Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht – Karls – University

Heidelberg

Presented by

Martina Varisco, M. Sc.

born in: Desio, Italy

Oral examination: July 19th, 2024

The impact of gene regulatory networks and genome topology on gene expression

Referees: Prof. Dr. Ingrid Lohmann

Dr. Oliver Stegle

I. Acknowledgements

I have always been fascinated by the 'Ph' in Ph.D. – *philosophiae*, 'love for knowledge'. A sacred quest to fill a tiny gap in our not so broad understanding of the world, just to reveal even more questions and to be surprised at how knowledge makes us aware of our ignorance, before moving on with the stubborn attitude of a curious and cheeky kid.

What my Ph.D. showed me is that in the P of Ph.D. there is much more than the passion for knowledge: the sense of pursuit, the urge to prove something and to prove myself, and most importantly the people who surrounded and supported me during this journey. Your scientific advice and mentorship, and all that we shared beyond the science, got me here, and I am privileged to have had you by my side.

First, I would like to express my deepest gratitude to Eileen Furlong for grasping my hunger for knowledge and giving me all the support and means to try to pursue it, and encouraging me to challenge my limits and to aim further than I could foresee. Thank you for guiding me in the research process and for helping me to grow into an independent researcher, while providing precious input to overcome the challenges in my projects and to open new directions when I could not find one.

I am very grateful to my thesis advisory committee members Prof. Dr. Ingrid Lohmann, Dr. Oliver Stegle, Dr. Justin Crocker and for their valuable advice and their inspiration during our yearly meetings and beyond. Sincere thanks also to Dr. Joachim Wittbrodt for joining in my defence examination committee.

I would like to thank the people who have worked most closely with me: thank you Gabriel Rodrigues Cavalheiro for helping me to get familiar with the techniques and the biology to investigate TADs and topology and for your friendship and the fun times, which brightened the long days in the lab. Thanks to Rebecca Rodriguez Viales and Charles Girardot for all your work on the TAD project, to Katharina Bender for your help in cloning and COVID times, to Christoph Schaub for sharing the hand-sorting sessions under the stereoscope, to Stefano Secchia for initiating me into the sci-ATAC-seq, and to Mattia Forneris and Simone Procaccia for your advice on data analysis. Thanks to Rosa Garcia Verdugo for the proof-reading and the feedback on my thesis.

To the entire Furlong lab, past and current members, thank you so much for being friends before lab mates, for sharing scientific discussion, advice and feedback, for creating a friendly atmosphere and for the small daily rituals like coffee and foosball. Alessandro, Valeria, Simone, Manos, Teresa, Baekgyu, Blanca, Ramile, Kurt, Babis, Karol, Mattia, Rebecca, Katharina, Christoph, Raquel, Matteo, Elena, Gabriel, Stefano, Olga, Adam, Tim, Yacine, Anna, it was and it is a pleasure to spend my time in the lab with you. A heartfelt thank you to Matteo, Babis, Gabriel and Valeria for supporting me through the challenges and pitfalls of my Ph.D., for cheering me up and for believing in me more than I would myself.

I would like to thank the EMBL core facilities for their valuable scientific support with a personal touch: thanks to Gene Core and especially to Vladimir Benes, to the Flow Cytometry Facility, to the Advanced Light Microscopy Facility and to Alessandra Reversi from the fly injection service.

Infine vorrei ringraziare chi mi ha sostenuto da lontano in questi anni e da vicino da molto più tempo: la mia famiglia, che mi ha sempre lasciato piena libertà nelle mie scelte e sostenuto nella mia crescita personale, permettendomi di cogliere nuove opportunità e restando sempre al mio fianco.

Grazie a te, Lorenzo, per non esserti scoraggiato davanti alla sfida della distanza durante questi anni di Ph.D., e per avermi accompagnato da lontano in questo percorso. Sono fortunata di poterti parlare non solo della mia giornata, ma anche di idee, di esperimenti e di progetti, e di poter condividere con te quel 'love for knowledge'. Grazie perché posso sempre contare su di te e su quanto tieni alla mia e alla nostra felicità.

II. Summary

Gene regulation is the result of the combined action of activators and repressors through a network of transcription factors (TFs) that bind to multiple enhancers, contained within topologically associating domains (TADs). Due to the high complexity of these systems, which contain many regulatory elements (including multiple enhancers, insulators, promoters) and highly interacting components in a multi-level organization, it is challenging to assess the molecular function of TFs, and how they interact with each other and with different regulatory elements, to give rise to cellular phenotypes. In my thesis, I employed perturbation approaches (in both *cis* and *trans*) to disentangle the contributions of individual TFs and regulatory elements to genome architecture and transcriptional output.

In my first project, I focused on the mesoderm gene regulatory network (GRN), i.e. the ensemble of TFs, enhancers and their connections responsible for the development of the *D. melanogaster* mesoderm. Mesodermal enhancers show extensive time- and tissue- specific co-binding of the TFs Twist (Twi), and of its direct targets Tinman (Tin) and Myocyte enhancer factor 2 (Mef2). By depleting each factor at the precise stages when they drive mesoderm specification and differentiation, and applying single-cell ATAC-seq, I assessed the effect of time-specific depletion on cell identity, and determined the enhancers and target genes most involved in the maintenance of the correct developmental trajectory. This revealed new relationships between different TFs that co-bind to mesodermal enhancers, in addition to an exciting discovery that one mutant leads to a conversion of cell fates across germ layers, from mesoderm to ectoderm/neuronal-like fate, which is extremely rare.

In my second project, I examined the elements contributing to the formation of TAD boundaries. In *Drosophila*, TAD boundaries have multiple insulator proteins binding in combination, and they generally coincide with transcriptional and epigenetic domains. Therefore, whether boundaries have intrinsic features driving their function, or if they result exclusively from the context, remained unclear. By inserting a selection of endogenous boundaries in ectopic locations inside other TADs, we showed that most boundaries had an effect, often orientation- or context- dependent, proving that boundaries have intrinsic features, but require contextual elements to function. Their orientation specificity implied that directional elements like promoters or motifs are required for boundary function, and that they need to be present on both sides, i.e. in both boundaries, to make a functioning pair. Since motifs of known insulators or promoters could not explain boundary pairing, I uncovered a set of conservation-based motif pairs that are associated with working boundary pairs with > 90% specificity and present at nearly 50% of endogenous genome-wide boundaries. This motif set includes known insulator motifs and motifs for TFs previously not associated with boundary function, in addition to new motifs with unknown binding proteins. The motifs in each pair have a preferred orientation, and many are convergent or divergent. This goes against the current dogma for how *Drosophila* TADs are formed, and suggests that the presence of oriented motifs for insulator proteins is a conserved feature to create functional boundary pairs from flies to humans.

In summary, using perturbation-based approaches, I could contribute to our functional understanding of two diverse, complex systems: I directly assessed the impact of single TFs to enhancer state, TF interactions, and their connection to cellular phenotype within the mesoderm GRN. We could also provide a new perspective on the logic governing what it takes to make a functioning TAD boundary pair in *Drosophila*, converging on the model of boundary formation through oriented motif pairs from mammals to *Drosophila*.

