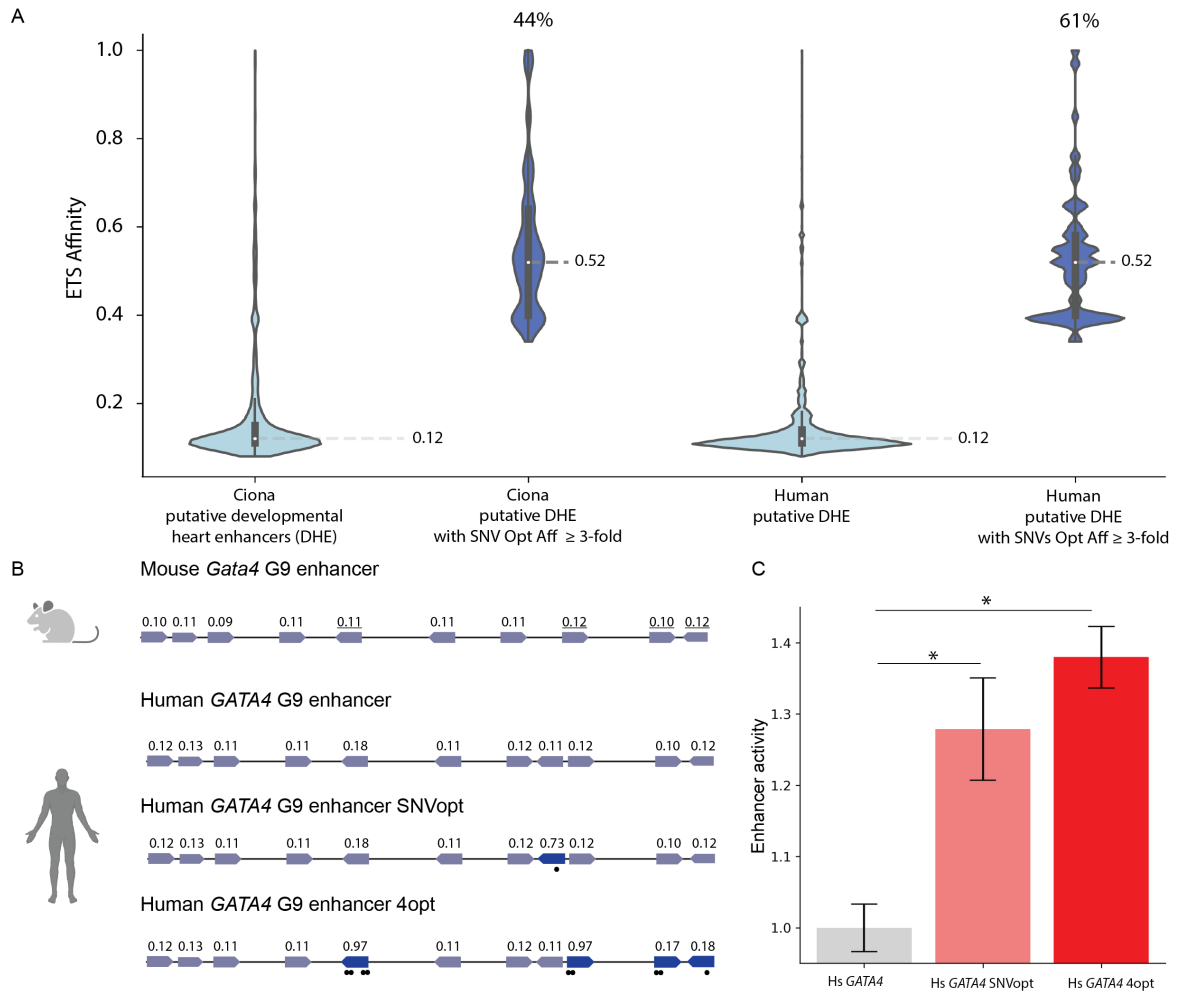


Figure 1.4. Low-affinity ETS binding sites are prevalent within developmental heart enhancers, and affinity-optimizing SNVs lead to gain-of-function expression in human iPSC-CMs (A) Violin plots in light blue of ETS site affinities within Ciona and human putative DHEs. 44% of Ciona DHEs and 61% of human DHEs contain an ETS site in which a SNV can increase ETS affinity ≥ 3 -fold. Dark blue violin plots show distributions of the ETS affinity within each enhancer upon addition of the highest affinity-optimizing SNV. (B) Schematics of human and mouse *Gata4* enhancers show high conservation of low-affinity ETS sites. Sites with underlined affinities mark functionally validated ETS sites in mice. *35* *GATA4* G9 enhancer SNVopt harbors a SNV that increases relative affinity from 0.11 to 0.73. *GATA4* G9 enhancer 4opt has 4 ETS sites with increased affinity. (C) Hs *GATA4*, Hs *GATA4* SNVopt, and Hs *GATA4* 4opt enhancer activity ($\log_2(\text{RNA RPM}/\text{DNA RPM})$) as measured by reporter assay in human iPSC-CMs. Data are represented as mean $\pm$ SEM. * $p < 1e-5$. See also Table S2.

1.3.7 Affinity-optimizing SNVs within the human GATA4 G9 enhancer cause GOF expression in human iPSC-derived cardiomyocytes

To determine if optimizing the affinity of the ETS sites within the GATA4 G9 enhancer leads to GOF expression, in this case an increase in expression within cardiomyocytes, we conducted a reporter assay in human iPSC-CMs to compare the expression of the reference human GATA4 G9 enhancer (Hs GATA4) with the same enhancer with a single (Hs GATA4 SNVopt) or four affinity-optimizing mutations (Hs GATA4 4Opt)^{105;106}. Increasing the affinity of a single ETS site from 0.11 to 0.73 (Hs GATA4 SNVopt) leads to a significant increase in expression. Increasing the affinity of four ETS sites (Hs GATA4 4opt) leads to an even larger increase in expression (Figure 4C). Thus, in the human GATA4 G9 heart enhancer, affinity-optimizing variants drive GOF gene expression, consistent with our findings in Ciona.

1.4 Discussion

1.4.1 Low-affinity ETS sites are a prevalent feature of DHEs

Our genome-wide analysis finds low-affinity sites are a common feature of DHEs across Ciona, mice, and humans. Ablation of low-affinity sites both within the Ciona FoxF enhancer and mouse Gata4 enhancer cause a significant reduction in enhancer activity, demonstrating that these low-affinity sites are functional. Low-affinity sites are also important within enhancers activated by other TFs downstream of signaling pathways and pleiotropic factors^{5;58;86;107;108}. Our findings highlight the importance and vastly underappreciated contribution of very low-affinity binding sites to regulation of gene expression and the advantages of scoring the affinity of sites.

1.4.2 Affinity-optimizing SNVs cause GOF enhancer activity, which can contribute to phenotypes

In Ciona, affinity-optimizing SNVs lead to loss of tissue-specific expression and migration defects that alter heart development. In the most extreme cases, we see two beating hearts. We propose that ETS affinity-optimizing SNVs likely contribute to congenital heart disease and cardiac traits. The ubiquitous role of FGF in other developmental programs^{109;110} and cancer¹¹¹ indicates that ETS affinity-optimizing SNVs are likely involved in other enhanceropathies. Searching for affinity increases as a method to filter for causal enhancer variants is not common;

yet, from our work and a few other studies, it appears that this could be a powerful approach. To our knowledge, only three studies search for an affinity increase and measure the impact on gene expression yet do not analyze phenotype^{112–114}. In a handful of examples, causal enhancer variants have been functionally validated via reporter assays. After the validation of the impact of these variants on gene expression, these sequence changes have been identified as increasing affinity^{101;115–118}. These other aforementioned studies have found sequence changes that increase affinity in TFs as diverse as Wnt effectors and Hox members that can alter gene expression and are statistically associated with a risk for certain phenotypes^{115;118?}. In another of our studies, we find affinity-optimizing SNVs for a variety of TFs including ETS, AP-1, Hox and IRX cause GOF enhancer activity. Within limb development, naturally occurring human SNVs and synthetic SNVs that increase the affinity of ETS cause extra digits in mice and humans¹¹⁹.

1.4.3 Affinity-optimizing SNVs may contribute to evolution of novel traits

There is a fine line between creation of novel phenotypes that are beneficial or detrimental¹. The Ciona heart, similar to all invertebrate chordates, is single chambered, whereas all extant vertebrates have at least two chambers. The evolution of a dual-chambered heart in vertebrates is thought to involve recruitment of additional precursor cells to the ventral midline⁸². Our study indicates that mutations within enhancers that increase binding affinity for ETS result in more cells migrating to the ventral midline and creation of another compartment within the pericardium. Although the animals in our study with two hearts did not survive, it is possible that some animals with multi-chambered hearts could survive and have a selective advantage. More generally, affinity-optimizing SNVs could contribute to the evolution of traits by affecting gene expression patterns and cell behavior.

1.4.4 Violations in regulatory principles governing enhancers can pinpoint causal enhancer variants

How enhancer variants contribute to phenotypes is a pressing challenge we need to solve if we want to causally link genomic variation to phenotypic diversity in health and disease. To date, only a handful of enhancer variants have been functionally validated as altering gene expression and phenotype. Yet, thousands of non-coding variants are associated with changes

in gene expression and phenotypic variation. We cannot possibly test all non-coding variants in the relevant cell types and time points to pinpoint causal enhancer variants. Finding rules that generalize across different enhancers relies on understanding the regulatory principles that govern enhancer function. One such regulatory principle is the use of low-affinity binding sites to encode tissue-specific expression and levels of expression. Here, we show that SNVs that violate this regulatory principle can cause GOF gene expression and organismal-level phenotypic changes. Identification of other regulatory principles, and subsequently variants that violate these principles, could provide further mechanistic approaches to pinpoint causal enhancer variants at scale.

1.4.5 Limitations of the study

Although many studies use position weight matrices to find matches to binding sites, here we use PBM data, an in vitro measurement of relative binding affinity that allows us to give a relative affinity score to every 8mer sequence. In vitro PBM measurements do not consider chromatin structure, protein-protein interactions, cooperative binding between multiple sites, or phase separation. Although this is a limitation, it also has some advantages because it enables us to have a fixed measurement of relative binding affinity that can be used across genomes, cell types, and species to annotate putative binding sites. As illustrated in our study, this approach can predict binding and the impact of SNVs in different cell types and organisms (e.g., FoxF Ciona TVC enhancer and human GATA4 cardiomyocyte enhancer). As such, it is a score that can be used to analyze genomic sequences and predict the impact of enhancer SNVs without the need to measure specific binding affinity of every site at every time point and cell type of an organism.

Our studies take advantage of disparate biological systems Ciona and human cardiomyocytes to detect GOF gene expression. In Ciona, we are able to assay expression in all cells of the embryo, but it is difficult to quantify expression levels within these cell types. While in human cells we see that the affinity-optimizing SNVs increase levels of expression, we cannot assay the impact of this SNV on specificity of gene expression within the entire human heart. Together, these complementary assays illustrate that affinity-optimizing SNVs can increase the

levels of expression within the endogenous location and also lead to expression in other cell types, both types of GOF expression could contribute to phenotypes.

Another limitation relates to how we study the role of enhancer variants on cell behavior. We characterize Ciona heart phenotypes by electroporation of a plasmid carrying a large 3 kb regulatory region driving FoxF cDNA into fertilized eggs. Although the WT 3 kb region driving FoxF mRNA does not alter migration, the same regulatory element with only a single SNV leads to migration defects. Indeed, we see consistent results for three different SNVs tested in this way. Ideally, we would show that a SNV within the endogenous locus has the same effect. However, targeted genome editing with homology-directed repair or base editing is not yet possible within Ciona. Once these techniques are established, testing the impact of the FoxF SNVs within the genome will be important. Although we have not tested the FoxF SNVs in the endogenous locus, we have done similar studies in mice and found that affinity-optimizing SNVs within the endogenous locus in mice lead to GOF expression and phenotypes¹¹⁹.

1.5 Methods

1.5.1 Tunicates

Adult *C. intestinalis* type A aka *Ciona robusta* (obtained from M-REP) were maintained under constant illumination in seawater (obtained from Reliant Aquariums) at 18 degrees C. Ciona filter feed on material in seawater and we do not add any food for the week we keep them for. 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.

1.5.2 iPSC-derived cardiomyocytes

Cell culture. iPSC-CMs from line UDID_109 and from line UDID_13947 were thawed and seeded onto gelatin-coated 48-well plates at a density of $1.3 \times 10^5/\text{cm}^2$. UDID_109 cell line is from an Asian male and UDID_139 cell line is from an Asian female. Cells were cultured using previously described culture conditions¹²⁰.

1.5.3 Mutagenesis and cloning

The FoxF enhancer construct was a gift from the Levine and Christiaen labs. The construct consists of the following elements in order: AscI, FoxF enhancer, XbaI, bpfog promoter, NotI, H2B-mCherry, EcoRI. Point mutations were introduced into the construct using mutagenesis with partially overlapping primers with 3'-overhangs. To clone the construct with all 6 ETS sites ablated from GGAW to GCAW, we ordered the FoxF enhancer synthesized from IDT. For experiments using the 3kb FoxF enhancer and FoxF cDNA, the 3kb FoxF regulatory region was amplified from the genome and cloned into our electroporation vector, the FoxF cDNA was also cloned into this vector. The FoxF cDNA sequence can be found in Aniseed with the Gene Model ID KH2012:KH.C3.170. The GATA4 G9 reporter plasmid for testing in iPSC-CMs consists of the following elements in order: GATA4 fragment, supercore promoter, unc76-EGFP, 30-nucleotide barcode and polyA.

1.5.4 Cional Electroporation

Electroporation. Constructs to be compared to each other are midipreped in the same batch using the NucleoBond Xtra Midi kit (Macherey-Nagel). For each construct, 70 ug of DNA in 100 uL is added to 400 uL of 0.96 M D-mannitol. Dechoriation, in vitro fertilization and electroporation were performed as described previously.⁷⁰ Typically for each electroporation, eggs and sperm were collected from 6 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. Differential interference contrast microscopy was used to obtain transmitted light micrographs with an Olympus FV3000, using the 40X objective. The same microscope was used to obtain all fluorescent images. Three biological replicates were analyzed for each construct unless otherwise noted.

Counting Embryos. In Ciona, as there are defined cell lineages, we can determine cell identity based on visually inspecting the nuclear signal and the region of the embryo in which the signal occurs. Strong, moderate, and weak expression levels through the eyepiece were determined as strong being visible at 25x of light source, moderate being visible at 50x of light

source, and weak being visible at 100x of light source. Fifty embryos were counted for each biological replicate, unless otherwise noted. For each experiment, slides were counted blind.

Image Acquisition. 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. Representative embryos on the slide were chosen and imaged prior to unblinding the slides. Two exposure times were taken for each construct. In each figure, the same exposure time for each image is shown to allow direct comparison, which occasionally leads to overexposed images being used for stronger constructs (e.g., Figure 1C).

Counting Larvae for Migration Defects. Embryos were transferred from gelatin-coated electroporation plates to 15cm uncoated plastic plates filled with filtered seawater at 7 hpf. All well-developed larvae on the plate were counted from 16-20 hpf and scored for presence or absence of normal ATM development. All larvae with migration defects were marked for heart morphology analysis in the juvenile stage. For each replicate, 10-20 larva with normal ATM development were selected for heart morphology analysis in the juvenile stage.

Juvenile Heart Morphology Analysis and Imaging. Juvenile hearts were analyzed for morphological defects starting at 48hrs after metamorphosis. Hearts with abnormal structure or function were scored as “deformed.” Images and video of hearts were taken on an Axiozoom microscope with a mounted iPhone 7. Video editing was done in Adobe Premiere Pro. Heartbeat animations were made in Adobe Animate.

TVC Migration Time-lapse Imaging. Embryos were electroporated with MesP 2kb>GFP as previously described and checked for GFP fluorescence at 6 hpf. Well-developed embryos with strong GFP expression were transferred to a glass-bottom microscopy plate in seawater. Z-stacks were taken of each embryo every 30 minutes for 8 hours with an Olympus FV3000 microscope with a 20x objective lens. Resulting time-lapses were analyzed in Fiji. A max projection was taken of the GFP channel and merged with the DIC. Time-lapses were cropped and rotated in Adobe Photoshop.

1.5.5 Electrophoretic mobility shift assay

EMSAs were performed using the LightShift™ Chemiluminescent EMSA Kit (ThermoFisher Scientific) with biotinylated and non-biotinylated double-stranded oligonucleotides corresponding to the ETS1 region of the FoxF enhancer and mutations that alter affinity to 0.97 and 1.00. Oligonucleotides were annealed according to the advanced protocol at this link (<https://tools.thermofisher.com/content/sfs/brochures/TR0045-Anneal-oligos.pdf>). Ciona ETS1 DBD protein was synthesized using the TNT Quick Coupled Transcription/Translation System (Promega) from the pTNT plasmid containing Ciona ETS1 DBD. Ciona ETS1 DBD was synthesized with flanking XhoI and NotI sites using Twist Bioscience and cloned into the pTNT-B18R vector (Addgene #58978). The Ciona ETS1 DBD sequence was determined as the sequence of Ciona ETS1 that aligns to Mouse ETS1 DBD^{121;122}. The binding reaction was carried out in a 20 uL volume containing 2 uL of 10X Binding Buffer (100mM Tris, 500mM KCl, 10mM DTT; pH 7.5), 50ng Poly(dI:dC), 4 uL of protein extract and 20 femtomol biotin-labeled probe. For competition experiments, a 200-fold molar excess of unlabeled probe was added. Binding reactions were pre-incubated for 10 minutes before adding the biotin-labeled probe. Binding reactions were then incubated at room temperature for 20 min and loaded onto a DNA retardation gel (6%). After sample electrophoresis with 0.5X TBE on ice and transfer to a 0.45-um Biotodyne™ B Nylon membrane (ThermoFisher Scientific) in the cold room, DNA was crosslinked for 15 minutes using 312 nm light, and the membrane was put between blank sheets of paper overnight. The next day, the biotinylated probes were detected using the Chemiluminescent Nucleic Acid Detection Module (ThermoFisher Scientific). Images of the resulting membrane were acquired using a Chemidoc MP imaging system (Bio-Rad).

1.5.6 Scoring Relative Affinities

Relative affinities were calculated using high-throughput binding datasets¹²³ (thebrain.bwh.harvard.edu/uniprobe/index.php). They were calculated using median signal intensities of mouse ETS1 universal protein binding microarray data from the UniProbe database⁹³. It has previously been shown that the binding specificity of ETS1 across 600 million years from flies to humans is conserved and thus the use of mouse ETS1 to calculate ETS affinity is a valid

approach⁹⁴. The relative affinity represents the ratio of the median signal intensity of the native 8-mer motif to the optimal 8-mer motif for ETS.

Computational Search for Motifs Altered by each SNV. For each SNV, a 30 base-pair region centered on the ETS site was put into JASPAR and scanned for all 1205 vertebrate profiles with 80% cutoff. Then, the same 30 base-pair region centered on the ETS site with the SNV was put into JASPAR74 and scanned for all 1205 vertebrate profiles with 80% cutoff. The list of sites destroyed and created by the SNV were found, we further filtered this list to focus on individual transcription factor sites with mutations that affect nucleotides critical for transcription factor binding. Of the ETS-family JASPAR profiles, we used MA0098.1.ETS1. For the GGAW to GCAW ablations, no TFBS for repressors expressed in the TVCs (ANISEED database) were created and no TFBS for activators expressed in the TVCs (ANISEED database) were destroyed. For the SNV opt mutations, no TFBS for repressors expressed in the ATMs (ANISEED database) were destroyed and no TFBS for activators expressed in the ATMs (ANISEED database) were created.

1.5.7 Computational Enhancer Analysis

Alignment of ATAC-seq reads. ATAC-seq alignment was performed following the methods in Racioppi et al¹⁰⁰. Raw reads from 6 hours post fertilization (hpf) (GEO accession GSE126691, LacZ_6hpf_1-3, LacZ_10hpf_1-4) were first preprocessed by FastQC (version 0.11.2, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>). Adaptors were trimmed using Trim Galore (version 0.4.4, http://www.bioinformatics.babraham.ac.uk/projects/trim_galore) and trimmed reads were aligned to the Ciona robusta genome63 using Bowtie2 (version 2.3.2¹²⁴) with the parameters `-very-sensitive -p 16 -X 1000`. Reads with mapping quality score > 30 were kept for downstream analysis using SAMtools (version 1.2¹²⁵). Mitochondrial reads were removed using bash command `egrep -v`. Following examination of read quality of the 6hpf data, replicate 2 was removed. Read pileup (BAM) files for replicates 1 and 3 at 6hpf were merged using the SAMtools merge function and peaks were called using MACS2 (version 2.7.9)¹²⁶ (`-nomodel -bdg -g 99000000 -f BAMPE -q 0.01`). To correct for nonspecific sequencing biases, we subtracted gDNA from these libraries¹²⁷.

Calculating top ATAC peaks. To find the highest confidence peaks in our dataset, we calculated the area under the curve (AUC) as the sum of the read depth at each position in the peak. Bedtools genomecov (-bga) and intersect functions were used to calculate the read depth at every genomic position within all the called peaks¹²⁸. Peaks with AUCs in the top 90% were kept and the rest were discarded.

Downloading putative enhancers. Human fetal heart putative enhancers were identified from epigenomic datasets from the key resources table. Any coordinates for hg18 or hg19 were lifted over to hg38 using pyliftover 0.4 (<https://pypi.org/project/pyliftover>).

Identifying affinity-optimizing SNVs in putative developmental heart enhancers. We first determined the distribution of ETS affinities in putative developmental heart enhancers. We collected affinities for all ETS sites defined as NNGGAWNN. Once all of the ETS binding site affinities were determined, we then performed all possible point mutations to each binding site and plotted those which resulted in ≥ 3 -fold change. We reported the number of putative heart enhancers which contained ≥ 1 ETS which could be affinity-optimized ≥ 3 fold by a SNV.

1.5.8 Enhancer Library

Library Construction. A library containing Hs GATA4, Hs GATA4 SNVopt and Hs GATA4 4opt enhancers upstream of the supercore promoter, GFP and barcode, was constructed as previously described in¹²⁹.

iPSC cell culture. iPSC-CMs from line UDID_109 and from line UDID_13947 were thawed and seeded onto gelatin-coated 48-well plates at a density of $1.3 \times 10^5/\text{cm}^2$. The iPSC-CMs were transfected with 325nM/well of plasmid containing the enhancer library driving GFP using ViaFect™ Transfection Reagent (Promega) at a 6:1 ratio in a medium containing 10% FBS and 5uM (Y27632) ROCK Inhibitor (Sigma). 16hrs post transfection and every other day, medium was changed. Cells were then collected 96hrs post transfection and stored in TriZol reagent frozen at -80°C . Library transfection was performed in duplicate for line 109 and triplicate for line 139.

iPSC RNA extraction, cDNA synthesis, DNA extraction, barcode extraction, sequencing. Total RNA was extracted using TriZol extraction and treated to remove contaminating DNA

(TURBO DNA-free DNase digestion kit, Ambion). Subsequent mRNA was isolated (Dynabeads mRNA Isolation kit, Invitrogen) and was used for cDNA synthesis (Transcriptor High Fidelity cDNA Synthesis kit, Sigma-Aldrich). Subsequent DNA extraction, barcode extraction, and sequencing was performed as previously described¹³⁰

Enhancer to Barcode Assignment, Dictionary Analysis, and SEL-Seq Data Analysis.

Assignment of barcodes to enhancers, dictionary setup and SEL-Seq data analysis was performed as previously described¹³⁰, with the only difference being that assignment of barcodes to enhancers was done using PacBio sequencing. Active enhancers will transcribe the mRNA barcode, thus the mRNA barcodes were measured to calculate enhancer activity. The library was transfected into human iPSC-CMs, samples were collected 96 hours after transfection, and we isolated barcode mRNA and DNA. To normalize the enhancer activity to differences in the amount of plasmid, we calculated $\log_2(\text{mRNA RPM} / \text{DNA RPM})$. This was done in 6 biological replicates across 2 cell lines, with an average Pearson correlation coefficient between replicates of 0.86.

Quantification and statistical analysis. For comparisons of counting data, the chi-squared test was used in Excel with the CHISQ.TEST function. For the post hoc test to avoid false positives, Bonferroni correction was used to adjust thresholds for significance. With N comparisons, the Bonferroni-adjusted P value needed for significance is $0.05/N$. Enhancer activity scores in Figure 4 were statistically analyzed using a standard two-sided T-test with Benjamini-Hochberg adjustment.

1.5.9 Data Availability

Sequencing data have been deposited at figshare and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.

Microscopy and scoring data reported in this paper will be shared by the lead contact upon request.

1.5.10 Code Availability

All original code has been deposited at GitHub and is publicly available as of the date of publication. DOI is listed in the key resources table.

Bed files of putative Ciona developmental heart enhancers and putative human developmental heart enhancers have been deposited at GitHub and DOI is listed in the key resources table

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

1.5.11 Acknowledgements

As this work was derived from the same manuscript, see *Jindal, Bantle*, Solvason*, et al. Developmental Cell 2023* for corresponding details.

Chapter 1, in full, is a reprint of the material as it appears in *Developmental Cell, 2023*. Granton Jindal, Alexis Bantle, Joe J. Solvason, Jessica L Grudzien, Agnieszka D’Antonio-Chronowska, Fabian Lim, Sophia H Le, Benjamin P Song, Michelle F Ragsac, Adam Klie, Reid O Larsen, Kelly A Frazer, & Emma K Farley. Thank you to all the co-authors for their contribution to the work, especially to Dr. Granton Jindal.

1.5.12 Author Contributions

As this work was derived from the same manuscript, see *Jindal, Bantle*, Solvason*, et al. Developmental Cell 2023* for corresponding details.

1.5.13 Competing Interests

As this work was derived from the same manuscript, see *Jindal, Bantle*, Solvason*, et al. Developmental Cell 2023* for corresponding details.

Chapter 2

Affinity-optimizing enhancer variants disrupt development

2.1 Abstract

Enhancers control the location and timing of gene expression and contain the majority of variants associated with disease¹⁻³. The ZRS is arguably the most well-studied vertebrate enhancer and mediates the expression of *Shh* in the developing limb¹³¹. Thirty-one human single-nucleotide variants (SNVs) within the ZRS are associated with polydactyly^{12;13;131}. However, how this enhancer encodes tissue-specific activity, and the mechanisms by which SNVs alter the number of digits, are poorly understood. Here we show that the ETS sites within the ZRS are low affinity, and identify a functional ETS site, ETS-A, with extremely low affinity. Two human SNVs and a synthetic variant optimize the binding affinity of ETS-A subtly from 15% to around 25% relative to the strongest ETS binding sequence, and cause polydactyly with the same penetrance and severity. A greater increase in affinity results in phenotypes that are more penetrant and more severe. Affinity-optimizing SNVs in other ETS sites in the ZRS, as well as in ETS, interferon regulatory factor (IRF), HOX, and activator protein (AP-1) sites within a wide variety of enhancers, cause gain-of-function gene expression. The prevalence of binding sites with suboptimal affinity in enhancers creates a vulnerability in genomes whereby SNVs that optimize affinity, even slightly, can be pathogenic. Searching for affinity-optimizing SNVs in genomes could provide a mechanistic approach to identify causal variants that underlie enhanceropathies.

2.2 Introduction

The human genome contains millions of enhancers¹³². These segments of the DNA act as switches to regulate where and when all the protein-coding genes in our genome are expressed. As such, enhancers encode the instructions for tissue-specific gene expression and thus successful development, adult homeostasis and cellular integrity¹³³. Most SNVs associated with phenotypic variation and disease are located in enhancers¹⁻³. Pinpointing which SNVs in an enhancer contribute to disease is a huge challenge because these causal variants are often embedded within a sea of inert variants^{1-3;134}. Our inability to pinpoint causal enhancer variants is a bottleneck in relating genotype to phenotype. Here we use mechanistic and generalizable principles governing enhancers to predict causal enhancer variants. Such an approach could enable systematic and scalable methods that harness the full potential of genomic data to improve human health.

To investigate the relationship between enhancer sequence and phenotypes, we focus on the ZRS enhancer¹³¹. This enhancer regulates the expression of *Shh* (*SHH* in humans) in the posterior of the developing limb buds in a region known as the zone of polarizing activity (ZPA), and is crucial for limb and digit development in vertebrates such as chicks, mice and humans¹³⁵⁻¹³⁷. This approximately 800-bp enhancer is highly conserved in sequence between mice and humans, and in both species it is located nearly 1 Mb away from the *Shh* promoter¹³¹. Although the ZRS is one of the most well-studied enhancers, how it encodes gene expression and how SNVs cause phenotypes are unclear. More than 30 SNVs in the ZRS found across vertebrates are associated with polydactyly and other limb defects such as tibial hemimelia (shortening of the tibia) (Supplementary Table 1). Several human families and mice have the same SNVs within the ZRS and show similar phenotypes, providing evidence of the robustness of polydactyly phenotypes across species and genetic backgrounds (Supplementary Table 1). The high degree of conservation in digit patterning and ZRS sequence across mice and humans makes the mouse an excellent system in which to study the genetic basis of polydactyly. Reporter assays analysing the effect of SNVs on ZRS enhancer activity in mice suggest that polydactyly is associated with gain-of-function (GOF) ectopic enhancer activity in the anterior limb bud. Eight human SNVs associated with polydactyly have been tested within the endogenous mouse locus,

and studies suggest that four of these cause Shh GOF expression in the anterior limb bud and extra digits^{13;138} (Supplementary Table 2). However, the mechanisms by which these SNVs alter enhancer function are poorly understood^{13;70;138}.

2.3 Results

2.3.1 Redundant low-affinity ETS sites regulate the ZRS

The ZRS is regulated by a combination of transcription factors, including HAND2, HOX, ETV4, ETV5, ETS-1 and GABPa^{96;138–141}. Five annotated sites known as ETS1–ETS5 are involved in the transcriptional activation of Shh from the ZRS and bind to the transcription factors ETS-1 and GABPa⁹⁶. Both ETS-1 and GABPa are activated downstream of fibroblast growth factor (FGF) signalling from the apical ectodermal ridge¹⁴². Deletion of all five ETS sites results in the complete loss of enhancer activity in the ZPA when tested by reporter assays in mice⁹⁶. Deletion of individual sites has no effect on expression; however, deleting combinations of these sites leads to a significant reduction of expression within the ZPA⁹⁶. These results show that the five ETS sites (ETS1–ETS5) are redundantly necessary for the activation of Shh expression in the ZPA. An emerging regulatory principle that governs enhancers—including ones regulated by ETS—is the use of suboptimal-affinity binding sites (also known as low-affinity or submaximal binding sites) to encode enhancer tissue specificity^{5;6;86}. This principle has been studied mainly in invertebrates^{5;6;86}.

To investigate whether the vertebrate ZRS also adheres to this regulatory principle, we measured the relative affinity of the five ETS sites (ETS1–ETS5) using protein binding microarray (PBM) data for the mouse transcription factor ETS-1^{93;123}. PBM measures the binding affinity of all possible 8-mer sequences for the transcription factor of interest to provide a direct measurement of binding¹²³. A relative affinity is then calculated by comparing the signal of all 8-mers to the signal of the highest-affinity site, which has a score of 1.00 or 100%. The DNA binding specificities of ETS-1 and other class I ETS transcription factors are conserved in mice and humans (Extended Data Fig. 1). Therefore, the binding affinities measured for ETS-1 convey the binding affinities of other class I ETS transcription factors that are expressed within the limb bud and which might also bind to this locus, such as GABPa^{93;96;140}. Although PBM is

an in vitro measurement, the relative affinity defined by PBM shows a strong correlation with in vivo chromatin immunoprecipitation (ChIP) peak intensity across several datasets (Extended Data Fig. 2).

The five previously annotated and functionally validated ETS sites (ETS1–ETS5)⁹⁶ in the ZRS have suboptimal binding affinities, ranging from 0.26 to 0.39 relative to consensus (Fig. 1a). We identify a total of 19 putative ETS sites in the human ZRS, and 18 in the mouse ZRS, all of which have suboptimal affinity. Fifteen of these sites are conserved in location and affinity between human and mouse (Extended Data Fig. 3a). One of these conserved sites is a newly identified ETS-A site that has an extremely low affinity of 0.15. We confirm that this ETS-A sequence binds the transcription factor ETS-1 using an electrophoretic mobility shift assay (EMSA) (Fig. 1a and Extended Data Fig. 4).

Figure 2.1 An ETS-A site in the ZRS enhancer contains two human variants that are associated with polydactyly, both of which subtly increase ETS binding affinity.

a, The human ZRS contains five known and functionally validated ETS sites, ETS1–ETS5, all of which have suboptimal affinity. We identify a new site, ETS-A, which has a relative affinity of 0.15. Six HOX sites (yellow) and one HAND2 site (pink) have also been previously identified. Thirty-one SNVs associated with polydactyly are found in humans (black bars), and SNVs are also found in other species such as cats, mice and chicks (green bars). b, Two human SNVs associated with polydactyly, denoted French 2 and Indian 2, occur within the ETS-A site. Both SNVs lead to a subtle increase in relative affinity of ETS-A to 0.24 and 0.26 respectively. c–f, The ETS-A sequence in a reference (Ref) mouse (c) drives the expression of Shh (d) and Ptch1 (e) restricted to the posterior domain of the developing limb bud in E11.75 and E12.0 embryos, respectively, as shown by in situ hybridization. f. Skeletal staining shows a WT mouse hindlimb with normal digit morphology. g–j, The French 2 SNV (g) drives the ectopic expression of Shh (h) and Ptch1 (i) in the anterior limb bud of homozygous embryos (arrow) in addition to the normal domain of posterior expression. j, A mouse hindlimb homozygous for French 2 has an extra triphalangeal thumb. k–n, The Indian 2 SNV (k) drives the ectopic expression of Shh (l) and Ptch1 (m) in homozygous embryos (arrow). n, A hindlimb from an Indian 2 homozygous mouse has an extra triphalangeal thumb. We did not calculate n for Shh because the expression is highly dynamic and thus hard to accurately capture; instead, we calculate the n of Ptch1 as a readout of Shh.

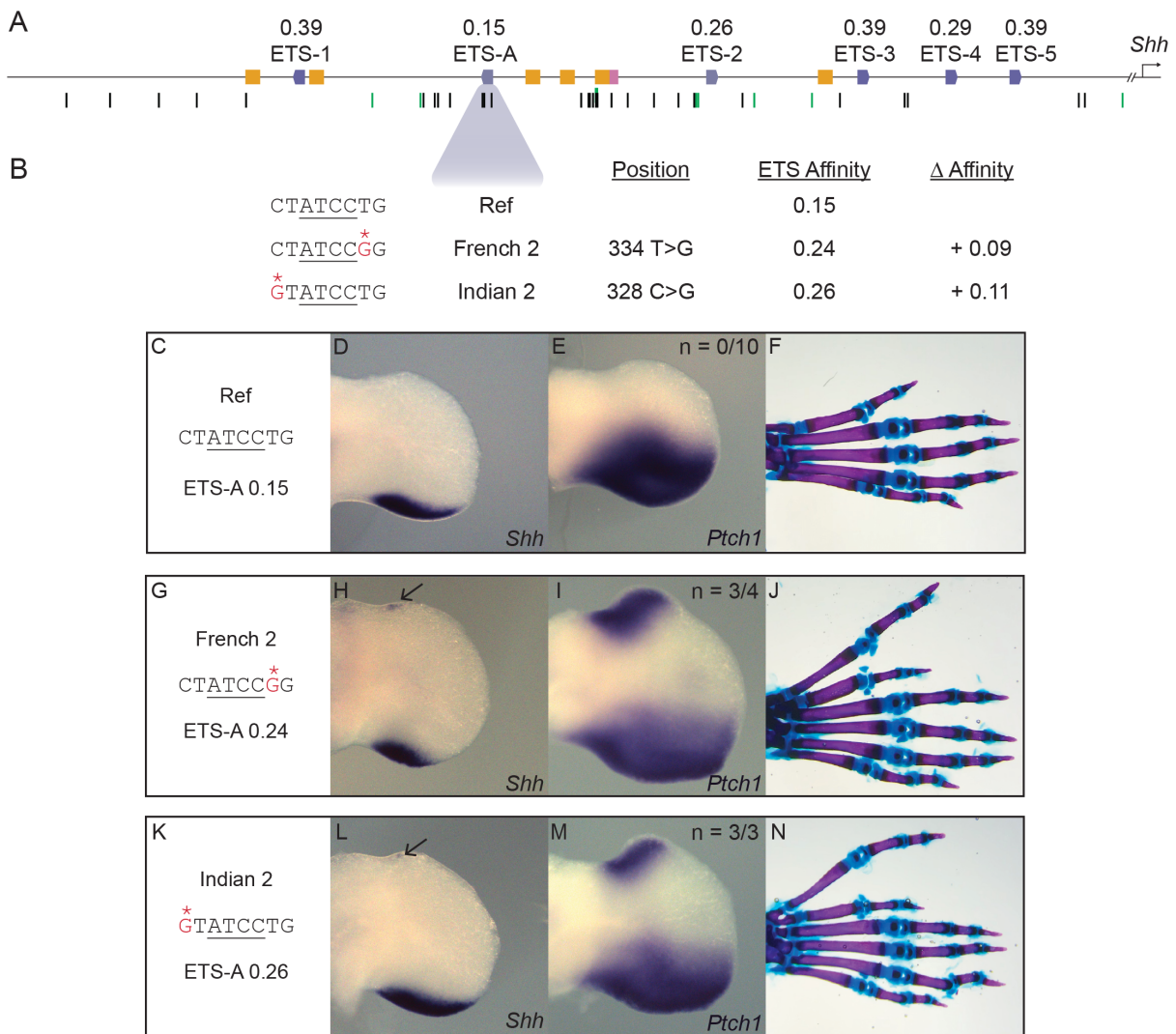


Figure 2.1. An ETS-A site in the ZRS enhancer contains two human variants that are associated with polydactyly, both of which subtly increase ETS binding affinity.

2.3.2 Human polydactyly SNVs subtly increase affinity

The ETS-A site lies in a region of the ZRS that is completely conserved between mice and humans (Extended Data Fig. 3b). Two human variants associated with polydactyly—the French 2 (334 T>G) and Indian 2 (328 C>G) variants—are located in the ETS-A site (Fig. 1b). The French 2 variant is found in a family that has preaxial polydactyly; it is incompletely penetrant because three out of four family members with this variant have an extra thumb¹². Only one individual with the Indian 2 variant has been identified and has preaxial polydactyly¹³. Notably, both human variants cause a similar subtle increase in the relative affinity of the ETS-A site from 0.15 in the reference to 0.24 in French 2 and 0.26 in Indian 2. EMSA confirms that both variants bind to ETS-1 more strongly than does the wild-type (WT) ETS-A (Extended Data Fig. 4). We hypothesized that this slight 1.6-fold increase in the relative affinity of the ETS-A site could be causing polydactyly. Although both variants have been studied using LacZ reporter assays in mouse, these assays suggest that only the French 2 variant alters enhancer activity¹³. Neither of these human variants have been studied within the endogenous ZRS locus, and the mechanism by which they alter gene expression is unknown. Therefore, we first sought to determine whether mice with the French 2 and Indian 2 variants exhibit ectopic expression of *Shh* and preaxial polydactyly.

French 2 and Indian 2 homozygous mice show ectopic expression of *Shh* in the anterior of the hindlimb at embryonic day (E)11.75 (Fig. 1h,l). The domain of ectopic *Shh* expression is tiny and highly dynamic. Therefore, we also looked at *Ptch1*, a direct downstream target of *Shh* that is commonly used as a readout for *Shh* signalling^{138;143}. *Ptch1* is ectopically expressed in the French 2 and Indian 2 E12.0 homozygotes (Fig. 1i,m). We observed no ectopic expression of *Shh* or *Ptch1* in the forelimb at E11.75 or E12.0, owing probably to differences in the regulation of *Shh* expression in the forelimb and hindlimb¹⁴⁴. Heterozygous and homozygous French 2 and Indian 2 mice have preaxial polydactyly in their hindlimbs, which indicates that very small and transient ectopic expression of *Shh* can have a strong effect on digit number (Fig. 1j,n). In humans, polydactyly occurs most commonly on the forelimbs, but in mice it typically occurs on the hindlimb^{12;13;138} (Supplementary Table 1). This is likely to be due to differences in forelimb and hindlimb specification between the two species^{145;146}. The additional anterior digit

in mouse hindlimbs resembles the extra triphalangeal thumb observed in the orthologous human congenital malformations, and we call this a triphalangeal toe. Thus, both variants are causal for polydactyly and phenocopy the observed human phenotype.

2.3.3 Affinity-optimizing SNVs cause polydactyly

The French 2 and Indian 2 variants have the same phenotype, which suggests that the mechanism that drives polydactyly could be the same subtle increase in affinity of the ETS-A site. To test this prediction, we generated two more mouse lines with manipulations within the ETS-A site. The first mouse line, Syn 0.25, contains a synthetically created ETS-A site with an affinity of 0.25—the same affinity as that of the French 2 and Indian 2 variants—but has a different sequence change (Fig. 2a). We further validate the binding of ETS-1 to each of these ETS-A sequences with approximately 0.25 affinity using EMSA, and see no significant difference in the binding of ETS-1 to the French 2, Indian 2 or Syn 0.25 sequences (Extended Data Fig. 4). It is possible that any disruption to the ETS-A sequence could lead to a phenotype. To show that this is not the case, we also made a mouse line that we predicted would have no effect on phenotype. We created a loss-of-function (LOF) mouse in which the ETS-A binding site is ablated by removing a crucial nucleotide required for binding (Fig. 2e and Extended Data Fig. 4). Owing to the redundancy of ETS sites within the ZRS¹⁴⁷, we predicted that the loss of the ETS-A site would have no effect on Shh expression or limb development.

Mice containing the Syn 0.25 ETS-A site show ectopic expression of Shh and Ptch1 in the anterior limb bud at E11.75 and E12.0, respectively, mirroring the expression patterns observed in the French 2 and Indian 2 mice (Fig. 2b,c). Syn 0.25 mice also have preaxial polydactyly in their hindlimbs (Fig. 2d), which suggests that the affinity change rather than the sequence change is driving the phenotype. As predicted, mice with the LOF mutation show no ectopic expression of Shh or Ptch1 (Fig. 2f,g) and have normal limb morphology (Fig. 2h). Together, these studies show that the GOF increase in ETS-A affinity within the ZRS enhancer is pathogenic, whereas the LOF variant is non-pathogenic. This work demonstrates our ability to successfully predict the relationship between genotype and phenotype for sequence variants in the ETS-A site.