III. Zusammenfassung

Die Genregulation ist das Ergebnis der kombinierten Wirkung von Aktivatoren und Repressoren durch ein Netzwerk von Transkriptionsfaktoren (TF), die an mehrere Enhancer binden, die innerhalb topologisch assoziierender Domänen (TADs) enthalten sind. Aufgrund der hohen Komplexität dieser Systeme, die viele regulatorische Elemente (einschließlich mehrerer Enhancer, Isolatoren, Promotoren) und stark interagierende Komponenten in einer mehrstufigen Organisation enthalten, ist es herausfordernd, die molekulare Funktion von TF zu bewerten und wie sie miteinander und mit verschiedenen regulatorischen Elementen interagieren, um zelluläre Phänotypen hervorzubringen. In meiner Dissertation habe ich als Ansatz Störungen (sowohl in *cis* als auch in *trans*) eingesetzt, um die Beiträge einzelner TF und regulatorischer Elemente zur Genomarchitektur und zum transkriptionellen Output zu entwirren.

In meinem ersten Projekt konzentrierte ich mich auf das mesodermale genregulatorische Netzwerk (GRN), d. h. das Ensemble der TF, Enhancer und ihr Zusammenspiel, das verantwortlich ist für die Entwicklung des Mesoderms in *D. melanogaster*. Mesodermale Enhancer zeigen umfangreiches zeit- und gewebespezifisches simultanes Binden der TF Twist (Twi) sowie seiner direkten Ziele Tinman (Tin) und Myocyte enhancer factor 2 (Mef2). Durch Depletion jedes Faktors in den präzisen Stadien, in denen sie die Spezifikation und Differenzierung des Mesoderms vorantreiben, und durch Anwendung von Single-Cell ATAC-Seq habe ich die Auswirkung der zeitlich spezifischen Depletion auf die Zellidentität bewertet und die Enhancer und Zielgene bestimmt, die hauptsächlich an der Aufrechterhaltung der korrekten Entwicklungsablaufs beteiligt sind. Dies offenbarte neue Beziehungen zwischen den verschiedenen TF, die an mesodermale Enhancern binden, sowie eine aufregende Entdeckung, dass eine Mutation zu einer Umwandlung von Zellschicksalen über Keimschichten hinweg führt, von Mesoderm zu einem ektodermalen/neuronalen Schicksal, was äußerst selten ist.

In meinem zweiten Projekt untersuchte ich die Elemente, die zur Bildung von TAD-Grenzen beitragen. In *Drosophila* haben TAD-Grenzen mehrere Isolatorproteine, die in Kombination binden, und sie fallen im Allgemeinen mit transkriptionellen und epigenetischen Domänen zusammen. Daher blieb unklar, ob Grenzen immanente Merkmale aufweisen, die ihre Funktion vorantreiben, oder ob sie ausschließlich aus dem Kontext resultieren. Durch Einfügen einer Auswahl endogener Grenzen an ektope Standorte innerhalb anderer TAD zeigten wir, dass die meisten Grenzen eine Wirkung hatten, die oft von der Orientierung oder dem Kontext abhängig war, was bewies, dass Grenzen immanente Merkmale haben, aber kontextabhängige Elemente benötigen, um zu funktionieren. Ihre Orientierungsspezifität deutete darauf hin, dass gerichtete Elemente wie Promotoren oder Motive für die Grenzfunktion erforderlich sind und dass sie auf beiden Seiten, d. h. in beiden Grenzen, vorhanden sein müssen, um ein funktionierendes Paar zu bilden. Da Motive bekannter Isolatoren oder Promotoren die Paarbildung von Grenzen nicht erklären konnten, identifizierte ich einen Satz von Motiven basierend auf Konservation, die mit funktionierenden Grenzpaaren mit einer Spezifität von > 90 % assoziiert und an fast 50 % der endogenen genomweiten Grenzen vorhanden sind. Dieser Satz an Motiven umfasst bekannte Isolatormotive und Motive für TF, die zuvor nicht mit der Grenzfunktion in Verbindung gebracht

wurden, sowie neue Motive mit unbekannten Bindungsproteinen. Die Motive in jedem Paar haben eine bevorzugte Orientierung, und viele sind konvergent oder divergent. Dies widerspricht dem aktuellen Dogma, wie *Drosophila*-TAD gebildet werden, und legt nahe, dass die Anwesenheit orientierter Motive für Isolatorproteine ein konserviertes Merkmal ist, um funktionale Paare von Grenzen von Fliegen bis Menschen zu schaffen.

Zusammenfassend konnte ich mit Ansätzen, die auf Störungen basieren, zu unserem funktionalen Verständnis zweier unterschiedlicher, komplexer Systeme beitragen: Ich habe direkt den Einfluss einzelner TFs auf den Status von Enhancern, TF-Interaktionen und deren Verbindung zum zellulären Phänotyp innerhalb des mesodermalen GRN bewertet. Wir konnten auch eine neue Perspektive auf die Logik geben, die erforderlich ist, um ein funktionierendes TAD-Grenzpaar in *Drosophila* zu bilden, in Übereinstimmung mit einem Modell für die Grenzbildung durch orientierte Motivpaare von Säugetieren bis *Drosophila*.

Table of contents

1 Introduction 19

1.1 Multiple elements contribute to transcriptional outputs 19

1.1.1 Enhancers 19

1.1.2 Silencers 21

1.1.3 Insulators 22

1.1.4 Promoters 22

1.1.4.1 Enhancer specificity is achieved through promoter- and insulator- dependent mechanisms 24

1.2 Gene regulatory networks integrate multiple inputs to achieve robust transcriptional control during development 25

1.2.1 A historical perspective on cis-regulatory networks 28

1.3 A primer on *Drosophila* development 30

1.4 The mesoderm GRN as a functional model 31

1.4.1 Mesoderm TFs have spatio-temporally restricted expression 32

1.4.2 Mesoderm TFs bind to different enhancers with different combinations of TFs across time and tissues 36

1.5 Limitations of current approaches to understanding the mesoderm GRN 37

1.5.1 Insights from KO or KD mutants 37

1.5.2 Insights from bulk and single-cell genome-wide assays 39

1.6 Fast perturbations and single-cell readouts could overcome current limitations 40

1.6.1 The iLEXY system enables fast stage-specific depletion of nuclear TFs 40

1.6.2 Molecular changes in the GRN over development can be captured by bulk and single-cell readouts 42

1.7 The genome is hierarchically organized in three dimensions 43

1.8 TADs are functionally important structures in 3D genome organization 44

1.8.1 TADs are associated with gene regulation and chromatin states 45

1.8.2 The functional importance of TADs is linked to the properties of the insulator proteins at their boundaries 47

1.8.3 Mammalian TADs are marked by convergent CTCF sites and formed by loop extrusion 49

1.8.4 Drosophila TAD boundaries show a combination of insulator proteins 51

1.8.5 Mechanisms other than loop extrusion could establish Drosophila TAD boundaries 52

1.8.6 TAD formation may rely on orthogonal evolutionary mechanisms 53

1.8.7 The CTCF-cohesin loop: far from a universal model 54

1.8.8 TAD prediction: learning the cause or the effect of 3D organization? 55

1.9 Aims 56

2 Materials and methods 57

2.1 Plasmid design and cloning 57

2.1.1 Generation of TF-iLEXY mutants by CRISPR cloning 57

2.1.2 DNA FISH probe cloning 57

2.1.3 TAD boundary cloning 58
2.2 Drosophila genetics and husbandry 58
2.2.1 CRISPR/Cas-mediated homology-directed repair 58
2.2.2 TAD boundary insertions 59
2.2.3 Handling of photosensitive embryos 59
2.2.4 Sensitized background 59
2.3 Genotyping 60
2.3.1 Dirty gDNA extraction for single-fly genotyping 60
2.3.2 Clean gDNA extraction for genotyping 60
2.3.3 Genotyping the orientation of TAD boundary insertions 60
2.3.4 Genotyping the identity and location of TAD boundary insertions 60
2.4 iLEXY depletion 61
2.4.1 Photoactivation in the LED illumination box 61
2.5 Embryo collection 61
2.5.1 Embryo collection for immunostaining and DNA FISH 61
2.5.2 Embryo collection for CUT&Tag, sci-ATAC-seq and capture Hi-C 61
2.5.3 Embryo collection of iLEXY lines 62
2.5.4 Hand sorting of iLEXY embryos with sensitized background 62
2.6 DNA Fluorescence *In Situ* Hybridization (DNA FISH) 62
2.6.1 Probe preparation 62
2.6.2 DNA FISH 63
2.6.3 Combined DNA/RNA FISH 63
2.6.4 Image acquisition 64
2.6.5 Quantification of 3D distance 64
2.7 RNA FISH 65
2.8 Immunostaining 66
2.8.1 Immunohistochemistry 66
2.8.2 Image acquisition 67
2.9 Nuclear isolation 67
2.9.1 Small scale nuclear isolation for CUT&Tag and sci-ATAC-seq 67
2.9.2 Large scale nuclear isolation for Hi-C 68
2.10 Spike-in CUT&Tag 68
2.10.1 Generation of spike-in CUT&Tag libraries 68
2.10.2 Spike-in CUT&Tag data analysis 69
2.11 sci-ATAC-seq 70
2.11.1 Generation of Tn5 transposomes 70
2.11.2 Generation of sci-ATAC-seq libraries 71
2.11.3 Library clean-up and sequencing 71
2.12 sci-ATAC-seq data analysis 72
2.12.1 Raw sequencing data processing, mapping and duplicate removal 72

2.12.2 Count matrix generation, LDA, clustering and UMAP 72
2.12.3 Gene activity matrix generation 73
2.12.4 Cell type annotation 73
2.12.5 Mesoderm reclustering 73
2.12.6 Calculation of TF deviation scores 73
2.12.7 Identification of differential ATAC peaks and genes 74
2.12.8 Integration with WT mesoderm time course 74
2.12.9 Analysis of DA peaks 75
2.12.10 Analysis of DA genes 75
2.13 Capture Hi-C 76
2.13.1 Design of capture Hi-C probes 76
2.13.2 Generation of capture Hi-C libraries 76
2.14 Capture Hi-C data analysis 78
2.14.1 Read separation 78
2.14.2 Capture Hi-C data processing 79
2.14.3 Quality control 80
2.14.4 TAD visualization 80
2.15 Identification and annotation of TAD boundary pairs 80
2.15.1 Manual annotation of TAD boundaries and boundary pairs from WT Hi-C data 80
2.15.2 Identification of boundary pairs by virtual 4C from capture Hi-C data 80
2.15.3 Annotation of features of TAD boundary pairs 81
2.16 Evolutionary conservation of motifs in TAD boundaries 82
2.16.1 Alignment of inserted and acceptor TAD boundaries across the D. melanogaster clade 82
2.16.2 Motif search and annotation in aligned boundaries 82
2.16.3 Finding motifs linked to working TAD boundaries in the experimental and genome-wide dataset 83
3 Results I 85
3.1 Motivation 85
3.1.1 Single-cell approaches overcome sensitivity limitations in the detection of aberrant developmental states 85
3.1.2 Time-resolved perturbations increase time resolution and prevent indirect effects 85
3.2 Using iLEXY to uncover the regulatory network of mesoderm development 86
3.3 Rationale for the choice of depletion and collection time points 87
3.4 iLEXY-tagged mesoderm TFs are depleted from the nucleus and from chromatin upon optogenetic stimulation 88
3.5 Depletion over the whole of embryogenesis partially recapitulates the loss-of-function mutant phenotypes 94
3.6 Investigating the contributions of TFs to cell fate over time by sci-ATAC-seq 97
3.6.1 Mef2 98
3.6.1.1 Mef2 depletion drives more than half of mesoderm cells towards two mutant cell states 98
3.6.1.2 Half of the mesoderm cells deviate from the wild-type trajectory upon Mef2 depletion 98

3.6.1.3 Mef2 depletion leads to two distinct mutant states 102
3.6.1.4 Mutant states do not regress to previous developmental state, but branch off the WT developmental trajectory at the right time 105
3.6.1.5 Accessibility is mostly reduced at Mef2-bound regions 109
3.6.1.6 Accessibility is mostly affected at co-bound sites 111
3.6.1.7 Affected enhancers and target genes mark the loss of mesodermal identity 112
3.6.2 Twi 117
3.6.2.1 Twi depletion obliterates the mesoderm, redirecting cells to an aberrant epidermis-like state, even after gastrulation 117
3.6.2.2 Twist mutants form a single cluster separated from the WT mesoderm trajectory 119
3.6.2.3 Loss of accessibility is prominent at Twi-bound sites 122
3.6.2.4 Twist depletion has many indirect effects, especially at the level of differential genes 126
3.6.3 Relationship between changes in TF binding, accessibility and transcription 128
3.6.3.1 Mef2 128
3.6.3.2 Twi 129
4 Discussion I 131
4.1 Mutant phenotypes in sc-ATAC-seq 131
4.2 Differential peaks and genes are consistent with Twi and Mef2 roles 132
4.3 Stage-specific perturbation dissects TF activity at precise time-points, enriching for direct effects 133
4.4 Incomplete depletion: from a limitation to a resource 134
4.4.1 Incomplete depletion in bulk may result from cell-cell heterogeneity, which can be disentangled in single-cell data 134
4.4.2 Incomplete depletion can be leveraged to infer enhancer affinity 135
4.5 The combination of binding and accessibility data suggests how TF perturbation impacts enhancer accessibility and transcription 136
4.6 Future directions 137
5 Results II 139
5.1 Genome-wide TAD boundaries show different insulator and promoter combinations 140
5.2 Rationale for the choice of inserted boundaries 142
5.3 Rationale for the choice of acceptor TADs 142
5.3.1 TAD boundary composition 143
5.3.2 Structural complexity 143
5.3.3 Gene regulation 143
5.4 Properties of acceptor TADs 143
5.4.1 TAD A 143
5.4.2 TAD B 145
5.5 Generation of fly lines with boundary insertions 145
5.6 Viability of boundary insertions 146
5.7 DNA FISH shows that most inserted boundaries have context- or orientation- specific effects 147
5.8 Capture Hi-C reveals insight into the topological changes induced by the new boundaries 151