We hypothesized that if all three of these ETS-A affinity-optimizing variants share the

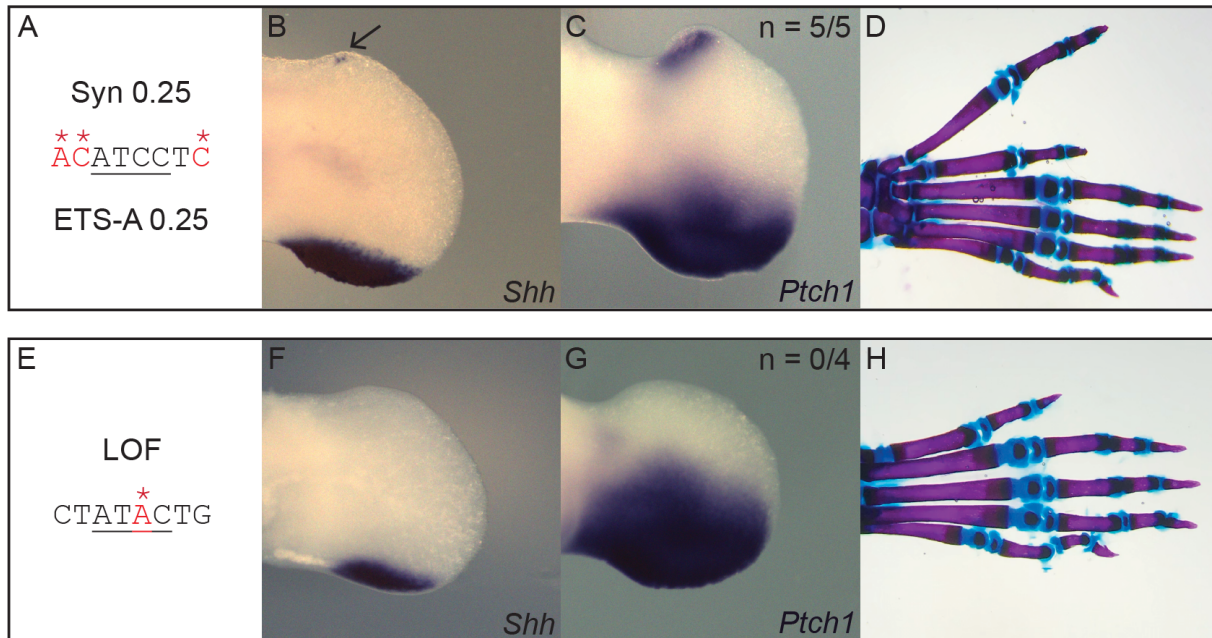


Figure 2.2. Synthetic changes to the ETS-A site that create a 0.25-affinity site and a LOF site cause predicted phenotypes. a–d, The Syn 0.25 ETS-A site (a), which has an affinity of 0.25, drives the ectopic expression of *Shh* (b) and *Ptch1* (c) in the anterior limb bud, in addition to the normal domain of expression in the posterior limb bud. d, Skeletal staining of a homozygous mouse hindlimb shows an extra triphalangeal thumb. e–h, The LOF ETS-A site (e) drives the normal expression of *Shh* (f) and *Ptch1* (g) in the posterior limb bud. h, Homozygous mice have normal digit morphology. We did not calculate n for *Shh* because the expression is highly dynamic and thus hard to accurately capture; instead, we calculate the n of *Ptch1* as a readout of *Shh*.

same mechanism of action then the penetrance, laterality and severity of polydactyly should be comparable between the three lines. Phenotyping of mice was done blind to genotype. French 2, Indian 2 and Syn 0.25 mice with the same affinity ETS-A site have similar penetrance and laterality in heterozygotes, with polydactyly occurring most frequently on the right hindlimb (Fig. 3a). Homozygotes exhibit phenotypes bilaterally and have a higher penetrance of polydactyly than do heterozygotes. Although all of the mice in this study are bred under identical conditions and have the same genetic background, there is a range of digit phenotypes (severity) in each line (Fig. 3b). Yet this distribution of digit phenotypes is identical across the lines. In heterozygotes the most common presentation is an extra digit that is either biphalangeal or triphalangeal, whereas homozygotes most commonly present with an extra digit that is triphalangeal. Thus, these three mouse lines with the same affinity increase show the same penetrance, laterality and severity with no statistical differences.

We next wanted to investigate whether the ETS-A site is indeed functional, which is challenging owing to redundancy within the enhancer⁹⁶. If the WT ETS-A site is contributing to enhancer activity, then the penetrance of polydactyly in mice containing one copy of the ETS-A Syn 0.25 allele and one copy of the WT ETS-A site should be higher than it is in mice with one allele of ETS-A Syn 0.25 and one allele of LOF ETS-A. The penetrance and severity of phenotypes in the ETS-A Syn 0.25/LOF ETS-A mice are significantly reduced relative to the ETS-A Syn 0.25/WT ETS-A mice (Fig. 3), thus demonstrating that the 0.15-affinity WT ETS-A site contributes to enhancer activity and phenotypes.

Our comprehensive analysis of phenotypes across 795 transgenic mice shows that all three variants that increase the affinity of ETS-A to around 0.25 have indistinguishable phenotypes in heterozygotes and homozygotes. This provides compelling evidence that the subtle affinity optimization of this ETS-A site is the mechanism that drives polydactyly. Although there is increasing recognition of the role of low-affinity sites within enhancers, sites with affinities as low as 0.15, and even those with affinities of 0.25, are still typically ignored¹³. Yet here we see that a 0.15-affinity site is functional, and that a 0.25-affinity site is not only functional but sufficient to disrupt normal limb development, indicating that subtle increases in low-affinity sites can be pathogenic.

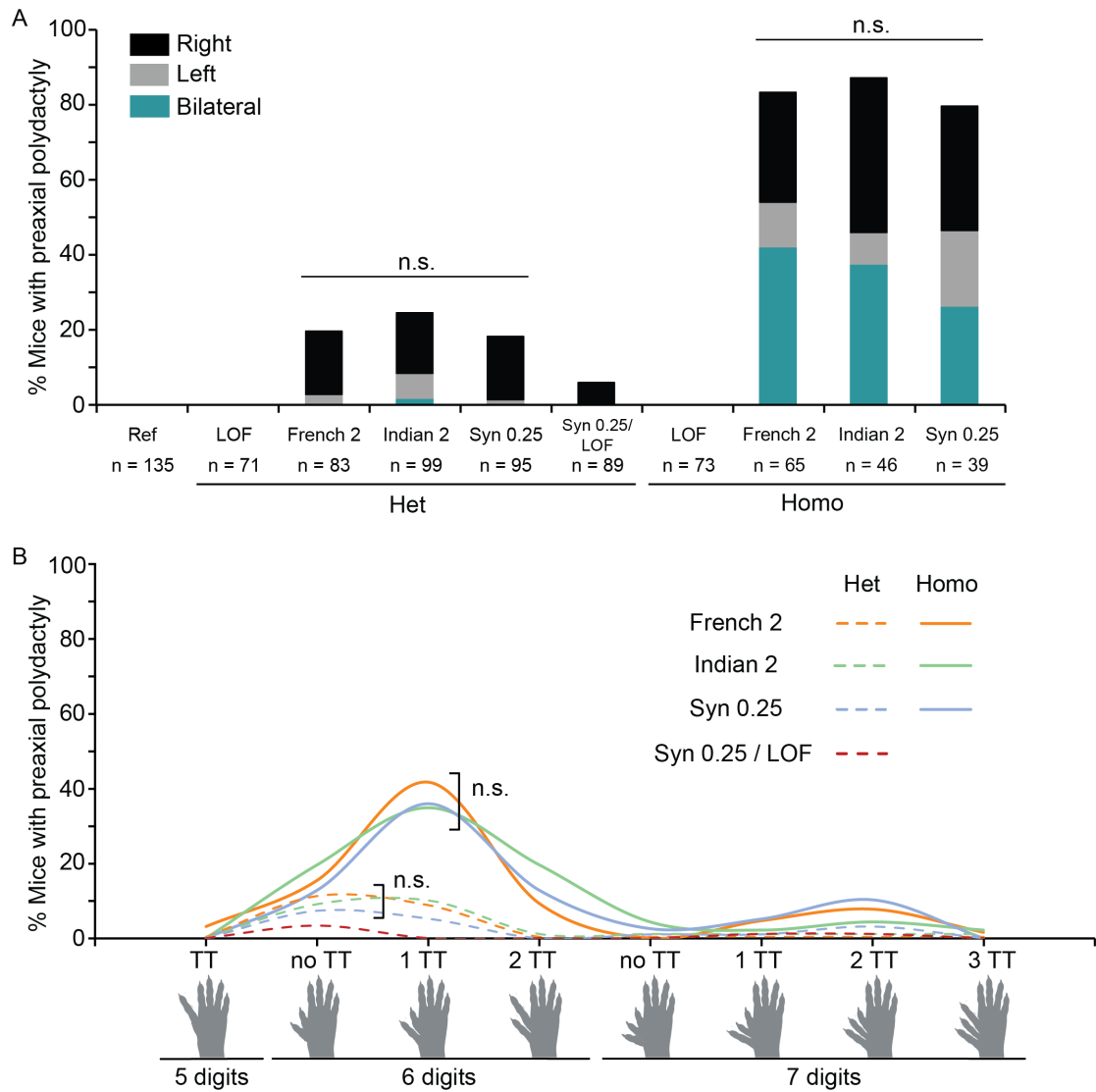


Figure 2.3. All mice with the approximately 0.25-affinity ETS-A are indistinguishable in terms of the penetrance, laterality and severity of polydactyly. a, Penetrance and laterality of phenotypes seen in French 2, Indian 2 and Syn 0.25 heterozygous (Het) and homozygous (Homo) mice. There is no significant difference in penetrance and laterality between heterozygotes and between homozygotes of French 2, Indian 2 and Syn 0.25 mice (Fisher's exact test). Polydactyly occurs more frequently on the right hindlimb in both heterozygotes and homozygotes. In Syn 0.25/LOF mice, polydactyly is less penetrant than in heterozygous Syn 0.25 mice (Syn 0.25/WT). NS, not significant. b, Polydactyly phenotypes seen in French 2, Indian 2 and Syn 0.25 heterozygotes (dashed lines) and homozygotes (solid lines), as well as in Syn 0.25/LOF mice. There is no significant difference in the phenotypes seen between the French 2, Indian 2 and Syn 0.25 lines (Fisher's exact test). TT, triphalangeal toe. Biorender.com was used to generate images in this figure.

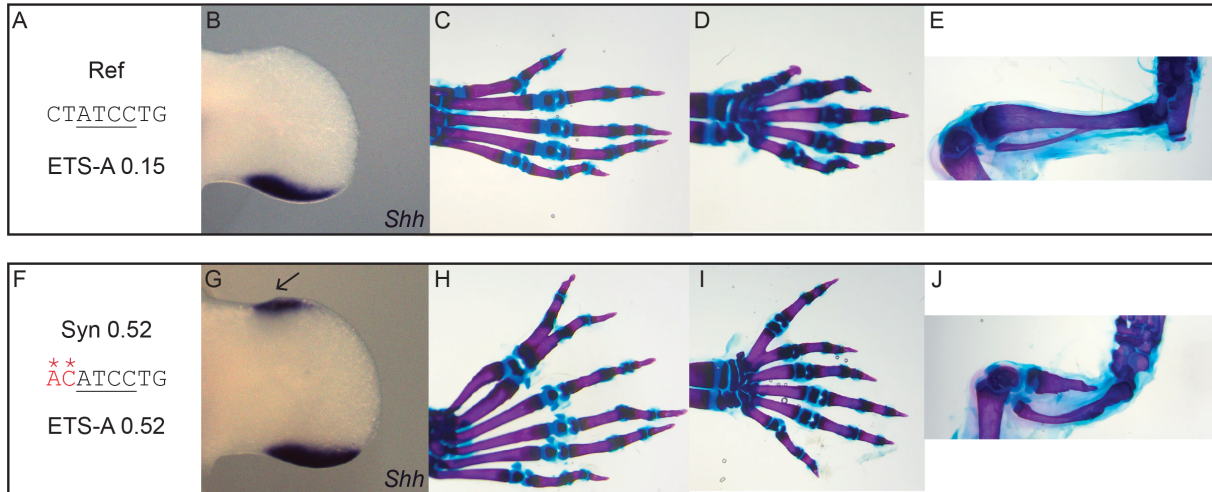


Figure 2.4. A greater increase in affinity at the ETS-A site causes more severe and penetrant polydactyly and long-bone defects. a, The ETS-A reference (Ref) sequence. b, In situ hybridization of *Shh* in WT hindlimb at E11.75 shows a domain of expression in the posterior of the limb bud. c, Skeletal staining of a WT mouse hindlimb showing five digits. d, WT forelimb showing five digits e, WT hindlimb focused on tibia and fibula. f, The Syn 0.52 ETS-A sequence. g, In situ hybridization of *Shh* in a Syn 0.52 homozygous mouse hindlimb bud at E11.75. Ectopic expression in the anterior of the limb bud is indicated by an arrow. h, Skeletal staining of a homozygous Syn 0.52 ETS-A mouse hindlimb showing six digits. The toe and the extra digit are both triphalangeal. i, Homozygous Syn 0.52 mouse forelimb showing six digits. Both the thumb and the extra digit are triphalangeal. j, Hindlimb focused on the tibia and fibula in a Syn 0.52 ETS-A homozygous mouse. The tibia is severely shortened. The hindlimbs in e,j were imaged with the same magnification. See Extended Data Fig. 6 for more details.

2.3.4 Predicting penetrance and severity

Having seen that a subtle increase in affinity can cause developmental defects, we wondered whether the degree of affinity change could predict the penetrance and severity of phenotypes. This mechanistic understanding could be valuable for diagnostic and treatment purposes. To test this hypothesis, we created a mouse line with a 0.52-affinity ETS-A site (Fig. 4f). EMSA confirms that this sequence binds ETS-1 more strongly than do the WT ETS-A or the 0.25 ETS-A sites (Extended Data Fig. 4).

French 2, Indian 2 and Syn 0.25 homozygotes have a small amount of ectopic Shh expression in the hindlimb, and the Syn 0.52 site, as predicted, causes a large domain of Shh and Ptch1 expression in the anterior of the hindlimb, as well as ectopic expression in the forelimb (Fig. 4g and Extended Data Fig. 5g,i). Consistent with these expression patterns, Syn 0.52 mice have polydactyly in both the forelimbs and the hindlimbs, whereas mice with 0.25 affinity have polydactyly only in the hindlimbs (Fig. 4h,i). Moreover, in Syn 0.52 mice, polydactyly is fully penetrant in the hindlimbs for both heterozygotes and homozygotes, and is almost fully penetrant in the forelimbs (Extended Data Fig. 6a). Polydactyly in Syn 0.52 mice is most commonly bilateral in both heterozygotes and homozygotes, whereas unilateral phenotypes are prevalent in 0.25 ETS-A affinity mice. Most Syn 0.52 mice have six digits that are all triphalangeal, but some have seven or even eight digits; this is more severe than the polydactyly seen in the 0.25 ETS-A affinity mice (Extended Data Fig. 6c). Syndactyly also occurs more frequently in the Syn 0.52 mice than in the French 2, Indian 2 and Syn 0.25 mice. In addition, Syn 0.52 mice have defects in the long bones. Tibial shortening or tibial hemimelia, a condition seen in humans¹³⁸, occurs in 95% of Syn 0.52 homozygotes (Fig. 4j and Extended Data Fig. 6b). Thus, as predicted, greater increases in affinity lead to more penetrant and severe phenotypes. This raises the possibility that affinity increases could be used to predict the severity and penetrance of phenotypes.

2.3.5 Affinity-optimizing SNVs prevalent across the ZRS

We next wanted to see whether our ability to predict causal enhancer variants could generalize to the other ETS sites within the ZRS. Seventeen SNVs cause an increase of at least 1.6-fold in ETS affinity; this is the fold change occurring in French 2, Indian 2 and Syn 0.25

(Fig. 5a). To see whether these variants drive gain of function, we analyse data from published ZRS reporter assays. In one of these assays that tests the effect of 2% sequence mutagenesis of the human ZRS enhancer on expression in mouse limb buds¹³, we find that enhancer variants containing affinity-optimizing ETS SNVs are significantly enriched for GOF expression. However, there are around 16bp changes within each enhancer, so we cannot definitively attribute the GOF expression to an individual SNV (Extended Data Fig. 7).

To more accurately relate SNVs to changes in gene expression, we analysed a massively parallel reporter assay (MPRA) that conducted saturation mutagenesis of a 485-bp region of the ZRS enhancer in a limb-like cell line¹⁴⁸. Because spatial expression cannot be assayed in a cell line, levels of reporter expression serve as a readout of GOF enhancer activity. There is a significant enrichment of GOF gene expression in enhancers that contain affinity-optimizing ETS SNVs, relative to all other SNVs within the dataset (Fig. 5b). By contrast, there is no significant enrichment of GOF activity within SNVs that do not alter ETS binding affinity. Four of the eleven affinity-optimizing SNVs in this region of the ZRS drive GOF gene expression; these SNVs occur in ETS-A, ETS-B, ETS2 and ETS3. In this cell-line MPRA, the French 2 and Indian 2 SNVs did not significantly increase expression¹⁴⁸. The false negative classification of French 2 and Indian 2 is likely to be due to the use of a cell line and the difficulty of assessing dynamic and very subtle changes in gene expression. The Indian 2 SNV was similarly misclassified in another reporter assay within the developing limb bud¹³, in which three embryos were screened for expression at E11.5. Our analysis of data from these two ZRS mutagenesis studies^{13;148}, and our in-depth study of variants within the ETS-A site, show that affinity-optimizing SNVs in four different ETS sites across the ZRS can cause GOF gene expression.

To see whether our findings generalize beyond ETS to other transcription factor binding sites (TFBSs) in the ZRS, we searched for HOX affinity-optimizing SNVs associated with polydactyly. Both HOXA13 and HOXD13 are expressed in the distal limb bud, and changes in their expression can affect digit and limb development¹⁴⁹. The Dutch 2 SNV (165A>G32) increases the affinity for both HOXA13 and HOXD13. EMSA confirms that the Dutch 2 SNV binds these HOX proteins more strongly than the WT sequence does, suggesting that an increase in HOX binding underlies the Dutch 2 polydactyly phenotype (Extended Data Fig. 8a,b and

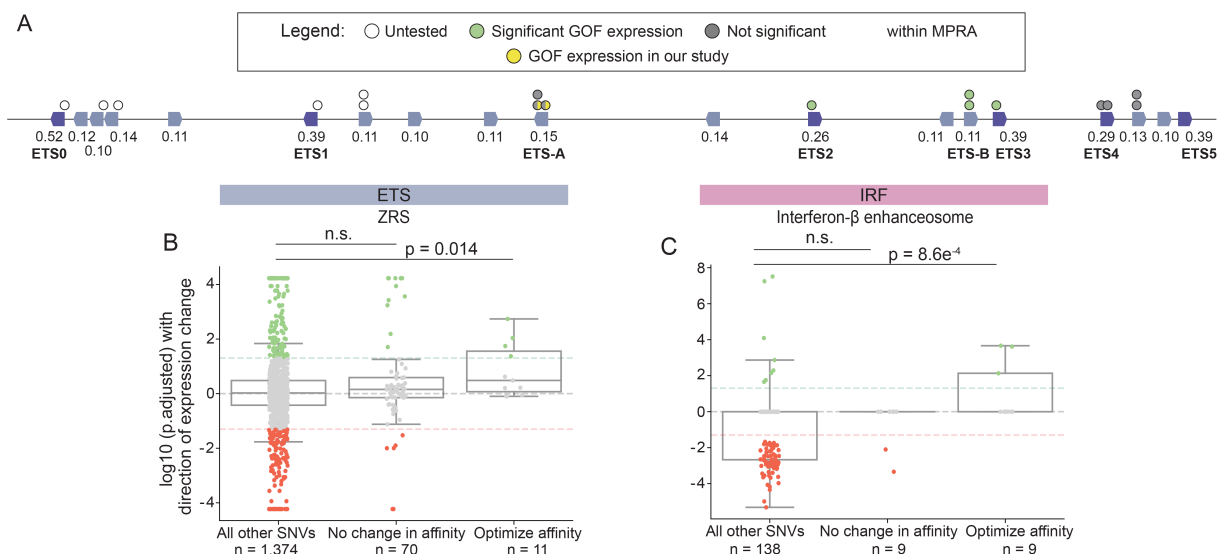


Figure 2.5. Affinity-optimizing SNVs drive GOF expression in the ZRS and IFN β enhanceosome. a, Schematic of the human ZRS showing 19 putative ETS sites and 17 SNVs that increase the affinity of ETS sites by 1.6-fold or more. ETS sites identified from previous studies (ETS0–ETS5) are labelled^{15,51}. We annotated changes in expression caused by these SNVs on the basis of results from a saturation mutagenesis MPRA and our study. b,c, Box plots showing all SNVs tested within MPRA mutagenesis experiments, their significance and their effects on expression. The bounds of the box plots define the 25th, 50th and 75th percentiles, and whiskers are 1.5 $\times$ the interquartile range. The y axis is the log of the adjusted P value (Padj) with the direction of expression change (positive values indicate an increase in expression and negative values indicate a decrease in expression). Dashed horizontal lines indicate significance thresholds at $P=.05$. Each dot represents a tested SNV. SNVs that have a significant increase in expression are shown in green, those with no significant change in expression are grey and those with a significant decrease in expression are red. For each plot, we show SNVs that increase the affinity of the transcription factor, SNVs within the TFBS that do not alter affinity and all other SNVs. b, Analysis of the ETS TFBS in a saturation mutagenesis MPRA on a 485-bp region of the ZRS enhancer. c, Analysis of the IRF TFBS in a saturation mutagenesis MPRA on the IFN β enhanceosome. We used one-tailed Mann–Whitney U tests to determine any significant enrichment for GOF enhancer activity

Supplementary Table 3).

2.3.6 The enhanceosome contains affinity-optimizing SNVs

The interferon- β (IFN β) enhanceosome is a well characterized enhancer that switches on IFN β gene expression as an immune response to viral infection^{43;150}. IRF-binding sites are necessary for enhancer activity¹⁵¹. We analysed MPRA assays that mutated every base pair within this enhancer⁶⁰. SNVs that increase the affinity of IRF binding sites by a fold change of at least 1.5 are significantly enriched in GOF expression relative to all other SNVs in the MPRA, whereas there is no significant enrichment for GOF expression in SNVs within IRF sites that did not alter IRF binding affinity (Fig. 5c). One of these IRF affinity-optimizing SNVs was independently tested in another assay and also showed GOF expression¹⁵¹. These findings provide further evidence of the general role of affinity-optimizing SNVs in GOF expression within another enhancer regulated by different transcription factors and active in different cells. Unlike the ZRS enhancer, the enhanceosome is a redundant enhancer^{152;153}. Therefore, within the context of reporter assays, the principle of affinity optimization applies to two classic enhancers and examples of a redundant and a non-redundant enhancer.

2.3.7 Other transcription factors and disease enhancers

To investigate the role of affinity-optimizing variants in other enhancers, we looked at MPRA assays conducted on 11 enhancers associated with a variety of diseases¹⁴⁸ (Extended Data Fig. 9). Saturation mutagenesis was used to assay the effect of each base on enhancer activity. Each MPRA was performed in a cell line relevant for each particular enhancer¹⁴⁸. Because FGF signalling is important in many cell types and aberrations to FGF signalling are implicated in a variety of disease contexts, we did not filter for enhancers responsive to ETS, but simply looked across all 11 enhancers. The median ETS affinity within these enhancers is 0.12. Enhancers with affinity-optimizing ETS SNVs (≥ 1.6 -fold) are significantly enriched for GOF expression, whereas SNVs that occur in ETS sites but do not change their affinity are not significantly enriched in GOF enhancer activity (Fig. 6a). There are many SNVs that slightly increase binding affinity and are associated with GOF enhancer activity, indicating that small increases in binding affinity can contribute to GOF gene expression across a wide range of

enhancers.

We also looked at AP-1, a transcription factor that is involved in many cellular processes, including differentiation and cell proliferation¹⁵⁴. Because AP-1 is a commonly used transcription factor, we again did not filter for enhancers known to be regulated by AP-1, but simply looked at all 11 enhancers. Similarly, we find that SNVs in the MPRA assay that increase the affinity of AP-1 (≥ 1.5 -fold) lead to GOF gene expression, whereas SNVs in AP-1 sites that do not alter affinity are not significantly enriched in GOF expression (Fig. 6b).

To see whether our findings generalize to other datasets, we analysed an MPRA screen that tested lymphoblastoid regulatory elements and variants within these elements that were identified in an expression quantitative trait locus (eQTL) study¹⁰. This study looked at the genomes and lymphoblastoid transcriptomes of 446 individuals of Yoruba and European descent, and individuals from the 1000 Genomes Project¹⁵⁵. The genomes and mRNAs from these lymphoblastoid cells were analysed to correlate genomic variation with changes in gene expression. Top-associated genomic variants, or variants in linkage disequilibrium with these eQTL variants, were chosen to be tested in the MPRA. In total, the effects of more than 3,000 SNVs were measured by MPRA in the same cell line as the eQTL study. As predicted, we see a significant enrichment in GOF enhancer activity within affinity-optimizing SNVs for both ETS and AP-1. SNVs occurring within ETS or AP-1 that do not alter affinity are not enriched for GOF enhancer activity (Fig. 6c,d). Together, these analyses on two orthogonal MPRA datasets show that, for two different transcription factors, over a massive variety of contexts, affinity-optimizing SNVs sites are a common mechanism driving GOF gene expression in reporter assays.

Figure 2.6 Affinity-optimizing SNVs drive GOF expression in a wide variety of disease-associated enhancers. a–d, Analysis of MPRA for a variety of enhancers in different cell types. Box plots showing all SNVs tested within MPRA mutagenesis experiments and their significance and effects on expression. The bounds of the box plots define the 25th, 50th and 75th percentiles, and whiskers are $1.5\times$ the interquartile range. One-tailed Mann–Whitney U test. a,b, Analysis of saturation mutagenesis MPRA assays of 11 disease-associated enhancers comparing all SNVs, SNVs within TFBSs that do not alter affinity and SNVs that increase the affinity of TFBSs for ETS (a) and AP-1 (b). c,d, Analysis of MPRA comparing the effect of SNVs within lymphoblastoid regulatory elements for the ETS TFBS (c) and the AP-1 TFBS (d). e, Filtering for affinity-optimizing (aff-opt) SNVs significantly increases our ability to predict causal GOF enhancer variants. The bar graph shows the percentage of all SNVs that lead to GOF expression relative to the percentage of affinity-optimizing SNVs that lead to GOF expression. Green bars indicate SNVs that cause GOF expression within analysed MPRA datasets; yellow bar indicates SNVs that cause GOF expression in our the current study—namely, French 2 and Indian 2. Fisher’s exact test was used to determine any significant enrichment for GOF expression in the all and aff-opt categories: $**P<0.001$, $*P<0.01$. Affinity-optimizing SNVs are those that lead to a fold change of at least 1.6 for ETS because this is the fold change for French 2, Indian 2 and Syn 0.25. Affinity-optimizing SNVs for AP-1 and IRF are those that cause a fold change of at least 1.5.

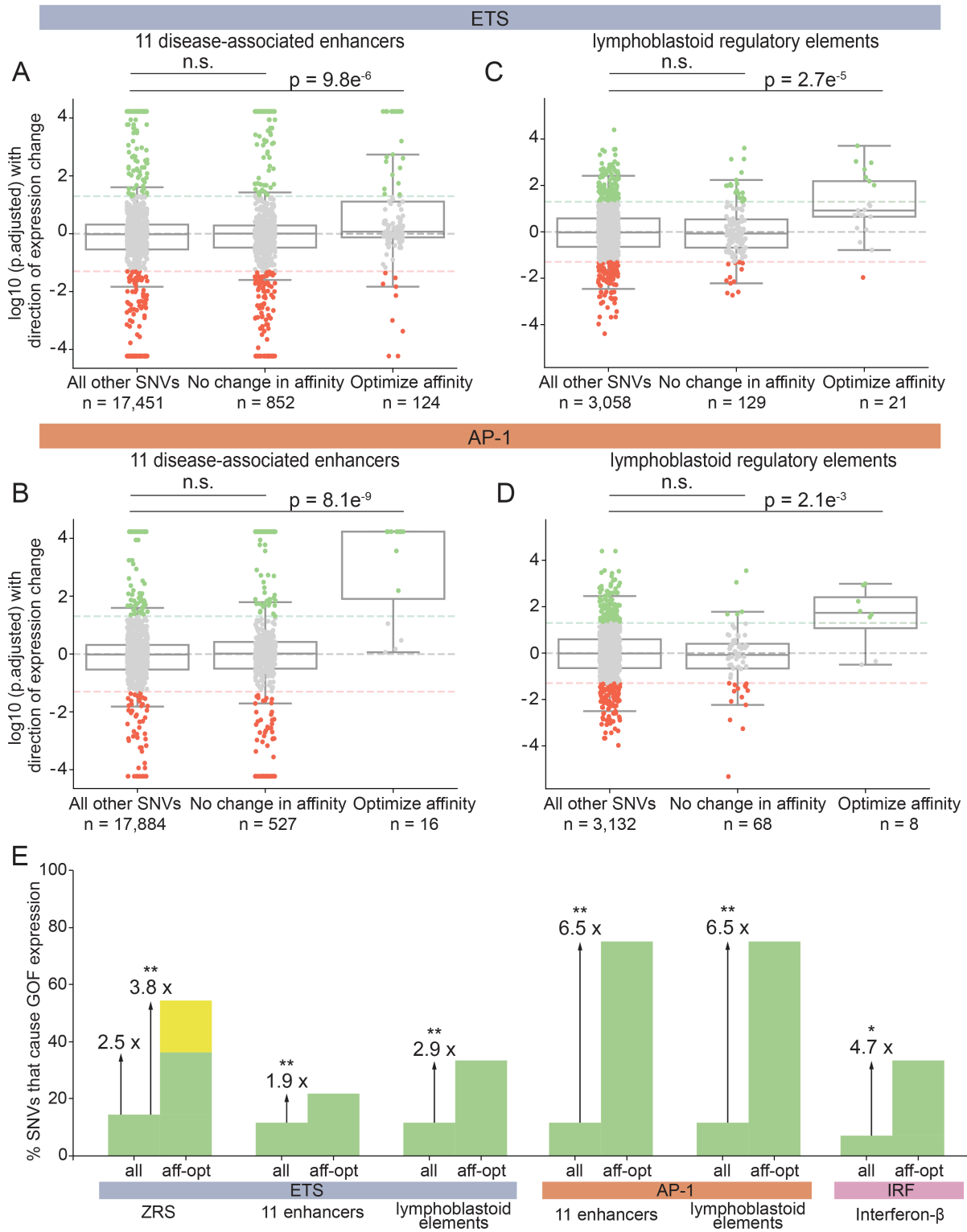


Figure 2.6. Affinity-optimizing SNVs drive GOF expression in a wide variety of disease-associated enhancers.

MPRAs allow us to study the effects of variants on expression in the context of a reporter assay, whereas eQTL infers the effects of variants on target gene expression in the context of the genome. Of the seven ETS affinity-optimizing GOF SNVs identified in the lymphoblastoid MPRA, five are significant eQTLs. All five of these eQTLs are associated with an increase in target gene expression (Extended Data Fig. 10a). Therefore, the increased reporter expression seen in the MPRA correlates with target gene expression in the endogenous locus. Indeed, when we analyse the eQTL signal of all 2,663 eQTL variants in the study¹⁰, we find that ETS affinity-optimizing SNVs are significantly enriched in GOF expression for target genes (Extended Data Fig. 10b). By contrast, eQTL variants in ETS sites that do not alter affinity have no significant enrichment in GOF expression for target genes.

Genome-wide, eQTL ETS affinity-optimizing SNVs show significant enrichment in the increased expression of target genes (positive beta values), whereas SNVs that do not alter the affinity of ETS sites show no enrichment in the increased expression of target genes¹⁵⁵ (Extended Data Fig. 10c). Furthermore, with the eQTL dataset, we find that higher fold changes of ETS show a more significant enrichment of GOF target gene expression (Extended Data Fig. 10d). We also see this enrichment for AP-1 (Extended Data Fig. 10e). Although not all changes in expression detected in eQTL analysis are direct, the enrichment we see suggests that affinity-optimizing SNVs drive the GOF expression of target genes in the endogenous locus and that this GOF activity is not buffered within the endogenous context.

2.3.8 Regulatory principles predict causal SNVs

Enhancers are littered with variation; therefore, predicting which SNVs are causal is a major challenge in relating genotype to phenotype. Within the ZRS MPRA dataset, 14.5% of all SNVs drive GOF expression, whereas 36% of affinity-optimizing SNVs drive GOF gene expression. Searching for affinity-optimizing ETS SNVs in the ZRS increases our ability to find causal GOF enhancer variants by 2.5 times when considering the MPRA data alone (Fig. 6e). The MPRA misses the two GOF variants that we identified in this study (French 2 and Indian 2). Thus, when taking into account all known affinity-optimizing SNVs in the ZRS, 6/11 or 55% of affinity-optimizing ETS SNVs drive GOF expression, which increases our ability to find causal

variants in the dataset by 3.8 times (Fig. 6e). Indeed, searching for SNVs that increase the affinity of ETS, AP-1 and IRF improves our ability to find causal GOF variants by twofold to sevenfold (Fig. 6e). Searching for affinity-optimizing SNVs is an effective and simple method to pinpoint putative causal enhancer variants, and could be a valuable filter to prioritize enhancer variants for further functional analysis.

2.4 Discussion

Suboptimal-affinity binding sites are prevalent in a variety of enhancers, including Otx-a, svb, the ZRS, the IFNb enhanceosome and many disease-associated enhancers and developmental heart enhancers^{5-7;86}. Here we show that single-nucleotide changes that subtly increase binding affinity cause the loss of tissue specificity and organismal phenotypes in the mouse and human limb. In a complementary study, we find that affinity-optimizing SNVs in a Ciona heart enhancer drive ectopic expression of the gene FoxF in non-heart cells, which causes abnormal cell migration and heart defects as severe as two beating hearts⁷. Our results suggest that the prevalence of suboptimal-affinity sites creates a vulnerability in genomes whereby affinity-optimizing SNVs can drive ectopic GOF expression and phenotypes.

Given our findings, a greater focus on low-affinity but highly degenerate sites is essential to identify and predict causal enhancer variants. In this study, we show that PBM is a highly effective method to measure affinity. Although in vivo binding is no doubt modulated by other in vivo factors, such as protein-protein interactions, IDRs and other molecular interactions, our results show that the affinity of transcription factor binding is a fundamental feature driving enhancer activity and gene expression. The use of such a simple in vitro measurement provides a systematic method to identify causal enhancer variants that does not rely on specific measurements within cell types. This enables a generalizable approach for pinpointing causal GOF enhancer variants that is applicable across genomes, cell types and even species.

The ETS-A site is an 8-bp sequence, and there are 65,536 possible sequence combinations within this region. Of these combinations, we successfully predicted the expression and phenotypes of five different ETS-A sequences (French 2, Indian 2, Syn 0.25, ETS-A LOF and Syn 0.52) on the basis of changes in binding affinity. These experiments demonstrate the power of mechanistic

rules to predict causal enhancer variants. In the future, we will want to move beyond just identifying causal enhancer variants to predicting severity and penetrance. Within the ZRS ETS-A site, a greater increase in affinity causes more severe and penetrant phenotypes. This is likely to be true for all changes to affinity that occur in the same binding site and at the same position within an enhancer, because all of these affinity increases happen within the same context or grammar. The effects of affinity-optimizing SNVs are likely to be modulated by the surrounding binding sites, such that variants at different positions in an enhancer could have different effects despite having the same affinity increase⁹. We see evidence of this in our study of affinity-optimizing SNVs within the Ciona FoxF heart enhancer⁷. Integrating an understanding of affinity-optimizing SNVs and enhancer grammar will refine our ability to predict the severity and penetrance of enhancer variants.

We find that enhancer variants that cause GOF—but not LOF—enhancer activity disrupt development. The redundancy of enhancers on multiple levels ensures robustness within an organism. Typically, multiple enhancers known as redundant or shadow enhancers regulate the same gene^{156;157}. Another layer of enhancer redundancy is encoded within enhancers^{97;158}, as exemplified by the five ETS sites (ETS1–ETS5) in the ZRS⁹⁶. Therefore, the loss of a single activator site, a reduction in the affinity of an activator site or even the loss of an entire enhancer can be compensated for by redundant sites or enhancers. By contrast, GOF variants that lead to increased levels of expression, or spatio-temporal ectopic expression, are harder to buffer and thus more likely to affect gene expression and development. This is exemplified by the LOF ETS-A site, which has no effect, whereas all four GOF variants drive ectopic gene expression and disrupt limb development. Focusing on variants that result in GOF expression could improve our ability to pinpoint causal enhancer variants.

Clusters of transcription factors are often found in close proximity to active enhancers. These have been described as hubs, and contain a large concentration of transcription factors that may be phase separated^{108;159}. In such an environment, it is counterintuitive that single-base-pair changes can have such a marked effect on expression. At a biochemical level, we speculate that the subtle increase in affinity allows a longer dwell time for an activator; this could increase the chances of all required factors binding and the formation of a functional complex that can

trigger transcription. Further investigations into the mechanisms by which one SNV that slightly increases binding affinity can nucleate transcriptional activation could help us to understand the driving forces behind transcriptional control.

Thirty-one human SNVs in the ZRS are associated with polydactyly. Only seven of these lie within validated binding sites; this is likely to be a result of the degenerate nature of binding sites and our poor annotation of functional binding sites in the ZRS. Of the seven SNVs in binding sites, three are affinity-optimizing SNVs. Two of these SNVs lie within the ETS-A site (French 2 and Indian 2) and one in a HOX site (Dutch 2). The ETS-A affinity-optimizing SNVs cause ectopic GOF expression in the anterior limb bud. Stem tetrapods were polydactylous with seven or eight digits¹⁶⁰. The expression of repressors in the anterior limb bud might have contributed to the derived five-digit state. We speculate that the ETS-A affinity-optimizing SNVs cause ectopic expression in the anterior limb bud because the increase in affinity disturbs the balance of activators and repressors acting on the enhancer in the anterior limb bud, and because of an evolutionary sensitivity.

Millions of variants that are significantly associated with phenotypes and disease are located in enhancers¹⁻³, and functionally testing all of these is a major challenge. MPRA-style experiments assay the effect of enhancer variants through reporter assays; however, such assays have limitations as they do not test variants in the endogenous locus, or the appropriate cellular or multicellular context^{10;148}. In addition, MPRA tend to highlight variants that lead to the largest and most significant changes in expression. However, these might not have the greatest effect on phenotype. Indeed, the French 2 and Indian 2 variants drive ectopic expression in just a handful of cells for a short time, but this is sufficient for the formation of extra digits. Although not all small dynamic domains of ectopic expression will cause a phenotype, small changes in the temporal and spatial expression of morphogens, signalling-pathway proteins and effectors—especially in cell types in which these would alter identity—are likely to lead to patterning defects and phenotypes. Weighing the degree of expression change on the basis of the type of target gene and the sensitivity of the cellular context, and by focusing on variants that cause GOF rather than LOF expression, could improve our ability to predict enhancer variants that alter phenotypes.

Enhancers are often categorized according to several characteristics: their mode of interaction with the promoter; their level of sequence conservation; their distance from the target promoter; their target gene; the tissue in which they are active; and the species in which they are found. Our studies of the ZRS, FoxF heart enhancer, IFNb enhanceosome and disease-associated enhancers, together with our eQTL analysis, show that the use of suboptimal-affinity sites to encode enhancer specificity, and the role of affinity-optimizing SNVs in GOF gene expression, transcend these categories. More broadly, the conservation of regulatory principles across diverse enhancers provides a framework for using violations of such rules to predict causal variants that underlie enhanceropathies.

2.5 Methods

2.5.1 Mice

All animal procedures and studies were approved by the University of California San Diego Institutional Animal Care and Use Committee according to the Association for Assessment and Accreditation of Laboratory Animal Care guidelines. Mice were maintained on a 12:12 light–dark cycle with ad libitum standard chow diet and water. Transgenic mouse assays were performed using *Mus musculus* C57BL/6NHsd strain (Envigo). Animals of both sexes were included in this study.

Generation of transgenic mice using CRISPR–Cas9. Cas9 protein (IDT, 1081058), trans-activating CRISPR RNA (tracrRNA) (IDT, 1072532), CRISPR RNA (crRNA) and single-stranded DNA (ssDNA) homology-directed repair template oligos were co-injected into one-cell embryos at the Moores UCSD Cancer Center Transgenic Mouse Shared Resource. Custom ssDNA repair oligonucleotides and crRNAs were synthesized by IDT. We designed and selected crRNA if the guide sequence is predicted to have high specificity on CRISPOR (<http://crispor.tefor.net/crispor.py>) and if the mutation introduced by homology-directed repair will ablate the protospacer adjacent motif (PAM) site. Because a PAM site is not available in the genomic locus for the human and synthetic mutations, we first generated a mouse line that contained a de novo PAM site within the ETS-A site. The French 2, Indian 2, Syn 0.25 and Syn 0.52 mouse lines were generated with CRISPR–Cas9, using one-cell embryos with the

new PAM background (Supplementary Table 4). The LOF mutation mouse line was generated using embryos with the WT background. All mouse lines were generated by homology-directed repair using ssDNA as a repair template. Genome-edited founders were identified by genotyping as described below. Wherever possible, multiple founders bearing the same desired allele were used to establish each line. Founders were crossed to WT C57BL/6N mice to generate the F1 generation for each mouse line.

Genotyping. Genomic tail DNA was obtained and used to genotype ETS-A transgenic mice with the following primers:

forward 5'-GGACAAGAGATTAGCGTGGCTGGTGATTTCCTTTCACCCAGC-3'

reverse 5'-GACACCAGACCAACTGGTAATGCATAATGACAGCAACATCC-3'.

The underlined sequences anneal to the ZRS, and the remaining sequences are overhangs used to clone ZRS PCR products into a vector containing an ampicillin resistance cassette by Gibson assembly. For all mice, including founders, PCR products were analysed by Sanger sequencing (sequencing primer: 5'-CATCCTAGAGTGTCCAGAACC-3') to identify ZRS genotypes. For all founder mice, PCR products were cloned, and individual clones were sequenced to confirm the initial genotyping results with single-allele resolution.

Phenotyping. Each mouse born into our colony has all four limbs inspected by an investigator blind to genotype at postnatal day (P)10–18 during routine ear clipping (for identification) and tail biopsy collection (for genotyping). Limb and/or digit phenotypes, or the absence thereof, are readily detectable in postnatal mice and recorded in detail. Each limb is inspected for the number of digits and the presence of other features, including triphalangeal first digit(s) and/or shortened limbs. After genotyping, phenotypic data for each genotype in each ETS-A transgenic line are collated to calculate penetrance (on the basis of the presence or absence of phenotype).

Statistical tests for mouse phenotypes. Fisher's exact test was used to measure any statistical difference in the penetrance and laterality of polydactyly across any pair of the approximately 0.25-affinity mouse lines (French 2, Indian 2 or Syn 0.25 mice). For penetrance, there are two factors: have phenotype or no phenotype. For laterality, there are three factors: bilateral, unilateral or no phenotype. To determine whether the occurrence of a unilateral

phenotype has a bias on the left or right, P values from chi square goodness of fit test were calculated. Fisher's exact test was used to measure any statistical difference in digit phenotypes across any pair of approximately 0.25-affinity mouse lines (French 2, Indian 2 or Syn 0.25 mice). P values measuring any difference in digit phenotypes across males and females in each mouse line are also calculated. These tests have nine factors: five digits no TT, five digits one TT, ... , seven digits three TT.