5.9 Well characterised insulators are not predictive of the inserted boundaries effect 153

5.10 The promoter does not establish the orientation specificity of inserted boundaries 156

5.11 Combinations of well characterised insulators binding to the inserted boundaries and endogenous boundaries are not predictive of a working boundary pair 157

5.12 Other known insulator features do not explain boundary function 158

5.13 Using conservation to identify new motifs in functional boundaries 159

5.14 Known and new motifs are retrieved among conserved sequences in inserted and acceptor TAD boundaries 161

5.15 Pairs of conserved sequences in a defined orientation are linked to working TAD boundaries in our experiment and genome-wide 162

5.16 The presence of non-oriented sequences is not sufficient to infer rules 168

5.17 Motif directionality explains most experimental boundaries 169

5.18 Three-way interactions contribute to TAD boundaries to a small extent 170

5.19 Motifs involved in boundary function include known insulators and transcriptional regulators 171

6 Discussion II 173

6.1 Pros and cons of using an endogenous setting 173

6.1.1 Why use endogenous boundaries? 173

6.1.2 Why use an endogenous landing site? 174

6.2 Readouts by DNA FISH and capture Hi-C focus on different aspects of the interaction and they are highly concordant 174

6.3 The orientation specificity of some boundaries points to a role of directional motifs in boundary pairing 175

6.4 Boundary pairs are not just consecutive boundaries 176

6.5 Experimental perturbations facilitate detection of features of boundary pairs by restricting the feature search space 176

6.6 Conservation reveals new motifs associated to boundary pairs 177

6.7 Limitations of the assignment of conserved sequences to TF motifs 178

6.8 Future directions 178

7 Conclusion and future perspectives 181

7.1 Prediction of gene transcription 181

7.2 Prediction of enhancer-to-gene connections 182

7.3 Prediction of TADs 182

7.4 How experimental data can inform predictions and vice versa 183

8 Appendix 185

9 References 193

List of figures

Figure 1: Multiple enhancers can control the same target, and a single enhancer may regulate multiple genes. 20
Figure 2: Chromatin marks at inactive, primed and active enhancers. 21
Figure 3: Structure of the core promoter. 23
Figure 4: GRNs integrate numerous inputs in cis and in trans. 26
Figure 5: TF-TF interactions at co-bound enhancers and transcriptional output. 27
Figure 6: Mesoderm development in D. melanogaster embryos. 31
Figure 7: Activation of the mesoderm GRN in time and tissue. 33
Figure 8: tin enhancers. 34
Figure 9: Mef2 enhancers. 35
Figure 10: The iLEXY system enables rapid and reversible TF nuclear export. 41
Figure 11: 3D genome organization. 44
Figure 12: TADs contribute to gene regulation by modulating E-P contacts, and they are associated to coregulation and chromatin states. 46
Figure 13: Insulator bypass. 48
Figure 14: Features of TAD boundaries in mammals and in Drosophila. 51
Figure 15: Timeline of the iLEXY depletion. 87
Figure 16: iLEXY-tagged TFs are depleted from the nucleus upon optogenetic stimulation. 90
Figure 17: Bin has higher retention in FCs. 90
Figure 18: iLEXY-tagged TFs are depleted from the chromatin upon optogenetic stimulation. 92
Figure 19: The iLEXY system partially recapitulates the phenotype of the LOF mutants upon optogenetic depletion for the whole embryogenesis. 96
Figure 20: The LVM fails to form upon Bin depletion. 96
Figure 21: tin-iLEXY homozygous mutants have mild defects. 97
Figure 22: Mef2-iLEXY mutants segregate in a central cluster in the whole-embryo dataset. 99
Figure 23: After mesoderm reclustering, Mef2-iLEXY mutants are separated into two clusters deviating from the normal mesoderm trajectory. 101
Figure 24: Both the Mef2 mutant clusters have increased accessibility at target sites of TF linked to gastrulation and specification of equivalence clusters, and they have specific enrichment for other regions. 103
Figure 25: Stage-specific clustering of Mef2 mutants shows that the mutant clusters Mut-Lexy-1 and Mut-Lexy2 are present both at stage 9 and stage 11, and more abundant at stage 11. 105
Figure 26: Comparison of the Mef2-iLEXY dataset to the WT and Mef2 KO dataset from Secchia et al. (2022). 108
Figure 27: Comparison of the Mef2-iLEXY dataset to the WT Calderon et al. (2022) dataset. 109
Figure 28: The Mef2-iLEXY mutant clusters mostly have reduced accessibility, especially at Mef2-bound sites. 110
Figure 29: Sites with reduced accessibility are mostly co-bound by multiple mesoderm TFs. 112

Figure 30: Sites with reduced accessibility are enriched in putative muscle enhancers. 113
Figure 31: DA genes are mostly downregulated and involved in development. 115
Figure 32: Mut-Lexy-1 has more indirect effects than Mut-Lexy-2. 116
Figure 33: twi-iLEXY mutants are depleted in the mesoderm and move towards an epidermal-like identity; they are also enriched in the nervous system and plasmatocytes. 117
Figure 34: After mesoderm reclustering, twi-iLEXY mutants cannot be fully separated from the mesoderm. 118
Figure 35: Comparison of the Twi-iLEXY dataset to the WT dataset from Secchia et al. (2022). 120
Figure 36: Comparison of the Twi-iLEXY dataset to the WT dataset from Cusanovich et al. (2018). 122
Figure 37: twi-iLEXY mutants have similar number of peaks with increased and decreased accessibility, and the down DA peaks are associated with strong Twi binding sites. 124
Figure 38: Sites with lower accessibility are enriched in putative muscle enhancers. 125
Figure 39: DA genes are mostly upregulated and enriched for general development and nervous system terms. 127
Figure 40: The selection of inserted boundaries recapitulates the diversity of TAD boundaries genome-wide. 141
Figure 41: Acceptor TADs are diverse in their architecture, boundary composition and gene regulation. 144
Figure 42: Most inserted boundaries have an effect, which can be orientation- and context- specific. 148
Figure 43: DNA FISH, capture Hi-C and virtual 4C for boundaries with different effects. 151
Figure 44: TAD disruption can be combined with loop strengthening, leading to no detectable change in DNA FISH. 153
Figure 45: The binding of insulator proteins does not explain the effect of inserted boundaries. 154
Figure 46: The pairing of boundaries with specific insulator proteins combinations does not explain boundary effect. 158
Figure 47: Conservation-based de novo motif discovery. 160
Figure 48: De novo motifs include known and new motifs enriched at genome-wide TAD boundaries and in working experimental boundaries. 161
Figure 49: Examples and composition of oriented motif pairs. 163
Figure 50: Motif pairs explain the effect of experimentally inserted boundaries and they are found in ~50% of genome-wide boundaries. 169
Figure 51: Combinations of 3 motifs in the left and right TAD boundary. 170
Figure 52: PWM of the two new conserved motifs. 171

1 Introduction

Cell survival and differentiation result from a well-concerted regulation of protein products throughout their whole life cycle, from gene expression to their translation, function and degradation. In highly dynamic systems such as the developing embryo, transcription sets the pace for cell differentiation, tightly controlling cell abundance and identity to ensure their stability, and consequently organism viability. Robust, precise gene expression is achieved by cumulative regulation at different levels: from the nature of the promoter, the set of enhancers and silencers acting on the promoter's activity, the regulatory connections between transcription factors (TFs) and their target sequences, and hierarchical activation of transcriptional targets, to higher-order constraints and genome topology. Although there is an extensive knowledge base on the enhancers and TFs involved in the regulation of gene expression, the causal link between their binding or co-binding and their impact on transcription remains poorly understood. Understanding the role of genome topology poses even bigger challenges, as the factors involved and their functional requirement are still under debate.

In my PhD, I investigated the functional contribution of enhancers, promoters and TFs to mesoderm development, using *Drosophila melanogaster* embryos as a model system. Moreover, I assessed what it takes to make a topological chromatin domain, quantifying the ability of genomic elements and their interactions to establish and maintain genome topology.

1.1 Multiple elements contribute to transcriptional outputs

The cell state, comprising its developmental identity, homeostasis and response to external and internal cues, is reflected in gene expression, which in turn relies on the contribution of multiple elements in, and binding to, a particular DNA sequence. Genomic regions which regulate gene expression, such as enhancers, silencers and insulators, are known as *cis*-regulatory modules (CRMs). Specifically, CRMs exert their action by recruiting different sets of TFs to drive transcriptional activation, repression or insulation.

1.1.1 Enhancers

Enhancers are sequences that can increase the transcriptional output of a gene irrespective of their distance, their orientation with respect to the gene's promoter, and their genomic location, for instance in the intron of another gene (Banerji, Rusconi and Schaffner, 1981; Moreau *et al.*, 1981).

Genes typically have many enhancers, which can be responsible for transcriptional activation in different tissues and at various timepoints, showcasing a modular control of gene expression, or be partially redundant by activating expression in the same tissue and at the same time (Cannavò *et al.*, 2016; Panigrahi and O'Malley, 2021). Stress conditions like elevated temperature can unmask the need for redundant enhancers, indicating a possible role in stabilizing gene expression (Frankel *et al.*, 2010; Tsai, Alves and Crocker, 2019). Overall, enhancer multiplicity ensures versatility and robustness of gene expression.

A classic example of this modular and combinatorial organization is the *even skipped* (*eve*) locus in flies: *eve* is a pair-rule gene expressed in seven vertical stripes after late stage 5 and it is involved in the segmentation of *Drosophila* embryos (Fujioka *et al.*, 1999) (Figure 1A). Its expression is driven by different enhancers in each domain, with each enhancer driving expression in 1 or 2 stripes. The single enhancers are bound by different sets of transcription factors (TFs), and the spatiotemporal confinement of TF expression controls the pattern of enhancer activation. The activity domain of each enhancer is further sharpened by the combined action of repressors and activators, confining *eve* expression to a small and reproducible number of cell rows.

On the other hand, to guarantee coordinated activation at developmental transitions or maintain a steady-state, the same enhancer may control multiple targets (Spitz and Furlong, 2012; Furlong and Levine, 2018). Such a shared enhancer may not be equally critical to all its targets: for instance, the shared intergenic enhancer of the paralogous genes *bab1* and *bab2* in the bric-à-brac complex is partly redundant for *bab1* expression in the leg, but essential for *bab2* expression (Bourbon *et al.*, 2022) (Figure 1B).

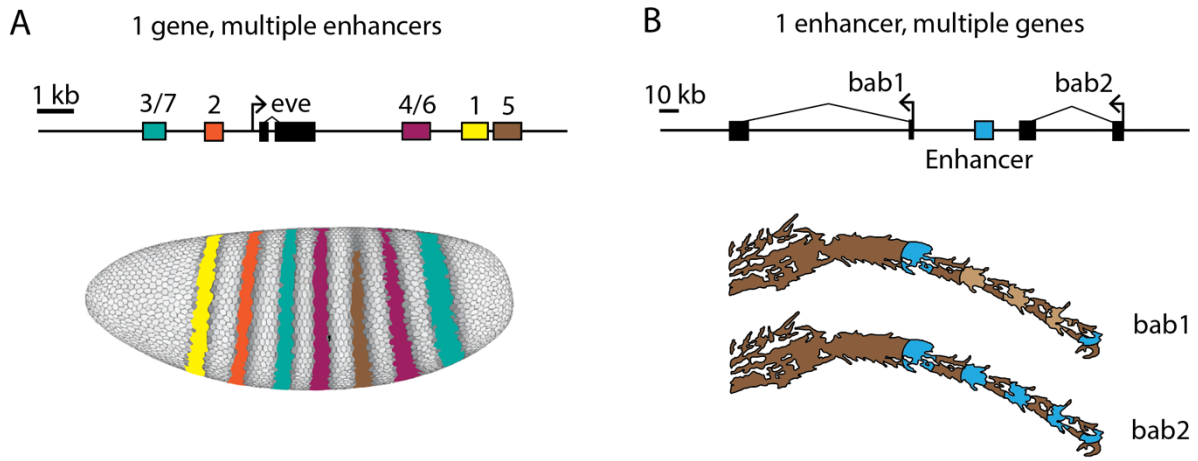


Figure 1: Multiple enhancers can control the same target, and a single enhancer may regulate multiple genes.

(A) Multiple enhancers regulate the expression of *eve* in different stripes. Schematic representation of the *eve* locus, with the *eve* gene in black and its enhancers in different colours, and of a *D. melanogaster* embryo with the eve stripes in the same colour as the enhancers that control them (Fujioka *et al.*, 1999). Adapted from Samee *et al.* (2017).

(B) In the *bab1-bab2* locus, a single enhancer regulates two distinct genes. Schematic representation of the *bab1-bab2* locus with the genes in black and the enhancer in blue, and of the adult *D. melanogaster* leg. *bab1* and *bab2* are expressed in all the joints; the domain where the enhancer is necessary for the expression of each gene is coloured in blue, and where it is redundant in light brown. Adapted from Bourbon *et al.* (2022).

To mediate transcriptional regulation, enhancers recruit multiple TFs via their TF binding sites (TFBS) or motifs, short DNA sequences with a precise set of bases recognized by the DNA-binding domain of each TF (Spitz and Furlong, 2012). Due to the high density of bound proteins, active enhancers are depleted from nucleosomes and highly accessible (He *et al.*, 2010), and thus easily detectable by chromatin accessibility profiling techniques such as DNAse I hypersensitivity (Thurman *et al.*, 2012) or ATAC-seq (Buenrostro *et al.*, 2013), based on enzymes – endonucleases and transposases – capable of cutting only in nucleosome-free DNA regions. Most TFs can only bind to already accessible

enhancers, but TFs with pioneering activity can bind to their motif within nucleosomes, open up the chromatin and establish accessibility to allow successive binding of other TFs (Zaret and Carroll, 2011). In addition, priming, an intermediate state between TF binding and enhancer activation where the enhancer is bound by one or more factor but not active, has been described in a number of developmental contexts (Rada-Iglesias *et al.*, 2011; Spitz and Furlong, 2012). The primed enhancer is typically activated at some subsequent stage of embryogenesis when the required additional TF(s) becomes expressed (Spitz and Furlong, 2012; A. Wang *et al.*, 2015).