Timed matings for embryo collection. Within each ETS-A transgenic mouse line, timed matings were set up and monitored each morning for vaginal plug formation. The date that plugs were observed was noted as E0.5. Females were removed from males on the plug date and embryos were staged at dissection. Embryos labelled as E11.75 have around 48 somites, and E12.0 embryos have around 53 somites. Pregnant females were humanely euthanized by isoflurane overdose. Embryos were dissected in ice-cold phosphate-buffered saline (PBS) pH 7.4 and then fixed in 4% paraformaldehyde in PBS pH 7.4 overnight with gentle rotation at 4°C. Embryos were then dehydrated through a graded methanol series at 4°C (25%, 50%, 75% methanol in PBS pH 7.4 plus 0.1% Tween-20, 100% methanol) and stored in 100% methanol at -20°C for up to six months until use. The yolk sac of each embryo was collected and used for genotyping as described above. The sex of embryos is unknown.

Probe cloning and synthesis for in situ hybridization. Shh and Ptch1 templates were amplified from mouse E11.5 cDNA using previously described primers¹⁶¹, and were ligated into a pCR BluntII TOPO vector, transformed into TOP10 competent cells and plated for selection on kanamycin plates. Colonies were selected for sequence verification and then plasmid prepped. Plasmid DNA was linearized with SpeI or NotI restriction enzyme and then used as a template for in vitro transcription using a digoxigenin labelling kit (Roche, 11175025910) with T7 (antisense) or Sp6 (sense) polymerase. After DNase treatment to digest template DNA, RNA probes were recovered using a RNeasy mini kit, and RNA concentration and purity were measured to confirm probe synthesis.

Whole-mount in situ hybridization. Embryos were treated with 6% H2O2 in methanol for one hour, and then rehydrated through a methanol series to PBS-T (1% Tween-20 in PBS pH 7.4). Embryos were washed 5×5min in PBS-T, and then treated with proteinase K (10ug/mlu1)

(Invitrogen, 1000005393) for 20min. After permeabilization, embryos were washed in PBS-T containing 2mg/ml glycine, then PBS-T, then post-fixed for 20min in 4% paraformaldehyde (PFA)/0.2% glutaraldehyde in PBS-T. Embryos were then washed 2×5min in PBS-T, followed by 10min in a 1:1 mixture of PBS-T and hybridization solution (50% formamide, 5× SSC pH 4.5, 1% SDS, 50ug/ml yeast tRNA, 50ug/ml heparin). Embryos were then allowed to sink (no rocking) in hybridization solution for 10min. They were then changed to new hybridization solution and incubated for at least one hour at 65°C. Hybridization solution was replaced with fresh hybridization solution containing 1ug/ml of antisense (all ETS-A embryos and WT control) or sense (WT control only) probe followed by overnight incubation at 65°C. Embryos were washed 3×30min in solution I (50% formamide, 5× SSC pH 4.5, 1% SDS) at 65°C followed by 3×30-min washes in solution III (50% formamide, 2× SSC pH 4.5) at 65°C. Embryos were then washed 3×5min in TBS-T (1% Tween-20 in Tris-buffered saline) and blocked for 1h in block solution (10% heat-inactivated sheep serum and 0.1% Roche blocking reagent in TBS-T). Roche blocking reagent (Roche, 11096176001) was dissolved in maleic acid buffer according to the manufacturer's recommendations. Embryos were then incubated in block solution containing 1:2,500 anti-digoxigenin-AP antibody (Roche, 11093274910) overnight at 4°C. Embryos were washed 3×5min in TBS-T and then 5×1h in TBS-T, followed by overnight incubation in TBS-T at 4°C. Embryos were then washed 3×10min in NTMT (100mM NaCl, 100mM Tris pH 9.5, 50mM MgCl₂, 1% Tween-20) before coloration in AP reaction mix (125ug/ml BCIP (Roche, 11383221001) and 250ug/ml NBT (Roche, 11383213001) in NTMT). Coloration was carried to completion in the dark. Embryos were washed 10min in NTMT followed by 3×10min in TBS-T and then overnight in TBS-T at 4°C. Embryos were imaged using the Leica M165 FC microscope with the Lumenera Infinity3 camera, then post-fixed in 4% PFA for 30min and stored in 1% PFA in 4°C. All steps were performed with gentle rocking and at room temperature unless otherwise specified.

Skeletal preparations. Young postnatal mice at age P10-12 were humanely euthanized by CO₂ inhalation before skeletal preparations. Dissected limbs and/or whole cadavers of representative homozygotes for each line were skinned and eviscerated, then fixed in 95% ethanol overnight. Samples were then stained over two nights in cartilage staining solution (75% ethanol,

20% acetic acid and 0.05% Alcian blue 8GX (Sigma-Aldrich, A3157)), rinsed overnight in 95% ethanol, cleared overnight in 0.8% KOH and stained overnight in bone staining solution (0.005% Alizarin red S (Sigma-Aldrich, A5533) in 1% KOH). After staining, samples were further cleared in 20% glycerol in 1% KOH until digits were free of soft tissue and long-bone morphology was visible. Samples were further processed through a graded series of 50% and 80% glycerol in 1% KOH and then into 100% glycerol for imaging and storage. All steps of the skeletal staining procedure were performed with gentle rocking at room temperature.

EMSA. EMSAs were performed using the LightShift Chemiluminescent EMSA Kit (Thermo Fisher Scientific) with biotinylated and non-biotinylated double-stranded oligonucleotides. Oligonucleotides were annealed according to an advanced protocol (<https://tools.thermofisher.com/content/sfs/brochures/TR0045-Anneal-oligos.pdf>).

DNA-binding domain (DBD) proteins were synthesized using the TNT Quick Coupled Transcription/Translation System (Promega) from the pTNT plasmid for each respective protein. ETS-1 DBD (residues 336–441, which is conserved in sequence from human and mouse) was amplified from the pET28b-Ets1-ETS vector (Addgene, 85735). DBDs for human HOXA13 (residues 322–389) and HOXD13 (residues 276–335) were amplified from human genomic DNA, and sequences were confirmed by Sanger sequencing. The coding sequences for these DBDs were amplified with flanking XhoI and NotI sites and cloned into the pTNT-B18R vector (Addgene, 58978). The binding reaction was performed in a 20- μ l volume containing 2 μ l of 10 $\times$ binding buffer (100mM Tris, 500mM KCl and 10mM DTT; pH 7.5), 50ng Poly(dI:dC), 20 femtomol biotin-labelled probe and protein extract. For competition experiments, a 200-fold molar excess of unlabelled probe was added. Binding reactions were pre-incubated for 10min before adding the biotin-labelled probe. Binding reactions were then incubated at room temperature for 20min and loaded onto a DNA retardation gel (6%). Electrophoresis with 0.5 $\times$ TBE on ice and transfer to a 0.45- μ m Biodyne B Nylon membrane (Thermo Fisher Scientific) was done in the cold room. DNA was cross-linked for 15min using 312-nm light, and the membrane was put between blank sheets of paper overnight. The next day, the biotinylated probes were detected using the Chemiluminescent Nucleic Acid Detection Module (Thermo Fisher Scientific). Images of the resulting membrane were acquired using a Chemidoc MP imaging system (Bio-Rad). For

quantification of ETS-1 binding to ETS-A variants, band quantifications were performed by taking the ratio of the volume (intensity) for the shifted band in the reaction with the ETS transcription factor to the volume (intensity) for the unshifted band in the reaction without the ETS transcription factor, using the Analysis Table in Image Lab 6.0.

2.5.2 Calculation of binding affinity

Relative binding affinity is calculated using high-throughput binding data from the UniProbe database (<http://thebrain.bwh.harvard.edu/uniprobe/index.php>)^{123;162}. Median intensity signals of 8-mers PBM data were measured as a percentage of their optimal 8-mer binding motif.

PBM-ChIP correlation analysis. BigWig files for ChIP-seq data were downloaded from Gene Expression Omnibus accessions for previously published data^{163–165}. BigWig files were chosen because they contain the most quantitative metric at base-pair resolution for the ChIP-seq signal. For each dataset, we predicted ETS sites using NNGAWNN across the reference genome used to create the bigWig. We then extracted the average bigWig ChIP-seq signal over all predicted ETS TFBS 8-mers. We associated each ETS TFBS with its predicted ETS affinity using PBM⁹³. We placed the ETS TFBS into bins of PBM affinity 0–0.1, 0.1–0.2, 0.2–0.3, ... 0.9–1.0. Within each bin, we took the average bigWig ChIP-seq signal across each chromosome and plotted the chromosomal averages. The Spearman correlation uses all points from all bins and all chromosomes.

2.5.3 Analysis of MPRA Data

Attaining the datasets. MPRA data were downloaded from previously published papers^{10;13;60;148}. Only single-base substitutions were considered across all datasets. We classified variants as significantly altering expression using P values that were provided in the published supplementary tables. If the study did not report adjusted P values, we adjusted all raw P values using Benjamini–Hochberg adjustment. We defined variants as having a significant change in expression if their adjusted P value was smaller than 0.05. We plot $\log_{10}(\text{Padj})$ with direction of change in gene expression, in which positive values depict a significant variant that leads to increased gene expression. We used the one-tailed Mann–Whitney U test to test for enrichments

in GOF enhancer activity in different groupings (that is, ‘All other SNVs’, ‘SNVs that do not change affinity’ and ‘SNVs that optimize affinity’). If there were more than 1,000 points in the dataset, we plotted a random 1,000 as dots over the box plots. MPRA data were analysed using standard Python libraries (pandas, numpy, scipy, seaborn, matplotlib). Processing, visualization and statistics were done using custom Python code.

ETS binding site definition. ETS-binding sites were defined as NNGGAWNN¹⁶⁶. We defined an ETS-optimizing variant as one that caused at least a 1.59-fold change (alt/ref) in all analyses with the exception of Extended Data Fig. 7. Analysis of the 2% sequence mutagenesis of the ZRS enhancer in Extended Data Fig. 7 defines ETS-optimizing variants as a fold change greater than 1.0, because only three variants within this dataset have a fold change of at least 1.59.

AP-1 binding site definition. The AP-1-binding site was defined as NTKANNMA. IRF binding affinity was defined as NWNNGANA. Motifs used for ETS, AP-1 and IRF were determined on the basis of crystal structure and mutational analysis data^{166–169}. We defined AP-1 and IRF-optimizing variants as one that caused at least a 1.5-fold change in binding affinity. We defined SNVs not changing the TFBS affinity as SNVs with a 0.8–1.25-fold change in affinity. For analyses on the IFNb enhanceosome, we excluded nucleotides that contributed to two overlapping binding sites, and only analysed the effects of affinity-optimization SNVs that affect a single binding site.

2.5.4 Analysis of eQTL data

We analysed eQTL data downloaded from the EBI eQTL catalogue GitHub page (<https://github.com/eQTL-Catalogue/eQTL-Catalogue-resources/blob/master/tabix/>) for a lymphoblastoid cell line generated by the Geuvadis consortium¹⁵⁵. These eQTL data were used to design the MPRA library used previously^{10;155}. For the eQTL data, we adjusted the raw P values using the Bonferroni procedure, in which the total number of tests is the total number of genotype–gene-expression associations tested. For the seven ETS affinity-optimizing SNVs that cause significant differential expression in the MPRA experiments¹⁰ and have eQTL values, we compare the MPRA expression and eQTL using adjusted log P values plotted with the

direction of differential expression for reporter assay and target gene expression (b value). For AP-1 affinity-optimizing SNVs that gave significant differential expression in the MPRA, only three overlapped with significant eQTL variants and so we did not study these. For larger-scale analysis of the relationship between significant variants and target gene expression, we analysed all eQTL variants in the MPRA library (n=2,663) in a previous report¹⁰, and their effects on target gene expression (b values). Finally, for genome-wide analysis of all eQTL variants in lymphoblastoid cell lines, we analysed all eQTLs from the Geuvadis consortium¹⁵⁵ with $P_{adj} < 0.01$. We categorized the eQTL variants into three categories: affinity-optimizing SNVs, SNVs that do not alter affinity and all other SNVs. For eQTL variants that have multiple genes associated, we plotted only the most significant association. A one-tailed Mann–Whitney U test was used to determine any significant enrichment in eQTL variants that do not change affinity or increase affinity with GOF target gene expression (b value).

2.5.5 Statistics

To assess any statistical differences in the penetrance and laterality percentages between French 2, Indian 2 and Syn 0.25 mice, we performed Fisher’s exact test using the `fisher.test` function in R (Supplementary Table 3). Statistical differences in digit phenotypes were measured using Fisher’s exact test using a 2×9 table (Supplementary Tables 3 and 4). To determine whether unilateral polydactyly deviates from the assumption that phenotype would occur at a 50%–50% rate on the right and left hindlimbs, a chi square goodness of fit test was used (Supplementary Tables 3 and 4). To measure any statistical difference between the band intensities in EMSAs across the French 2, Indian 2 and Syn 0.25 sequences, we performed a one-way ANOVA test and found no significant difference with $P=0.18$.

2.5.6 Data Availability

All data supporting the findings of this study are available within the paper and its Supplementary Information. ChIP–seq data analysed in this paper were downloaded using accession codes GSM2218592, GSM3520734 and GSM4110116. eQTL data analysed in this paper were downloaded from the EBI eQTL catalogue GitHub page (<https://github.com/eQTL-Catalogue/eQTL-Catalogue-resources/blob/master/tabix/>) for the lymphoblastoid cell line gen-

erated by the Geuvadis consortium. The UniProbe database (http://thebrain.bwh.harvard.edu/pbms/UniPROBE_staging/browse.php) was used to access PBM data. Source data are provided with this paper.

2.5.7 Code Availability

All custom code used in the analyses has been deposited at GitHub (<https://github.com/jsolvason/nature2023-limb>) and Zenodo (<https://doi.org/10.5281/zenodo.10368918>).

2.5.8 Acknowledgements

As this work was derived from the same manuscript, see *Lim**, *Solvason**, *Ryan* et al. Nature 2024* for corresponding details.

Chapter 2, in full, is a reprint of the material as it appears in Nature, 2024. Fabian Lim, Joe J. Solvason, Genevieve E. Ryan, Sophia H. Le, Granton A. Jindal, Paige Steffen, Simran K. Jandu & Emma K. Farley. Thank you to all the co-authors for their contribution to the work, especially Fabian Limb and Dr. Genevieve Ryan.

2.5.9 Author Contributions

As this work was derived from the same manuscript, see *Lim**, *Solvason**, *Ryan* et al. Nature 2024* for corresponding details.

2.5.10 Competing Interests

As this work was derived from the same manuscript, see *Lim**, *Solvason**, *Ryan* et al. Nature 2024* for corresponding details.

Chapter 3

Dependency grammar of developmental enhancers.

3.1 Abstract

Enhancers direct precise tissue-specific expression and as such provide the instructions for successful development and homeostasis of an organism. A major question in the field is how DNA sequence encodes this tissue-specific expression. To investigate this we made 460,800 versions of a neural enhancer and tested all of these for activity in developing embryos. Each enhancer has exactly the same number and affinity binding sites, and the same sequence content, yet only a fraction of these elements are able to direct transcription illustrating enhancer activity requires more than a combination of binding sites. We find that some organization or syntaxes, robust syntaxes, are functional no matter what affinity sites are located there, while other syntaxes are more nuanced, only working when certain affinity sites are located near each other. Small changes to enhancers can convert them from inactive to active, with 36% of non-functional elements becoming active enhancers by flipping the orientation of a single site. These flips also move enhancers into more functional organizational families. We identified 910 syntax motifs enriched within the active enhancers. Synthetic enhancers containing these motifs are sufficient to drive tissue-specific expression in whole developing embryos. We developed a syntax score and applied to every cluster of ETS and GATA in the Ciona genome. We find that active enhancers have higher syntax scores, and that high syntax score can compensate for low affinity sites illustrating a dependency between syntax and affinity to encode tissue specific enhancers. We created synthetic enhancers with high syntax scores and moderate affinity and find that these

can be used to design tissue specific enhancers. Collectively this study provides a comprehensive illustration of the importance of dependency grammar to encode tissue specific expression, and a framework for decoding the regulatory genome and building synthetic enhancers grammatical rules.

3.2 Introduction

Enhancers are genomic elements that regulate where and when genes are expressed, they provide the instructions for successful development and homeostasis of an organism. Changes in enhancer sequence and function can lead to dramatic changes in morphology^{7;8;13;71}. While we can find regions of the genome likely to be enhancers using accessibility or epigenetic marks^{14;170}, we do not know where these open regions of the genome are indeed active enhancers, nor do we understand which types of changes within an enhancer can alter gene expression. Identifying the functional features within enhancers and understanding the mechanisms behind how these functional features interact with each other to drive gene expression patterns is fundamental in decoding the instructions for development and homeostasis within our genomes⁹. Advances in this field can improve our ability to pinpoint disease-associated variants that alter enhancer activity and potentially phenotypes, and to design tissue specific enhancers.

While we know that transcription factors bind to motifs within enhancers, finding these features within the genome is challenging as functional features within endogenous enhancers are typically low affinity sites⁷. Furthermore, we do not understand what combination of sites encode active enhancers. The higher order rules governing the combinatorial code of enhancers, and specifically the interplay of the affinity and syntax of sites (order, orientation and affinity) is referred to as enhancer grammar⁹. While there are examples of enhancers in which grammar is important for encoding enhancer activity^{9;39;46;52} there are also plenty of studies that cannot find grammatical constraints on enhancers⁴⁷.

Here we set out to determine if grammar plays a role in tissue specific expression, and if grammatical constraints can be used to predict tissue specific enhancers within the genome and to design tissue specific enhancers. To achieve this we focus the neural Otx-a enhancer, a well characterized enhancer that has a pervasive logic, integrating signals from the FGF signaling

pathway (ETS) and a tissue specific determinant (GATA) to encode gene expression^{91;171;172}. We create every possible organization of this enhancer and test these for activity in whole developing *Ciona robusta* embryos. The marine chordate, *Ciona robusta*, is ideal to study the relationship between enhancer sequence and function as we can electroporate libraries of a million elements into whole developing embryos to study the relationship between enhancer activity and tissue specificity.

3.3 Results

3.3.1 ETS and GATA are necessary and sufficient for neural enhancer activity

The Otx-a enhancer drives expression in the neural a6.5 lineage which gives rise to the anterior sensory vesicle and palps, and the b6.5 lineage which gives rise to the dorsal epidermis and dorsal nerve cord^{5;91;171;173} (Fig. 1A,B,C,D,E). The Otx-a enhancer contains 2 ETS and 3 GATA sites¹⁷¹. We have previously shown that the suboptimal affinity of these sites is critical for tissue-specific expression⁵. We measure affinity of these sites using PBM data^{94;123;174} and find that the ETS sites have affinities of 0.58 and 0.39 relative affinities, while the GATA sites have relative affinities of 0.85, 0.32 and 0.45⁵. Ablation of these 5 sites (Otx-a ABL) using GGAW to GCAW for ETS¹⁰⁴ and GATA to AATA^{92;171} results in complete loss of expression (Fig. 1D,G), demonstrating the necessity of these GATA and ETS sites. To test if these 5 sites are sufficient for expression, we randomize the Otx-a linker sequences to create a library 5 million unique sequences (Otx-a R). Across these 5 million sequences, the only constants are the five 8mer binding sites (Fig. 1D). The Otx-a R enhancer activates neural expression that is indistinguishable from the WT Otx-a enhancer (Fig. 1F, S1, S2), demonstrating the sufficiency of these 5 sites to drive expression. As the 3 GATA and 2 ETS sites are sufficient for neural expression, the Otx-a enhancer is an ideal system in which to understand how the syntax (order, orientation and spacing) of these binding sites can contribute to tissue-specific expression.

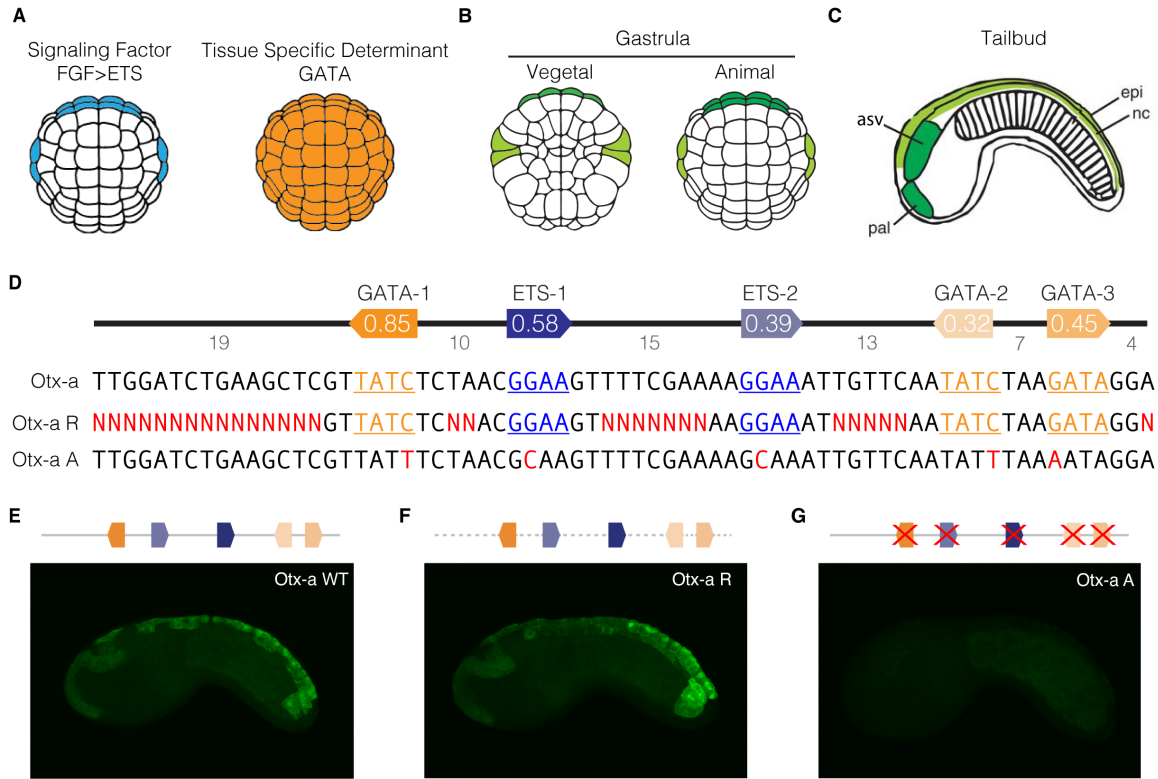


Figure 3.1. Otx-a neural enhancer is regulated by a common logic: GATA and ETS. (A) Ciona gastrula embryo showing the animal side which contains ectodermal cells. FGF signaling activates ETS in cells shown in blue. GATA is expressed in all cells on the ectodermal side. (B) Otx-a expression shown in green occurs in the a6.5 and b6.5 lineages. (C) Tailbud image shows the expression at a later time point with the dark green cells from the a6.5 lineage giving rise to the anterior sensory vesicle (asv) and palps (pal). The light green cells from the b6.5 lineage give rise to the dorsal epidermis (epi), nerve cord (nc), and two tail muscle cells (not shown). (D) Otx-a contains 3 GATA and 2 ETS sites. Relative affinity of each site is labeled within the site. Spacing between sites is labeled in grey. The GATA and ETS core sequences required for binding labeled in orange and blue font, respectively. Otx-a Randomized (Otx-a R) library contains fixed 8mers to maintain the affinity of the GATA and ETS binding sites and Ns at linkers representing an equal chance of A,T,G or C. Otx-a R library contains 5 million unique sequences (see detailed explanation on randomized library in Fig. S1). Otx-a Ablated (Otx-a A) contains one point mutation per GATA or ETS core known to ablate binding of the TF (GATA to AATA, GGAW to GCAW). (E) Otx-a activates expression in the anterior sensory vesicle (asv) and palps (pal), nerve cord (nc), dorsal epidermis (epi), and two tail muscle cells. (F) Otx-a R activates the same neural expression as Otx-a WT. (G) Otx-a A activates no expression. See details on expression data for Otx-a WT, Otx-a R and Otx-a A in Fig. S2.

3.3.2 Testing every possible organization of Otx-a enhancer in developing embryos

To determine if there are constraints on the organization of the five sites within the Otx-a enhancer, or if the presence of these sites regardless of organization is sufficient for neural expression, we made a library of elements in which the order, orientation and spacing between the ETS and GATA sites were altered. We generated every possible organization of GATA and ETS sites using the five binding sites and five inert linkers within the Otx-a enhancer. The scrambled elements are constructed by creating elements with any of the alternating linkers and followed by any binding sites (Fig. 2A). Once a linker or binding site is added to an element it cannot be used again in that element. Linkers are always inserted in the forward direction; binding sites can be orientated in either direction. In total there are 460,800 possible combinations of these features (Fig. 2B). Importantly, every element has the same number and affinity of sites and the same sequence content. The differences between elements are the binding site organization (the order, orientation and spacing between the binding sites) and the placement of the high- and low-affinity sites and thus the interplay of syntax and affinity in the elements. Therefore, differences in expression patterns can only be explained by changes in syntax and placement of different affinity sites.

To ensure that any active elements are dependent on GATA and ETS, each element has a paired ablated element in which all 5 sites are ablated with single base-pair changes (Fig. 2B). In total, with the scrambled elements and ablated controls, the library contains 921,600 elements. All elements are cloned upstream of a minimal promoter, GFP and a barcode (Fig. 2B). Each element has on average 54 unique barcodes, and the total library contains 49 million barcodes. The plasmid library is electroporated into fertilized *Ciona* eggs and mRNA and plasmid DNA is collected from embryos 5.5hpf when the endogenous Otx-a enhancer is active. We conducted three biological replicates. If the element is successfully electroporated into embryos then we will recover the plasmid DNA. For elements that are active enhancers, the enhancer will transcribe its barcode and we will be able to isolate mRNA for this barcode. 888,800 (96%) of the elements were successfully electroporated into the embryos in triplicate.

3.3.3 50,000 active elements regulated by GATA and ETS

To determine how many of these elements can activate transcription and are thus active enhancers, we normalize the mRNA barcode signal to the number of copies of the plasmid DNA for a particular barcode. We then average the signal for each barcode associated with the same element to get an enhancer activity score. Of the 460,800 scrambled elements with functional ETS and GATA sites, 17.5% (80,965) of the elements have an activity score of above one, these are defined as active enhancers (Fig. 2C). Otx-a is within this list of active enhancers with an activity score of 1.2 (Figure 2). The Otx-a ablated sequence is not active. 31% (143,799) of the elements have an activity of 0 or below and thus are classified as inert elements. Elements with activity between 0-1 are likely very weak enhancers and were not considered in our further analysis. To validate the MPRA activity score and our thresholds for active enhancers and inert elements we selected 22 elements that spanned a broad range of enhancer activity from the library to test orthogonally in a reporter assay. We electroporated each of these individually into embryos and quantified the GFP expression within at least 50 embryos per replicate. All 9 elements with MPRA score ≥ 1 show expression in the GFP reporter assay. Of these, 8 of 9 show only neural activity while 1 has expression in neural and weak expression in other cell types responding to FGF signaling. All 13 sequences with MPRA values below 0 show no GFP expression in reporter assays thus validating the inert classification. The abundance of both neural enhancers and inert sequences produced from the same number and affinity of binding sites suggests the way in which these sites are organized is important for mediating enhancer activity.

To identify enhancers dependent on ETS and GATA, we compare the expression of the 80,965 active enhancers with their ablated control. 52,139 active enhancers are entirely dependent on ETS and GATA as their ablated counterparts have activity values less than 0 (Fig. 2E). These 52,139 enhancers represent a set of active enhancers regulated by GATA and ETS that can be mined for functional grammatical signatures that contribute to their activity. We selected two enhancers, OLS 19 and 87, and confirmed that upon ablation of the five sites that there is complete loss of GFP expression (Fig. S7). While our enhancers all have the same sequence content, binding site count and affinity, only a small proportion are active enhancers,

demonstrating that organization of sites matters. It is possible that the inactive elements contain de novo repressors at the junctions of the linkers and binding sites and that render the elements inactive. To test this hypothesis, we selected two inert elements and randomized the linker sequence (Fig. S6). Randomizing the linkers between binding sites is not sufficient to convert these inert elements into active enhancers, demonstrating that the non-functional sequence is inert due to the syntax and/or placement of binding sites rather than any de novo repressors.

3.3.4 Not all clusters of sites are equal: Despite having the same sequence content, binding sites and affinities, only a fraction of elements are functional.

Despite having the same binding sites and linker sequences, only 52,136 elements are active enhancers and dependent on ETS and GATA. These results suggest that the organization of sites within the elements is critical for determining if an element is an active enhancer or inert sequence. Visual inspection of organizations that encode active enhancers elucidates no clear patterns shared among the neural enhancers, highlighting the degenerate nature of encoding enhancer activity and tissue specificity. The lack of congruency between the related active enhancers has previously led people to declare the absence of grammar within enhancers, however it is likely that the degenerate nature of binding sites coupled with a higher order grammar that is not linear has obscured our ability to see these grammatical constraints. Furthermore, the interplay of affinity and grammar makes the grammatical encoding of enhancers more flexible, this has been described as dependency grammar and soft syntax^{9;48}. Revealing grammatical rules may require mining of thousands of active enhancers that have the same regulatory inputs to find syntax and grammatical rules associated with active enhancers. Our dataset is ideal for finding such constraints. We therefore wondered if there were grammatical rules governing the active enhancers within our enhancer screen.

Figure 3.2 Characterizing the activity of 460,800 versions of Otx-a enhancer. (A) Generation of the Otx-a scrambled library. Binding sites and inert linker sequences are parsed. Every possible sequence made by placing the linkers and binding sites into scaffold that requires alternation of linker and binding sites. Linkers are always in the forward direction while binding sites can be in the forward or reverse orientations. We exhaustively generate all 460,800 possible elements using this scaffold rule. (B) To test if GATA and ETS are regulatory inputs of active enhancers in the library, a paired ablated version of each variant (ABL) is included which has a point mutation at each GATA and ETS core known to ablate binding (GATA $\geq$ AATA, GGAW $\geq$ GCAW). Enhancers are cloned upstream of a promoter, GFP and a transcribable barcode. All 921,600 sequences are electroporated into fertilized eggs. RNA and DNA are extracted from the mid gastrula-stage embryos (5.5hpf). The ratio of barcode mRNA to barcode DNA gives the enhancer activity score. (C) The enhancer activity distribution of all detected 393,554 (85%) WT elements with Otx-a plotted as a single green dot. Activity scores ≥ 1 are active enhancers and activity scores ≤ 0 are inert elements. Microscope validation of 28 inert sequences (red), neural enhancers (green), and sequences activating neural and ectopic expression (purple). See details on expression data for microscope validated elements in Fig. S3-5). (D) Neural enhancers activate expression in the same neural tissues as Otx-a at the tailbud stage (anterior sensory vesicle, palps, dorsal epidermis, nerve cord, two tail muscle cells. (G) 52,139 active enhancers detected with paired ablated sequence activity ≤ 0 demonstrating the active sequences are dependent on the GATA and ETS sites.

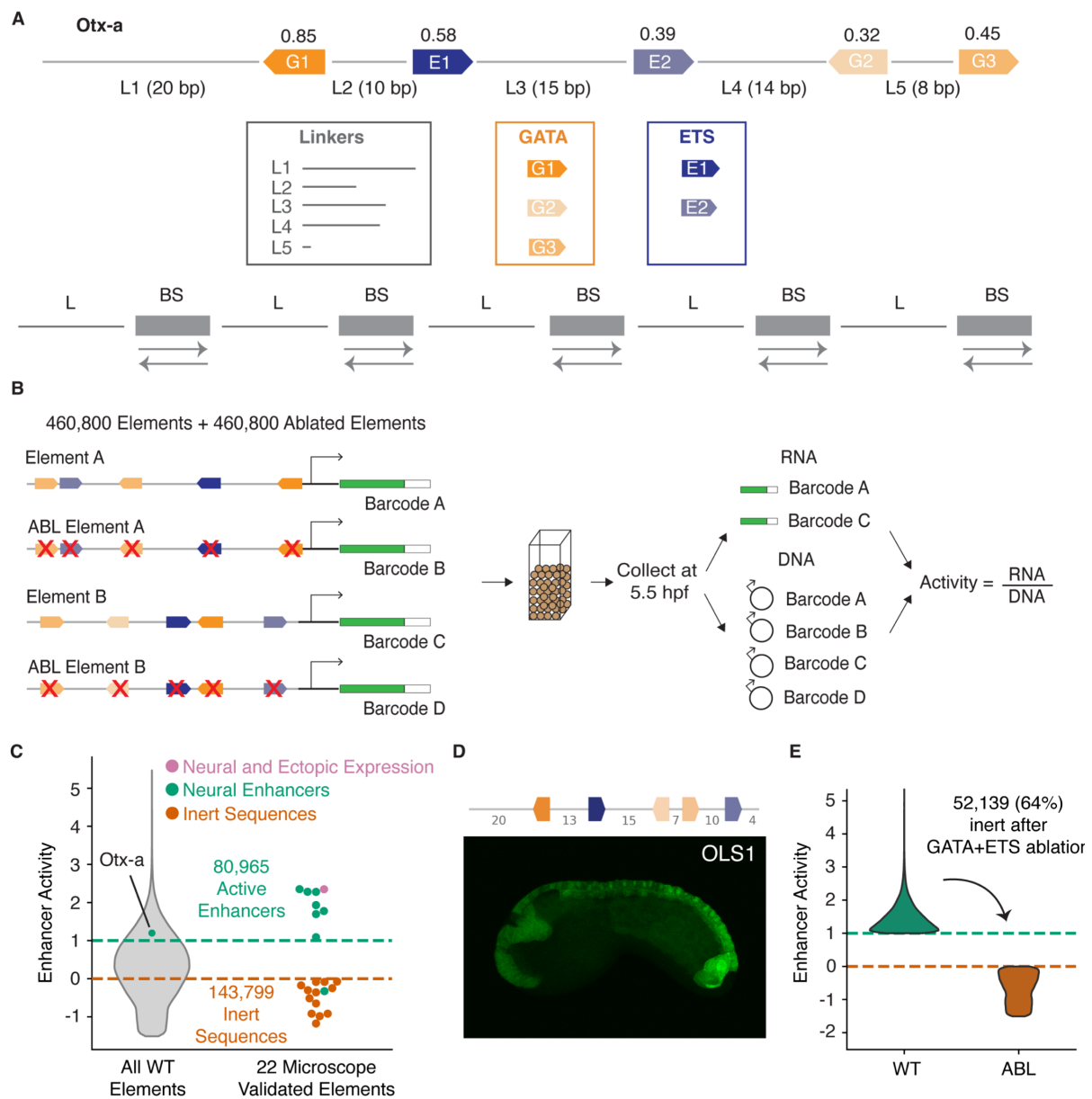


Figure 3.2. Characterizing the activity of 460,800 versions of Otx-a enhancer.

3.3.5 Syntax Contributes to Enhancer Activity

An important aspect of grammar is syntax, which is the order, orientation and spacing of binding sites. To see if syntax is important for encoding enhancer activity, we group elements with shared organization of GATA and ETS. Across the 460,800 elements there are 38,400 “organizational families”. Within each family are 12 elements which have the same order, orientation and spacing of their GATA and ETS sites; the only difference between syntax members is the placement of the different affinity binding sites (Fig. 3A). Many organizational families are enriched for active enhancers while other organizational families contain few or no active enhancers. (Fig. 3B,C,D).

We defined organizational families in which 7 or more active siblings with no inactive siblings as robust syntaxes (Fig. 3C, dark green). We define organizational families in which 7 or more members are inactive and none are active as inert organizations (Fig. 3C, dark red). Organizational families where there are a mix of active and inert enhancers are defined as moderate syntaxes. We require moderate syntaxes to contain at least 4 elements classified as either active or inactive. The *Otx-a* organizational family is an example of a moderate syntax as 5 of the 12 members are neural enhancers while the remaining family members are not active (Fig. 3B,C). In the most extreme examples of robust syntax, all 12 siblings are active enhancers (Fig. 3B,C). There are 9 of these extremely robust syntaxes. The most extreme inert organization contain no active enhancers (Fig. 3B,C). There are 1,349 examples of these extremely inert organizations.

To validate the MPRA results, we randomly selected 4 members from a robust syntax and 3 members from an inert organization which we orthogonally tested by reporter assay. We electroporated the siblings of these organizational families and quantified their GFP expression. All four members of the robust organizational family are active enhancers that drive strong expression in the neural tissue (Fig. 3E, S9). Interestingly all these enhancers show very low levels of ectopic expression in other cell types indicating that the robust syntax also causes ectopic enhancer activity. All 3 inert syntax siblings show no expression in reporter assays (Fig. 3F, S9). Hundreds of robust syntaxes and inert sequences exist, and they are enriched over what would be expected by chance alone. We can generate a null distribution for the syntax analysis

by shuffling the element labels and enhancer activity values. The shuffling and organizational family analysis is repeated 100 times to generate a null distribution from which Z scores can be calculated for observed values (See Methods). Strikingly, in none of our permutations did we find any organizational families with all 12 neural enhancer siblings were detected, while 9 extremely robust organizational families contained 12 active enhancers in the real dataset (Fig. 3C). Moreover, we observe a statistically significant enrichment of robust syntaxes (top left of heat map) and inert organizations (bottom right of heat map) with extreme Z-scores greater than 150 representing Bonferroni-adjusted p-values of $<1e-5$. We find active enhancers from robust syntaxes drive higher levels of expression than active enhancers in moderate syntaxes (Fig. 3G), which suggests that these robust syntaxes are more optimally organized for transcriptional output.

Together, these results provide the most compelling evidence to date that binding site syntax contributes to enhancer activity. Robust syntaxes are able to drive higher levels of enhancer activity regardless of the position of the different affinity binding site. However, moderate syntaxes have varied outcomes in gene expression suggesting that there is less flexibility in the requirement of certain affinity sites in particular positions. It is likely that the more robust syntaxes that function regardless of affinity represent more optimal organization of sites that enable binding of the TFs more easily and thus changes in affinity are less important. While for moderate syntaxes the organization of sites is less optimal and thus higher energy interactions are required to offset the less optimal organization of sites, and thus binding affinity preferences occur. In line with this, the robust syntaxes have higher levels of enhancer activity within the MRPA than the moderate syntaxes and show high neural expression in GFP reporter assays, with some weak ectopic expression (Fig. 3E) indicating that the robust syntaxes are stronger enhancers despite having the same affinity binding sites.

Figure 3.3 Robust and moderate syntaxes drive enhancer activity. (A) Example of the organizational family for Otx-a. Each organizational family contains 12 sequences which we call syntax siblings. All siblings within a organizational family have the same order, orientation and spacing between sites. The only difference is the placement of affinity of each site. (B) Enhancer activities of the siblings of 3 organizational families: Otx-a organizational family (an example of a moderate organizational family), a robust organizational family, and an inert organizational family. Vertical dashed lines at 1 and 0 to indicate cutoffs for active enhancers and inert sequences, respectively. The Otx-a organizational family is of moderate strength with 4 active enhancers and 2 inert sequences. All siblings of the robust organizational family are active enhancers. All siblings of the inert organizational family are inert sequences. (C) All organizational families can be sorted into a heat map based on the number of siblings which are active enhancers (Y-axis) versus inert sequences (x-axis). Boxes shown in purple exhibit a significant enrichment of organizational families over the null distribution. The null distribution is generated by 100 random permutation of the number of organizational families within each bin (see Fig. S8). We define families containing ≤ 7 active siblings to have a robust syntax, and families with ≥ 7 inert siblings to have an inert syntax. Moderate syntaxes are organizational families that are not inert nor active, and with at least 4 elements classified as either active or inert. (D) Distribution of enhancer activities within inert, moderate, and robust organizational families. Robust organizational families have significantly higher expression than moderate and inert organizational families (***) ($p < 1e-10$, Mann-Whitney U test).

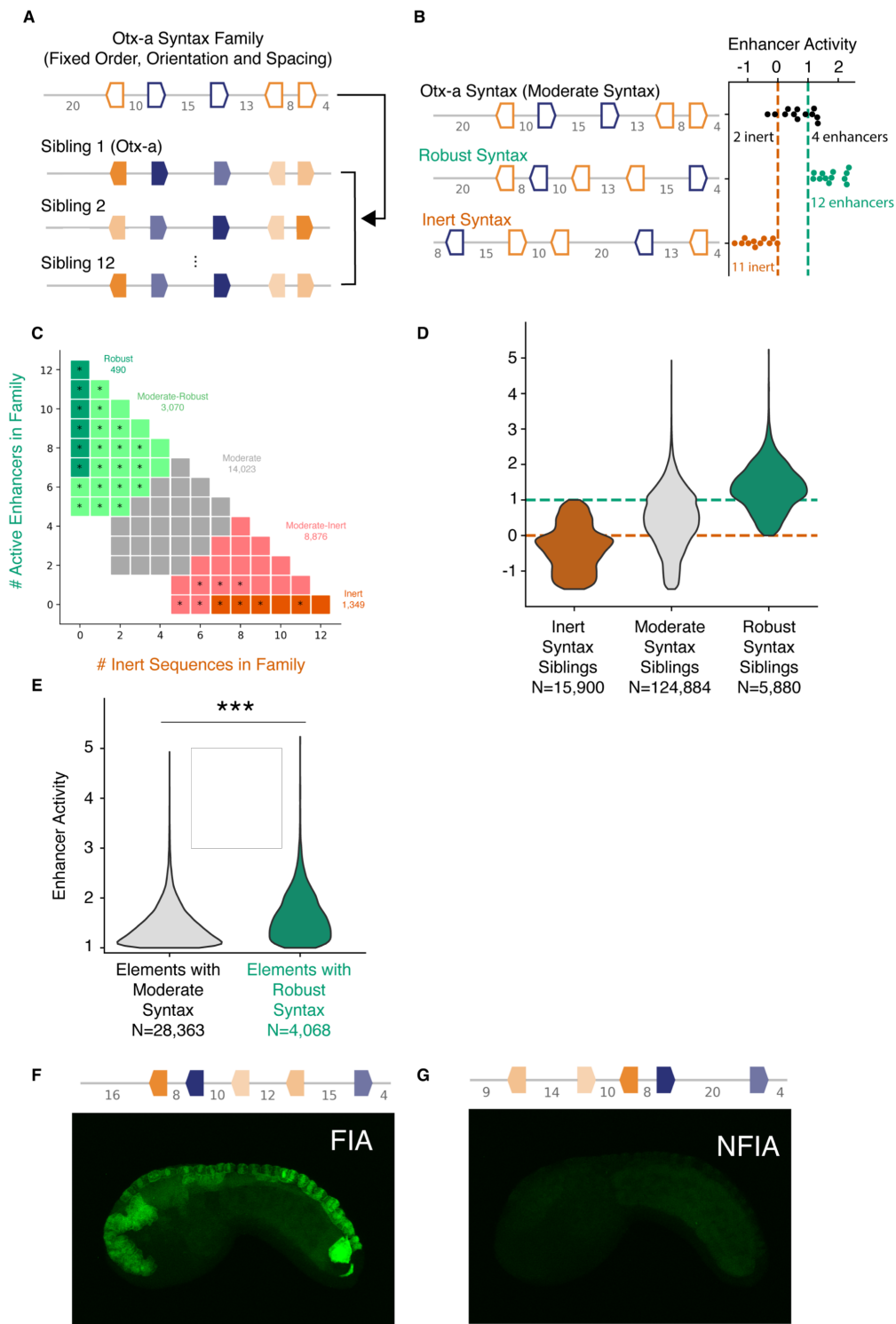


Figure 3.3. Robust and moderate syntaxes drive enhancer activity.