Enhancers have different chromatin marks according to their activity: active enhancers are marked by deposition of the chromatin mark H3K27ac, inactive enhancers by H3K27me3, and primed enhancers by both, pointing to a bivalent chromatin state (Creyghton *et al.*, 2010; Rada-Iglesias *et al.*, 2011; Bonn, Zinzen, Girardot, *et al.*, 2012) (Figure 2).

TF-bound enhancers mechanistically activate their target gene by coming into physical proximity to their promoter and by modulating RNA polymerase II (Pol II) transcriptional bursting through a transient interaction (Fukaya, Lim and Levine, 2016), where the enhancer is displaced from the promoter by the nascent RNA (hit-and-run and kiss-and-kick model) (Benabdallah *et al.*, 2019; Pownall *et al.*, 2023).

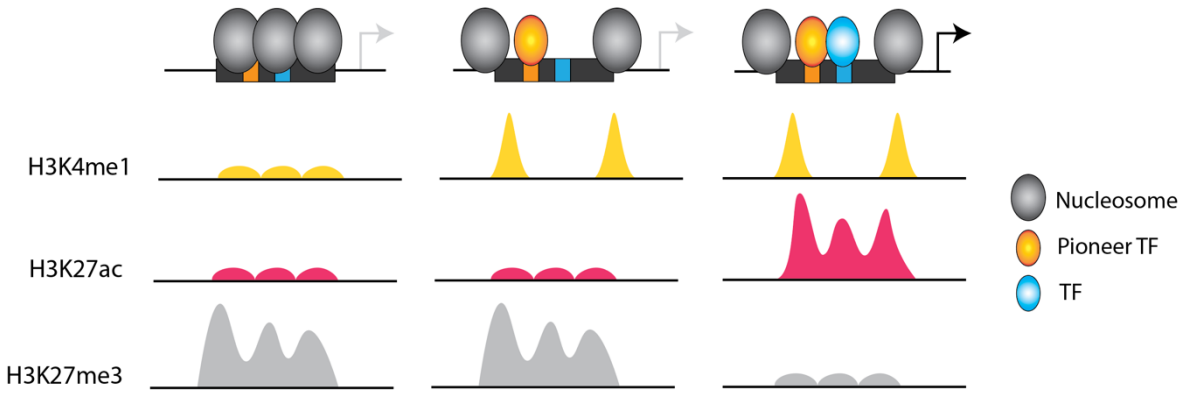


Figure 2: Chromatin marks at inactive, primed and active enhancers.
Schematic representation of inactive, primed and active enhancers (top, left to right) and associated genomic tracks of chromatin marks H3K4me1, H3K27ac, H3K27me3.

1.1.2 Silencers

Similar to enhancers, silencers are distance-, position- and orientation- independent elements with multiple TF binding sites but responsible for transcriptional repression (Abraham *et al.*, 1984; reviewed in Pang *et al.*, 2023). However, in comparison to enhancers, silencers are less well understood. Given their extensive protein binding and open chromatin state, silencers cannot be distinguished from enhancers based on chromatin accessibility assays, and require an additional transcriptional readout or the detection of chromatin marks (Pang *et al.*, 2023). The mechanism whereby silencers affect transcription may be through direct contact with the transcriptional machinery at the promoter, similar

to enhancers (Courey and Jia, 2001; Friedman *et al.*, 2021), or by epigenetic modifications which suppress transcription (Huang *et al.*, 2019).

1.1.3 Insulators

Insulators are elements which restrict the range of action of enhancers and silencers, and establish domains of gene regulation (Wallace and Felsenfeld, 2007; Özdemir *et al.*, 2019). They were first described in *D. melanogaster*: early observations showed that a chromosomal deletion between two enhancer domains that drove expression of the Hox gene *Abdominal-B* (*abd-B*) in separate segments caused aberrant *abd-B* activation, as if the enhancers could not act independently (Gyurkovics *et al.*, 1990). Other hints came from the studies on the *gypsy* transposable element, which could prevent gene expression in a selected tissue if integrated between the gene promoter and a distal tissue-specific enhancer (Peifer and Bender, 1986; Jack *et al.*, 1991; Geyer and Corces, 1992). The first loci tested directly for insulator function using a reporter assay were the *specialized chromosome structure* (*scs*) and *scs'* (Udvardy, Maine and Schedl, 1985; Kellum and Schedl, 1991, 1992), which flank the *87A* heat-shock locus and were first thought to locally restrict DNA decompaction upon heat-shock, and later to limit regulatory interaction within this locus (Kuhn, Hart and Geyer, 2004). Insulators were later discovered in other species, starting from the locus control region *H19* (Lyko *et al.*, 1997) in human and *5'HS4* in the chicken *beta-globin* domain (Chung, Whiteley and Felsenfeld, 1993). In such contexts, insulators blocked E-P communication (Kellum and Schedl, 1991, 1992), or prevented the spreading of chromatin marks (Kahn *et al.*, 2006; Fujioka, Sun and Jaynes, 2013; Bowman *et al.*, 2014).

The function of insulators, the insulator proteins and their role in 3D genome topology will be discussed more in detail in sections 1.8.2 and 1.8.4.

1.1.4 Promoters

Promoters are located at the 5' end of each gene, around the transcription start site (TSS), which is the precise position where transcription starts. The core promoter is defined as a DNA sequence from about 50 bp upstream to about 50 bp downstream of the TSS (Haberle and Stark, 2018). It may include multiple motifs (reviewed in Juven-Gershon and Kadonaga, 2010), among which the main ones are (Figure 3A):

- The initiator motif (Inr), which encompasses the TSS and is the most frequent core promoter motif (Ohler *et al.*, 2002; FitzGerald *et al.*, 2006). Its activity is correlated to the binding of the general transcription factor TFIID, and its consensus sequence is TCAKTY in *Drosophila* (Purnell, Emanuel and Gilmour, 1994), and YYANWYY in mammals (Chalkley and Verrijzer, 1999). The third nucleotide (A) in the consensus sequence is conventionally indicated as the +1 position (Burke and Kadonaga, 1997)

- The TATA motif, which is conserved in all metazoans and has the consensus motif TATAWAAR. It is usually located at -31/-30 bp relative to the +1 position (Carninci *et al.*, 2006), and bound by the TATA-binding protein subunit of TFIID (Horikoshi *et al.*, 1988, 1989)
- The Downstream core Promoter Element (DPE), which is a TFIID recognition site starting at +28 from the A +1 position (Burke and Kadonaga, 1997), conserved from *Drosophila* to human (Burke and Kadonaga, 1997). It is generally mutually exclusive with the TATA box (Juven-Gershon and Kadonaga, 2010).