3.3.6 Single syntax changes convert inert elements into active enhancers

To further investigate the importance of syntax we looked at the importance of binding site orientation. 38% (54,164) of the inert elements can be converted to active enhancers by flipping the orientation of just a single binding site (Fig. 4A). Of all the flips, 58% flipped a GATA and 42% flipped an ETS site. We tested two single flips which convert an inert sequence into an active enhancer via standard reporter assays (Fig. 4B,C, Fig. S11A,B). In OLS 93, flipping the orientation of an ETS site converts an inert sequence into a neural enhancer (Fig. 4B,C). In OLS 89, flipping a GATA site converts an inert sequence into a neural enhancer (Fig. S11A,B). To more definitively rule out the presence of de novo activators or repressors in these single flips, we randomized the backbone of a single flip OLS 89 and 90 (Fig. S6A,B). For these two enhancers we see that randomization of the linkers does not alter enhancer activity, with OLS89R showing indistinguishable expression from OLS89, while the same element with a flipped ETS site has no activity even with randomized backbones. We next wondered how these flips rendering enhancers active relate to organizational families. Single flips that make enhancers functional do so by moving elements into functional organizational families (Fig. 4D). The most extreme examples of this are single flips that moves an enhancer from an inert organizational family to a robust syntax (Fig. 4E).

3.3.7 Searching for signatures of functional grammar

Having seen that syntax is critical for enhancer activity we next wanted to discover if there were particular syntax features that are enriched within the active enhancers. Although we discovered hundreds of robust unique organizational families, it is likely that some of these 5-site syntaxes share smaller 2- or 3-site syntax sub-features that contribute to gene expression activation. If such functional syntax motifs exist, these syntax features should be enriched in active enhancers both within our screen and potentially the genome.

To determine the enrichment of syntax features in active enhancers, we first enumerated all possible 2-site and 3-site syntax features that are present across all 460,800 library elements. We encode these syntax features using order, orientation and spacing for 2-site and 3-site features. For example, g_15_E is a 2-site syntax feature (Fig. 5A). We use uppercase G to represent

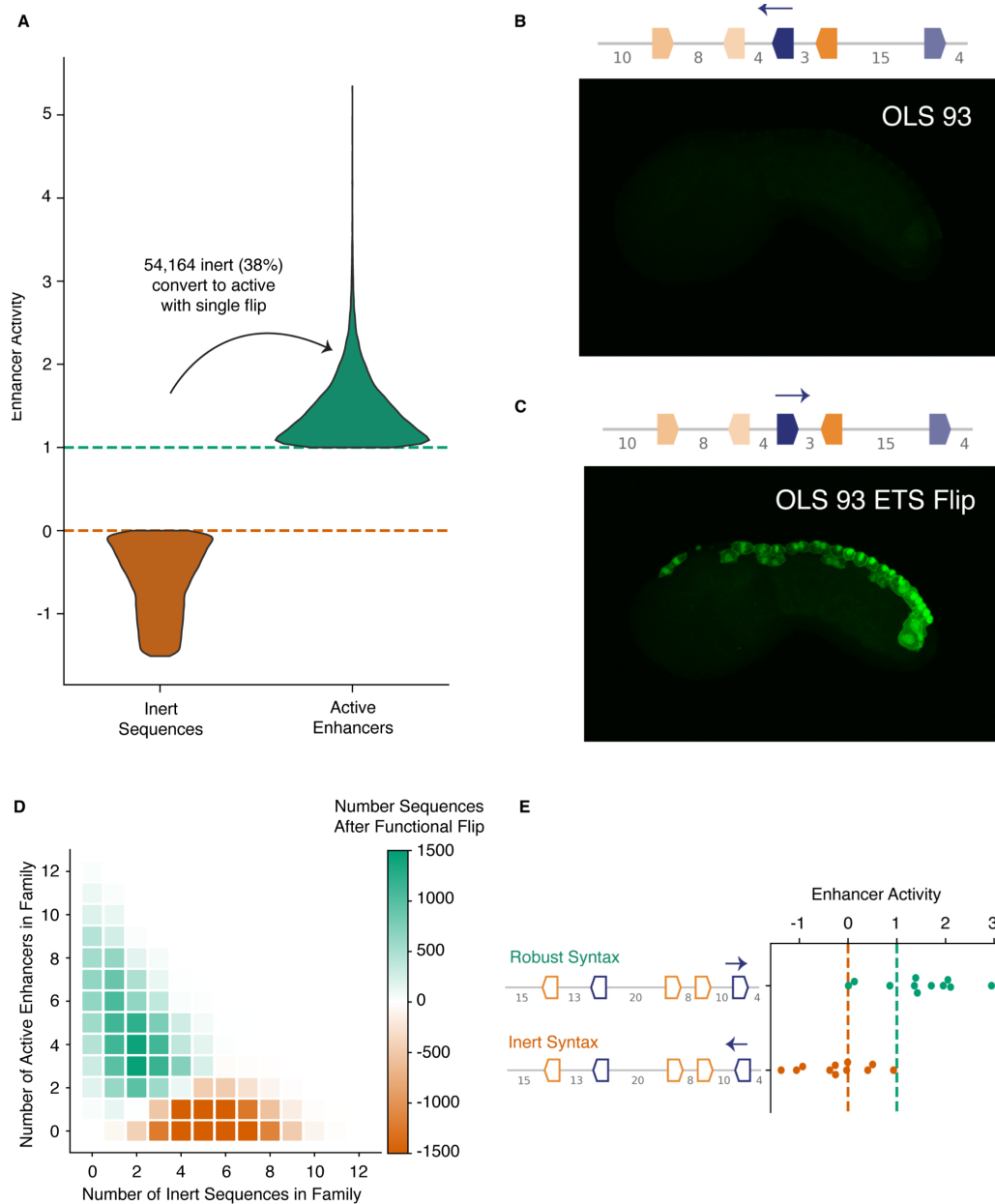


Figure 3.4. Orientation of individual sites convert inert elements into active enhancers. (A) 54,163 inert sequences (38%) can be converted to an active enhancer with a single flip of a binding site. (B) The inert sequence OLS 93 activates no expression. (C) A single flip of an ETS site in OLS 93 converts the inert sequence into a neural enhancer. (D) Heatmap showing that flips lead to elements moving into more functional organizational families. (E) An inert and a robust organizational family which only differ due to the orientation of a single ETS site.

forward GATA site, while lowercase g represents reverse GATA site. g_7_E therefore represents a reverse GATA followed by a forward ETS which are spaced by 7bp. Distance is measured from the middle of the GATA or GGAW core. Exhaustive generation of all 2-site and 3-site syntax features found across the 460k elements results in 9,022 possible syntax features.

A syntax feature is present in many enhancers. For example, E_35_G is present in 2,556 active enhancers (26%) and 7,025 inert sequences (74%) (Fig. 5B). To determine if there are syntax features enriched within active enhancers, we test for enrichment of a syntax feature by comparing the proportion of syntax-containing elements that are active enhancers compared with the proportion of active enhancers across all 460,800 elements (Fig. 5B). Across all classifiable elements of all 460,800 library elements, we observe 52,139k (27%) active enhancers dependent on GATA-ETS and 143,799k (73%) inert sequences. We use the hypergeometric distribution to calculate a p-value which signifies if a syntax feature is enriched in active enhancers. We find that although E_35_G is not enriched, we find g_15_E and g_15_E_35_G are significantly enriched in active enhancers. Interestingly, the effect size of g_15_E_35_G as calculated by the odds ratio is greater than g_15_E (Fig. 5B). After Bonferroni adjustment of p-values, 931 of 9,022 syntax features are found to be significantly enriched in active enhancers ($p < .05$, Fig. 5C). To assess the false positive rate, we permuted the element labels and enhancer activities values, and repeat the search for syntax features. In the permuted dataset, no syntax features are significantly enriched in active enhancers (Fig. 5C). Looking at 2-site syntax features, we observe that there are certain spacing between sites that are more enriched within active enhancers compared to other spacings, this presents as a periodicity of spacing preferences between the sites. Using these datasets we were able to impute spacing not present in our library. Within the 931 activator syntax features enriched in active enhancers we see a wide variety of spacings, combinations and orientations. The diversity of syntax features we observe suggests that these 2-site and 3-site features represent many functional ways in which GATA and ETS can be put together to form an active enhancer.

3.3.8 Syntax features and syntax scores with active enhancers.

Many of the enhancers contain several enriched syntax motifs. We therefore developed a methodology to score the different GATA-ETS syntax motifs within any DNA sequence (Fig.

5D). First, the GATA and ETS sites are located by searching for GATA and GGAW TFBS core sequences. Given the relative organization of GATA and ETS in the query DNA sequence, we determine if any of the 931 activator syntax features are present in this organization. The syntax score is the sum of the effect sizes for all the activator syntax feature within the query sequence. The effect size of an activator syntax feature is the odds-ratio which is a quantification of enrichment of the syntax feature in active enhancers. Thus for any synthetic or genomic element we can create a syntax score. We do not know how these syntax features work together, but by adding their odds ratio together, we assume that each of these syntax features are contributing to the expression of the elements, and we want to assess if the syntax score correlates with expression.

Figure 3.5 Syntax features and syntax scores with active enhancers. (A) The Otx-a enhancer contains many possible syntax features, some inert and some that are enriched in active enhancers. Three examples of syntax features E_35_G, g_10_E and g_10_E_35_G are shown. (B) Enrichment of syntax features in active enhancers. Across all elements tested within the Otx-a grammar library, (27%) are active enhancers (shown in green) and (73%) are inert sequences (shown in red), this represents the null distribution. 27% of sequences containing E_35_G are active enhancers, representing no significant enrichment. g_10_E (37% sequences of are active enhancers) and g_10_E_35_G (53% of sequences are active enhancers) are significantly enriched in active enhancers (**=Bonferroni adjusted $p < 1e-10$). Because g_10_E_35_G is enriched in a higher percentage of enhancers, it has a higher effect size. (C) All tested syntaxes and the corresponding log-transformed Bonferroni-adjusted p-value. Only syntax features with enrichments in active enhancers of $\geq 10\%$ were tested. Statistics of real data in black dots, while statistics of permuted dataset plotted with grey dots. No syntax features are significantly enriched in active enhancers in the permuted dataset. Red dashed line indicates significance cutoff p-value of .05. (D) The syntax score of any arbitrary DNA sequences is calculated by first determining GATA and ETS positions using GATA and GGAW, respectively. The GATA and ETS positions are then used to determine presence of activator syntax features. Syntax score is calculated by summing the effect size of all present activator syntax features. Effect size of each syntax feature is the odds ratio enrichment in active enhancers. (E) Elements with syntax score ≥ 30 are highly enriched in active enhancers compared to elements with syntax score ≤ 5 . (F) Depiction of the enhancer activities of the elements in panel (E). (G) Active enhancers with syntax score ≥ 30 have significantly higher levels of enhancer activity than enhancers with syntax score of ≤ 5 . (H) Robust syntaxes tend to have higher syntax scores than inert syntaxes.

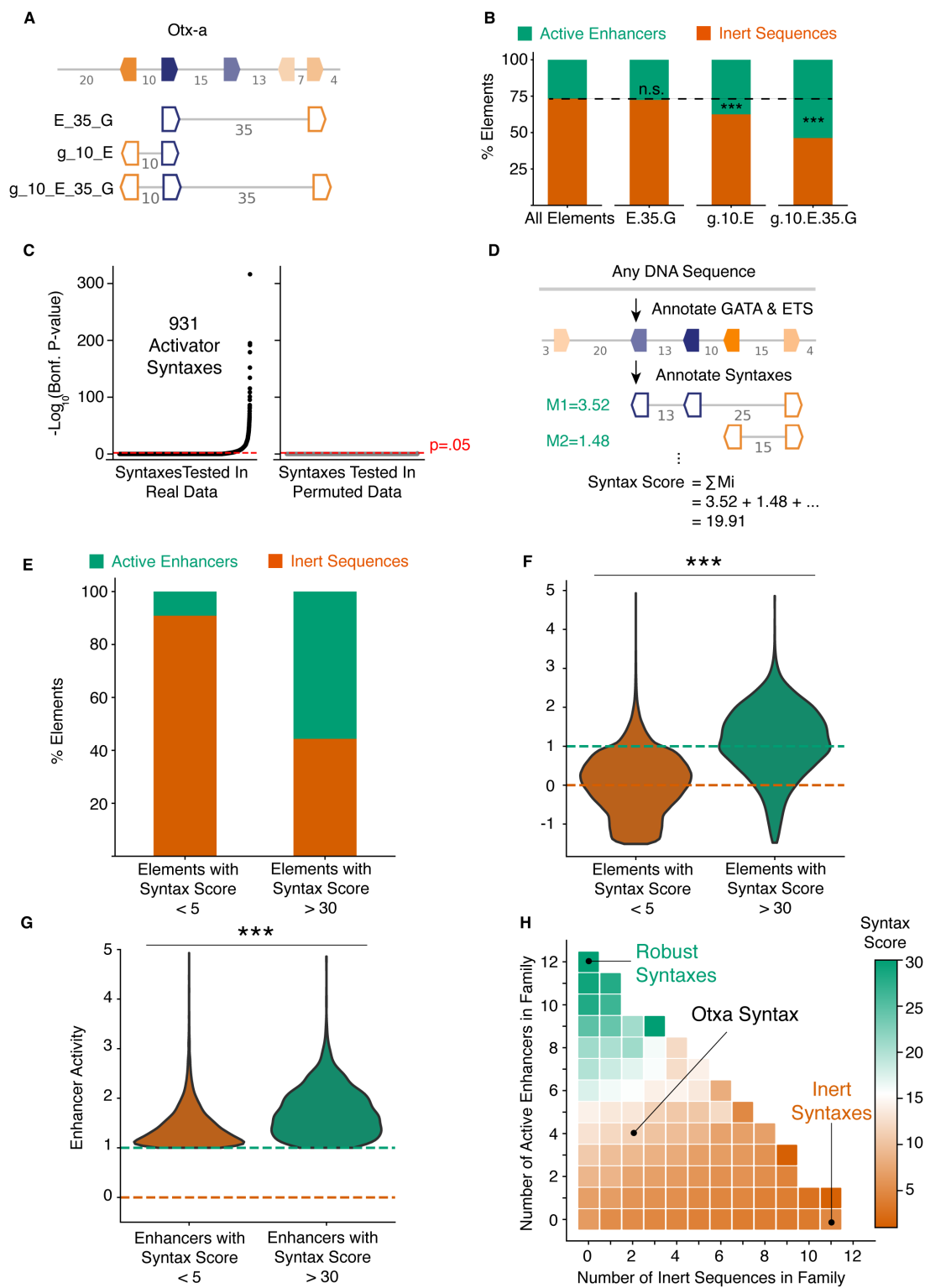


Figure 3.5. Syntax features and syntax scores with active enhancers.

3.3.9 Syntax score defines functional and non-active enhancers.

Although the activator syntax features are enriched within active enhancers we did not know if elements with combining syntaxes and our syntax score would be an effective method to find active enhancers. We measured the syntax score for all 52,139k active enhancers and 143,799 inert elements. We find that 10% of elements with syntax score ≤ 5 are active enhancers, compared to 45% of elements with syntax score ≥ 30 that are active enhancers. This suggests that syntax can be used as a feature to predict active enhancers. Enhancers with low syntax have significantly lower MPRA activity levels than enhancers with high syntax (Fig. 5G). These results indicate that syntax score is an effective way to identify active enhancers, and combinations of these syntax features do lead to more functional enhancers.

3.3.10 Genomic enhancers encode activity using dependency grammar

We wanted to see if searching for syntax features could help us find active enhancers within the genome. We looked for elements with at least 3 sites within the genome which contains at least one ETS and one GATA. We only considered ETS and GATA sites with ≥ 0.1 for ETS and ≥ 0.2 relative affinity for GATA (Fig. 6A). We found 2,247 3-site elements and 907 4-site elements that contain ETS and GATA clusters. These genomic elements have a wide distribution of ETS and GATA affinities, the median relative affinity of these elements is 0.3 for GATA and 0.12 for ETS. None of the elements contain consensus ETS or GATA sites. To see how affinity contributes to enhancer activity we looked at the relationship between affinity and enhancer activity. We calculate a single affinity for each TF GATA and ETS by summing the relative activities of the sites (Fig. 6C). We found summing the affinity to be most effective, likely because it takes into account valence. Active enhancers have subtle increase in summed ETS and GATA affinity (Fig. S16 B,C), although the median affinity of these sites within active enhancers is still low at .25 for ETS and .7 for GATA. To integrate the affinities of ETS and GATA we created a heatmap in which each bin represents the affinity of both GATA and ETS (Fig. 6D). We find that there are active enhancers within many different affinity bins, indicating that enhancers with a range of affinities are active.

We next looked at the importance of syntax for enhancer activity. Regardless of their

affinity, active elements have a subtle but significantly higher syntax score than inactive elements, indicating that active enhancers contain more of the syntax features found within our synthetic screen (Fig. S16A). Interestingly, the effect of syntax is more enriched than that of affinity of either binding site alone suggesting syntax to be an important feature when identifying active enhancers.

Enhancer grammar is the interplay of affinity and syntax. We have previously seen within notochord enhancers that there is an interplay between affinity and syntax of binding sites⁶, we therefore wondered if looking at both syntax score and affinity could help identify active enhancer. We binned the elements based on the syntax score (low versus high syntax) and the affinity of ETS and GATA. To exclude the possibility of the number of sites confounding the analysis, we first only considered 3-site clusters. There are 1,007 3-site elements with no syntax motifs (no syntax group, Fig. 6E). We compared these no syntax elements to the elements with the top 500 syntax scores (with syntax group, Fig. 6F). We can test for significant enrichment in each bin by comparing the observed enrichment in activity labels with the enrichment observed after permuting the enhancer labels and activity values. None of the bins in the no syntax group were significantly enriched for active enhancers. In the high syntax group, elements with 0.5-2.1 ETS and 0.55-1.0 GATA were significantly enriched for active enhancers (Fig. 6E). We found the elements in the high syntax groups to be significantly more enriched in active enhancers when comparing similar-affinity bins (Fig. 6G). We observed increases in percent active enhancers in all bins except the elements containing the highest affinity GATA sites. This suggests that the effect of syntax may not be observed in the highest affinity clusters. However, these high-affinity clusters only account for 15% of the total clusters, indicating that the capability of the vast majority of clusters to be an active enhancer will be influenced by the strength of their syntax. These results illustrate that syntax can compensate for affinity to encode tissue specific expression and that the dependency between these features encodes enhancer activity.

3.3.11 Designing tissue specific enhancers using grammar

We only looked at ETS and GATA sites within these genomic clusters and did not look at factor in the contribution of any other factors, despite this we observe evidence of enhancer

grammar where the affinity and syntax together are important in driving expression. As syntax features can predict active enhancers within the genome we next wanted to see if syntax features are sufficient to create synthetic tissue specific enhancers. To reflect the types of sites found within the genome we chose elements with moderate affinity ETS and GATA sites and good syntax. We selected syntax features found within the 0.5-1 bin of ETS and GATA affinities, which we found to be significantly enriched in active enhancers. We placed these syntax features onto inert backbones. Our synthetic element with a high grammar score drives expression similar to the WT Otx-a enhancer. Thus indicating that not only are syntax features helpful to find genomic enhancers, but they can also be used to design tissue specific enhancers.

Figure 3.6 Dependency grammar encode genomic and synthetic enhancers. (A) GATA and ETS genomic clusters library is generated by extracting all 66bp tiles with 3 to 4 non-overlapping sites. At least one GATA and at least one ETS is required with affinity of 0.2 and 0.1 respectively. (B) The enhancer activity of 3,144 GATA and ETS genomic clusters taken from the Ciona genome. Otx-a enhancer activity plotted with green dot. Enhancer activity of 8 microscope validated neural enhancer and inert sequences annotated with green and red dots, respectively. Purple dot represents enhancers with neural and ectopic expression. (C) GATA and ETS elements can be grouped by their aggregated GATA and ETS affinity by summing the affinity of the individual sites. The syntax of these elements is quantified using the same syntax activator features detected in the synthetic screen. (D) The association of enhancer activity with aggregate affinity of GATA and ETS within elements. The X-axis is the summed affinity of GATA; the Y-axis is the summed affinity of ETS. The labels for each bin are as follows: (# active enhancers) / (total # elements) (% of elements active). Color intensity of the square is determined by the percent of elements that are active. Strongest activity bins observed at the highest percentile of affinity. Some element activity observed in lower affinity bins. (E,F) Same heatmaps as panel D but separating elements into low and high syntaxes. (E) Most elements without syntax are 35-38% active, except the higher affinity elements which are 44-47% active (F) The 500 elements containing the highest syntax scores (corresponding to a syntax score (SS) of greater than 2.81). Most elements with high syntax have higher percent active elements than elements in the low syntax bin, ranging from 45-58%, even for low affinity elements. (G) Elements with syntax have significantly higher percent active elements than compared to elements with no syntax. Lines connect bins which contain similar affinity elements, only differing in the syntax of sites. (H) Generation of synthetic enhancers using activator syntax features with similar strengths to high syntax scoring genomic elements. Activator syntax features used were detected in the screen but not in the genomic clusters library. ETS and GATA 8mer sequences chosen using PBM data which have affinity similar to wild-type Otx-a. Inert backbone taken from ablated sequences known to activate no gene expression. An example of a synthetic enhancer SE1 with syntax score of 11.8 driving neural expression in the asv, dorsal epidermis, palps, and two tail muscles.

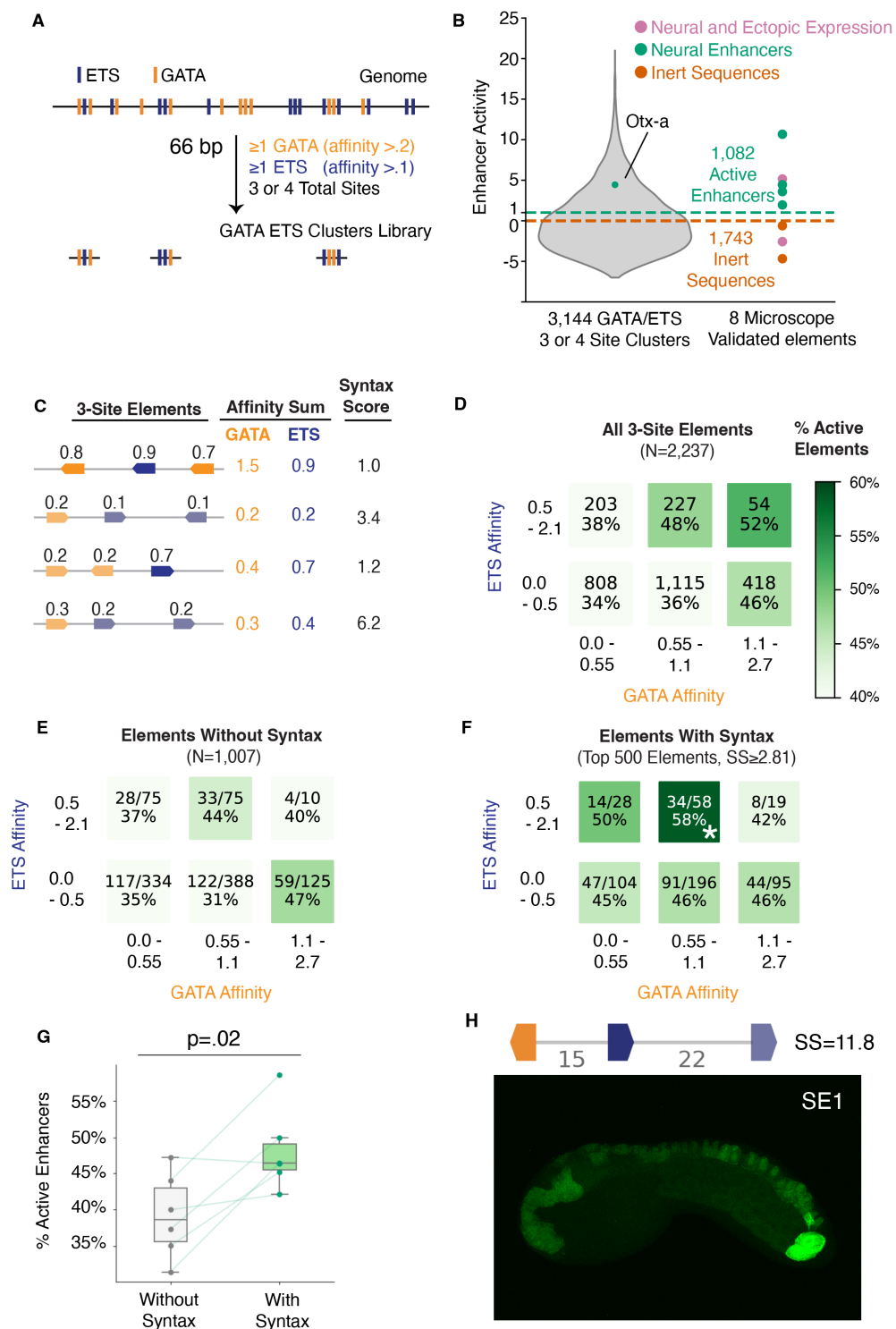


Figure 3.6. Dependency grammar encode genomic and synthetic enhancers.

3.4 Discussion

Enhancers are a critical feature of our genomes, as they provide the instructions for when and where genes are transcribed. Yet in contrast to coding regions of the genome we have very limited understanding of how DNA sequences encode for active enhancers. To address this problem we did an exhaustive screen of the well characterized Otx-a enhancer. We made every possible order, orientation and spacing of this enhancer and tested these for activity within *Ciona* embryos. Strikingly, only a fraction of the elements are functional despite the fact that they all have the same number and affinity binding sites and same sequence content. The active enhancers have seemingly very different organization of sites. Why all this disparate elements are functional is a remaining mystery, but these diverse elements likely all adhere to higher order grammatical rules governing construction of a functional complex that can activate transcription.

We found that these elements could be separated into organizational families all of which share the same order, orientation and spacing of sites but have affinities at different positions. We find that some syntaxes are robust and drive expression regardless of where the different affinity sites are located, interestingly these robust syntaxes had the highest levels of enhancer activity and also showed low levels of ectopic expression in non-neural cells raising the possibility that syntax most permissive to building a functioning complex that can activate transcription are also closer to driving ectopic expression. We determined that some syntaxes are not permissive to active enhancer activity and are inert. The majority of enhancers lie within moderate syntaxes where affinity of binding sites influences if the element is active. For moderate enhancers the organization of sites is likely less optimal and thus the energy from the affinity interactions are important for the assembly of the functional complex, whilst for the robust elements the organization is so permissive to the assembly of the complex the organization of sites does not matter.

Very small changes to elements can have dramatic effects, we find more than a third of inert elements can be converted into active enhancers by flipping the orientation of a single binding site and this moves elements into more robust organizational families. This has important implications for assessing if an element is an active enhancer within the genome, and if a de

novo site will likely contribute to changes in enhancer activity and organismal level phenotypes.

We analyzed syntax in a rigorous manor by comparing how presence of syntax contributes to differential enhancer activity for GATA and ETS genomic elements with similar affinities. This results in a heat map, where you physically separate elements by their affinity, and then quantify how the presence of syntax relates to enhancer activity. Future efforts will focus on collapsing all elements to a single axis and predict their enhancer activity by integrating the affinity and syntax of binding sites. The relative contributions of affinity and syntax is unclear, making this single quantification of enhancer grammar difficult. Additionally, it is also unclear how to encode the affinity of a binding site cluster with multiple types and numbers of binding sites (e.g. 2 GATA + 1 ETS versus 1 GATA + 2 ETS). One approach is to do this explicitly, assigning weights to affinity and syntax and assess how well this predicts enhancer activities. Another approach is to encode the type, number, affinity and syntax of an element's binding sites into a machine learning model to see if it can learn how to combine the relative contribution of the features which optimize prediction accuracy.

Machine learning models rely on a numerical encoding of enhancers, almost exclusively using one-hot encoding to represent the precise sequence of the enhancer. This approach results in a $4 \times N$ matrix where each of the 4 rows encodes presence of A/T/G or C and N is the length of the enhancer^{11;49;61;175;176}. This is where models had to begin because we still do not understand the enhancer features encoding gene expression, so we must rely on encoding the precise sequence and hoping the model will learn the features. There is no doubt these models have advanced the field of enhancer predictions. However, these sequence based models may be limited in their ability to make predictions based on the linear encoding of precise sequence. Although some organizational preferences have been detected using machine learning models as an "oracle" to predict how sequence changes enhancer function^{48;49;177}, we do not yet know if these models can pick up on nuances of site number, affinity and syntax given the limitation of how many training sequences can be inputted. Indeed, low affinity sites are degenerate and there are multiple functional syntaxes which can put together the same tissue-specific expression pattern^{9;178}. We hypothesize that the more we understand about the functional features governing enhancer levels and tissue specificity, the more precisely we can encode these features for the machine learning

models to learn and predict on. Given the importance of enhancer grammar which is encoded by the site type, number, affinity, and syntax, we hypothesize that these features will be critical to encode.

Together these results illustrate that there are clear grammatical constraints on how proteins assemble on enhancers and that even very subtle changes to an enhancer sequence can cause a dramatic change in activity. While there is clearly grammatical constraints on enhancers we have not yet been able to full unravel these rules. Using our synthetic enhancer screen we were able to determine syntax feature enriched within active enhancers and developed a syntax score based on this enrichment. This syntax score enabled us to enrich for functional enhancers within the genome demonstrating that the synthetic screen could inform us within the context of the genome. Beyond the genome, we were able to use syntax features to design tissue specific enhancers. Together this work demonstrates that an understanding of the grammatical constraints on enhancers can unlock the instructions for tissue specific genome expression embedded within our genomes.

3.5 Methods

3.5.1 Ciona Electroporations

Counting Embryos. For enhancers being compared, GFP expression were counted on the same day along with Otx-a wild-type as a reference in each set of experiment. Fifty embryos were counted for each biological replicate. For all experiments, slides were counted blind. When comparing % embryos with expression, we used Fisher’s Exact Test to determine if two constructs produced significantly different expression in a tissue.

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 in the same way to an appropriate size. In each figure, the same exposure time for each image is shown to allow direct comparison.

Library Construction for Fluorescence Microscopy . Randomized linker constructs were ordered from IDT with Ns (equal chance of A/T/C/G) in place of the linker regions. The

enhancer library was transformed using MegaX DHB10 electrocompetent cells with 1,000,000 to 5,000,000 total randomized linker variants.

Electroporation for Fluorescence Microscopy. All constructs whose expression patterns are to be compared to each other were midiprepmed in the same batch along with a Otx-a WT control, using the NucleoBond Xtra Midi kit (Macherey-Nagel). Dechoriation, in vitro fertilization, and electroporation were performed as described previously in Farley et al., 2016. 70 ug DNA was resuspended in 100 uL water and added to 400 uL of 0.96 M D-mannitol. Typically for each electroporation, eggs and sperm were collected from 6 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 Antifade Mountant. GFP images were obtained with an Olympus FV3000, using the 40X objective. All constructs were electroporated in at least two biological replicates.

3.5.2 Scoring Relative Affinities of Binding Sites

We calculated the relative ETS and GATA binding affinity using the median signal intensity of the universal protein binding microarray (PBM) data for mouse ETS-1 (Wei et al., 2010) and GATA-6 (Badis et al. 2009) proteins from the UniProbe database (<http://thebrain.bwh.harvard.edu/uniprobe/index.php>)¹²³. Previous studies have shown that the specificity of ETS family members is highly conserved 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. The binding preference for *Ciona* Gata.a is highly similar to the orthologous human and mouse GATA4/5/6^{94;174}.

The relative affinity score represents the ratio of the median signal intensities of the native 8-mer motifs compared to the optimal 8-mer motifs for optimal TF binding. For ETS, the consensus site with a GGAW core is CCGGAAGT. For GATA, the consensus site with a GATA core is GAGATAAG. The optimal or consensus sites have a relative affinity score of 1.0.

3.5.3 MPRA Libraries

Gata and Ets Grammar Library Generation . The Otx-a Gata and Ets Grammar Library was designed by reorganization of the Otx-a enhancer with minimal changes to the sequence content. Since there is an ETS core of GGAW in Linker 1 of Otx-a, we deleted the first 3 bases of linker one to get rid of the ETS core. We also found that Linker 4 has a TTC which could form a de novo ETS TTCC core if joined with a binding site starting with a C. We therefore performed a point mutation in Linker 4 to avoid this (TGTTTC to TGCTC). We exhaustively generating every sequence following this formula: any order of linker sequences in the forward direction with any order and orientation of binding sites. This generated $25 * 5! * 5! = 460,800$ grammar variants. We additionally generated another 460,800 variants with point mutations to each Gata (GATA to AATA) and Ets (GGAW to GCAW) site known to ablate binding. This totaled to 921,600 sequences in the Gata and Ets Grammar Library. The oligo library was synthesized by Twist.

Gata Ets Genomic Clusters Library Generation. We extracted all 66bp tiles which include 3 or four binding sites, of which at least 1 Gata (matching GATA with affinity ≥ 0.2) and 1 Ets (matching GGAW with affinity ≥ 0.1) sites must be present. We filtered to only look at non-overlapping Gata and Ets sites. For dense clusters of Gata and Ets, we ordered every 66bp tiles every stepping 15bp. This results in a library of 3,144. The oligo library was synthesized by TWIST Biosciences.

Library Cloning . Both the Gata and Ets Grammar Library and the Genomic Clusters library were cloned with the following steps. The enhancer variants 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 BseRI⁵. After cloning, the library was transformed into MegaX DHB10 electrocompetent cells, and the culture was grown up until an OD of 1 was reached. DNA was extracted using the Macherey-Nagel Nucleobond Xtra Midi kit. A 30bp barcode with adapters containing Esp3I sites was cloned into this library using type II restriction enzyme Esp3I. The library was transformed into 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. The Gata and

Ets Grammar Library members contain 50 barcodes per enhancer. The Gata and Ets Genomic Clusters Library members contain 12 barcodes per enhancer.

Enhancer to Barcode assignment. To determine which enhancer was associated with which barcode tag, we removed the intervening sequence containing promoter and GFP by restriction digest with XbaI and SacI and re-ligated the remaining vector to get the enhancer and barcode in close proximity. We amplified the region of interest with 10 cycles of PCR, and size selected the PCR product using Agencourt AMPure XP beads. We ran the purified sample on a bioanalyzer to check the product was the correct size then sequenced the library using NovaSeq S4 PE150 mode (Illumina).

Enhancer to Barcode Dictionary Analysis . Raw sequencing data was first deduplicated to get a file with unique reads and associated read counts. The location of the barcode and enhancer in the read is known based on the vector design. For each unique read, we searched for the enhancer in the expected location -5 to +1 bp. If the sequence at the enhancer location is a perfect match to a sequence we ordered, then we extract the enhancer from the read and the first 25bp as the barcode. We filtered out any barcodes that were mapped to multiple enhancers.

MPRA electroporation . Each library was electroporated into 5000 fertilized eggs. Embryos were developed until 5hrs 15 min at 22°C. Embryos were then put into TRIzol, and RNA was extracted following the manufacturer’s instructions (TURBO DNA-free Kit, Invitrogen). 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 magnetic beads as per instructions (Dynabeads mRNA Purification Kit, Life Technologies). The mRNA was used in an RT reaction that was specifically selected for the barcoded mRNA (Transcriptor High Fidelity cDNA Synthesis Kit, Roche). The RT product was PCR amplified and size selected using Agencourt AMPure XP 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). At least three biological replicates were sequenced.

The DNA was extracted by mixing the phenol-chloroform and interphase of TRIzol extraction with 500uL of Back Extraction Buffer (4M guanidine thiocyanate, 50mM sodium citrate, and 1M Tris-base). DNA was treated with RNase A (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 XP beads (Beckman Coulter), then checked for quality and size on the 2100 Bioanalyzer (Agilent) before sequencing on the NovaSeq S4 PE100 mode (Illumina). At least three biological replicates were sent for sequenced.

Gata and Ets Grammar MPRA Analysis . We extracted the barcode as the first 25 bases of the read. We determined the read counts of each library in the FASTQ file and normalized the read counts to reads per million (RPM). We generated a table of all dictionary barcodes and associated RPMs from each replicate. We aggregated barcode RPM measurements within enhancers using the barcode-enhancer dictionary. For each replicate, enhancer activity is calculated by summing associated barcode RPM in RNA and dividing by the sum of the RPM in DNA. Replicate activities < 0.5 are considered not detected and filtered out. The remaining replicate activities for an enhancer are summed to calculate the experiment-wide enhancer activity score. Finally, the activity scores are log2 transformed, and rescaled such that 0 is the 40th percentile of activity and 1 is the 75th percentile of activity. This rescaling of the data to these percentiles transforms the scores into a more intuitive scale where elements with activity less than 0 are inert, and elements greater than 1 are active enhancers. The set of 52,139 active enhancers dependent on Gata and Ets were determined by requiring the element enhancer activity to be ≥ 1 and the matched ablated control to be ≤ 0 .

Gata and Ets Genomic Clusters MPRA Analysis. We extracted the barcode as the first 25 bases of the read. We determined the read counts of each library in the FASTQ file and normalized the read counts to reads per million (RPM). We generated a table of all dictionary barcodes and associated RPMs from each replicate. We aggregated barcode RPM measurements within enhancers using the barcode-enhancer dictionary. For each replicate, enhancer activity is calculated by taking the average RNA:DNA ratio across all barcodes associated with the sequence. Replicate activities < 1.5 are considered not detected and filtered out. The remaining replicate activities for a sequence are summed to calculate the experiment-wide enhancer activity score. Sequences were removed from the analysis if they were not detected in ≥ 3 replicates. Finally, the activity scores are log2 transformed, and rescaled such that 0 is the 55th percentile of activity and 1 is the 65th percentile of activity. This rescaling of the data to these percentiles

transforms the scores into a more intuitive scale where elements with activity less than 0 are inert, and elements greater than 1 are active enhancers and less than 0 are inert sequences.

3.5.4 Analysis of grammar

Syntax Family Analysis. Syntax families are defined by a fixed order, orientation and spacing. Sequences are grouped into syntax families if their (1) order of linkers and (2) order and orientation of Gata/Ets binding sites are the same. Since there are 2 Ets sites and 3 Gata sites, there are $3! * 2! = 12$ possible sequences per syntax family. For each family we determined how many active enhancers and inert sequences were detected in the syntax family. We defined inert syntaxes as families 7 or more inert, with 0 active enhancers. We defined robust families as having 7 or more active enhancers, with 0 inert sequences. We determined the average syntax score in each bin of the syntax family by taking the average syntax score of all sequences in that bin.

Single Flip Analysis. For every inert sequence (activity ≤ 0), we determined if any of sequences created by any of the possible 5 single flips were active enhancers dependent on GATA and ETS (wt variant activity ≥ 1 , abl version ≤ 0). We determined which syntax family was associated with the inert sequence and active enhancer associated with all single flips that resulted in converting an inert sequence to a functional enhancer. Using the syntax family heatmap from Fig. 3C, we determined how many inert sequences were in each bin. We then did the same for active enhancers, determining how many active enhancers (products of the single flips) were in each bin of the syntax family heatmap. We then subtracted the count of active enhancer counts from inert sequences to get the transition between each bin. We find negative values in inert syntaxes and positive values in robust syntaxes, indicating that the active enhancers from single flips typically moved from inert syntaxes to robust syntaxes.

Detection of Syntax Motif Enrichment. Syntax features were encoded using order, orientation and spacing of Gata and Ets binding sites. We enumerated exhaustively all possible orders, orientations and spacings of which a 2-mer or 3-mer binding site syntax feature could be arranged. We then calculated the enrichment of each syntax feature in active enhancers and calculated a p-value using the hypergeometric distribution.

For a given possible syntax motif, we calculated the following parameters of the hypergeometric distribution to test the significance:

Hypergeometric Parameters of the null distribution N = population size = total number of sequences in the library (460,800)

K = number of success states in the population = number of neural enhancers dependent on GATA/ETS (50,643)

Hypergeometric Parameters calculated for each syntax feature n = number of draws = number sequences (of 460,800) which contain the grammar motif

k = number of observed successes = number of neural enhancers dependent on GATA/ETS (of 50,643) which contain the grammar motif

P-value for hypergeometric test is defined as $P(X \geq k)$. Statistical tests only performed for enrichments $\geq 10\%$. Bonferroni correction applied to adjust for multiple tests.

To assess the rate of false positives due to noise, we performed the same statistical test for enrichments of grammar motifs in a shuffled dataset. The shuffled dataset was created by simply shuffling the activity label for each sequence. No significant motifs are detected in the shuffled dataset.

Since the Gata and Ets grammar library contains fixed length linker sequences, there are some spacings between Gata and Ets not observed in our library. To estimate the enrichment of the syntax feature we would observe if it were in the library. To do this, we first determined what the observed enrichment is for +1/-1 of that spacing. We then imputed the enrichment of that spacing by averaging the two adjacent enrichments of +1/-1 spacing. Finally, we calculated the p-value using imputed enrichment with the same procedure as above.

Syntax Score Calculation. The location and orientation of sites within a DNA sequence were determined through matches to the core sequences: GATA for Gata and GGAW for Ets. From the order, orientation and spacing of those annotated sites, we then determine which syntax motifs are present in the sequence. We then sum within each sequence all syntax motif odds ratios as a measurement of the syntax motif's enrichment in active enhancers as observed in the screen. Sequence with no motifs are given a summed odds ratio of 1 as an odds ratio of 1 is no enrichment at all.

Analysis of GATA and ETS genomic clusters syntax and affinity. GATA (≥ 0.2 affinity) and ETS cores (≥ 0.1 affinity) excluded due to potentially being nonfunctional due to ultra low affinity. The remaining sites were summed to calculate the aggregate summed affinity. Syntax score of sites remaining using the same process as outlined in “Syntax Score Calculation” above.

3.5.5 Design of Synthetic Enhancers

Synthetic enhancers were generated starting with the 3mer syntax motifs most strongly enriched in active enhancers. Gata and Ets binding site sequences were determined by using PBM data to find 8-mer sequences which are similar to Otx-a binding sites. Inert backbone sequences used to link binding sites at specified spacing were taken from sequences with ablated TFBS shown not to activate any transcription from Song et al 2022¹³⁰ (ZEE36 and ZEE29).

3.5.6 Data Availability

Data will be made available upon publication

3.5.7 Code Availability

Code will be made available upon publication

3.5.8 Acknowledgements

Chapter 3 is a study that is in review at the time this thesis was written. I acknowledge Fabian Lim (Co-first), Jessica Grudzien, Ben Song, Sophia Le, and Emma Farley.

3.5.9 Author Contributions

Joe J Solvason performed all the computational analysis and lead the direction of the paper. Fabian Lim, Jessica Grudzien, and Emma Farley were critical for the direction that this work took. Ben Song generated the synthetic library and performed the screen in embryos. Fabian limb generated the GATA/ETS clusters library and performed the screen in embryos. Fabian Lim, Jessica Grudzien, Ben Song, and Sophia Le performed electroporations where a single construct was counted.

3.5.10 Competing Interests

No competing interests.

Epilogue

4.1 Discussion

4.1.1 The prevalence of low affinity sites creates a vulnerability in genomes

The first demonstration of the importance of low-affinity binding sites comes from studies of hox patterning in the fly embryo⁴ and the nervous system of *Ciona robusta* (*Ciona*)^{5;6}. Optimizing the affinity of transcription factor (TF) binding in these enhancers causes loss of tissue-specificity, though the downstream phenotypic consequences were not studied. In this thesis, we demonstrate that single nucleotide changes to DNA sequence which drive affinity optimizations cause gain-of-function gene expression and developmental defects in the *Ciona* heart⁷ and vertebrate limb bud⁸. Importantly, more than 61% of predicted human heart enhancers from contain an ETS binding site that can be affinity-optimized by a SNV at levels known to cause loss of enhancer tissue-specificity⁷. This reveals a vulnerability in genomes in which genetic variation can easily cause affinity optimizations in enhancers. Indeed, we find that affinity optimizations of ETS, AP1 and IRF1 are all enriched for causal variants driving increased enhancer activity levels across hundreds of enhancers⁸. Importantly, affinity optimizations of ETS enriched for genetic variants that are associated with increased gene expression (cis-expression quantitative trait loci, cis-eQTL). The effect of affinity optimizations in ETS cis-eQTLs correlated well with the effect in reporter assays, suggesting MPRA can recapitulate effect of gene expression at the endogenous locus. Although massively parallel reporter assays (MPRA) stands the ideal method to test millions of variants for their effect on enhancer activity levels, they cannot yet determine the effects on tissue specificity. However, there is evidence that suboptimality of enhancers is associated with tissue-specificity, suggesting that higher levels of expression in one tissue may correlate with loss of tissue-specificity^{5;6}.

Although we focus on the importance of affinity optimizing genetic variation in causing disease, we also hypothesize this is a mechanism through which traits evolve. Enhancers are known to contain genetic variation driving evolutionary adaptations¹³³. In relation to vertebrate limb digit development, we speculate that these types of mutations can drive changes in the limb digits which diversify the kinds of interactions the hand can make with the environment (e.g. grabbing objects, manipulating tools). In relation to the chordate heart, we speculate these kinds of mutations could contribute to the evolution of extra chambers in the heart.

Enhancers are the most common genomic location of disease associated mutations¹⁻³. Pinpointing disease-causing genetic variation in enhancers is a current major challenge. Across this work we show that in Ciona, mice, and humans, affinity optimizing genetic variants represent pathogenic gain-of-function (GOF) mutations. We do not expect loss-of-function genetic variants (decreases in affinity or deletion of the site) to be pathogenic as often because of enhancer redundancy. Within enhancers, there are multiple transcription factor binding sites which provide robust activation^{97;158}. Indeed, in our work we find that deleting the ETS-A results in no phenotype in the ZRS, likely due because there are 19 ETS binding cores in the ZRS, six of which are functionally validated. The large number of ETS cores in the ZRS not mentioned in the literature highlights that we may be vastly underappreciating the number of sites within enhancers. In addition to redundancy built within enhancers, genes are typically regulated by also multiple enhancers^{156;157}. We know that other enhancers can compensate in cases in which a single enhancer is lost. Therefore, when pinpointing causal variants, we propose focusing on GOF variants will be most successful as these are likely to drive loss of tissue-specificity and downstream phenotypes.

Importantly, this perspective of GOF vs LOF came from mechanistic studies of the rules governing enhancer function. This highlights the importance of approaches that study enhancers in a mechanistic and interpretable way. Although machine learning approaches have been successful in making genomic enhancer predictions and generating synthetic enhancers, it is difficult to understand how to improve these enhancer models if we do not understand the underlying mechanisms that govern enhancer sequence to function relationships.

Machine learning models rely on a numerical encoding of enhancers, almost exclusively

using one-hot encoding to represent the precise sequence of the enhancer. This approach results in a 4xN matrix where each of the 4 rows encodes presence of A/T/G or C and N is the length of the enhancer^{11;49;61;175;176}. This is where models had to begin because we still do not understand the enhancer features encoding gene expression, so we must rely on encoding the precise sequence and hoping the model will learn the features. There is no doubt these models have advanced the field of enhancer predictions. However, these sequence based models may be limited in their ability to make predictions based on the linear encoding of precise sequence. Although some organizational preferences have been detected using machine learning models as an “oracle” to predict how sequence changes enhancer function^{48;49;177}, we do not yet know if these models can pick up on nuances of site number, affinity and syntax given the limitation of how many training sequences can be inputted. Indeed, low affinity sites are degenerate and there are multiple functional syntaxes which can put together the same tissue-specific expression pattern^{9;178}. We hypothesize that the more we understand about the functional features governing enhancer levels and tissue specificity, the more precisely we can encode these features for the machine learning models to learn and predict on. Given the importance of enhancer grammar which is encoded by the site type, number, affinity, and syntax, we hypothesize that these features will be critical to encode.

4.1.2 Towards increasing prediction accuracy of causal mutations using grammar

Beyond the affinity of binding sites, the organization of sites is another component of functional enhancers. The physical properties of TFs and the DNA to which they bind are thought to constrain the way in which binding sites within a functional enhancer can be organized. The interplay of binding site syntax (the order, orientation and spacing of sites) and the affinity of sites is known as enhancer grammar⁹. The features encoding grammar are binding site type, number, affinity and syntax (organization). Previous studies of grammar have only found the “smoking gun” of grammar, or traces of grammatical features which are enriched in predicted functional regions^{48–50;52;177}. We sought to study enhancer grammar with a more explicit and interpretable approach to provide a foundation for future studies to investigate grammar and use grammatical rules to make genomic enhancer predictions.

We study enhancer grammar using the Ciona neural enhancer Otx-a as a model, which is known to be regulated only by GATA and ETS^{9;91;171;173}. We generated a synthetic library using the Otx-a well defined sequence components: the functional GATA and ETS sites and the inert linker sequences. Therefore, these clusters have the same sequence content, binding site type, number and affinity. The difference between these sites is the relative syntax of the sites (their relative order, orientation and spacing). We find syntax is critical for a cluster of sites being an active enhancer. Thousands of sequences are rendered inert from insufficient syntax, while thousands of sequences are active enhancers due to their syntax. We found that 2-site and 3-site syntax features were significantly enriched in active enhancers. We developed a method to use predict functional enhancers using these activator syntaxes. Using this method, we found syntax could increase the accuracy of enhancer predictions when integrated with affinity. This is the first time syntax has been explicitly used successfully to improve genomic enhancer predictions.

We analyzed syntax in a rigorous manor by comparing how presence of syntax contributes to differential enhancer activity for GATA and ETS genomic elements with similar affinities. This results in a heat map, where you physically separate elements by their affinity, and then quantify how the presence of syntax relates to enhancer activity. Future efforts will focus on collapsing all elements to a single axis and predict their enhancer activity by integrating the affinity and syntax of binding sites. The relative contributions of affinity and syntax is unclear, making this single quantification of enhancer grammar difficult. Additionally, it is also unclear how to encode the affinity of a binding site cluster with multiple types and numbers of binding sites (e.g. 2 GATA + 1 ETS versus 1 GATA + 2 ETS). One approach is to do this explicitly, assigning weights to affinity and syntax and assess how well this predicts enhancer activities. Another approach is to encode the type, number, affinity and syntax of an element's binding sites into a machine learning model to see if it can learn how to combine the relative contribution of the features which optimize prediction accuracy.

Given the power of syntax to enrich for functional enhancers, it will be an important next step to test the effect of affinity optimizing genetic variant in clusters predicted to be functional using enhancer grammar (affinity and syntax). We expect genetic variation optimizing affinity in functional binding site clusters to be more likely to be causal than in non-functional clusters.

Additionally, the likelihood of causality of a de novo TFBS could also be scaled based on the syntax of that site with known cofactors. Specifically, we would determine if the de novo binding site emerged in an active syntax with nearby co-factor binding sites relevant to that cell type. In sum, grammar gives us a methodology to further increase the accuracy of predicted causal affinity optimizations and de novo binding sites.

Another finding from our synthetic enhancer screen of Otx-a is that active enhancers with robust syntaxes drove significantly higher levels of enhancer activity than active enhancers with moderate syntaxes. This provides substantial evidence to previous experiments which hinted that syntax is not only important for forming a functional enhancer, but can also contribute to the levels of expression. This represents another feature that could be used to identify genetic variation associated with disease. Specifically, indels which change spacing between sites could optimize combinatorial TF binding similar to the effect of increasing TF affinity. Indeed, indels between sites were observed to dramatically effect binding of transcription factors as measured by ChIP-seq⁵². Although these observations are an important first step, this thesis suggests we may need to study syntax in depth using synthetic screens to be able to make predictions of which indels will cause changes to enhancer function.

This thesis focuses on supplementing current efforts to reveal genotype-phenotype relationships with a mechanistic layer. We formulate our mechanistic understanding of enhancers from decades of research as enhancer regulatory rules which govern sequence to function relationships of enhancers. Genetic variants which break these regulatory rules are predicted to cause changes to enhancer levels and tissue-specificity, and these changes in gene expression are thus likely to drive phenotypic changes. The mechanistic low-affinity regulatory principle alone was able to enrich for causal variants in high-throughput assays. Integrating mechanism with epigenetics and fine-mapping will likely generate the variants most likely to be causal. In the future, we can improve prediction accuracy of the low-affinity regulatory principle by prioritizing variants within functional clusters as defined through enhancer grammar. Additionally, our finding that syntax can increase enhancer activity levels suggest that it is another “optimization knob” through which genetic variation could changes in gene expression levels or tissue-specificity. Adoption of these principles in studying other TFs and other diseases will increase our collective power to

uncover the genetic underpinnings of enhanceropathies.

Appendix A

Supplemental Figures for Chapter 1

All Supplemental Tables can be found in the version 2 bioRxiv preprint at the following location: <https://www.biorxiv.org/content/10.1101/2022.06.10.495510v2.supplementary-material>.

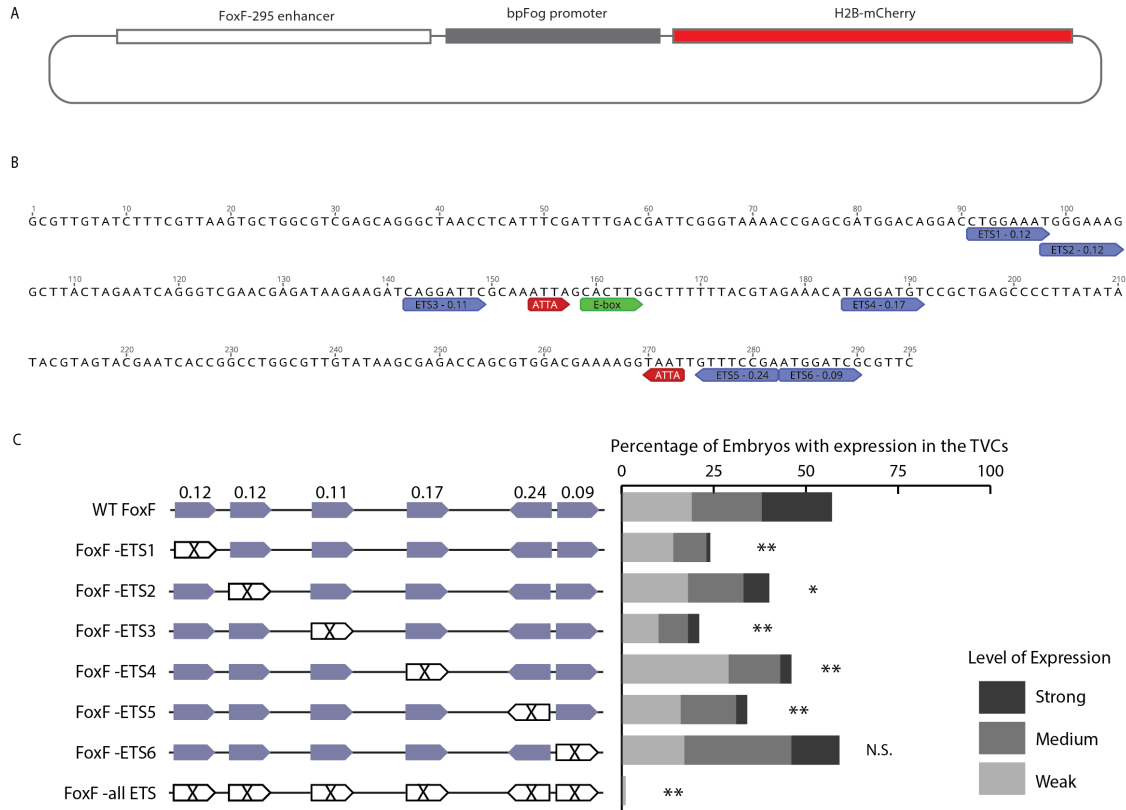


Figure A.1. Counting data for ETS site ablation within the FoxF enhancer. Related to Figure 1. A. Schematic of FoxF enhancer reporter construct. B. Sequence of the 295bp FoxF enhancer with transcription factor binding sites found by previous studies and the current study. Previously studied EtsA, EtsB, and EtsC sites are described as the ETS3, ETS4, and ETS5 sites. C. Schematics show the enhancers that were electroporated into embryos. Diagrams indicate the WT FoxF fragment and X indicates GGAW > GCAW mutation. Graph shows the proportions of embryos and levels of expression in TVCs (3 replicates, $n \geq 50$ embryos for each replicate). P values are obtained by using Chi-square test comparing each condition to the WT FoxF enhancer with Bonferroni correction: ** $P < 0.0007$, * $P < 0.007$, N.S. not significant. Data in stacked bar chart are represented as the mean. Table S1 contains sequences of ETS sites in each construct. Figure S2 contains representative images for each of these constructs.

Figure A.2 Representative images of each of the FoxF constructs tested in this study. The first nine images are alternate representative images to those included in the main figures. The final six images are representative images for the single-site ablation constructs in Figure S1. The counting data for FoxF -ETS1 and FoxF -ETS3 shows fluorescence in less than 25% of embryos, therefore we included embryos with no fluorescence as representative images. Scale bars, 50 μ m.

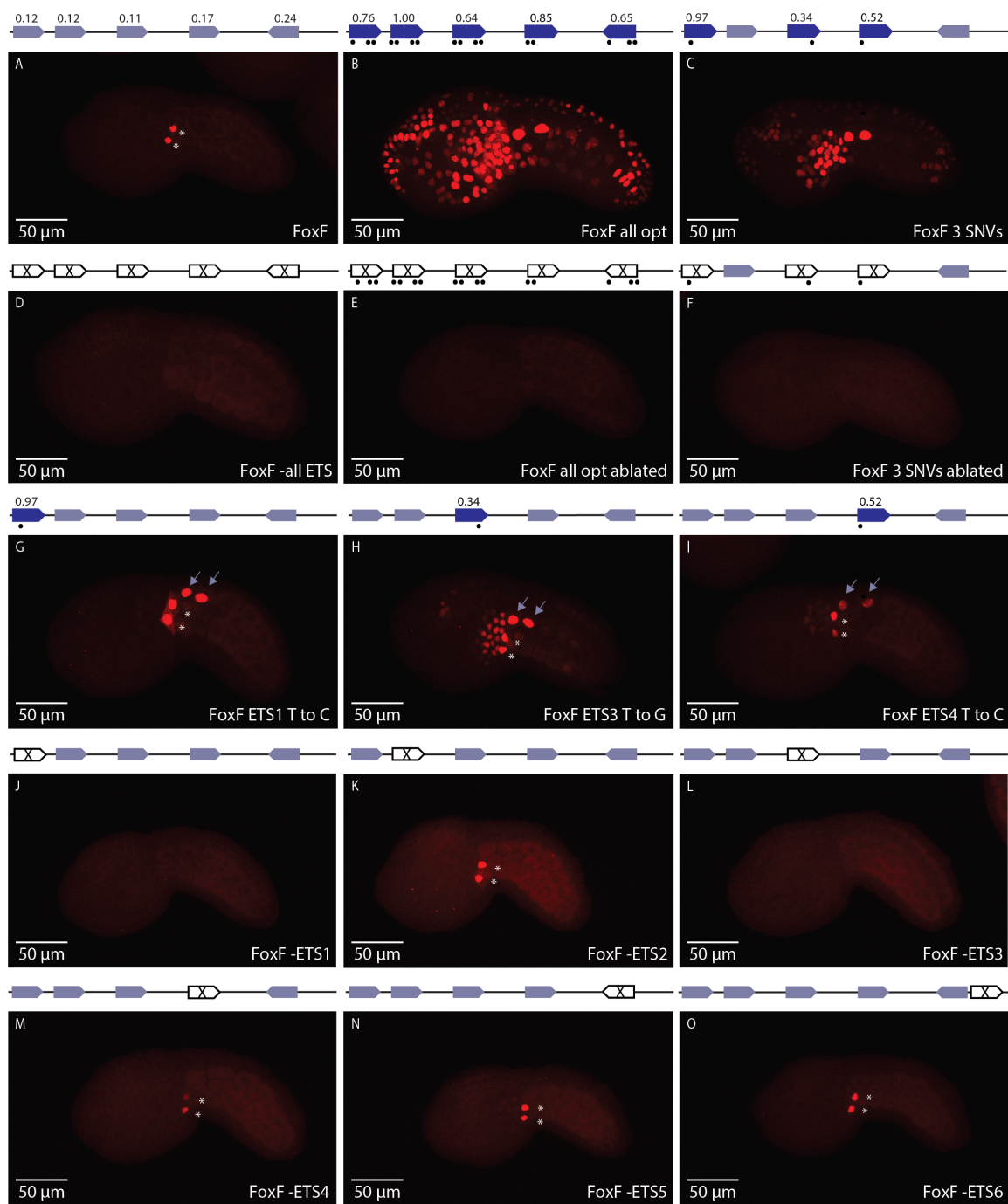


Figure A.2. Representative images of each of the FoxF constructs tested in this study.

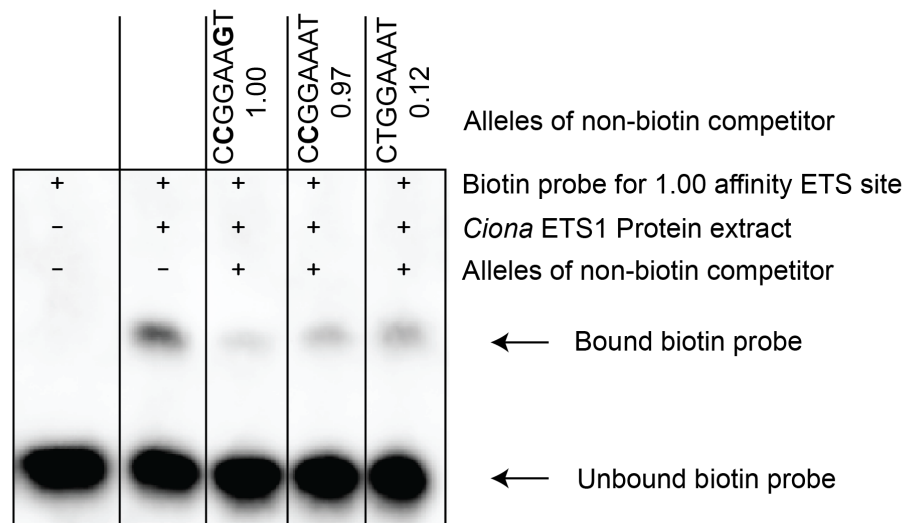


Figure A.3. Differential binding affinity with ETS1 single nucleotide variant. Related to Figure 2. We performed an EMSA to detect binding of *Ciona* ETS1 DNA Binding Domain (DBD) to a biotin labelled 30bp sequence of the FoxF enhancer centered on ETS1 with a 1.0 affinity site. We compared the impact of unbound competitor alleles containing the ETS site of 1.0 affinity, the 0.97 ETS1-T-to-C site and the WT ETS1 site. *Ciona* ETS1 DBD bound to the 1.0 ETS binding site in the control probe (lane 2) and this specific binding was competed most effectively by the unlabeled competitor with the 1.0 affinity site (lane 3). The unlabeled competitor probe harboring the single nucleotide variant ETS1-T-to-C competed almost as efficiently as the probe harboring a 1.0 affinity ETS site (lane 4), while probe harboring the wild-type ETS1 site was the least efficient at competing binding to the labelled probe (lane 5).

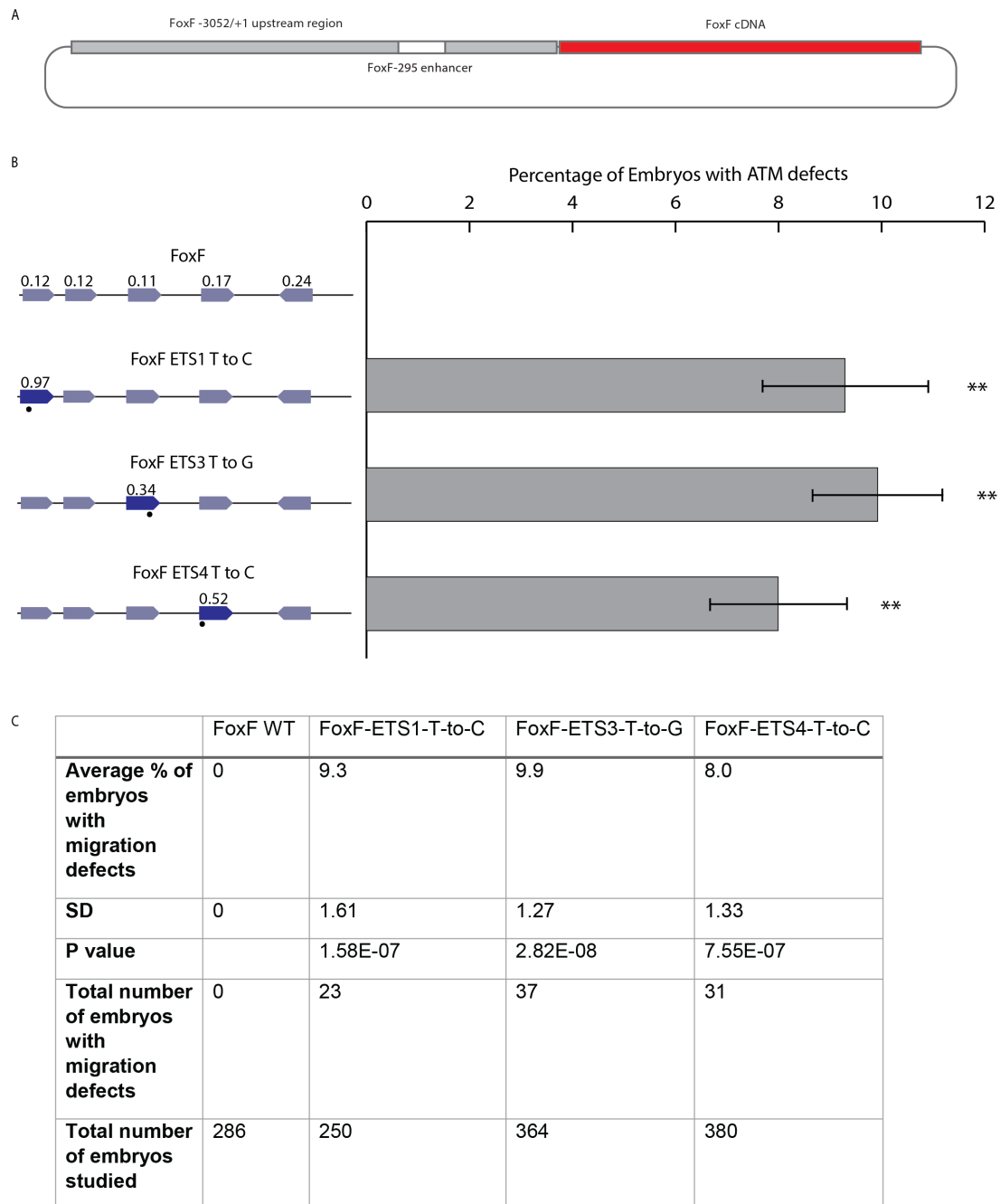


Figure A.4. Counting of migration defects in tailbud embryos with affinity-optimizing mutations. Related to Figure 3. A. Schematic of FoxF enhancer FoxF cDNA construct. B. Bar chart of percentage of embryos with ATM migration defects. We scored the migration of ATMs into the ventral midline in between 250 and 380 embryos per condition. P-values are obtained by using Chisquare test comparing to WT FoxF enhancer with Bonferroni correction: ** $P < 0.0017$. The single nucleotide variants made are shown by the dots in the enhancer schematic. Data are represented as mean $\pm$ standard deviation. C. Table with number of embryos.

Appendix B

Supplemental Figures for Chapter 2

All Supplemental Tables can be found in the version 2 bioRxiv preprint at the following location: <https://www.biorxiv.org/content/10.1101/2022.06.10.495510v2.supplementary-material>.

		Helix 1		Helix 2		Helix 3						Strand 4				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Human	ETS-1	LWQF		WGKRK		YEKLSRGLRYYY						RYVY				
Mouse	ETS-1	LWQF		WGKRK		YEKLSRGLRYYY						RYVY				

Figure B.1. Class I ETS family members have conserved DBDs. DBDs of ETS-1 transcription factors. Numbers show the location of contacts between the DNA-binding site and the protein in human and mouse^{24,61,62}. The DBDs are highly conserved across human, mouse and flies across other class I ETS family members^{24,63,64}.

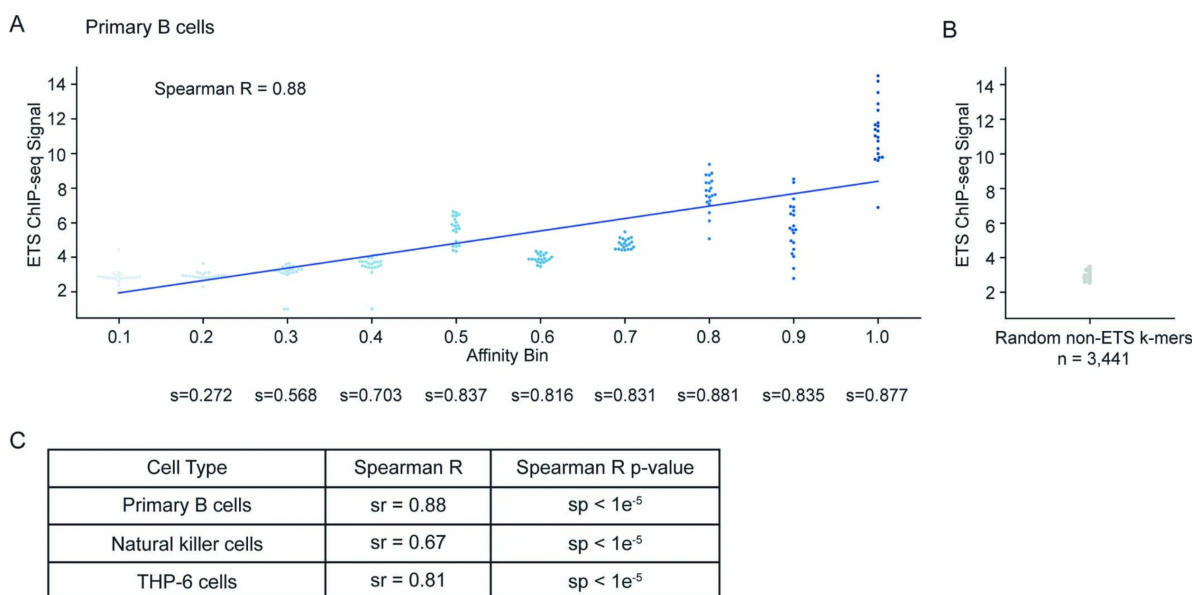


Figure B.2. PBM binding affinities correlate with the in vivo ETS-1 ChIP signal in various cell types. a, ETS-1 ChIP-seq signals from mouse primary B cells (y axis) and PBM mouse ETS-1 affinities (x axis) show a strong Spearman's rank correlation of 0.88 (ref. 58). To show that the correlation is not heavily dependent on values from affinity bins 0.1 and 1.0, we show the Spearman's rank correlation between cumulative affinity bins below the graph. For example, the Spearman's correlation using bins 0.1, 0.2, 0.3, 0.4 is 0.703. b, The ETS-1 ChIP signal is low for all random non-ETS k-mers c, Correlation of ChIP signal and PBM affinity for other ETS-1 ChIP assays in mouse primary B cells, human natural killer cells and human THP-6 cell lines^{58,59,60}.

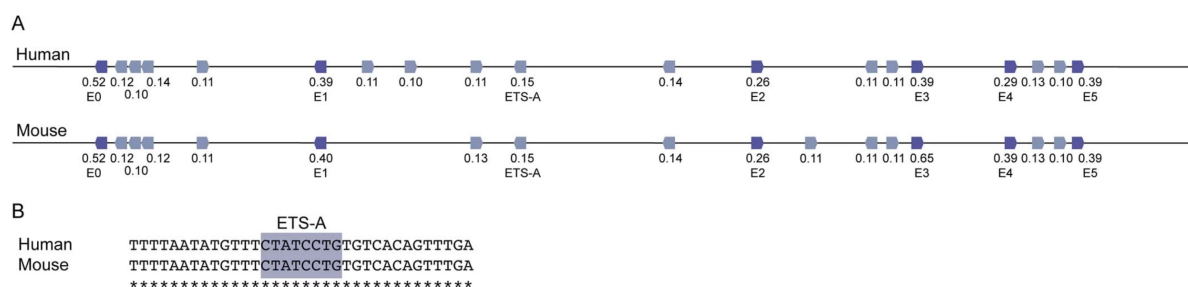


Figure B.3. Conservation of ZRS ETS sites between humans and mice. a, The human and mouse ZRS sequence is highly conserved. The human ZRS has 19 putative ETS-binding sites, all with affinities equal to or lower than 0.52. The mouse ZRS has 6 functionally validated and 12 putative ETS sites; 15 of these are conserved in location and affinity with the human ZRS. b, The ETS-A site and surrounding sequence show perfect conservation between mouse and human, as indicated by asterisks. Blue box highlights the ETS-A sequence within the human and mouse ZRS.

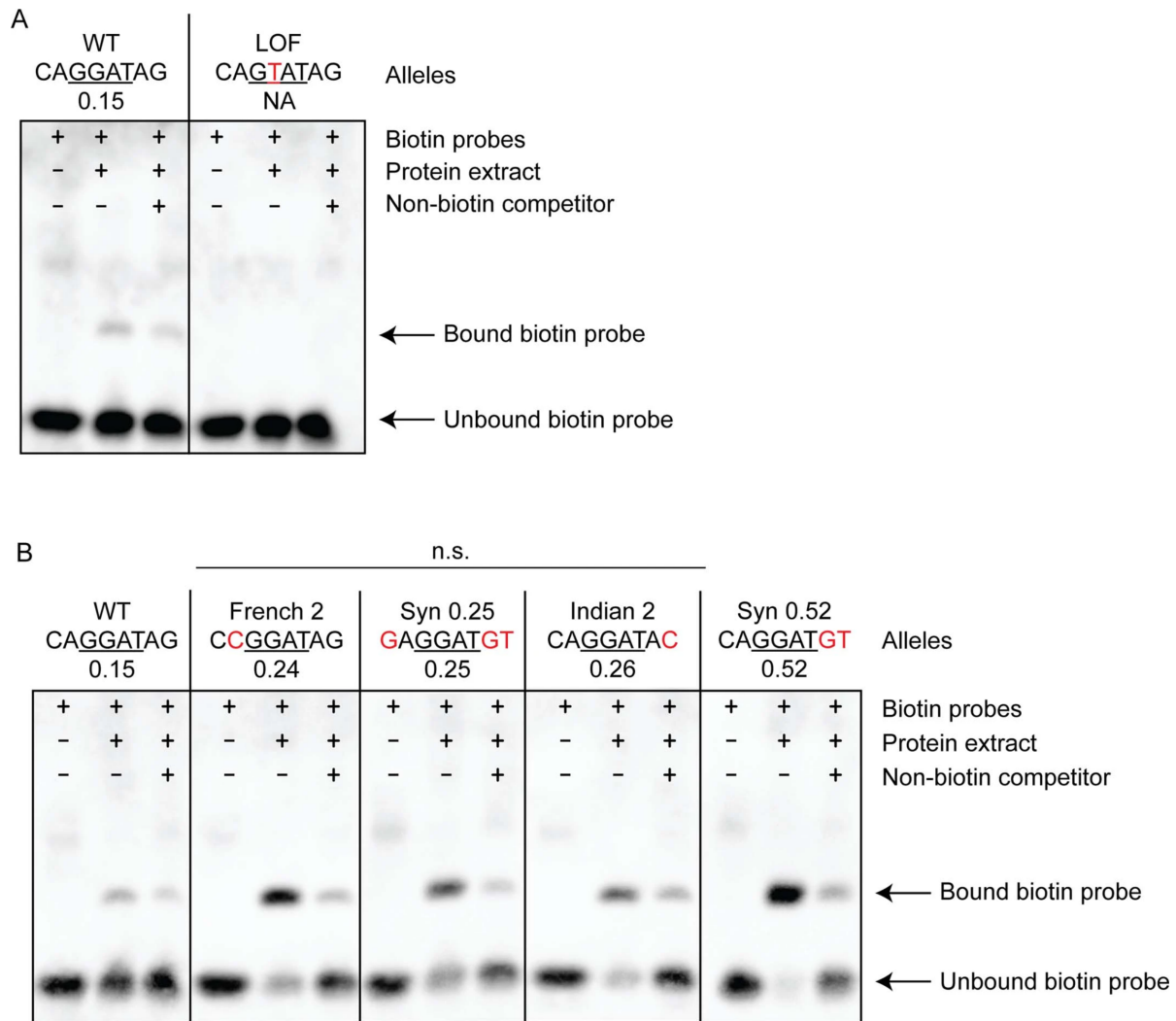


Figure B.4. EMSA shows the binding of human and mouse ETS-1 to the ETS-A site and ETS-A variants. a, WT ETS-A probe sequence can bind to the ETS-1 DBD to generate a band, the intensity of which decreases as the biotin-labelled probe is outcompeted by a non-biotin competitor probe. A single-bp change in the ETS-A site (LOF) leads to a loss of binding signal. b, EMSAs for WT, French 2, Syn 0.25, Indian 2 and Syn 0.52 sequences. The total amount of bound probe relative to the unbound probe is not statistically different between French 2, Indian 2 and Syn 0.25 sequences, suggesting that all three sequences have the same affinity. $P = 0.18$, one-way ANOVA. The Syn 0.52 sequence binds more strongly than the 0.25 or 0.15 sequences do. EMSAs were performed independently twice and both replicates show similar results. For gel source data, see Supplementary Fig. 1.

Figure B.5 Ptch1 in situ hybridization in the hindlimb bud and forelimb bud of transgenic mice. Embryos were collected at around the 53-somite stage for Ptch1 in situ hybridization. a, Ptch1 expression is restricted to the posterior domain in WT hindlimb buds. Ectopic Ptch1 expression can be seen in the anterior domain of homozygous French 2 (b), Indian 2 (c), Syn 0.25 (d) hindlimb buds. LOF homozygotes (e) and embryos with Syn0.25/LOF alleles (f) do not have ectopic Ptch1 expression as predicted. g, Syn 0.52 hindlimb buds have larger domain of ectopic Ptch1 expression than embryos with approximately 0.25 affinity ETS-A sites. h, In WT forelimb buds, Shh and Ptch1 expression is restricted to the posterior domain. i, Syn 0.52 homozygotes display ectopic Shh expression and Ptch1 expression in the forelimb buds. The Shh in situ hybridizations shown were performed on embryos at around the 48-somite stage. All limb bud images were acquired and cropped using the same settings.

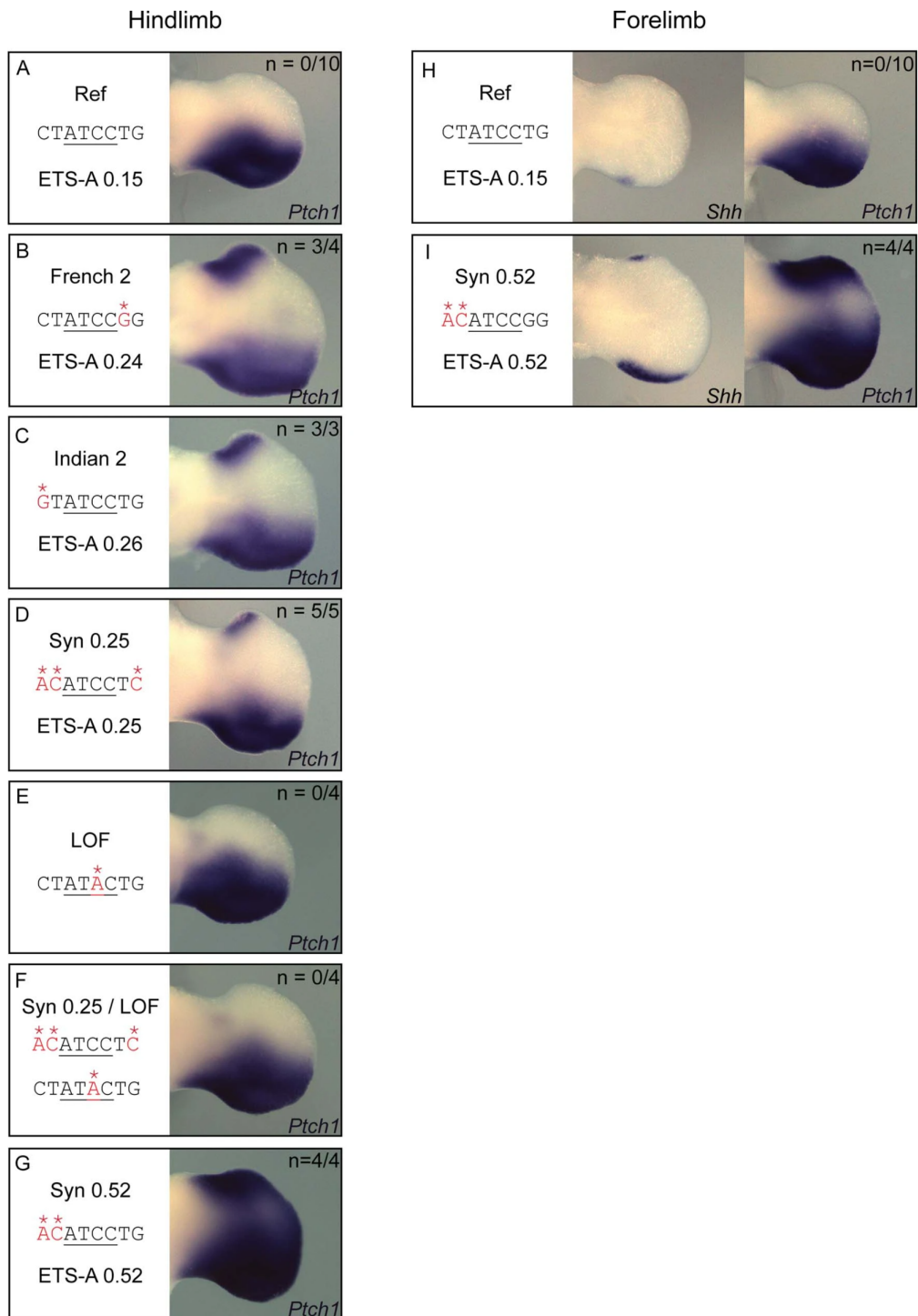


Figure B.5. *Ptch1* in situ hybridization in the hindlimb bud and forelimb bud of transgenic mice.

Figure B.6 Ptch1 in situ hybridization in the hindlimb bud and forelimb bud of transgenic mice. a, In the hindlimb, heterozygotes and homozygotes have 100% penetrance in polydactyly. In the forelimb, homozygotes have 100% penetrance whereas heterozygotes have 93.6% penetrance. b, Tibial hemimelia is observed in 95.3% of homozygotes but no heterozygotes. c, Digit phenotypes on the hindlimbs of all mice studied. d, Digit phenotypes on the forelimbs of Syn 0.52 mice. WT, LOF, French 2, Indian 2 and Syn 0.25 have no forelimb phenotypes. TT denotes triphalangeal toe or thumb.

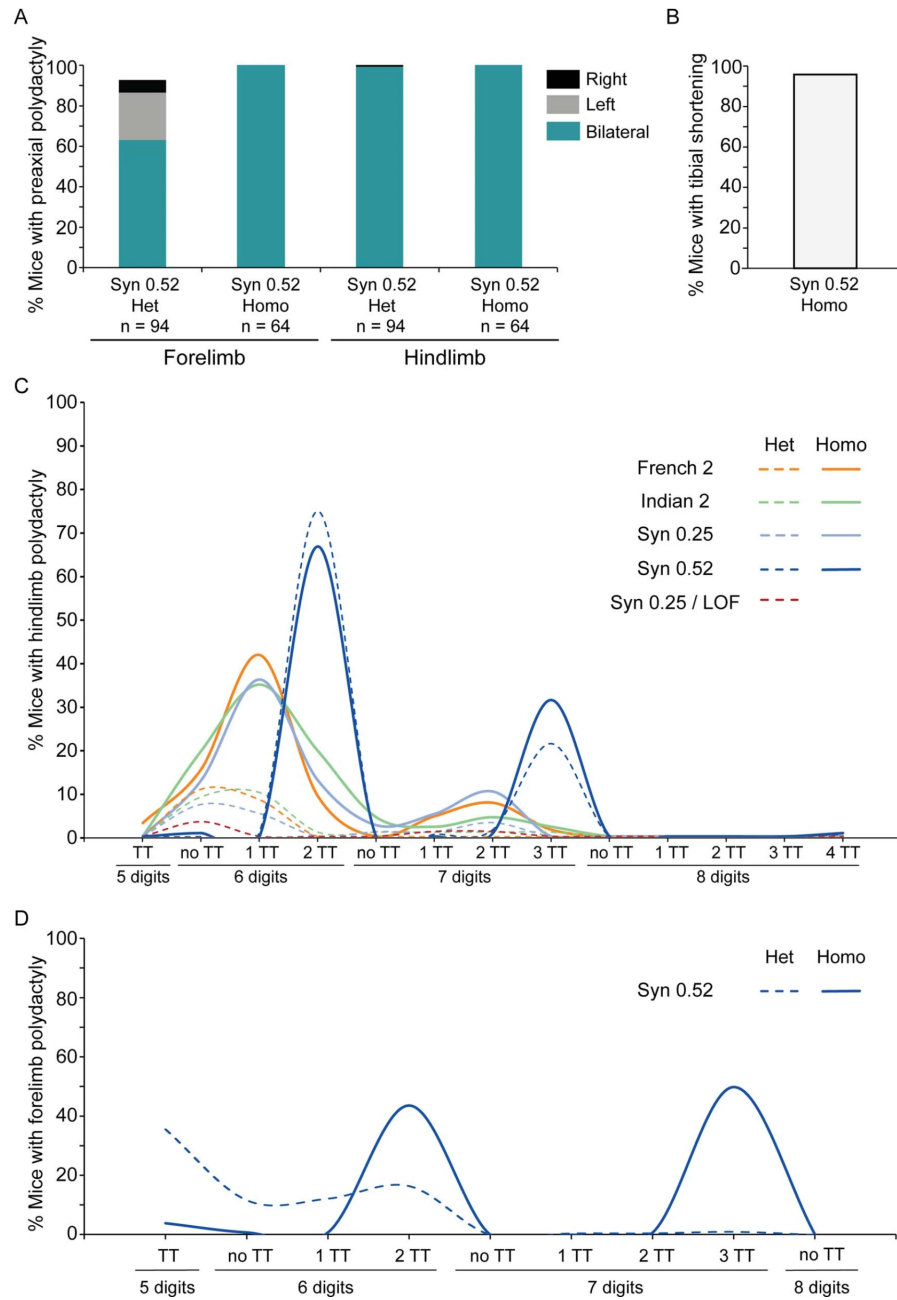


Figure B.6. Syn 0.52 mice show highly penetrant polydactyly in both forelimb and hindlimb, as well as tibial hemimelia.