The core promoter recruits general transcription factors (among which TFIID) and the RNA polymerase II and is sufficient to promote transcription initiation, albeit with low efficiency (Kadonaga, 2012) and ensures productive transcription in one direction (reviewed in Haberle and Stark, 2018). Initiator (Inr) motifs within enhancers can also initiate transcription, but in a bidirectional and non-productive manner, resulting in transient short non-coding RNAs (Kim *et al.*, 2010; Andersson *et al.*, 2014; Core *et al.*, 2014). Bidirectional transcription is also present at active promoters (Core *et al.*, 2014), but it is outweighed by the stabilized gene transcript. The integration of inputs from enhancers and silencers further modulates transcriptional activation (Spitz and Furlong, 2012).

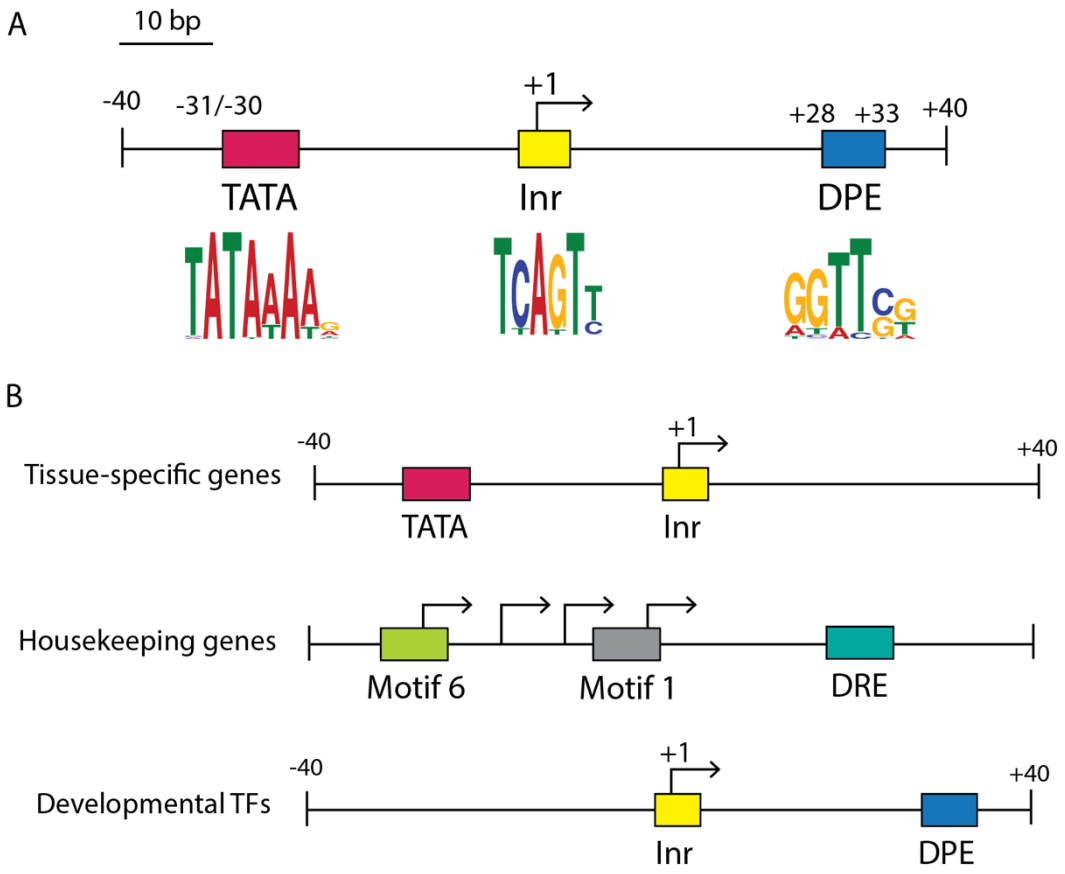


Figure 3: Structure of the core promoter.

(A) Schematic representation of the position of the main core promoter motifs conserved from flies to mammals. The Inr motif contains the TSS (+1), and the start and end of the other motifs are relative to the TSS. Below, position-weight matrices (PWMs) of the core promoter motifs in *D. melanogaster*.

(B) Scheme of the composition of different promoter types in *D. melanogaster*.

According to the function of the regulated gene, promoters can have different shapes and motifs (Rach *et al.*, 2009; Lenhard, Sandelin and Carninci, 2012) (Figure 3B):

- Tissue- and cell type-specific genes with restricted expression domains generally have a sharp TSS and comprise the TATA-box and Inr motifs in mammals (Carninci *et al.*, 2006; Roider *et al.*, 2009; Ngoc *et al.*, 2017) and flies (Rach *et al.*, 2009; Danino *et al.*, 2015)
- Housekeeping genes have a broad nucleosome-depleted region with multiple TSS, followed by an array of precisely positioned nucleosomes (Carninci *et al.*, 2006; Rach *et al.*, 2011). In mammals, they are in CpG islands (Bajic *et al.*, 2006). In flies, they are enriched in motif-1 bound by M1BP, and the Ohler-6 and DRE bound by Dref (Ohler, 2006)
- Developmental TFs responsible for patterning and morphogenesis in flies typically have a sharp TSS as well as Inr and DPE motifs (Ohler, 2006; Lenhard, Sandelin and Carninci, 2012; Zehavi *et al.*, 2014). In mammals they are generally similar to housekeeping genes promoters in differentiated tissues, whereas in embryonic stem cells they have a bivalent chromatin state, showing the repressive mark H3K27me3 alongside the standard promoter chromatin marks (Bernstein *et al.*, 2006).

1.1.4.1 Enhancer specificity is achieved through promoter- and insulator-dependent mechanisms

Limiting enhancer action as to establish the correct enhancer–gene connection(s) is important for proper gene regulation. In principle, enhancers can activate genes across large genomic distances, albeit with lower efficiency (Zuin *et al.*, 2022; Balasubramanian *et al.*, 2024). In such cases of long-range regulation, there are often (many) other genes in between the enhancer and its target gene's promoter. If the enhancer also regulated those genes, such undiscriminated activation would result in highly overlapping patterns of gene expression, obliterating tissue specificity and being incompatible with bifurcating cell fates. This is especially problematic in species with a more compact genome, like *D. melanogaster*, as enhancers are in physical proximity to many genes in the linear genome but they should only activate their correct target. To ensure enhancer specificity, cells leverage multiple mechanisms (Galouzis and Furlong, 2022):

- promoter accessibility: if the promoter of the closest gene is in a closed chromatin state, the enhancer could regulate the next open promoter. This was observed in spermatocyte differentiation, where a set of promoters is only accessible at a specific differentiation stage (Lu *et al.*, 2020)
- enhancer-promoter (E-P) compatibility: the sequence of the core promoter contains the specificity to interact with selected enhancers in a dense neighbourhood and to selectively activate paralogous genes, as shown for *Mrf4* and *Myf5* in mouse (Chang *et al.*, 2004; Carvajal, Keith and Rigby, 2008), and for *gooseberry* (*gsb*) and *gooseberry neuro* (*gsbn*) in *Drosophila* (Li and Noll, 1994). Moreover, in *Drosophila* developmental and housekeeping promoters are preferentially activated by distinct enhancer sets (Ohtsuki, Levine and Cai, 1998; Butler and Kadonaga, 2001; Zabidi *et al.*, 2014): for instance, developmental TFs like Caudal (Juven-