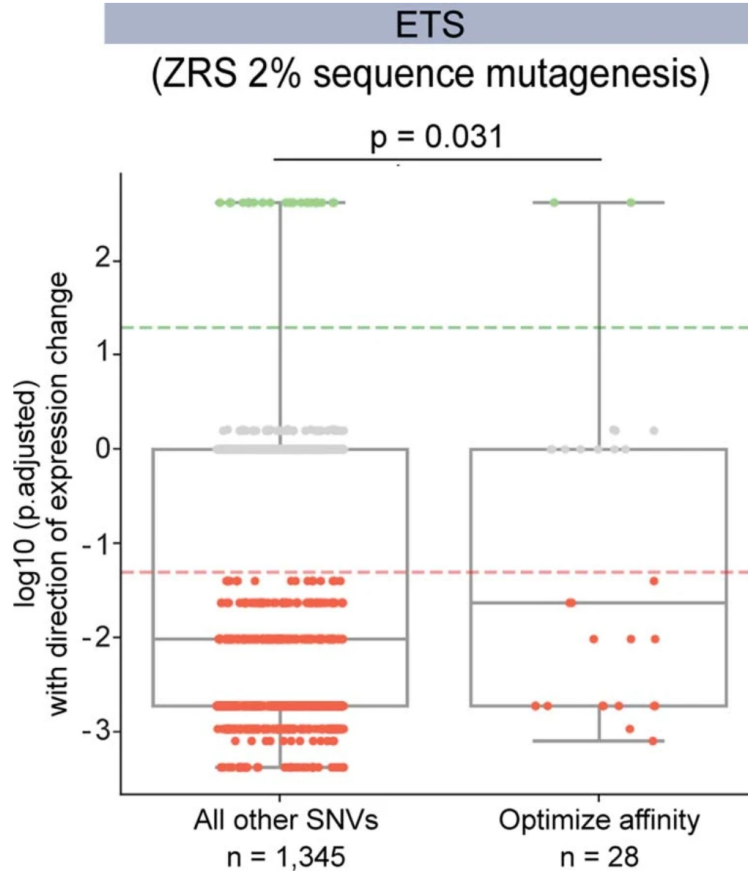


Figure B.7. Affinity-optimizing SNVs are significantly associated with GOF enhancer activity. Enhancers containing SNVs that increase ETS affinity are significantly enriched in GOF expression relative to all other SNVs within a 2% sequence mutagenesis assay of the ZRS enhancer⁵. The bounds of the box plots define the 25th, 50th and 75th percentiles, and whiskers are $1.5 \times$ the interquartile range. One-tailed Mann–Whitney U test.

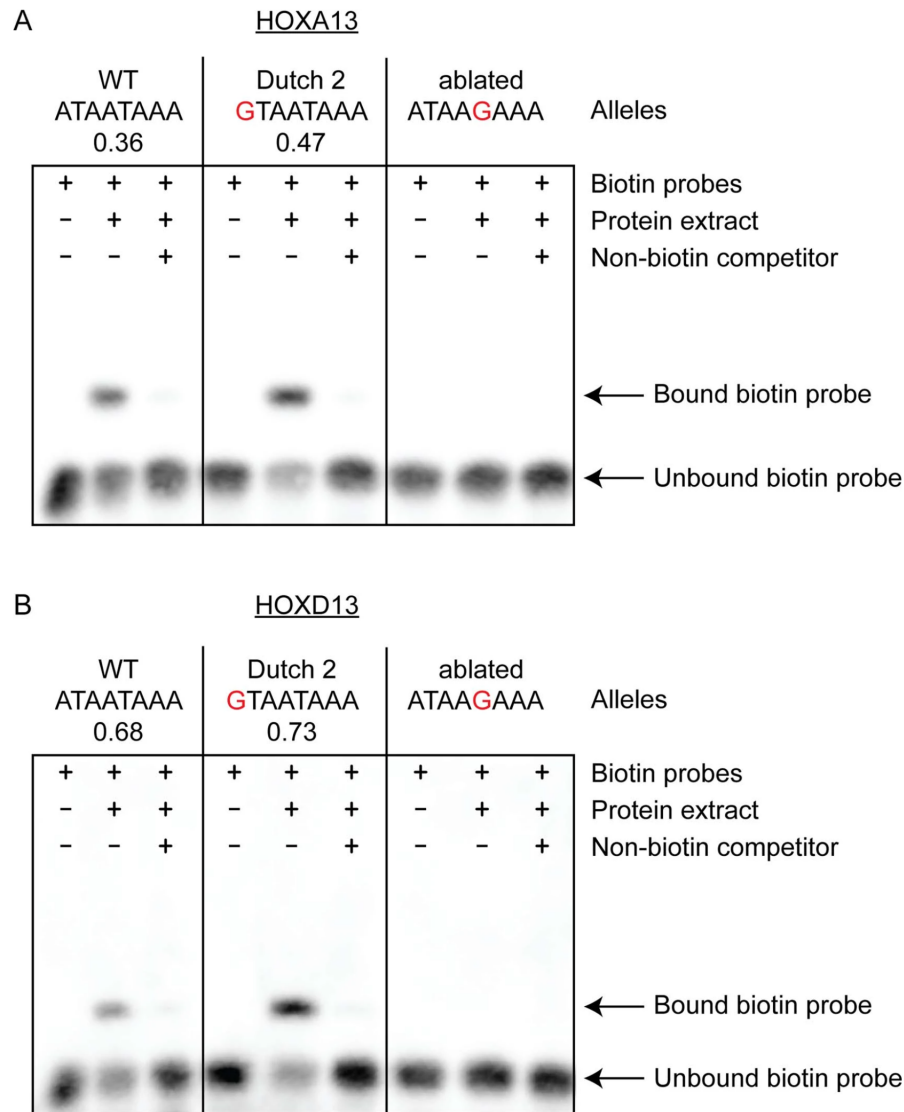


Figure B.8. EMSA shows stronger binding of the human HOXA13 and HOXD13 DBDs to the Dutch 2 variant relative to the WT sequence. a, WT probe sequence can bind to the human HOXA13 DBD to generate a band, the intensity of which decreases in the presence of the non-biotin competitor probe. The Dutch 2 variant binds to HOXA13 more strongly than does the WT allele. A single-bp change that ablates HOXA13 binding shows no binding. b, EMSA with the same WT and Dutch 2 probe sequences as in a, performed with the human HOXD13 DBD. The Dutch 2 variant binds to HOXD13 more strongly than does the WT allele. EMSAs were performed independently twice and both replicates show similar results. For gel source data, see Supplementary Fig. 1.

Enhancer	Disease associated	Cell line	
BCL11A	Sickle cell disease	HEL92.1.7	Erythroblast cell line
IRF4	Human pigmentation	SK-MEL-28	Melanoma cell line
IRF6	Cleft lip	HaCat	Keratinocyte cell line
MYC (rs6983267)	Various types of cancer (SNP is associated with colorectal cancer development)	HEK293T + 20mM LiCl	Derived embryonic kidney cell line (activated Wnt pathway)
MYC (rs11986220)	Various types of cancer (SNP is associated with prostate cancer)	LNCaP + 100nM DHT	Prostate adenocarcinoma cell line (stimulated androgen activity)
RET	Hirschsprung's disease	Neuro-2a	Mouse neuroblastoma cell line
SORT1	Plasma low-density lipoprotein cholesterol & myocardial infarction	HepG2	Hepatocellular carcinoma cell line
TCF7L2	Type 2 diabetes	MIN6	Mouse pancreatic beta cell line
UC88	(ultraconserved enhancer with strong forebrain activity)	Neuro-2a	Mouse neuroblastoma cell line
ZFAND3	Type 2 diabetes	MIN6	Mouse pancreatic beta cell line
ZRS	Polydactyly and other limb malformations	NIH/3T3 (+ HOXD13 + HAND2)	Mouse fibroblast cells transfected with limb bud TFs to give a limb-like cell identity

Figure B.9. Experimental details for MPRA performed with 11 disease-associated enhancers. Eleven disease-associated enhancers tested in saturation mutagenesis MPRA; table modified from a previous study³⁰. Two different MYC enhancers were assayed. UC88 is an ultraconserved enhancer. The MPRA for each enhancer is performed within cell lines relevant to the disease studied, as detailed.

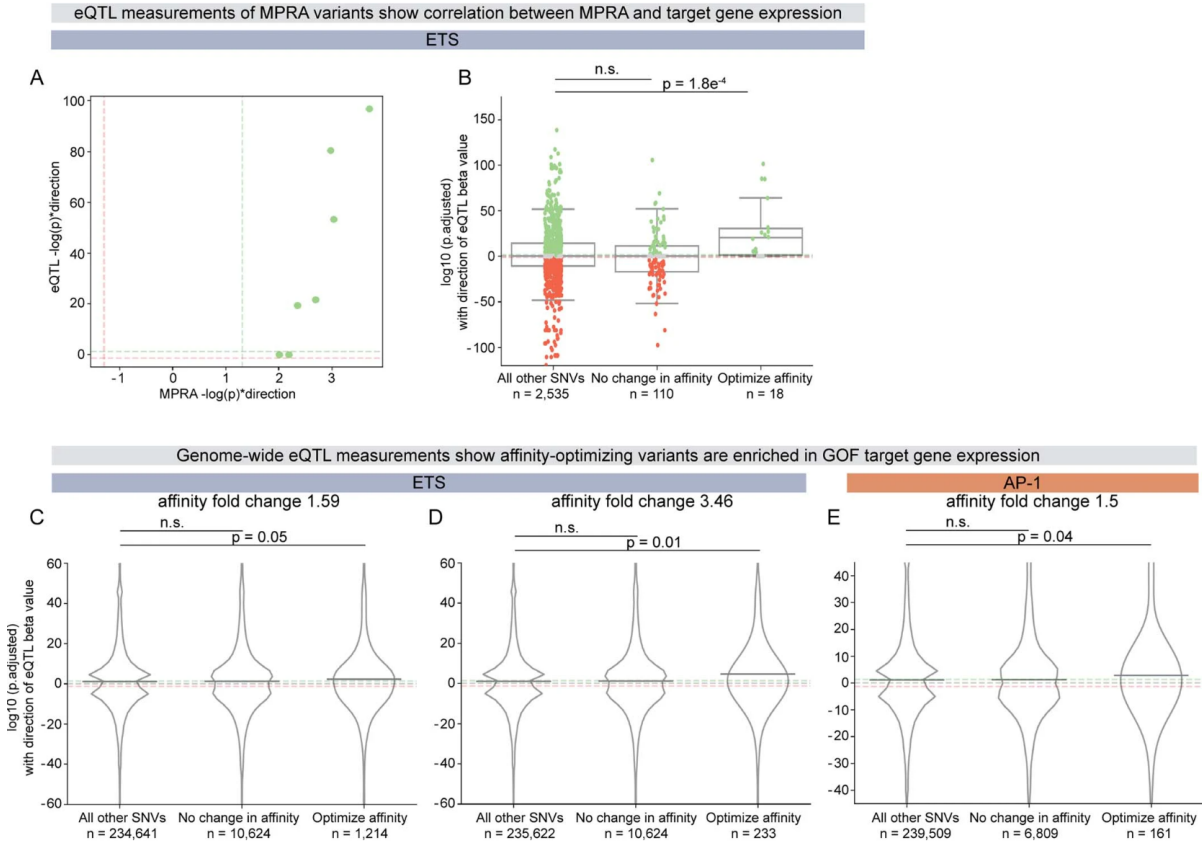


Figure B.10. Affinity-optimizing eQTL variants are enriched in GOF target gene expression. a, Seven ETS affinity-optimizing variants within lymphoblastoid regulatory element MPRA drive significant GOF reporter expression⁴⁰. Five of these are associated with significant eQTL differential expression and all of these SNVs drive increased target gene expression⁴¹. Dotted lines indicate thresholds for significance at $P < 0.05$ (significant GOF in green, and significant LOF in red). b, eQTL analysis for all MPRA variants tested within the lymphoblastoid regulatory elements (regardless of variant effect within MPRA assay)^{40,41}. ETS affinity-optimizing SNVs are enriched in GOF target gene expression. The bounds of the box plots define the 25th, 50th and 75th percentiles, and whiskers are $1.5 \times$ the interquartile range. One-tailed Mann–Whitney U test. c–e, Genome-wide analyses for all significant eQTLs ($P < 0.01$) within the lymphoblastoid cell line⁴¹ show that affinity-optimizing variants are enriched for GOF expression. Line indicates mean values. One-tailed Mann–Whitney U Test. c, ETS SNVs with ≥ 1.59 affinity fold change. d, ETS SNVs with ≥ 3.46 affinity fold change; this is the fold change of 0.52 mice relative to WT 0.15 ETS-A. e, AP-1 SNVs ≥ 1.5 affinity fold change.

Appendix C

Supplemental Figures for Chapter 3

All Supplemental Tables can be found in the version 2 bioRxiv preprint at the following location: <https://www.biorxiv.org/content/10.1101/2022.06.10.495510v2.supplementary-material>.

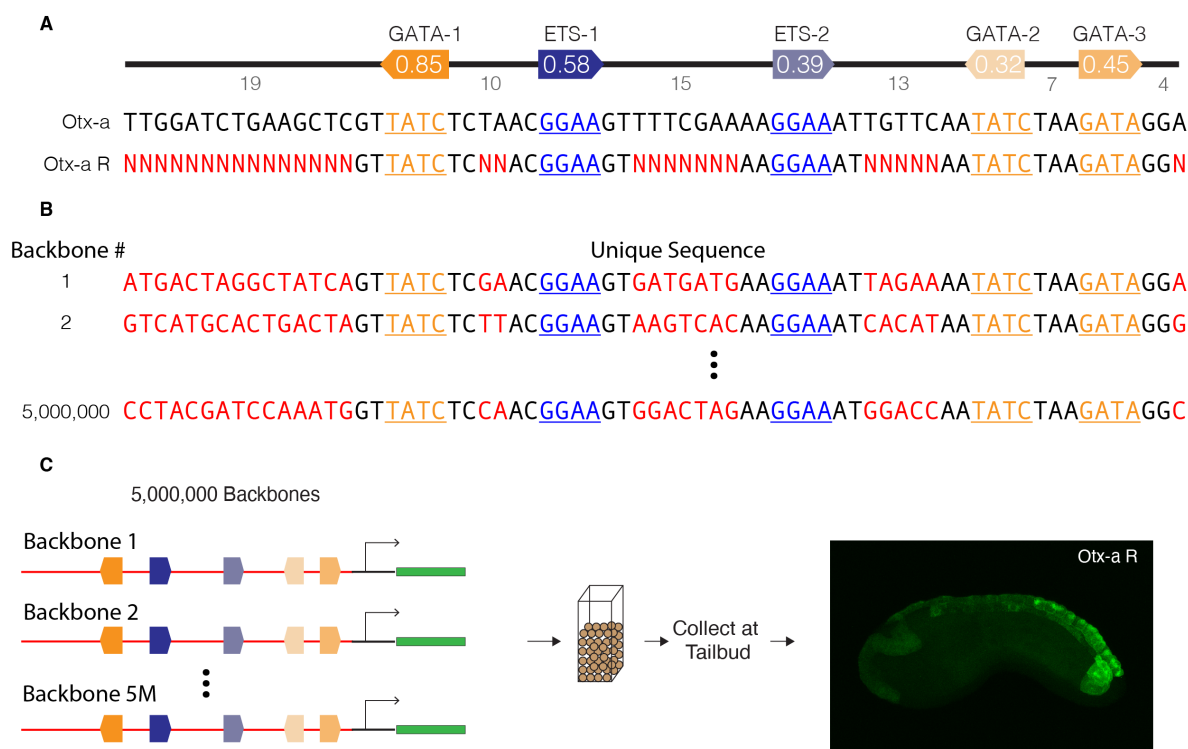


Figure C.1. GATA and ETS are sufficient for the neural expression pattern of Otx-a. (A) Randomization strategy ablates all linker sequences. The features shared across all Otx-a R unique sequences are the order, orientation, spacing, affinity and 8-mer sequence of the GATA and ETS binding sites. (B) Examples of unique sequences in the Otx-a R library. 5 million total sequences are present in the library. (C) Otx-a R library is electroporated into embryos and the GFP expression pattern analyzed at the tailbud stage.

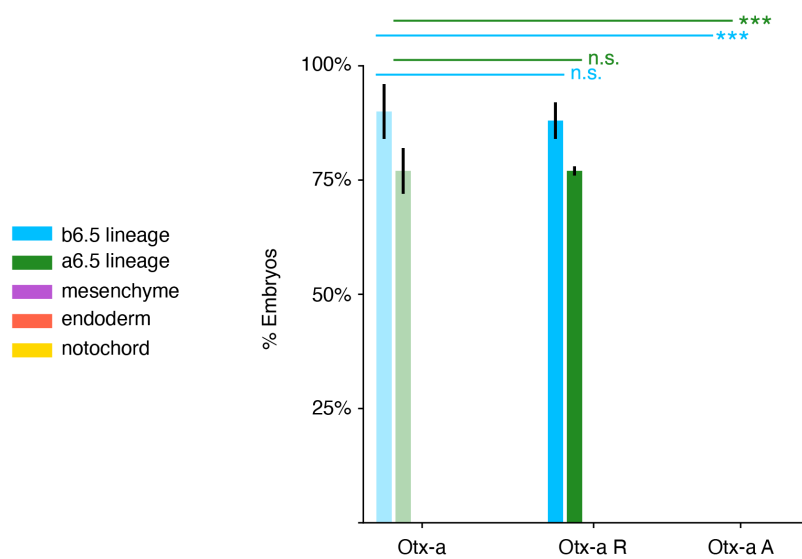


Figure C.2. The GATA and ETS binding sites are necessary and sufficient to drive the Otx-a neural expression pattern. No significant difference between Otx-a and Otx-a R in the percent of embryos with expression in neural tissues derived from a6.5 and b6.5 lineage cells (two replicates of 50 embryos for each construct; Fisher's Exact Test). Complete loss of expression observed in neural tissues in Otx-a A. (n=50 each for two replicates). Error bars indicate the standard error of the mean. No non-neural tissue expression observed in any construct.

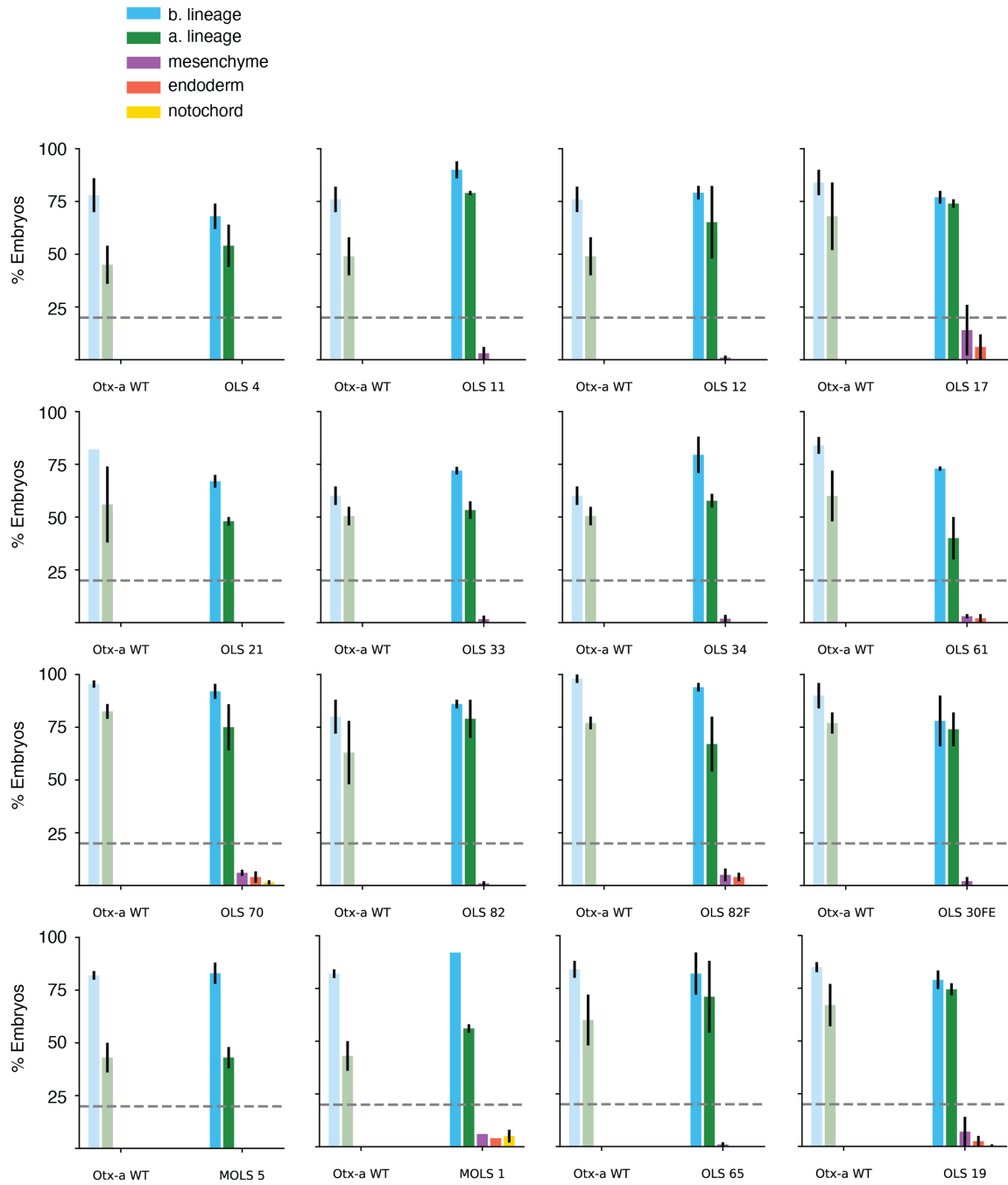


Figure C.3. GATA ETS elements validated to be neural enhancers by fluorescent microscopy. These elements activate Otx-a WT-like expression in a6.5 and b6.5 lineage neural tissues and expression in ectopic tissues (mesenchyme, endoderm or notochord) of $\leq 10\%$. n=50 each for two replicates.

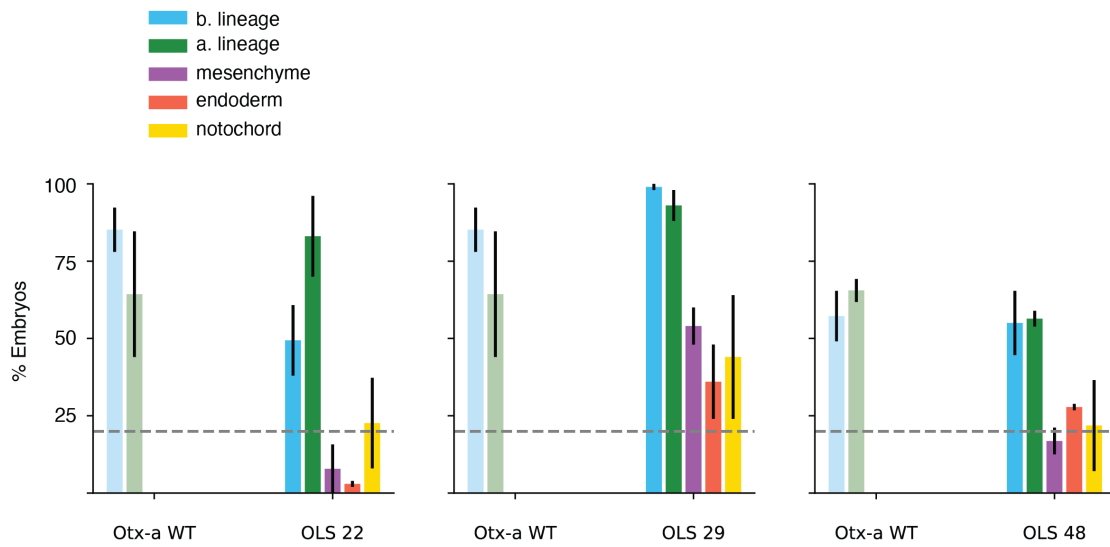


Figure C.4. GATA ETS elements validated to activate neural and ectopic expression by fluorescent microscopy. These elements activate neural expression in both a6.5 and b6.5 lineage cells at lg20% and expression in at least one ectopic tissue (mesenchyme, endoderm or notochord) at $\geq 20\%$ at the tailbud stage. n=50 each for two replicates.

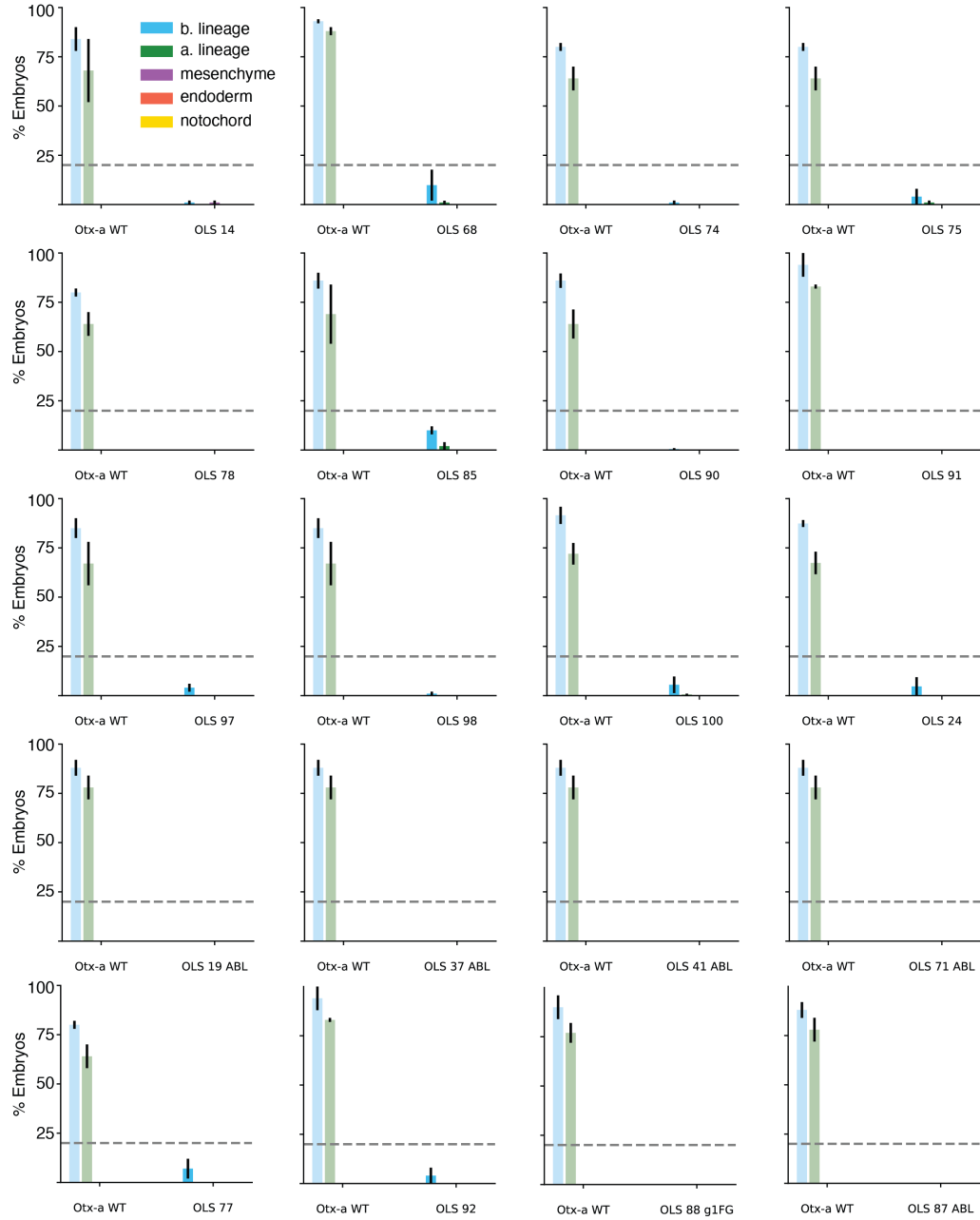


Figure C.5. GATA ETS elements validated to be inert sequences by fluorescent microscopy. These variants have $\leq 20\%$ expression in both a6.5 and b6.5 neural lineages at the tailbud stage. n=50 each for two replicates.

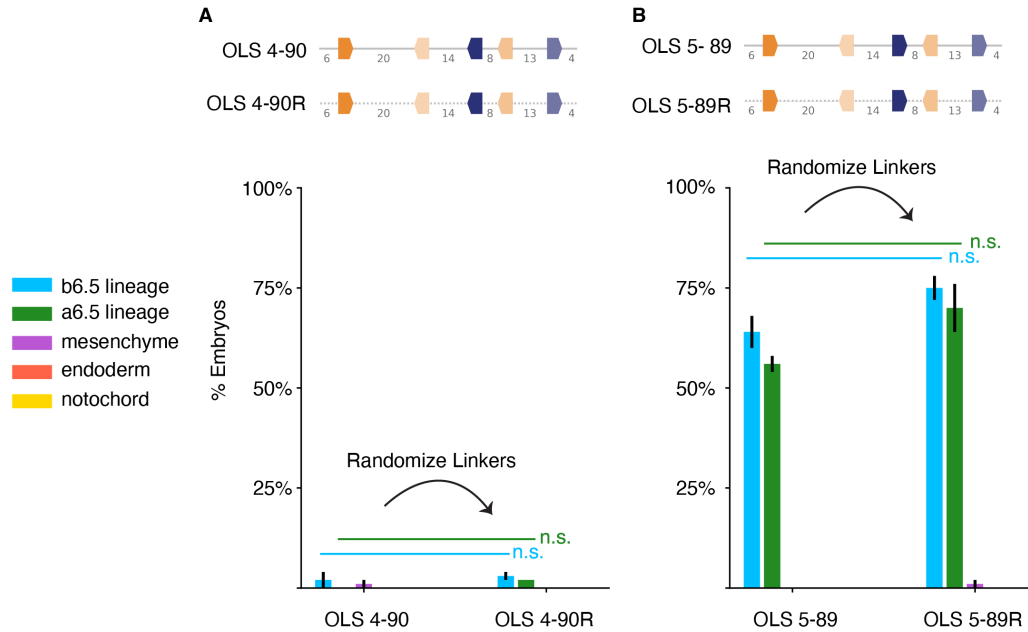
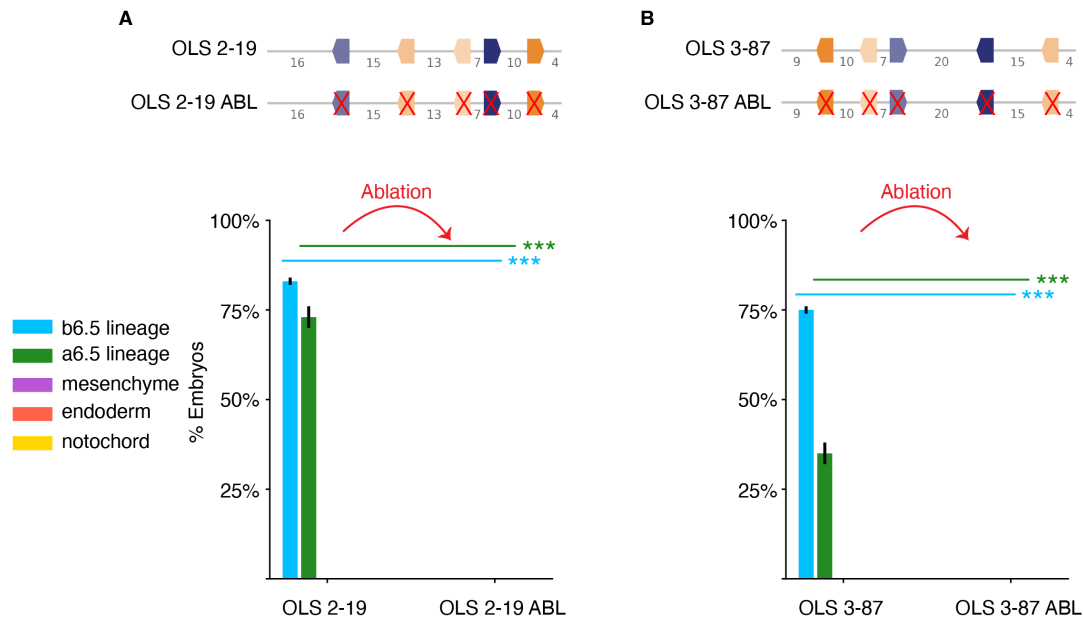


Figure C.6. The GATA and ETS binding sites are the drivers of neural enhancer activity in the Otx-a grammar library members. (A) Randomizing the linker sequences of an inert sequence OLS 90. Dashed grey bars in the linkers of the enhancer schematic indicate randomized N nucleotides in the linker sequences, with an equal chance of A,T,G or C. 5 million unique sequences in the OLS90R library. Randomization of the linker sequences in OLS90 causes no significant change in percent embryos with expression (Fisher's exact test). (B) Randomizing the linker sequences of a neural enhancer OLS 89. 5 million unique sequences in the OLS89R library. Randomization of the linker sequences in OLS89 causes no significant change in percent embryos with expression (Fisher's exact test). n=50 each for two replicates.



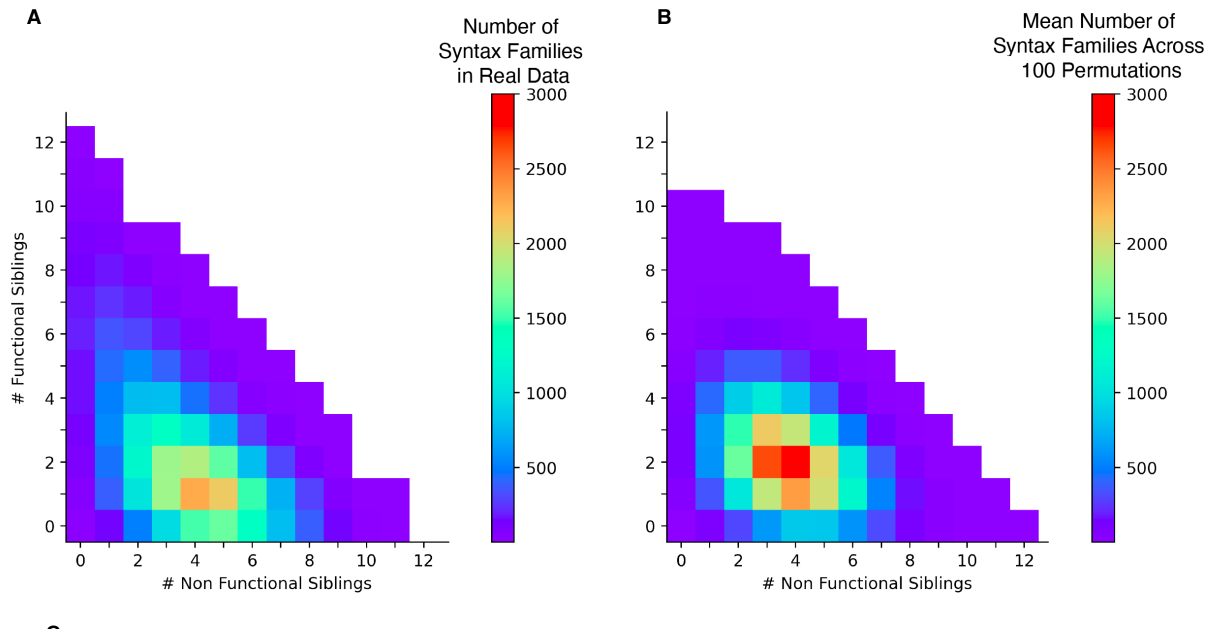


Figure C.8. Examples of Otx-a grammar library members. (A) Observed counts of organizational family members sorted into the heatmap by the number of active enhancers and number of inert sequences. (B) Mean count over 100 permutations of the number of organizational families in each bin.

Figure C.9 Sequences from robust and inert organizational families reproduce MPRA enhancer activity under the microscope. (A) Expression patterns of four siblings within a robust organizational family, with Otx-a expression pattern as a reference. All four siblings activate expression in a6.5 and b6.5 lineage neural tissues in more embryos than Otx-a. A fraction of the embryos have expression in ectopic tissues including the mesenchyme ($\leq 5\%$ embryos), notochord ($\leq 10\%$ embryos), and endoderm ($\leq 10\%$ embryos). For each construct, two biological replicates each with 50 embryos assayed. (B) The four robust syntax siblings drive strong neural expression. (C) Expression patterns of three siblings of an inert organizational family, with Otx-a expression pattern as a reference. All three siblings have no expression in the a6.5 neural lineage and $\leq 5\%$ embryos expressing in the b6.5 lineage. (D) No observable expression of the three inert syntax siblings at the tailbud stage. For each construct, two biological replicates each with 50 embryos assayed.

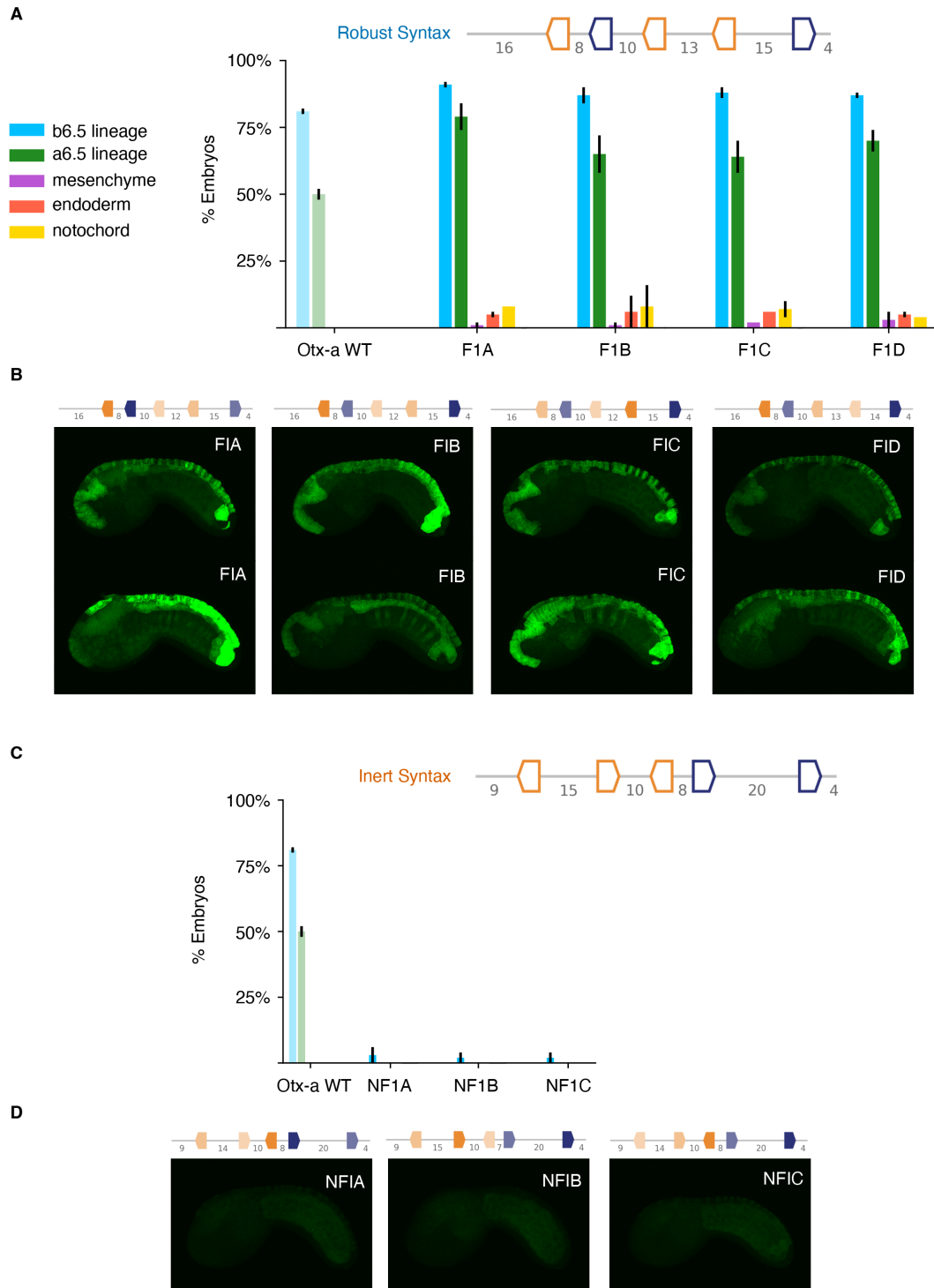


Figure C.9. Sequences from robust and inert organizational families reproduce MPRA enhancer activity under the microscope.

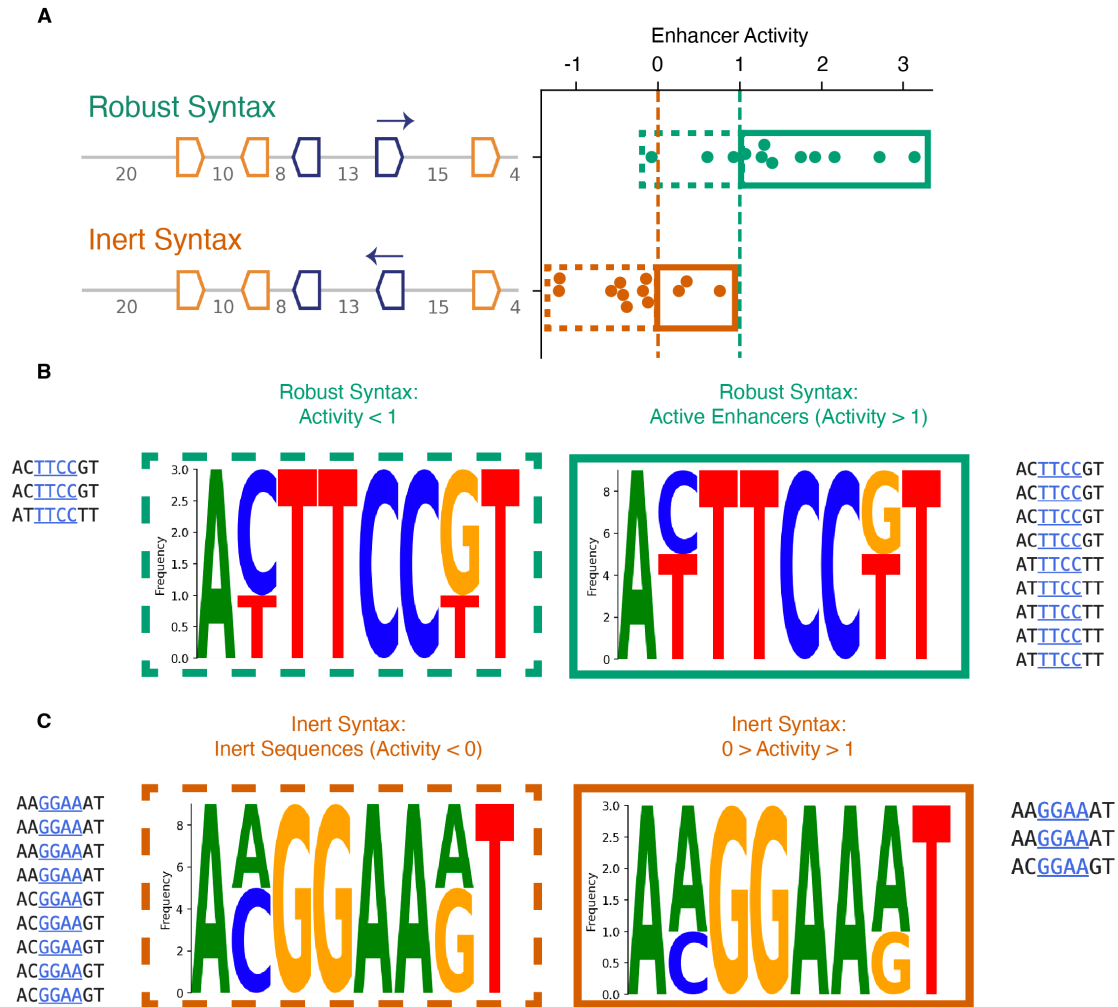


Figure C.10. In a single flip comparison, there are no sequence features which distinguish inert and active enhancers. Motif logos for the robust syntax (top) and the inert syntax (bottom) for two organizational families separated by a single flip. (A) In the robust syntax, the motif logo is shown for the non-active sequences (left, dashed outline) and the active enhancers (right, solid outline). Individual 8mer sequences which make up the logo are also shown with ETS core underlined and in blue. Sequence content of non-active and active sequences are shared, there are no distinguishing features across the site-linker junction which can explain the difference in non-functionality. Therefore it is unlikely that a de novo motif is created at the site-linker junction (B) In the inert syntax, the motif logo is shown for the inert sequences (left, dashed outline) and the non-inert sequences (right, solid outline). Sequence content is shared, therefore it is unlikely that a de novo motif causing the change in function has been created at the site-linker junction.

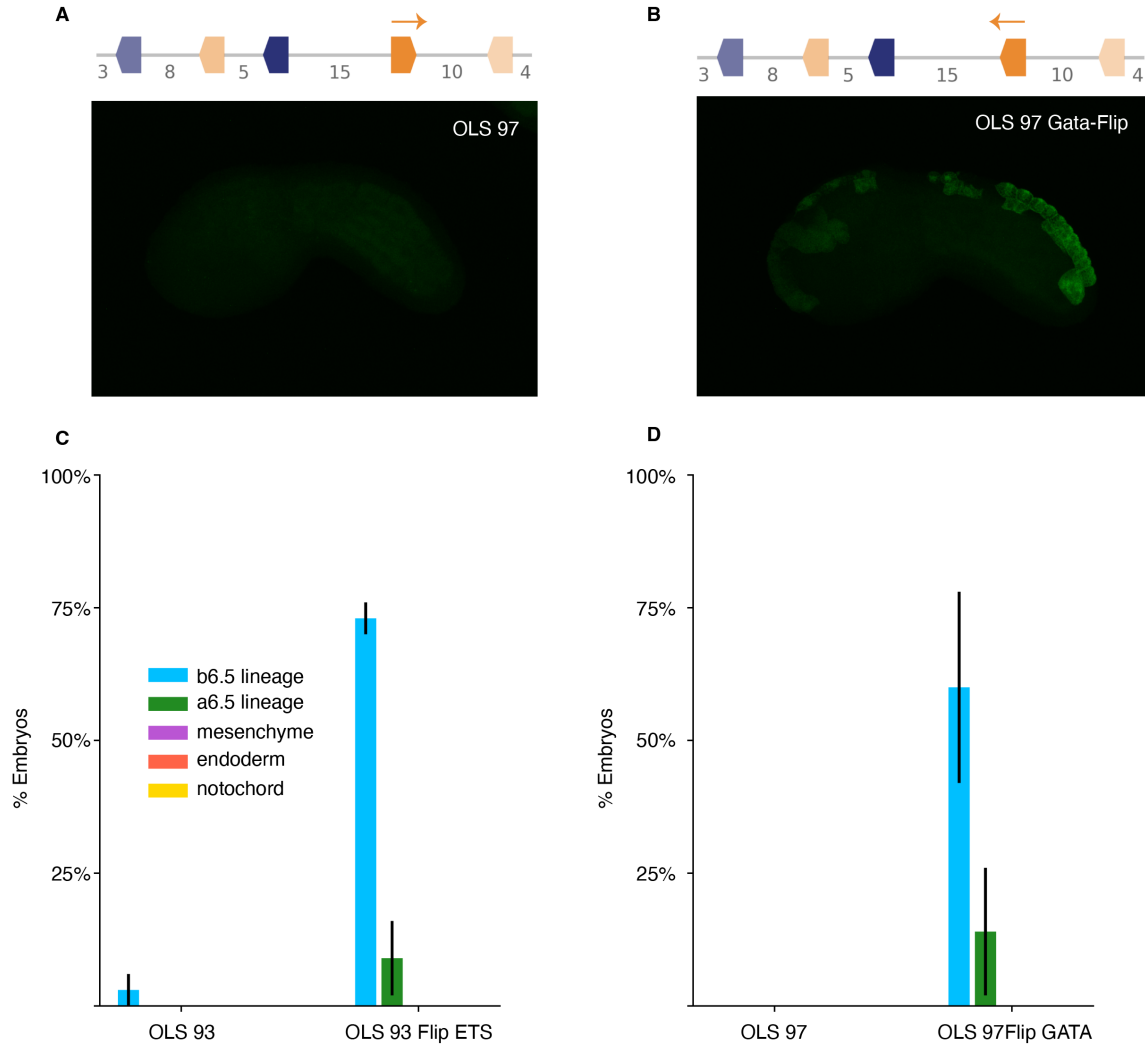


Figure C.11. Single flips cause inert sequences to convert to neural enhancers. A single flip of a GATA site from the forward direction (A) to the reverse direction (B) causes an inert sequence to become functional. (C) Expression data for OLS 93 show that a single flip of an ETS site causes an inert sequence to become an active enhancer driving strong expression in b6.5 neural tissues and some expression in a6.5 neural tissues. (Fisher's exact test) $n=50$ each across two replicates. (D) Expression data for OLS 97 show that a single flip of a GATA site causes an inert sequence to become an active enhancer. (Fisher's exact test) $n=50$ each across two replicates.

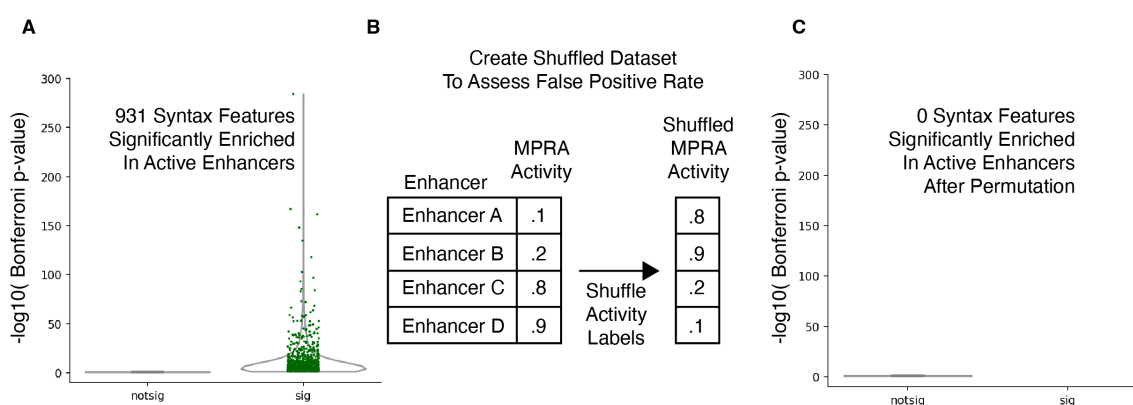


Figure C.12. No syntax features are significantly enriched in active enhancers after permuting enhancer activity and sequence labels. (A) Distribution of log-transformed Bonferroni-adjusted p-values for syntaxes that are significantly enriched in active enhancers (significance level of $p \leq .05$). (B) Permuted dataset created by permuting the enhancer activity and sequence labels. Note that no activities were created or destroyed. In the toy example, there are 4 enhancers and 4 activities (.1, .2, .8, .9). The procedure simply reassigns the activity to a randomly chosen sequence. This way, the distribution of activities is the same in the real dataset as the permuted dataset. (C) Performing the same search for syntax features yields no significantly enriched syntaxes in active enhancers.

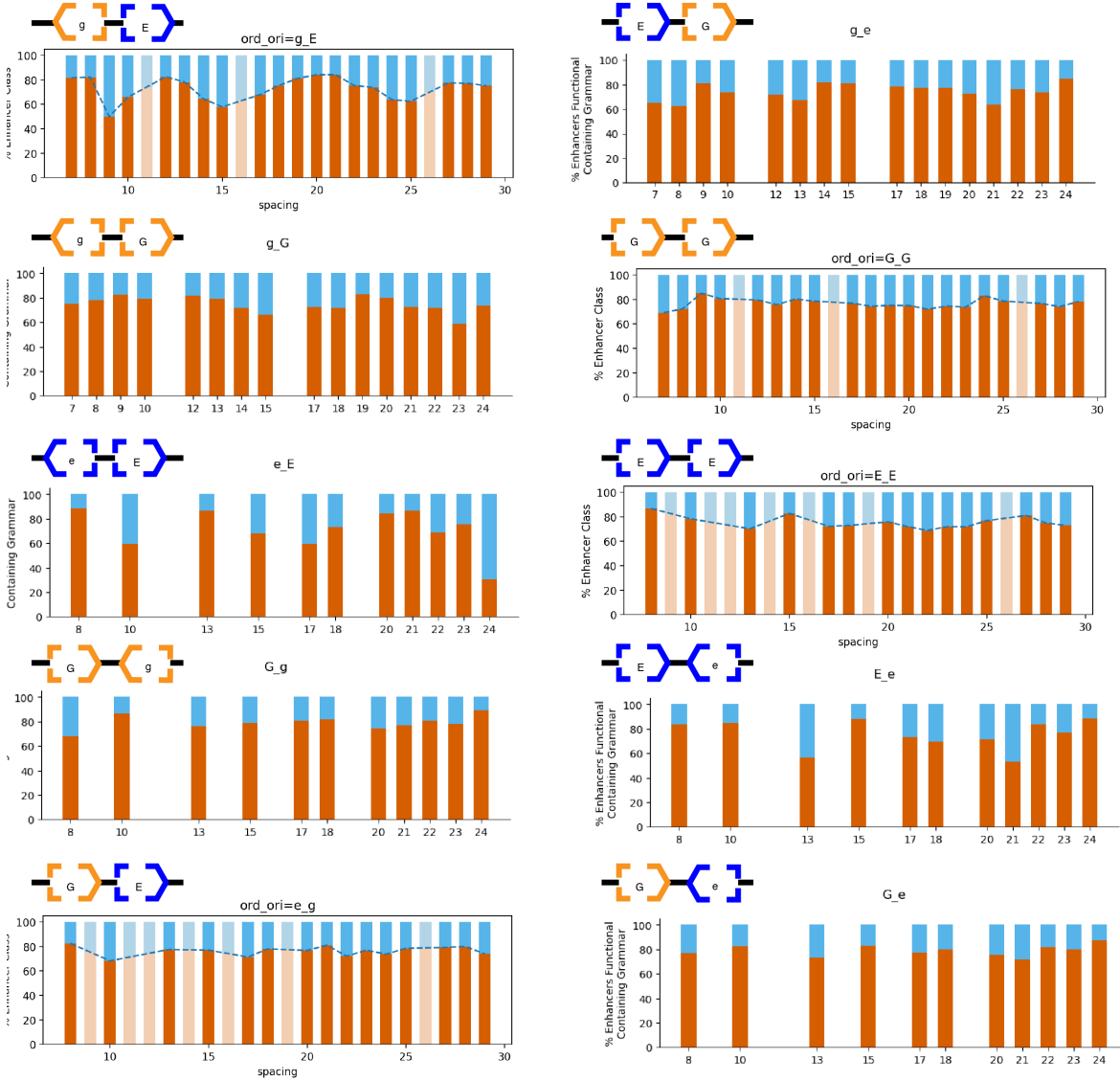


Figure C.13. Optimal spacing across doublet syntax features. Percent enrichment in active enhancers for different spacings of all possible doublet motifs. X-axis is the spacing between the two sites of the doublet. Some doublets exhibit clear spacing preferences (g_E, g_G, e_E, E_e) as indicated by differential enrichment in active enhancers which is dependent on spacing. Some doublets do not exhibit spacing preferences as indicated by all spacings having roughly similar enrichment in active sequences (G_G, G_E). Light grey bars indicate imputed enrichments in active enhancers as estimated by averaging the left- and right-most spacing. For example, g₁₁_E is not observed in the library, but we can estimate it by taking the average of g₁₀_E and g₁₂_E.

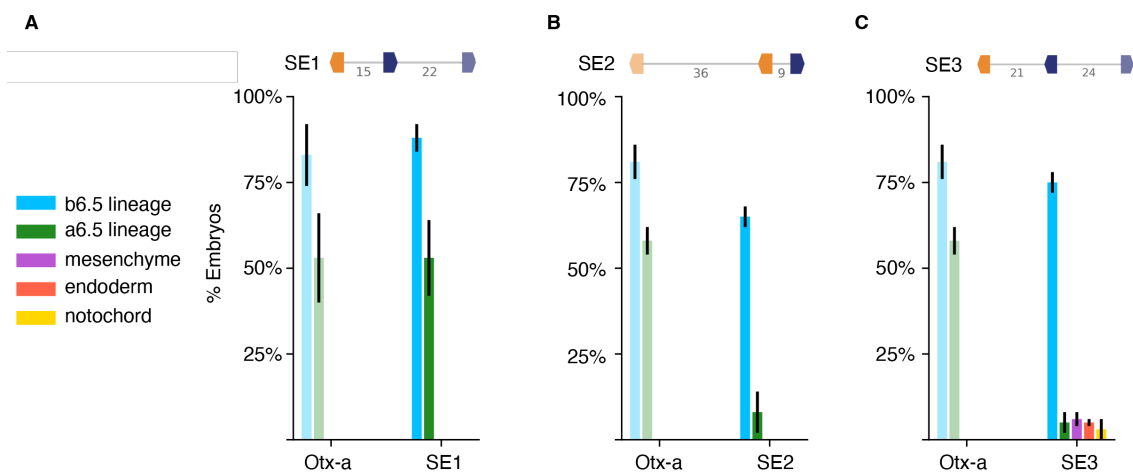


Figure C.14. Percent of embryos with neural expression activated by synthetic neural enhancers Four synthetic enhancers activate expression in neural tissues.

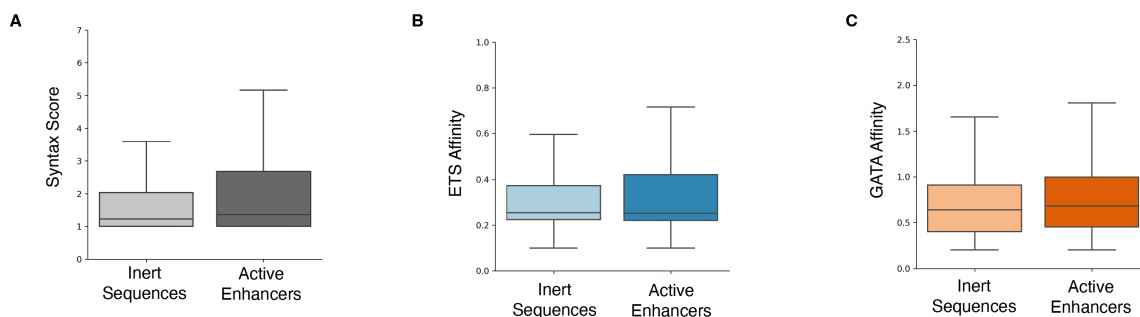


Figure C.15. Syntax and affinity alone slightly enrich for active enhancers. (A) Active enhancers are slightly enriched for higher syntax scores. (B,C) Active enhancers are slightly enriched for increased ETS and GATA affinity.

Bibliography

- [1] 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 Kutyaev, 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(6099):1190–1195, September 2012. doi: 10.1126/science.1222794. URL <https://www.science.org/doi/10.1126/science.1222794>. Publisher: American Association for the Advancement of Science.
- [2] 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. URL <https://doi.org/10.1186/s13072-015-0050-4>.
- [3] Axel Visel, Edward M. Rubin, and Len A. Pennacchio. Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205, September 2009. ISSN 1476-4687. doi: 10.1038/nature08451. URL <https://www.nature.com/articles/nature08451>. Number: 7261 Publisher: Nature Publishing Group.
- [4] Andrea I. Ramos and Scott Barolo. Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1632), 2013. ISSN 14712970. doi: 10.1098/rstb.2013.0018. Publisher: Royal Society.
- [5] 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. URL <https://www.science.org/doi/10.1126/science.aac6948>.
- [6] 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 of the United States of America*, 113(23):6508–6513, 2016. ISSN 10916490. doi: 10.1073/pnas.1605085113.
- [7] Granton A. Jindal, Alexis T. Bantle, Joe J. Solvason, Jessica L. Grudzien, Agnieszka

- D’Antonio-Chronowska, Fabian Lim, Sophia H. Le, Benjamin P. Song, Michelle F. Ragsac, Adam Klie, Reid O. Larsen, Kelly A. Frazer, and Emma K. Farley. Single-nucleotide variants within heart enhancers increase binding affinity and disrupt heart development. *Developmental Cell*, 58(21):2206–2216.e5, November 2023. ISSN 1534-5807. doi: 10.1016/j.devcel.2023.09.005. URL <https://www.sciencedirect.com/science/article/pii/S1534580723004926>.
- [8] Fabian Lim, Joe J. Solvason, Genevieve E. Ryan, Sophia H. Le, Granton A. Jindal, Paige Steffen, Simran K. Jandu, and Emma K. Farley. Affinity-optimizing enhancer variants disrupt development. *Nature*, 626(7997):151–159, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06922-8. URL <https://www.nature.com/articles/s41586-023-06922-8>. Number: 7997 Publisher: Nature Publishing Group.
- [9] 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. URL <https://linkinghub.elsevier.com/retrieve/pii/S1534580721001568>.
- [10] Ryan Tewhey, Dylan Kotliar, Daniel S. Park, Brandon Liu, Sarah Winnicki, Steven K. Reilly, Kristian G. Andersen, Tarjei S. Mikkelsen, Eric S. Lander, Stephen F. Schaffner, and Pardis C. Sabeti. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, 165(6):1519–1529, 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.04.027. URL <https://doi.org/10.1016/j.cell.2016.04.027>. Publisher: Elsevier Inc.
- [11] Felix Richter, Sarah U Morton, Seong Won Kim, Alexander Kitaygorodsky, Lauren K Wasson, Kathleen M Chen, Jian Zhou, Hongjian Qi, Nihir Patel, Steven R Depalma, Michael Parfenov, Jason Homsy, and Joshua M Gorham. Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nature Genetics*, pages 0–1, 2020. ISSN 1546-1718. doi: 10.1038/s41588-020-0652-z. URL <http://dx.doi.org/10.1038/s41588-020-0652-z>. Publisher: Springer US.
- [12] J Albuisson, B Isidor, M Giraud, O Pichon, T Marsaud, A David, C Le Caignec, and S Bezieau. Identification of two novel mutations in Shh long-range regulator associated with familial pre-axial polydactyly. *Clinical Genetics*, 79(4):371–377, 2011. ISSN 00099163. doi: 10.1111/j.1399-0004.2010.01465.x. Publisher: John Wiley & Sons, Ltd.
- [13] Evgeny Z. Kvon, Yiwen Zhu, Guy Kelman, Catherine S. Novak, Ingrid Plajzer-Frick, Momoe Kato, Tyler H. Garvin, Quan Pham, Anne N. Harrington, Riana D. Hunter, Janeth Godoy, Eman M. Meky, Jennifer A. Akiyama, Veena Afzal, Stella Tran, Fabienne Escande, Brigitte Gilbert-Dussardier, Nolwenn Jean-Marçais, Sanjarbek Hudaiberdiev, Ivan Ovcharenko, Matthew B. Dobbs, Christina A. Gurnett, Sylvie Manouvrier-Hanu, Florence Petit, Axel Visel, Diane E. Dickel, and Len A. Pennacchio. Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants. *Cell*, 180(6):1262–1271.e15, March 2020. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2020.02.031. URL [https://www.cell.com/cell/abstract/S0092-8674\(20\)30208-7](https://www.cell.com/cell/abstract/S0092-8674(20)30208-7). Publisher: Elsevier.

- [14] Genevieve E. Ryan and Emma K. Farley. Functional genomic approaches to elucidate the role of enhancers during development. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, 12(2):e1467, March 2020. ISSN 1939-005X. doi: 10.1002/wsbm.1467.
- [15] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nat Rev Methods Primers*, 1(1):1–21, August 2021. ISSN 2662-8449. doi: 10.1038/s43586-021-00056-9. URL <https://www.nature.com/articles/s43586-021-00056-9>. Number: 1 Publisher: Nature Publishing Group.
- [16] Sarah L. Spain and Jeffrey C. Barrett. Strategies for fine-mapping complex traits. *Hum Mol Genet*, 24(R1):R111–R119, October 2015. ISSN 0964-6906. doi: 10.1093/hmg/ddv260. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4572002/>.
- [17] Matteo D’Antonio, Jennifer P. Nguyen, Timothy D. Arthur, Hiroko Matsui, Agnieszka D’Antonio-Chronowska, and Kelly A. Frazer. Fine mapping spatiotemporal mechanisms of genetic variants underlying cardiac traits and disease. *Nat Commun*, 14(1):1132, February 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36638-2. URL <https://www.nature.com/articles/s41467-023-36638-2>. Number: 1 Publisher: Nature Publishing Group.
- [18] Qingbo S. Wang, David R. Kelley, Jacob Ulirsch, Masahiro Kanai, Shuvom Sadhuka, Ran Cui, Carlos Albors, Nathan Cheng, Yukinori Okada, Francois Aguet, Kristin G. Ardlie, Daniel G. MacArthur, and Hilary K. Finucane. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat Commun*, 12(1):3394, June 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23134-8. URL <https://www.nature.com/articles/s41467-021-23134-8>. Number: 1 Publisher: Nature Publishing Group.
- [19] Joon-Yong An, Kevin Lin, Lingxue Zhu, Donna M. Werling, Shan Dong, Harrison Brand, Harold Z. Wang, Xuefang Zhao, Grace B. Schwartz, Ryan L. Collins, Benjamin B. Currall, Claudia Dastmalchi, Jeanselle Dea, Clif Duhn, Michael C. Gilson, Lambertus Klei, Lindsay Liang, Eirene Markenscoff-Papadimitriou, Sirisha Pochareddy, Nadav Ahituv, Joseph D. Buxbaum, Hilary Coon, Mark J. Daly, Young Shin Kim, Gabor T. Marth, Benjamin M. Neale, Aaron R. Quinlan, John L. Rubenstein, Nenad Sestan, Matthew W. State, A. Jeremy Willsey, Michael E. Talkowski, Bernie Devlin, Kathryn Roeder, and Stephan J. Sanders. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science*, 362(6420):eaat6576, December 2018. doi: 10.1126/science.aat6576. URL <https://www.science.org/doi/10.1126/science.aat6576>. Publisher: American Association for the Advancement of Science.
- [20] Mohiuddin Mohiuddin, R. Frank Kooy, and Christopher E. Pearson. De novo mutations, genetic mosaicism and human disease. *Frontiers in Genetics*, 13, 2022. ISSN 1664-8021. URL <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.983668>.
- [21] Augustine Kong, Michael L. Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A. Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adal-

- bjorg Jonasdottir, Wendy S. W. Wong, Gunnar Sigurdsson, G. Bragi Walters, Stacy Steinberg, Hannes Helgason, Gudmar Thorleifsson, Daniel F. Gudbjartsson, Agnar Helgason, Olafur Th Magnusson, Unnur Thorsteinsdottir, and Kari Stefansson. Rate of de novo mutations and the importance of father’s age to disease risk. *Nature*, 488(7412):471–475, August 2012. ISSN 1476-4687. doi: 10.1038/nature11396. URL <https://www.nature.com/articles/nature11396>. Number: 7412 Publisher: Nature Publishing Group.
- [22] Sigurgeir Olafsson and Carl A. Anderson. Somatic mutations provide important and unique insights into the biology of complex diseases. *Trends in Genetics*, 37(10):872–881, October 2021. ISSN 0168-9525. doi: 10.1016/j.tig.2021.06.012. URL [https://www.cell.com/trends/genetics/abstract/S0168-9525\(21\)00168-2](https://www.cell.com/trends/genetics/abstract/S0168-9525(21)00168-2). Publisher: Elsevier.
- [23] Felix Dietlein, Alex B. Wang, Christian Fagre, Anran Tang, Nicolle J. M. Besselink, Edwin Cuppen, Chunliang Li, Shamil R. Sunyaev, James T. Neal, and Eliezer M. Van Allen. Genome-wide analysis of somatic noncoding mutation patterns in cancer. *Science*, 376(6589):eabg5601, April 2022. doi: 10.1126/science.abg5601. URL <https://www.science.org/doi/full/10.1126/science.abg5601>. Publisher: American Association for the Advancement of Science.
- [24] Wei Zhang, Ana Bojorquez-Gomez, Daniel Ortiz Velez, Guorong Xu, Kyle S. Sanchez, John Paul Shen, Kevin Chen, Katherine Licon, Collin Melton, Katrina M. Olson, Michael Ku Yu, Justin K. Huang, Hannah Carter, Emma K. Farley, Michael Snyder, Stephanie I. Fraley, Jason F. Kreisberg, and Trey Ideker. A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat Genet*, 50(4):613–620, April 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0091-2. URL <https://www.nature.com/articles/s41588-018-0091-2>. Number: 4 Publisher: Nature Publishing Group.
- [25] Julian Banerji, Sandro Rusconi, and Walter Schaffner. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2, Part 1):299–308, December 1981. ISSN 0092-8674. doi: 10.1016/0092-8674(81)90413-X. URL <https://www.sciencedirect.com/science/article/pii/009286748190413X>.
- [26] Sreejith J. Nair, Tom Suter, Susan Wang, Lu Yang, Feng Yang, and Michael G. Rosenfeld. Transcriptional enhancers at 40: evolution of a viral DNA element to nuclear architectural structures. *Trends in Genetics*, 38(10):1019–1047, October 2022. ISSN 0168-9525. doi: 10.1016/j.tig.2022.05.015. URL <https://www.sciencedirect.com/science/article/pii/S016895252200141X>.
- [27] Jian Yan, Yunjiang Qiu, André M. Ribeiro dos Santos, Yimeng Yin, Yang E. Li, Nick Vinckier, Naoki Nariai, Paola Benaglio, Anugraha Raman, Xiaoyu Li, Shicai Fan, Joshua Chiou, Fulin Chen, Kelly A. Frazer, Kyle J. Gaulton, Maike Sander, Jussi Taipale, and Bing Ren. Systematic analysis of binding of transcription factors to noncoding variants. *Nature*, 591(7848):147–151, March 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03211-0. URL <https://www.nature.com/articles/s41586-021-03211-0>. Number: 7848 Publisher: Nature Publishing Group.

- [28] Chandler Zuo, Sunyoung Shin, and Sündüz Keleş. atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics*, 31(20):3353–3355, October 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv328. URL <https://doi.org/10.1093/bioinformatics/btv328>.
- [29] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W. Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(suppl_1):D91–D94, January 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh012. URL <https://doi.org/10.1093/nar/gkh012>.
- [30] Timothy L. Bailey, James Johnson, Charles E. Grant, and William S. Noble. The MEME Suite. *Nucleic Acids Research*, 43(W1):W39–W49, July 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv416. URL <https://doi.org/10.1093/nar/gkv416>.
- [31] 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. *Mol Cell*, 38(4):576–589, May 2010. ISSN 1097-4164. doi: 10.1016/j.molcel.2010.05.004.
- [32] Sebastian J. Maerkl and Stephen R. Quake. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science*, 315(5809):233–237, January 2007. doi: 10.1126/science.1131007. URL <https://www.science.org/doi/10.1126/science.1131007>. Publisher: American Association for the Advancement of Science.
- [33] Daniel D. Le, Tyler C. Shimko, Arjun K. Aditham, Allison M. Keys, Scott A. Longwell, Yaron Orenstein, and Polly M. Fordyce. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proc Natl Acad Sci U S A*, 115(16):E3702–E3711, April 2018. ISSN 1091-6490. doi: 10.1073/pnas.1715888115.
- [34] Christophe Jung, Peter Bandilla, Marc von Reutern, Max Schnepf, Susanne Rieder, Ulrich Unnerstall, and Ulrike Gaul. True equilibrium measurement of transcription factor-DNA binding affinities using automated polarization microscopy. *Nat Commun*, 9(1):1605, April 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03977-4. URL <https://www.nature.com/articles/s41467-018-03977-4>. Number: 1 Publisher: Nature Publishing Group.
- [35] Michael F. Berger and Martha L. Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the dna-binding specificities of transcription factors. *Nature Protocols*, 4(3):393–411, 2009. ISSN 17502799. doi: 10.1038/nprot.2008.195.
- [36] Daniel E. Newburger and Martha L. Bulyk. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Res*, 37(Database issue):D77–D82, January 2009. ISSN 0305-1048. doi: 10.1093/nar/gkn660. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2686578/>.
- [37] Trevor Siggers, Michael H Duyzend, Jessica Reddy, Sidra Khan, and Martha L Bulyk. Non-

- DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Molecular Systems Biology*, 7(1):555, January 2011. ISSN 1744-4292. doi: 10.1038/msb.2011.89. URL <https://www.embopress.org/doi/full/10.1038/msb.2011.89>. Publisher: John Wiley & Sons, Ltd.
- [38] Yuning Zhang, Tiffany D. Ho, Nicolas E. Buchler, and Raluca Gordân. Competition for DNA binding between paralogous transcription factors determines their genomic occupancy and regulatory functions. *Genome Res*, 31(7):1216–1229, July 2021. ISSN 1088-9051. doi: 10.1101/gr.275145.120. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8256859/>.
 - [39] 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. *Dev Cell*, 18(3):359–370, March 2010. ISSN 1878-1551. doi: 10.1016/j.devcel.2009.12.026.
 - [40] Michal Levo and Eran Segal. In pursuit of design principles of regulatory sequences. *Nat Rev Genet*, 15(7):453–468, July 2014. ISSN 1471-0064. doi: 10.1038/nrg3684. URL <https://www.nature.com/articles/nrg3684>. Number: 7 Publisher: Nature Publishing Group.
 - [41] Maria I. Arnone and Eric H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864, May 1997. ISSN 0950-1991. doi: 10.1242/dev.124.10.1851. URL <https://doi.org/10.1242/dev.124.10.1851>.
 - [42] Scott Barolo. How to tune an enhancer. *Proceedings of the National Academy of Sciences*, 113(23):6330–6331, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1606109113.
 - [43] Dimitris Thanos and Tom Maniatis. Virus induction of human IFN β 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. Publisher: Cell Press.
 - [44] K. M. Scully, E. M. Jacobson, K. Jepsen, V. Lunyak, H. Viadiu, C. Carrière, D. W. Rose, F. Hooshmand, A. K. Aggarwal, and M. G. Rosenfeld. Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. *Science*, 290(5494):1127–1131, November 2000. ISSN 0036-8075. doi: 10.1126/science.290.5494.1127.
 - [45] Y. J. Passamaneck, L. Katikala, L. Perrone, M. P. Dunn, I. Oda-Ishii, and A. Di Gregorio. Direct activation of a notochord cis-regulatory module by Brachyury and FoxA in the ascidian *Ciona intestinalis*. *Development*, 136(21):3679–3689, 2009. ISSN 0950-1991. doi: 10.1242/dev.038141.
 - [46] 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. URL <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002796>. Publisher: Public Library of Science.

- [47] 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. URL [https://www.cell.com/cell/abstract/S0092-8674\(12\)00096-7](https://www.cell.com/cell/abstract/S0092-8674(12)00096-7). Publisher: Elsevier.
- [48] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet*, 53(3):354–366, March 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00782-6. URL <https://www.nature.com/articles/s41588-021-00782-6>. Number: 3 Publisher: Nature Publishing Group.
- [49] Bernardo P. De Almeida, Franziska Reiter, Michaela Pagani, and Alexander Stark. Deep-STARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet*, 54(5):613–624, May 2022. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-022-01048-5. URL <https://www.nature.com/articles/s41588-022-01048-5>.
- [50] Mette Bentsen, Vanessa Heger, Hendrik Schultheis, Carsten Kuenne, and Mario Looso. TF-COMB – Discovering grammar of transcription factor binding sites. *Computational and Structural Biotechnology Journal*, 20:4040–4051, 2022. ISSN 20010370. doi: 10.1016/j.csbj.2022.07.025. URL <https://linkinghub.elsevier.com/retrieve/pii/S2001037022003051>.
- [51] Ibrahim I. Taskiran, Katina I. Spanier, Hannah Dickmanken, Niklas Kempynck, Alexandra Pančíková, Eren Can Ekşi, Gert Hulselmans, Joy N. Ismail, Koen Theunis, Roel Vandepoel, Valerie Christiaens, David Mauduit, and Stein Aerts. Cell-type-directed design of synthetic enhancers. *Nature*, 626(7997):212–220, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06936-2. URL <https://www.nature.com/articles/s41586-023-06936-2>. Number: 7997 Publisher: Nature Publishing Group.
- [52] Zeyang Shen, Rick Z Li, Thomas A Prohaska, Marten A Hoeksema, Nathan J Spann, Jenhan Tao, Gregory J Fonseca, Thomas Le, Lindsey K Stolze, Mashito Sakai, Casey E Romanoski, and Christopher K Glass. Systematic analysis of naturally occurring insertions and deletions that alter transcription factor spacing identifies tolerant and sensitive transcription factor pairs. *eLife*, 11:e70878, January 2022. ISSN 2050-084X. doi: 10.7554/eLife.70878. URL <https://elifesciences.org/articles/70878>.
- [53] Peter J Skene and Steven Henikoff. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*, 6:e21856. ISSN 2050-084X. doi: 10.7554/eLife.21856. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5310842/>.
- [54] Mette Bentsen, Philipp Goymann, Hendrik Schultheis, Kathrin Klee, Anastasiia Petrova, René Wiegandt, Annika Fust, Jens Preussner, Carsten Kuenne, Thomas Braun, Johnny Kim, and Mario Looso. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun*, 11(1):4267, August 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18035-1. URL <https://www.nature.com/articles/>

s41467-020-18035-1. Number: 1 Publisher: Nature Publishing Group.

- [55] J Klein, V Agarwal, F Inoue, A Keith, B Martin, M Kircher, N Ahituv, and J Shendure. A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays. pages 1–71, 2019. doi: 10.1101/576405. URL <http://europepmc.org/abstract/PPR/PPR73179>.
- [56] Cosmas D. Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryń, Martina Rath, and Alexander Stark. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science*, 339(6123):1074–1077, March 2013. doi: 10.1126/science.1232542. URL <https://www.science.org/doi/10.1126/science.1232542>. Publisher: American Association for the Advancement of Science.
- [57] Felix Muerdter, Łukasz M. Boryń, and Cosmas D. Arnold. STARR-seq — Principles and applications. *Genomics*, 106(3):145–150, September 2015. ISSN 0888-7543. doi: 10.1016/j.ygeno.2015.06.001. URL <https://www.sciencedirect.com/science/article/pii/S0888754315300100>.
- [58] Justin Crocker, Ella Preger-Ben Noon, and David L. Stern. Chapter Twenty-Seven - The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. In Paul M. Wassarman, editor, *Current Topics in Developmental Biology*, volume 117 of *Essays on Developmental Biology, Part B*, pages 455–469. Academic Press, January 2016. doi: 10.1016/bs.ctdb.2015.11.018. URL <https://www.sciencedirect.com/science/article/pii/S0070215315001842>.
- [59] H N Cai, D N Arnosti, and M Levine. Long-range repression in the Drosophila embryo. *Proc Natl Acad Sci U S A*, 93(18):9309–9314, September 1996. ISSN 0027-8424. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC38424/>.
- [60] 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. *Nat Biotechnol*, 30(3): 271–277, March 2012. ISSN 1546-1696. doi: 10.1038/nbt.2137. URL <https://www.nature.com/articles/nbt.2137>. Number: 3 Publisher: Nature Publishing Group.
- [61] Jasper Janssens, Sara Aibar, Ibrahim Ihsan Taskiran, Joy N. Ismail, Alicia Estacio Gomez, Gabriel Aughey, Katina I. Spanier, Florian V. De Rop, Carmen Bravo González-Blas, Marc Dionne, Krista Grimes, Xiao Jiang Quan, Dafni Papasokrati, Gert Hulselmans, Samira Makhzami, Maxime De Waegeneer, Valerie Christiaens, Tony Southall, and Stein Aerts. Decoding gene regulation in the fly brain. *Nature*, 601(7894):630–636, January 2022. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-04262-z. URL <https://www.nature.com/articles/s41586-021-04262-z>.
- [62] S. J. Gosai, R. I. Castro, N. Fuentes, J. C. Butts, S. Kales, R. R. Noche, K. Mouri, P. C. Sabeti, S. K. Reilly, and R. Tewhey. Machine-guided design of synthetic cell type-specific

- cis-regulatory elements, August 2023. URL <https://www.biorxiv.org/content/10.1101/2023.08.08.552077v1>. Pages: 2023.08.08.552077 Section: New Results.
- [63] 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. URL <https://doi.org/10.7554/eLife.41279>. Publisher: eLife Sciences Publications, Ltd.
 - [64] 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, 2006. ISSN 14764687. doi: 10.1038/nature04336.
 - [65] Paramvir Dehal, 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, Stephen Chin-bow, Rosaria Desantis, Sharon Doyle, 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, Miki Tokuoka, Kasumi Yagi, Fumiko Yoshizaki, Frank Larimer, Chris Detter, Norman Doggett, Susan Lucas, Yuji Kohara, Michael Levine, and Nori Satoh. into Chordate and Vertebrate Origins The Draft Genome of *Ciona intestinalis* : Insights into Chordate and Vertebrate Origins. 2157(2002), 2007. doi: 10.1126/science.1080049.
 - [66] Alberto Stolfi, Kerriane Ryan, Ian A. Meinertzhagen, and Lionel Christiaen. Migratory neuronal progenitors arise from the neural plate borders in tunicates. *Nature*, 527(7578): 371–374, 2015. ISSN 14764687. doi: 10.1038/nature15758. URL <http://dx.doi.org/10.1038/nature15758>. Publisher: Nature Publishing Group.
 - [67] Yutaka Satou and Nori Satoh. Gene regulatory networks for the development and evolution of the chordate heart. *Genes and Development*, 20(19):2634–2638, 2006. ISSN 08909369. doi: 10.1101/gad.1485706.
 - [68] Kyle Kai-How Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shores, Holly Whitton, Russell J. H. Ryan, Alexander A. Shishkin, Meital Hatan, Marlene J. Carrasco-Alfonso, Dita Mayer, C. John Luckey, Nikolaos A. Patsopoulos, Philip L. De Jager, Vijay K. Kuchroo, Charles B. Epstein, Mark J. Daly, David A. Hafler, and Bradley E. Bernstein. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, February 2015. ISSN 1476-4687. doi: 10.1038/nature13835. URL <https://www.nature.com/articles/nature13835>. Number: 7539 Publisher: Nature Publishing Group.
 - [69] Mike Levine. Transcriptional Enhancers in Animal Development and Evolution. *Current*

- Biology*, 20(17):R754–R763, September 2010. ISSN 09609822. doi: 10.1016/j.cub.2010.06.070. URL <https://linkinghub.elsevier.com/retrieve/pii/S0960982210008560>.
- [70] Laura A. Lettice, Alison E. Hill, Paul S. Devenney, and Robert E. Hill. Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. *Hum Mol Genet*, 17(7):978–985, April 2008. ISSN 1460-2083. doi: 10.1093/hmg/ddm370.
 - [71] Scott Smemo, Juan J. Tena, Kyoung-Han Kim, Eric R. Gamazon, Noboru J. Sakabe, Carlos Gómez-Marín, Ivy Aneas, Flavia L. Credidio, Débora R. Sobreira, Nora F. Wasserman, Ju Hee Lee, Vijitha Puvindran, Davis Tam, Michael Shen, Joe Eun Son, Niki Alizadeh Vakili, Hoon-Ki Sung, Silvia Naranjo, Rafael D. Acemel, Miguel Manzanares, Andras Nagy, Nancy J. Cox, Chi-Chung Hui, Jose Luis Gomez-Skarmeta, and Marcelo A. Nóbrega. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507(7492):371–375, March 2014. ISSN 1476-4687. doi: 10.1038/nature13138. URL <https://www.nature.com/articles/nature13138>. Publisher: Nature Publishing Group.
 - [72] Christophe Tournamille, Yves Colin, Jean Pierre Cartron, and Caroline Le Van Kim. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet*, 10(2):224–228, June 1995. ISSN 1546-1718. doi: 10.1038/ng0695-224. URL <https://www.nature.com/articles/ng0695-224>. Number: 2 Publisher: Nature Publishing Group.
 - [73] Michael D. Gallagher and Alice S. Chen-Plotkin. The Post-GWAS Era: From Association to Function. *Am J Hum Genet*, 102(5):717–730, May 2018. ISSN 1537-6605. doi: 10.1016/j.ajhg.2018.04.002.
 - [74] Richard P. Harvey. Patterning the vertebrate heart. *Nature Reviews Genetics*, 3(7):544–556, July 2002. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg843. URL <http://www.nature.com/articles/nrg843>.
 - [75] Stéphane Zaffran and Manfred Frasch. Early Signals in Cardiac Development. *Circulation Research*, 91(6):457–469, September 2002. ISSN 0009-7330, 1524-4571. doi: 10.1161/01.RES.0000034152.74523.A8. URL <https://www.ahajournals.org/doi/10.1161/01.RES.0000034152.74523.A8>.
 - [76] Andrew D. Sharrocks. The ETS-domain transcription factor family. *Nat Rev Mol Cell Biol*, 2(11):827–837, November 2001. ISSN 1471-0080. doi: 10.1038/35099076. URL <https://www.nature.com/articles/35099076>. Number: 11 Publisher: Nature Publishing Group.
 - [77] M Beiman, B Z Shilo, and T Volk. Heartless, a Drosophila FGF receptor homolog, is essential for cell migration and establishment of several mesodermal lineages. *Genes & Development*, 10(23):2993–3002, December 1996. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.10.23.2993. URL <http://genesdev.cshlp.org/lookup/doi/10.1101/gad.10.23.2993>.

- [78] Margaret Buckingham, Sigolène Meilhac, and Stéphane Zaffran. Building the mammalian heart from two sources of myocardial cells. *Nature Reviews Genetics*, 6(11):826–835, November 2005. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg1710. URL <http://www.nature.com/articles/nrg1710>.
- [79] Frank Reifers, Emily C Walsh, Sophie Leger, Didier Y. R. Stainier, and Michael Brand. Induction and differentiation of the zebrafish heart requires fibroblast growth factor 8 (fgf8/acerebellar). *Development*, 127(2):225–235, January 2000.
- [80] Stacey L House, Brian E House, Betty Glascock, Thomas Kimball, Eyad Nusayr, El J Schultz, and Thomas Doetschman. Fibroblast Growth Factor 2 Mediates Isoproterenol-induced Cardiac Hypertrophy through Activation of the Extracellular Regulated Kinase. *Mol Cell Pharmacol*, 2(4):143–154, 2010.
- [81] Maoqing Ye, Chris Coldren, Xingqun Liang, Teresa Mattina, Elizabeth Goldmuntz, D. Woodrow Benson, Dunbar Ivy, M.B. Perryman, Lee Ann Garrett-Sinha, and Paul Grossfeld. Deletion of ETS-1, a gene in the Jacobsen syndrome critical region, causes ventricular septal defects and abnormal ventricular morphology in mice. *Human Molecular Genetics*, 19(4):648–656, February 2010. ISSN 1460-2083, 0964-6906. doi: 10.1093/hmg/ddp532. URL <https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddp532>.
- [82] Brad Davidson, Weiyang Shi, Jeni Beh, Lionel Christiaen, and Mike Levine. FGF signaling delineates the cardiac progenitor field in the simple chordate, *Ciona intestinalis*. *Genes & Development*, 20(19):2728–2738, October 2006. ISSN 0890-9369. doi: 10.1101/gad.1467706. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1578698/>.
- [83] Brad Davidson. *Ciona intestinalis* as a model for cardiac development. *Seminars in Cell & Developmental Biology*, 18(1):16–26, February 2007. ISSN 10849521. doi: 10.1016/j.semcdb.2006.12.007. URL <https://linkinghub.elsevier.com/retrieve/pii/S1084952106001303>.
- [84] 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*, 134(18):3297–3305, September 2007. ISSN 1477-9129, 0950-1991. doi: 10.1242/dev.010140. URL <https://journals.biologists.com/dev/article/134/18/3297/64471/FoxF-is-essential-for-FGF-induced-migration-of>.
- [85] Lionel Christiaen, Brad Davidson, Takeshi Kawashima, Weston Powell, Hector Nolla, Karen Vranizan, and Michael Levine. The Transcription/Migration Interface in Heart Precursors of *Ciona intestinalis*. *Science*, 320(5881):1349–1352, June 2008. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1158170. URL <https://www.science.org/doi/10.1126/science.1158170>.
- [86] Justin Crocker, Namiko Abe, Lucrezia Rinaldi, Alistair P. McGregor, Nicolás Frankel, Shu Wang, Ahmad Alsawadi, Philippe Valenti, Serge Plaza, François Payre, Richard S. Mann, and David L. Stern. Low affinity binding site clusters confer HOX specificity and regulatory robustness. *Cell*, 160(1-2):191–203, 2015. ISSN 10974172. doi: 10.1016/j.cell.2014.11.041. Publisher: Cell Press.

- [87] Rebecca K. Delker, Vikram Ranade, Ryan Loker, Roumen Voutev, and Richard S. Mann. Low affinity binding sites in an activating CRM mediate negative autoregulation of the *Drosophila* Hox gene *Ultrabithorax*. *PLoS genetics*, 15(10):e1008444, October 2019. ISSN 1553-7404. doi: 10.1371/journal.pgen.1008444.
- [88] Judith F. Kribelbauer, Chaitanya Rastogi, Harmen J. Bussemaker, and Richard S. Mann. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annual Review of Cell and Developmental Biology*, 35(1):357–379, 2019. doi: 10.1146/annurev-cellbio-100617-062719. URL <https://doi.org/10.1146/annurev-cellbio-100617-062719>. eprint: <https://doi.org/10.1146/annurev-cellbio-100617-062719>.
- [89] Christina I. Swanson, David B. Schwimmer, and Scott Barolo. Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Curr Biol*, 21(14):1186–1196, July 2011. ISSN 1879-0445. doi: 10.1016/j.cub.2011.05.056.
- [90] Florian Razy-Krajka, Basile Gravez, Nicole Kaplan, Claudia Racioppi, Wei Wang, and Lionel Christiaen. An FGF-driven feed-forward circuit patterns the cardiopharyngeal mesoderm in space and time. *eLife*, 7:e29656, February 2018. ISSN 2050-084X. doi: 10.7554/eLife.29656.
- [91] Ute Rothbächer, Vincent Bertrand, Clement Lamy, and Patrick Lemaire. A combinatorial code of maternal GATA, Ets and beta-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. *Development*, 134(22):4023–4032, November 2007. ISSN 0950-1991. doi: 10.1242/dev.010850.
- [92] Arielle Woznica, Maximilian Haeussler, Ella Starobinska, Jessica Jemmett, Younan Li, David Mount, and Brad Davidson. Initial deployment of the cardiogenic gene regulatory network in the basal chordate, *Ciona intestinalis*. *Developmental Biology*, 368(1):127–139, August 2012. ISSN 1095564X. doi: 10.1016/j.ydbio.2012.05.002. Publisher: Academic Press Inc.
- [93] Gong-Hong Wei, Gwenaél 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. *EMBO J*, 29(13):2147–2160, July 2010. ISSN 0261-4189, 1460-2075. doi: 10.1038/emboj.2010.106. URL <http://emboj.embopress.org/cgi/doi/10.1038/emboj.2010.106>.
- [94] 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. URL <https://doi.org/10.7554/eLife.04837>. Publisher: eLife Sciences Publications, Ltd.

- [95] 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. URL <https://www.nature.com/articles/s41586-020-2816-5>.
- [96] Laura A. Lettice, Iain Williamson, John H. Wiltshire, Silvia Peluso, Paul S. Devenney, Alison E. Hill, Abdelkader Essafi, James Hagman, Richard Mort, Graeme Grimes, Carlo L. DeAngelis, and Robert E. Hill. Opposing Functions of the ETS Factor Family Define Shh Spatial Expression in Limb Buds and Underlie Polydactyly. *Developmental Cell*, 22(2):459–467, February 2012. ISSN 1534-5807. doi: 10.1016/j.devcel.2011.12.010. URL <https://www.sciencedirect.com/science/article/pii/S1534580711005764>.
- [97] Mikhail Spivakov. Spurious transcription factor binding: Non-functional or genetically redundant? *BioEssays*, 36(8):798–806, 2014. ISSN 1521-1878. doi: 10.1002/bies.201400036. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201400036>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.201400036>.
- [98] James Cooley, Stacia Whitaker, Sarah Sweeney, Scott Fraser, and Brad Davidson. Cytoskeletal polarity mediates localized induction of the heart progenitor lineage. *Nature Cell Biology*, 13(8):952–957, August 2011. ISSN 1465-7392, 1476-4679. doi: 10.1038/ncb2291. URL <http://www.nature.com/articles/ncb2291>.
- [99] Christina D. Cota and Brad Davidson. Mitotic Membrane Turnover Coordinates Differential Induction of the Heart Progenitor Lineage. *Developmental Cell*, 34(5):505–519, September 2015. ISSN 15345807. doi: 10.1016/j.devcel.2015.07.001. URL <https://linkinghub.elsevier.com/retrieve/pii/S1534580715004529>.
- [100] Claudia Racioppi, Keira A Wiechecki, and Lionel Christiaen. Combinatorial chromatin dynamics foster accurate cardiopharyngeal fate choices. *eLife*, 8:e49921, November 2019. ISSN 2050-084X. doi: 10.7554/eLife.49921. URL <https://elifesciences.org/articles/49921>.
- [101] Paola Benaglio, Agnieszka D’Antonio-Chronowska, Wubin Ma, Feng Yang, William W. Young Greenwald, Margaret K. R. Donovan, Christopher DeBoever, He Li, Frauke Drees, Sanghamitra Singhal, Hiroko Matsui, Jessica van Setten, Nona Sotoodehnia, Kyle J. Gaulton, Erin N. Smith, Matteo D’Antonio, Michael G. Rosenfeld, and Kelly A. Frazer. Allele-specific NKX2-5 binding underlies multiple genetic associations with human electrocardiographic traits. *Nat Genet*, 51(10):1506–1517, October 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0499-3. URL <https://www.nature.com/articles/s41588-019-0499-3>. Publisher: Nature Publishing Group.
- [102] Dalit May, Matthew J Blow, Tommy Kaplan, David J McCulley, Brian C Jensen, Jennifer A Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Veena Afzal, Paul C Simpson, Edward M Rubin, Brian L Black, James Bristow, Len A Pennacchio, and Axel Visel. Large-scale discovery of enhancers from human heart tissue. *Nature Genetics*, 44(1):89–93, January 2012. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.1006.

URL <http://www.nature.com/articles/ng.1006>.

- [103] Marc Sylva, Maurice J. B. van den Hoff, and Antoon F. M. Moorman. Development of the human heart. *American Journal of Medical Genetics. Part A*, 164A(6):1347–1371, June 2014. ISSN 1552-4833. doi: 10.1002/ajmg.a.35896.
- [104] William Schachterle, Anabel Rojas, Shan Mei Xu, and Brian L. Black. ETS-dependent regulation of a distal Gata4 cardiac enhancer. *Developmental Biology*, 361(2):439–449, January 2012. ISSN 1095564X. doi: 10.1016/j.ydbio.2011.10.023. Publisher: Academic Press Inc.
- [105] Agnieszka D’Antonio-Chronowska, Margaret K. R. Donovan, William W. Young Greenwald, Jennifer Phuong Nguyen, Kyohei Fujita, Sherin Hashem, Hiroko Matsui, Francesca Soncin, Mana Parast, Michelle C. Ward, Florence Coulet, Erin N. Smith, Eric Adler, Matteo D’Antonio, and Kelly A. Frazer. Association of Human iPSC Gene Signatures and X Chromosome Dosage with Two Distinct Cardiac Differentiation Trajectories. *Stem Cell Reports*, 13(5):924–938, November 2019. ISSN 2213-6711. doi: 10.1016/j.stemcr.2019.09.011.
- [106] Athanasia D. Panopoulos, Matteo D’Antonio, Paola Benaglio, Roy Williams, Sherin I. Hashem, Bernhard M. Schuldt, Christopher DeBoever, Angelo D. Arias, Melvin Garcia, Bradley C. Nelson, Olivier Harismendy, David A. Jakubosky, Margaret K. R. Donovan, William W. Greenwald, KathyJean Farnam, Megan Cook, Victor Borja, Carl A. Miller, Jonathan D. Grinstein, Frauke Drees, Jonathan Okubo, Kenneth E. Diffenderfer, Yuriko Hishida, Veronica Modesto, Carl T. Dargitz, Rachel Feiring, Chang Zhao, Aitor Aguirre, Thomas J. McGarry, Hiroko Matsui, He Li, Joaquin Reyna, Fangwen Rao, Daniel T. O’Connor, Gene W. Yeo, Sylvia M. Evans, Neil C. Chi, Kristen Jepsen, Naoki Nariai, Franz-Josef Müller, Lawrence S. B. Goldstein, Juan Carlos Izpisua Belmonte, Eric Adler, Jeanne F. Loring, W. Travis Berggren, Agnieszka D’Antonio-Chronowska, Erin N. Smith, and Kelly A. Frazer. iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. *Stem Cell Reports*, 8(4):1086–1100, April 2017. ISSN 2213-6711. doi: 10.1016/j.stemcr.2017.03.012.
- [107] Sheldon Rowan, Trevor Siggers, Salil A. Lachke, Yingzi Yue, Martha L. Bulyk, and Richard L. Maas. Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. *Genes Dev.*, 24(10):980–985, May 2010. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.1890410. URL <http://genesdev.cshlp.org/content/24/10/980>. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.
- [108] Albert Tsai, Anand K Muthusamy, Mariana RP Alves, Luke D Lavis, Robert H Singer, David L Stern, and Justin Crocker. Nuclear microenvironments modulate transcription from low-affinity enhancers. *eLife*, 6:e28975, November 2017. ISSN 2050-084X. doi: 10.7554/eLife.28975. URL <https://doi.org/10.7554/eLife.28975>. Publisher: eLife Sciences Publications, Ltd.

- [109] Granton A. Jindal, Yogesh Goyal, Rebecca D. Burdine, Katherine A. Rauen, and Stanislav Y. Shvartsman. RASopathies: unraveling mechanisms with animal models. *Dis Model Mech*, 8(8):769–782, August 2015. ISSN 1754-8411. doi: 10.1242/dmm.020339.
- [110] Yangli Xie, Nan Su, Jing Yang, Qiaoyan Tan, Shuo Huang, Min Jin, Zhenhong Ni, Bin Zhang, Dali Zhang, Fengtao Luo, Hangang Chen, Xianding Sun, Jian Q. Feng, Huabing Qi, and Lin Chen. FGF/FGFR signaling in health and disease. *Sig Transduct Target Ther*, 5(1):1–38, September 2020. ISSN 2059-3635. doi: 10.1038/s41392-020-00222-7. URL <https://www.nature.com/articles/s41392-020-00222-7>. Number: 1 Publisher: Nature Publishing Group.
- [111] Richard Grose and Clive Dickson. Fibroblast growth factor signaling in tumorigenesis. *Cytokine Growth Factor Rev*, 16(2):179–186, April 2005. ISSN 1359-6101. doi: 10.1016/j.cytogfr.2005.01.003.
- [112] Gareth L. Bond, Wenwei Hu, Elisabeth E. Bond, Harlan Robins, Stuart G. Lutzker, Nicoleta C. Arva, Jill Bargonetti, Frank Bartel, Helge Taubert, Peter Wuerl, Kenan Onel, Linwah Yip, Shih-Jen Hwang, Louise C. Strong, Guillermina Lozano, and Arnold J. Levine. A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell*, 119(5):591–602, November 2004. ISSN 0092-8674. doi: 10.1016/j.cell.2004.11.022.
- [113] Pouya Kheradpour, Jason Ernst, Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Jessica Alston, Tarjei S. Mikkelsen, and Manolis Kellis. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.*, 23(5):800–811, May 2013. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.144899.112. URL <https://genome.cshlp.org/content/23/5/800>. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.
- [114] David N. Arnosti, Scott Barolo, Michael Levine, and Stephen Small. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development*, 122(1):205–214, January 1996. ISSN 0950-1991. doi: 10.1242/dev.122.1.205. URL <https://doi.org/10.1242/dev.122.1.205>.
- [115] Li Jia, Gilad Landan, Mark Pomerantz, Rami Jaschek, Paula Herman, David Reich, Chunli Yan, Omar Khalid, Phil Kantoff, William Oh, J. Robert Manak, Benjamin P. Berman, Brian E. Henderson, Baruch Frenkel, Christopher A. Haiman, Matthew Freedman, Amos Tanay, and Gerhard A. Coetzee. Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS genetics*, 5(8):e1000597, August 2009. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000597.
- [116] Sari Tuupanen, Mikko Turunen, Rainer Lehtonen, Outi Hallikas, Sakari Vanharanta, Teemu Kivioja, Mikael Björklund, Gonghong Wei, Jian Yan, Iina Niittymäki, Jukka-Pekka Mecklin, Heikki Järvinen, Ari Ristimäki, Mariachiara Di-Bernardo, Phil East, Luis

- Carvajal-Carmona, Richard S. Houlston, Ian Tomlinson, Kimmo Palin, Esko Ukkonen, Auli Karhu, Jussi Taipale, and Lauri A. Aaltonen. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nature Genetics*, 41(8):885–890, August 2009. ISSN 1546-1718. doi: 10.1038/ng.406.
- [117] Juliet D. French, Maya Ghoussaini, Stacey L. Edwards, Kerstin B. Meyer, Kyriaki Michailidou, Shahana Ahmed, Sofia Khan, Mel J. Maranian, Martin O'Reilly, Kristine M. Hillman, Joshua A. Betts, Thomas Carroll, Peter J. Bailey, Ed Dicks, Jonathan Beesley, Jonathan Tyrer, Ana-Teresa Maia, Andrew Beck, Nicholas W. Knoblauch, Constance Chen, Peter Kraft, Daniel Barnes, Anna González-Neira, M. Rosario Alonso, Daniel Herrero, Daniel C. Tessier, Daniel Vincent, Francois Bacot, Craig Luccarini, Caroline Baynes, Don Conroy, Joe Dennis, Manjeet K. Bolla, Qin Wang, John L. Hopper, Melissa C. Southey, Marjanka K. Schmidt, Annegien Broeks, Senno Verhoef, Sten Cornelissen, Kenneth Muir, Artitaya Lophatananon, Sarah Stewart-Brown, Pornthep Siriwanarangsang, Peter A. Fasching, Christian R. Loehberg, Arif B. Ekici, Matthias W. Beckmann, Julian Peto, Isabel dos Santos Silva, Nichola Johnson, Zoe Aitken, Elinor J. Sawyer, Ian Tomlinson, Michael J. Kerin, Nicola Miller, Frederik Marme, Andreas Schneeweiss, Christof Sohn, Barbara Burwinkel, Pascal Guénel, Thérèse Truong, Pierre Laurent-Puig, Florence Menegaux, Stig E. Bojesen, Børge G. Nordestgaard, Sune F. Nielsen, Henrik Flyger, Roger L. Milne, M. Pilar Zamora, Jose Ignacio Arias Perez, Javier Benitez, Hoda Anton-Culver, Hermann Brenner, Heiko Müller, Volker Arndt, Christa Stegmaier, Alfons Meindl, Peter Lichtner, Rita K. Schmutzler, Christoph Engel, Hiltrud Brauch, Ute Hamann, Christina Justenhoven, GENICA Network, Kirsimari Aaltonen, Päivi Heikkilä, Kristiina Aittomäki, Carl Blomqvist, Keitaro Matsuo, Hidemi Ito, Hiroji Iwata, Aiko Sueta, Natalia V. Bogdanova, Natalia N. Antonenkova, Thilo Dörk, Annika Lindblom, Sara Margolin, Arto Mannermaa, Vesa Kataja, Veli-Matti Kosma, Jaana M. Hartikainen, kConFab Investigators, Anna H. Wu, Chiu-chen Tseng, David Van Den Berg, Daniel O. Stram, Diether Lambrechts, Stephanie Peeters, Ann Smeets, Giuseppe Floris, Jenny Chang-Claude, Anja Rudolph, Stefan Nickels, Dieter Flesch-Janys, Paolo Radice, Paolo Peterlongo, Bernardo Bonanni, Domenico Sardella, Fergus J. Couch, Xianshu Wang, Vernon S. Pankratz, Adam Lee, Graham G. Giles, Gianluca Severi, Laura Baglietto, Christopher A. Haiman, Brian E. Henderson, Fredrick Schumacher, Loic Le Marchand, Jacques Simard, Mark S. Goldberg, France Labrèche, Martine Dumont, Soo Hwang Teo, Cheng Har Yip, Char-Hong Ng, Eranga Nishanthie Vithana, Vessela Kristensen, Wei Zheng, Sandra Deming-Halverson, Martha Shrubsole, Jirong Long, Robert Winqvist, Katri Pylkäs, Arja Jukkola-Vuorinen, Mervi Grip, Irene L. Andrulis, Julia A. Knight, Gord Glendon, Anna Marie Mulligan, Peter Devilee, Caroline Seynaeve, Montserrat García-Closas, Jonine Figueroa, Stephen J. Chanock, Jolanta Lissowska, Kamila Czene, Daniel Klevebring, Nils Schoof, Maartje J. Hooning, John W. M. Martens, J. Margriet Collée, Madeleine Tilanus-Linthorst, Per Hall, Jingmei Li, Jianjun Liu, Keith Humphreys, Xiao-Ou Shu, Wei Lu, Yu-Tang Gao, Hui Cai, Angela Cox, Sabapathy P. Balasubramanian, William Blot, Lisa B. Signorello, Qiuyin Cai, Paul D. P. Pharoah, Catherine S. Healey, Mitul Shah, Karen A. Pooley, Daehee Kang, Keun-Young Yoo, Dong-Young Noh, Mikael Hartman, Hui Miao, Jen-Hwei Sng, Xueling Sim, Anna Jakubowska, Jan Lubinski, Katarzyna Jaworska-Bieniek, Katarzyna Durda, Suleeporn Sangrajang, Valerie Gaborieau, James McKay, Amanda E. Toland, Christine B. Ambrosone, Drakoulis Yannoukakos, Andrew K. Godwin, Chen-Yang Shen, Chia-Ni Hsiung, Pei-Ei Wu, Shou-Tung Chen, Anthony Swerdlow, Alan Ashworth,

- Nick Orr, Minouk J. Schoemaker, Bruce A. J. Ponder, Heli Nevanlinna, Melissa A. Brown, Georgia Chenevix-Trench, Douglas F. Easton, and Alison M. Dunning. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *American Journal of Human Genetics*, 92(4):489–503, April 2013. ISSN 1537-6605. doi: 10.1016/j.ajhg.2013.01.002.
- [118] Qilai Huang, Thomas Whittington, Ping Gao, Johan F. Lindberg, Yuehong Yang, Jielin Sun, Marja-Riitta Väisänen, Robert Szulkin, Matti Annala, Jian Yan, Lars A. Egevad, Kai Zhang, Ruizhu Lin, Arttu Jolma, Matti Nykter, Aki Manninen, Fredrik Wiklund, Markku H. Vaarala, Tapio Visakorpi, Jianfeng Xu, Jussi Taipale, and Gong-Hong Wei. A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nature Genetics*, 46(2):126–135, February 2014. ISSN 1546-1718. doi: 10.1038/ng.2862.
- [119] Fabian Lim, Genevieve E Ryan, Sophia H Le, Joe J Solvason, Paige Steffen, and Emma K. Farley. Affinity-optimizing variants within the ZRS enhancer disrupt limb development. *bioRxiv*. doi: <https://doi.org/10.1101/2022.05.27.493789>. URL <https://www.biorxiv.org/content/10.1101/2022.05.27.493789v1>.
- [120] Agnieszka D’Antonio-Chronowska, Matteo D’Antonio, and Kelly A. Frazer. In vitro Differentiation of Human iPSC-derived Cardiovascular Progenitor Cells (iPSC-CVPCs). *Bio-Protocol*, 10(18):e3755, September 2020. ISSN 2331-8325. doi: 10.21769/BioProtoc.3755.
- [121] Sarah De Val, Joshua P. Anderson, Analeah B. Heidt, Dustin Khiem, Shan-Mei Xu, and Brian L. Black. Mef2c is activated directly by Ets transcription factors through an evolutionarily conserved endothelial cell-specific enhancer. *Developmental Biology*, 275(2): 424–434, November 2004. ISSN 0012-1606. doi: 10.1016/j.ydbio.2004.08.016.
- [122] J. A. Nye, J. M. Petersen, C. V. Gunther, M. D. Jonsen, and B. J. Graves. Interaction of murine ets-1 with GGA-binding sites establishes the ETS domain as a new DNA-binding motif. *Genes & Development*, 6(6):975–990, June 1992. ISSN 0890-9369. doi: 10.1101/gad.6.6.975.
- [123] Maxwell A. Hume, Luis A. Barrera, Stephen S. Gisselbrecht, and Martha L. Bulyk. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res*, 43(Database issue): D117–122, January 2015. ISSN 1362-4962. doi: 10.1093/nar/gku1045.
- [124] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, April 2012. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.1923. URL <http://www.nature.com/articles/nmeth.1923>.
- [125] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp352. URL <https://academic.oup>.

com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352.

- [126] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoutte, David S Johnson, Bradley E Bernstein, Chad Nussbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008. ISSN 1465-6906. doi: 10.1186/gb-2008-9-9-r137. URL <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-9-r137>.
- [127] Christa Geeke Toenhake, Sabine Anne-Kristin Fraschka, Mahalingam Shanmugiah Vijayabaskar, David Robert Westhead, Simon Jan van Heeringen, and Richárd Bártfai. Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying Plasmodium falciparum Blood-Stage Development. *Cell Host & Microbe*, 23(4):557–569.e9, April 2018. ISSN 19313128. doi: 10.1016/j.chom.2018.03.007. URL <https://linkinghub.elsevier.com/retrieve/pii/S1931312818301367>.
- [128] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btq033. URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033>.
- [129] Yong Huan Song, Yu Ting Zhu, Jian Ding, Fei Ya Zhou, Ji Xin Xue, Jin Hee Jung, Zhi Jie Li, and Wei Yang Gao. Distribution of fibroblast growth factors and their roles in skin fibroblast cell migration. *Molecular Medicine Reports*, 14(4):3336–3342, October 2016. ISSN 1791-2997. doi: 10.3892/mmr.2016.5646. URL <https://www.spandidos-publications.com/10.3892/mmr.2016.5646>. Publisher: Spandidos Publications.
- [130] 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. *Cell Reports*, 42(2), February 2023. ISSN 2211-1247. doi: 10.1016/j.celrep.2023.112052. URL [https://www.cell.com/cell-reports/abstract/S2211-1247\(23\)00063-3](https://www.cell.com/cell-reports/abstract/S2211-1247(23)00063-3). Publisher: Elsevier.
- [131] Laura A. Lettice, Simon J.H. Heaney, Lorna A. Purdie, Li Li, Philippe de Beer, Ben A. Oostra, Debbie Goode, Greg Elgar, Robert E. Hill, and Esther de Graaff. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, 12(14):1725–1735, July 2003. ISSN 0964-6906. doi: 10.1093/hmg/ddg180. URL <https://doi.org/10.1093/hmg/ddg180>.
- [132] 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*, 584(7820):244–251, August 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2559-3. URL <https://www.nature.com/articles/s41586-020-2559-3>. Number: 7820 Publisher: Nature Publishing Group.
- [133] Hannah K. Long, Sara L. Prescott, and Joanna Wysocka. Ever-Changing Landscapes:

- Transcriptional Enhancers in Development and Evolution. *Cell*, 167(5):1170–1187, November 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.09.018. URL <https://www.sciencedirect.com/science/article/pii/S009286741631251X>.
- [134] Noboru Jo Sakabe, Daniel Savic, and Marcelo A. Nobrega. Transcriptional enhancers in development and disease. *Genome Biology*, 13(1):238, January 2012. ISSN 1474-760X. doi: 10.1186/gb-2012-13-1-238. URL <https://doi.org/10.1186/gb-2012-13-1-238>.
 - [135] Robert D. Riddle, Randy L. Johnson, Ed Laufer, and Cliff Tabin. Sonic hedgehog mediates the polarizing activity of the ZPA. *Cell*, 75(7):1401–1416, December 1993. ISSN 0092-8674. doi: 10.1016/0092-8674(93)90626-2. URL <https://www.sciencedirect.com/science/article/pii/0092867493906262>.
 - [136] Iain Williamson, Laura A. Lettice, Robert E. Hill, and Wendy A. Bickmore. Shh and ZRS enhancer colocalisation is specific to the zone of polarising activity. *Development*, 143(16):2994–3001, August 2016. ISSN 0950-1991. doi: 10.1242/dev.139188. URL <https://doi.org/10.1242/dev.139188>.
 - [137] Tomoko Sagai, Masaki Hosoya, Youichi Mizushima, Masaru Tamura, and Toshihiko Shiroishi. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development*, 132(4):797–803, February 2005. ISSN 0950-1991. doi: 10.1242/dev.01613. URL <https://doi.org/10.1242/dev.01613>.
 - [138] Laura A. Lettice, Paul Devenney, Carlo De Angelis, and Robert E. Hill. The Conserved Sonic Hedgehog Limb Enhancer Consists of Discrete Functional Elements that Regulate Precise Spatial Expression. *Cell Reports*, 20(6):1396–1408, August 2017. ISSN 2211-1247. doi: 10.1016/j.celrep.2017.07.037. URL <https://www.sciencedirect.com/science/article/pii/S2211124717310033>.
 - [139] Marco Osterwalder, Dario Speziale, Malak Shoukry, Rajiv Mohan, Robert Ivanek, Manuel Kohler, Christian Beisel, Xiaohui Wen, Suzie J. Scales, Vincent M. Christoffels, Axel Visel, Javier Lopez-Rios, and Rolf Zeller. HAND2 Targets Define a Network of Transcriptional Regulators that Compartmentalize the Early Limb Bud Mesenchyme. *Developmental Cell*, 31(3):345–357, November 2014. ISSN 1534-5807. doi: 10.1016/j.devcel.2014.09.018. URL [https://www.cell.com/developmental-cell/abstract/S1534-5807\(14\)00624-8](https://www.cell.com/developmental-cell/abstract/S1534-5807(14)00624-8). Publisher: Elsevier.
 - [140] Silvia Peluso, Adam Douglas, Alison Hill, Carlo De Angelis, Benjamin L Moore, Graeme Grimes, Giulia Petrovich, Abdelkader Essafi, and Robert E Hill. Fibroblast growth factors (FGFs) prime the limb specific Shh enhancer for chromatin changes that balance histone acetylation mediated by E26 transformation-specific (ETS) factors. *eLife*, 6:e28590, September 2017. ISSN 2050-084X. doi: 10.7554/eLife.28590. URL <https://doi.org/10.7554/eLife.28590>. Publisher: eLife Sciences Publications, Ltd.
 - [141] Terence D. Capellini, Giuseppina Di Giacomo, Valentina Salsi, Andrea Brendolan, Elisabetta Ferretti, Deepak Srivastava, Vincenzo Zappavigna, and Licia Selleri. Pbx1/Pbx2 requirement for distal limb patterning is mediated by the hierarchical control of Hox gene

- spatial distribution and Shh expression. *Development*, 133(11):2263–2273, June 2006. ISSN 0950-1991. doi: 10.1242/dev.02395. URL <https://doi.org/10.1242/dev.02395>.
- [142] H. Ohuchi, T. Nakagawa, A. Yamamoto, A. Araga, T. Ohata, Y. Ishimaru, H. Yoshioka, T. Kuwana, T. Nohno, M. Yamasaki, N. Itoh, and S. Noji. The mesenchymal factor, FGF10, initiates and maintains the outgrowth of the chick limb bud through interaction with FGF8, an apical ectodermal factor. *Development*, 124(11):2235–2244, June 1997. ISSN 0950-1991. doi: 10.1242/dev.124.11.2235. URL <https://doi.org/10.1242/dev.124.11.2235>.
- [143] V. Marigo, M. P. Scott, R. L. Johnson, L. V. Goodrich, and C. J. Tabin. Conservation in hedgehog signaling: induction of a chicken patched homolog by Sonic hedgehog in the developing limb. *Development*, 122(4):1225–1233, April 1996. ISSN 0950-1991. doi: 10.1242/dev.122.4.1225.
- [144] Jianjian Zhu, Eiichiro Nakamura, Minh-Thanh Nguyen, Xiaozhong Bao, Haruhiko Akiyama, and Susan Mackem. Uncoupling Sonic Hedgehog Control of Pattern and Expansion of the Developing Limb Bud. *Developmental Cell*, 14(4):624–632, April 2008. ISSN 1534-5807. doi: 10.1016/j.devcel.2008.01.008. URL <https://www.sciencedirect.com/science/article/pii/S1534580708000361>.
- [145] Keith Moore, Vid Persaud, and Mark Torchia. *The Developing Human. Clinically Oriented Embryology*. January 2013. ISBN 978-1-4377-2002-0.
- [146] P. Martin. Tissue patterning in the developing mouse limb. *Int J Dev Biol*, 34(3):323–336, September 1990. ISSN 0214-6282.
- [147] Laura A. Lettice, Taizo Horikoshi, Simon J. H. Heaney, Marijke J. van Baren, Herma C. van der Linde, Guido J. Breedveld, Marijke Joosse, Nurten Akarsu, Ben A. Oostra, Naoto Endo, Minoru Shibata, Mikio Suzuki, Eiichi Takahashi, Toshikatsu Shinka, Yutaka Nakahori, Dai Ayusawa, Kazuhiko Nakabayashi, Stephen W. Scherer, Peter Heutink, Robert E. Hill, and Sumihare Noji. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proceedings of the National Academy of Sciences*, 99(11):7548–7553, May 2002. doi: 10.1073/pnas.112212199. URL <https://www.pnas.org/doi/10.1073/pnas.112212199>. Publisher: Proceedings of the National Academy of Sciences.
- [148] Martin Kircher, Chenling Xiong, Beth Martin, Max Schubach, Fumitaka Inoue, Robert J. A. Bell, Joseph F. Costello, Jay Shendure, and Nadav Ahituv. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun*, 10(1):3583, August 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11526-w. URL <https://www.nature.com/articles/s41467-019-11526-w>. Number: 1 Publisher: Nature Publishing Group.
- [149] C. Fromental-Ramain, X. Warot, N. Messadecq, M. LeMeur, P. Dollé, and P. Chambon. Hoxa-13 and Hoxd-13 play a crucial role in the patterning of the limb autopod. *Development*, 122(10):2997–3011, October 1996. ISSN 0950-1991. doi: 10.1242/dev.122.10.2997.

- [150] S. Goodbourn, K. Zinn, and T. Maniatis. Human beta-interferon gene expression is regulated by an inducible enhancer element. *Cell*, 41(2):509–520, June 1985. ISSN 0092-8674. doi: 10.1016/s0092-8674(85)80024-6.
- [151] Carlos R. Escalante, Estanislao Nistal-Villán, Leyi Shen, Adolfo García-Sastre, and Aneel K. Aggarwal. Structure of IRF-3 Bound to the PRDIII-I Regulatory Element of the Human Interferon- β Enhancer. *Molecular Cell*, 26(5):703–716, June 2007. ISSN 1097-2765. doi: 10.1016/j.molcel.2007.04.022. URL <https://www.sciencedirect.com/science/article/pii/S1097276507002596>.
- [152] Martin Klar and Juergen Bode. Enhanceosome Formation over the Beta Interferon Promoter Underlies a Remote-Control Mechanism Mediated by YY1 and YY2. *Mol Cell Biol*, 25(22): 10159–10170, November 2005. ISSN 0270-7306. doi: 10.1128/MCB.25.22.10159-10170.2005. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1280260/>.
- [153] A. Raja Banerjee, Yoon Jung Kim, and Tae Hoon Kim. A novel virus-inducible enhancer of the interferon- β gene with tightly linked promoter and enhancer activities. *Nucleic Acids Research*, 42(20):12537–12554, November 2014. ISSN 0305-1048. doi: 10.1093/nar/gku1018. URL <https://doi.org/10.1093/nar/gku1018>.
- [154] Eitan Shaulian and Michael Karin. AP-1 in cell proliferation and survival. *Oncogene*, 20(19):2390–2400, April 2001. ISSN 1476-5594. doi: 10.1038/sj.onc.1204383. URL <https://www.nature.com/articles/1204383>. Number: 19 Publisher: Nature Publishing Group.
- [155] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, Peter A. C. ’t Hoen, Jean Monlong, Manuel A. Rivas, Mar González-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P. J. Buermans, Ismael Padioleau, Thomas Schwarzmayer, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy, Paul Flicek, Tim M. Strom, Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E. Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G. Gut, Xavier Estivill, and Emmanouil T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, September 2013. ISSN 1476-4687. doi: 10.1038/nature12531.
- [156] Enrico Cannavò, Pierre Khoeiry, David A. Garfield, Paul Geeleher, Thomas Zichner, E. Hilary Gustafson, Lucia Ciglar, Jan O. Korbel, and Eileen E. M. Furlong. Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Current Biology*, 26(1):38–51, January 2016. ISSN 0960-9822. doi: 10.1016/j.cub.2015.11.034. URL <https://www.sciencedirect.com/science/article/pii/S0960982215014281>.
- [157] 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 0960-9822. doi: 10.1016/j.cub.2010.07.043. URL <https://www.sciencedirect.com/science/article/pii/S0960982210009450>.
- [158] François Spitz and Eileen E M Furlong. Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics*, 13(9):613–26, 2012. ISSN 1471-0064. doi: 10.1038/nrg3207. URL <http://www.ncbi.nlm.nih.gov/pubmed/22868264>. Publisher: Nature Publishing Group.
- [159] Ann Boija, Isaac A. Klein, Benjamin R. Sabari, Alessandra Dall’Agnese, Eliot L. Coffey, Alicia V. Zamudio, Charles H. Li, Krishna Shrinivas, John C. Manteiga, Nancy M. Hannett, Brian J. Abraham, Lena K. Afeyan, Yang E. Guo, Jenna K. Rimel, Charli B. Fant, Jurian Schuijers, Tong Ihn Lee, Dylan J. Taatjes, and Richard A. Young. Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell*, 175(7):1842–1855.e16, December 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.10.042. URL <https://www.sciencedirect.com/science/article/pii/S0092867418313989>.
- [160] M. I. Coates and J. A. Clack. Polydactyly in the earliest known tetrapod limbs. *Nature*, 347(6288):66–69, September 1990. ISSN 1476-4687. doi: 10.1038/347066a0. URL <https://www.nature.com/articles/347066a0>. Number: 6288 Publisher: Nature Publishing Group.
- [161] Kimberly L. Cooper, Karen E. Sears, Aysu Uygur, Jennifer Maier, Karl-Stephan Baczkowski, Margaret Brosnahan, Doug Antczak, Julian A. Skidmore, and Clifford J. Tabin. Patterning and post-patterning modes of evolutionary digit loss in mammals. *Nature*, 511(7507):41–45, July 2014. ISSN 1476-4687. doi: 10.1038/nature13496. URL <https://www.nature.com/articles/nature13496>. Number: 7507 Publisher: Nature Publishing Group.
- [162] Martha L. Bulyk. Protein Binding Microarrays for the Characterization of DNA–Protein Interactions. In Harald Seitz, editor, *Analytics of Protein–DNA Interactions*, Advances in Biochemical Engineering/Biotechnology, pages 65–85. Springer, Berlin, Heidelberg, 2007. ISBN 978-3-540-48150-8. doi: 10.1007/10_025. URL https://doi.org/10.1007/10_025.
- [163] Prontip Saelee, Alyssa Kearly, Stephen L. Nutt, and Lee Ann Garrett-Sinha. Genome-Wide Identification of Target Genes for the Key B Cell Transcription Factor Ets1. *Frontiers in Immunology*, 8, 2017. ISSN 1664-3224. URL <https://www.frontiersin.org/articles/10.3389/fimmu.2017.00383>.
- [164] Sylvie Taveirne, Sigrid Wahlen, Wouter Van Looke, Laura Kiekens, Eva Persyn, Els Van Ammel, Katrien De Mulder, Juliette Roels, Laurentijn Tilleman, Marc Aumercier, Patrick Matthys, Filip Van Nieuwerburgh, Tessa C. C. Kerre, Tom Taghon, Pieter Van Vlierberghe, Bart Vandekerckhove, and Georges Leclercq. The transcription factor ETS1 is an important regulator of human NK cell development and terminal differentiation. *Blood*, 136(3):288–298, July 2020. ISSN 0006-4971. doi: 10.1182/blood.2020005204. URL <https://www.sciencedirect.com/science/article/pii/S0006497120618808>.
- [165] Anna C. McCarter, Giusy Della Gatta, Ashley Melnick, Erin Kim, Cher Sha, Qing

- Wang, Jahnavi K. Nalamolu, Yiran Liu, Theresa M. Keeley, Ran Yan, Mengxi Sun, Rohan Kodgule, Nicholas Kunnath, Alberto Ambesi-Impiombato, Rork Kuick, Arvind Rao, Russell J.H. Ryan, Barbara L. Kee, Linda C. Samuelson, Michael C. Ostrowski, Adolfo A. Ferrando, and Mark Y. Chiang. Combinatorial ETS1-Dependent Control of Oncogenic NOTCH1 Enhancers in T-cell Leukemia. *Blood Cancer Discovery*, 1(2):178–197, September 2020. ISSN 2643-3230. doi: 10.1158/2643-3230.BCD-20-0026. URL <https://doi.org/10.1158/2643-3230.BCD-20-0026>.
- [166] C. Wasylyk, A. Gutman, R. Nicholson, and B. Wasylyk. The c-Ets oncoprotein activates the stromelysin promoter through the same elements as several non-nuclear oncoproteins. *EMBO J*, 10(5):1127–1134, May 1991. ISSN 0261-4189.
- [167] 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. *EMBO J*, 27(14):2006–2017, July 2008. ISSN 1460-2075. doi: 10.1038/emboj.2008.117.
- [168] Gregory J. Fonseca, Jenhan Tao, Emma M. Westin, Sascha H. Duttke, Nathanael J. Spann, Tobias Strid, Zeyang Shen, Joshua D. Stender, Mashito Sakai, Verena M. Link, Christopher Benner, and Christopher K. Glass. Diverse motif ensembles specify non-redundant DNA binding activities of AP-1 family members in macrophages. *Nat Commun*, 10(1):414, January 2019. ISSN 2041-1723. doi: 10.1038/s41467-018-08236-0. URL <https://www.nature.com/articles/s41467-018-08236-0>. Number: 1 Publisher: Nature Publishing Group.
- [169] Daniel Panne, Tom Maniatis, and Stephen C. Harrison. An atomic model of the interferon-beta enhanceosome. *Cell*, 129(6):1111–1123, June 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.05.019.
- [170] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. ISSN 1476-4687. doi: 10.1038/nature11247.
- [171] Vincent Bertrand, Clare Hudson, Danielle Caillol, Cornel Popovici, and Patrick Lemaire. Neural Tissue in Ascidian Embryos Is Induced by FGF9/16/20, Acting via a Combination of Maternal GATA and Ets Transcription Factors. *Cell*, 115(5):615–627, November 2003. ISSN 00928674. doi: 10.1016/S0092-8674(03)00928-0. Publisher: Cell Press.
- [172] Géraldine Guillaume, Sophie de Buyl, Cathy Sirour, Nicolas Haupaix, Rossana Bettoni, Kaoru S. Imai, Yutaka Satou, Geneviève Dupont, Clare Hudson, and Hitoyoshi Yasuo. Cell geometry, signal dampening, and a bimodal transcriptional response underlie the spatial precision of an ERK-mediated embryonic induction. *Developmental Cell*, 56(21):2966–2979.e10, November 2021. ISSN 1534-5807. doi: 10.1016/j.devcel.2021.09.025. URL <https://www.sciencedirect.com/science/article/pii/S1534580721007693>.
- [173] Pierre Khoueiry, Ute Rothbächer, Yukio Ohtsuka, Fabrice Daian, Eric Frangulian, Agnès

- Roure, Inna Dubchak, and Patrick Lemaire. A cis-Regulatory Signature in Ascidians and Flies, Independent of Transcription Factor Binding Sites. *Current Biology*, 20(9):792–802, 2010. ISSN 09609822. doi: 10.1016/j.cub.2010.03.063.
- [174] Gwenael Badis, Michael F. Berger, Anthony A. Philippakis, Shaheynoor Talukder, Andrew R. Gehrke, Savina A. Jaeger, Esther T. Chan, Genita Metzler, Anastasia Vedenko, Xiaoyu Chen, Hanna Kuznetsov, Chi Fong Wang, David Coburn, Daniel E. Newburger, Quaid Morris, Timothy R. Hughes, and Martha L. Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723, 2009. ISSN 00368075. doi: 10.1126/science.1162327.
- [175] Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*, 12(10):931–934, October 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3547. URL <https://www.nature.com/articles/nmeth.3547>. Publisher: Nature Publishing Group.
- [176] Adam Klie, David Laub, James V. Talwar, Hayden Stites, Tobias Jores, Joe J. Solvason, Emma K. Farley, and Hannah Carter. Predictive analyses of regulatory sequences with EUGENE. *Nat Comput Sci*, 3(11):946–956, November 2023. ISSN 2662-8457. doi: 10.1038/s43588-023-00544-w. URL <https://www.nature.com/articles/s43588-023-00544-w>. Publisher: Nature Publishing Group.
- [177] Bernardo P. de Almeida, Christoph Schaub, Michaela Pagani, Stefano Secchia, Eileen E. M. Furlong, and Alexander Stark. Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo. *Nature*, 626(7997):207–211, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06905-9. URL <https://www.nature.com/articles/s41586-023-06905-9>. Number: 7997 Publisher: Nature Publishing Group.
- [178] Emily S. Wong, Dawei Zheng, Siew Z. Tan, Neil I. Bower, Victoria Garside, Gilles Vanwalleghe, 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):eaax8137, November 2020. doi: 10.1126/science.aax8137. URL <https://www.science.org/doi/10.1126/science.aax8137>. Publisher: American Association for the Advancement of Science.