Gershon, Hsu and Kadonaga, 2008) and Ftz (Shir-Shapira *et al.*, 2019) have a preference to activate DPE- containing promoters, which are associated to developmental genes. This specificity could be mediated by different cofactors (Haberle *et al.*, 2019)

- enhancer sequestration or promoter competition: the gain of a promoter can interfere with the activation of the normal enhancer target, and its loss can induce enhancer-mediated activation of other promoters (Carvajal, Keith and Rigby, 2008)
- promoter-proximal binding of insulator proteins at insulators. On the one hand, insulator proteins can insulate genes from the action of neighbouring enhancers, and on the other hand they promote E-P interactions via looping, bringing into proximity two distant regions on the linear DNA. Notably, insulator proteins are often in close proximity to promoters (Lee *et al.*, 2008; Bushey, Ramos and Corces, 2009; Jiang *et al.*, 2009), and some of them have a preference for a specific promoter type (Jiang *et al.*, 2009) or enhancer set (Dong *et al.*, 2020). This preference, combined with the high number of insulator proteins in *Drosophila*, may significantly contribute to E-P specificity (Galouzis and Furlong, 2022).

1.2 Gene regulatory networks integrate multiple inputs to achieve robust transcriptional control during development

To initiate and maintain specific developmental programs, CRMs must regulate their target genes in a stereotypical, robust and reproducible pattern (Spitz and Furlong, 2012; Long, 2016). In principle, a single enhancer could contain all the information for the spatial and temporal regulation of its target gene. However, often multiple regulatory elements act together to establish gene expression patterns, thus achieving tighter regulation and higher robustness (Perry, Boettiger and Levine, 2011; Osterwalder *et al.*, 2018), preventing stochastic activation (Waymack *et al.*, 2020) and integrating a larger number of signals, while limiting the number of sequences encoding for information (Buecker and Wysocka, 2012). Enhancers themselves are bound by the products of other enhancers' activity, as their target genes may be TFs which in turn bind to a downstream enhancer set (Davidson and Erwin, 2006). These connections and the single elements within – CRMs, TFs, target genes – are activated in reproducible time- and tissue-specific patterns (Bonn and Furlong, 2008; Spitz and Furlong, 2012; Wilczynski *et al.*, 2012). By leveraging a robust network of factors and genes, this tightly regulated differential use of regulatory inputs leads to intrinsic symmetry-breaking events, from pattern formation to cell-fate specification and differentiation and the maintenance of alternative cell fates (Leifer *et al.*, 2020).

A gene regulatory network (GRN) is the ensemble of regulatory connections among genes coding for TFs, their products and their target CRMs (Davidson, 2006). Its underlying structure is genetically encoded and constant across cell types and time, however its deployment is modelled as a circuit with inputs, switches and connections that can be modulated across time and tissues (Davidson, 2006). Certain subcircuits within developmental GRNs are reserved for particular body parts and tissues and highly conserved through evolution (kernels), whereas some are reused for similar purposes during different developmental stages (plug-ins), and others act as a switch to control the activation of downstream modules (Davidson, 2006; Davidson and Erwin, 2006; Oliveri and Davidson, 2007). The reliable and precise progression of development, aside from the necessary external inputs, mainly relies

on sub-networks, particularly on the necessity of integrating the input from multiple TFs to activate each of the enhancers within. The time and tissue specificity of kernel activity is defined by switches, which can integrate internal and external cues through enhancer elements to determine the on/off state of GRN modules. The imbalance of inputs across cell types over time causes differential activation of switches and subcircuits in the network, leading to progressive cell specification and differentiation.

GRNs integrate various input levels (Figure 4). In *trans*, the main inputs are the abundance and activity of each TF. TF abundance results from the transcription of the corresponding gene and the balance between translation and degradation, while TF activity depends on post-translational modifications, sequestration in cellular compartments and multimerization or cooperativity with different partners, which in turn may result in different abundance and activity at various times and tissues (Hobert, 2008).

In *cis*, enhancer activation results from a combination of local variables within the DNA sequence, the chromatin and the site-specific TF accumulation (MacNeil and Walhout, 2011; Brignall *et al.*, 2019). Within the DNA sequence, properties of the TF binding site such as the affinity and the presence of multiple motifs strongly influence binding kinetics and binding probability, ultimately determining the binding strength and magnitude (Gaudet and Mango, 2002; Sandmann *et al.*, 2006; Wagh *et al.*, 2023). However, the TF motif is not sufficient for binding: a pre-existing permissive open chromatin state is required for most TFs to access their motif within an enhancer (Inukai, Kock and Bulyk, 2017), whereas very few TFs can open chromatin and bind regardless of the initial enhancer accessibility, referred to as pioneering activity (Iwafuchi-Doi and Zaret, 2014). TF occupancy does not always coincide with

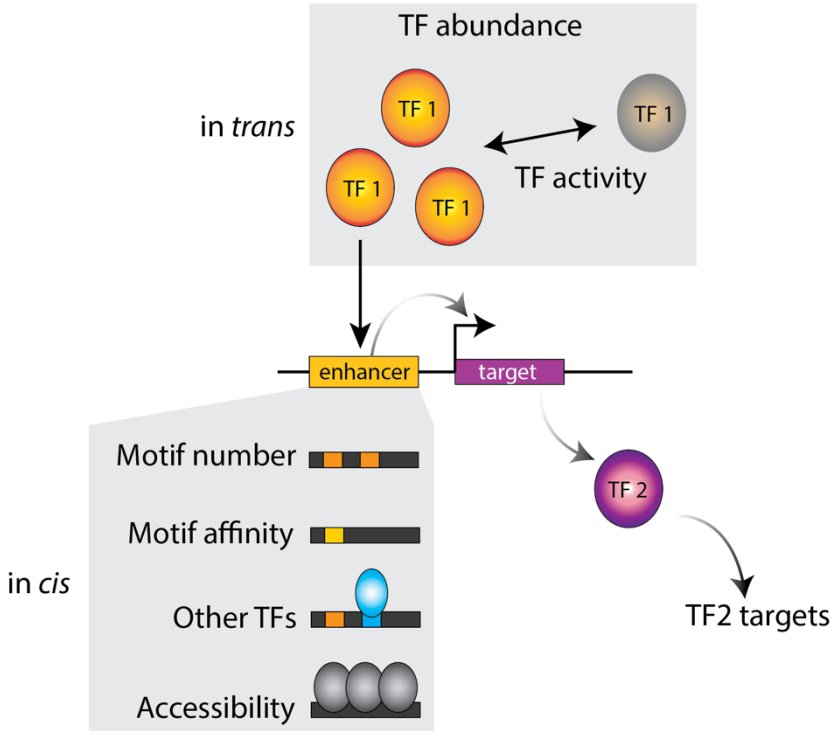


***Figure 4: GRNs integrate numerous inputs in* cis *and in* trans.**
Schematic representation of a genomic locus with an enhancer and its target gene, and of the inputs that converge on the enhancer: the binding of active TFs in *trans*, and the number of motifs, motif affinity, binding of other TFs and accessibility in *cis*. If the target gene encodes a TF, its product will contribute to the activation of downstream genes with the same logic.

enhancer activity: in fact, it may either predispose the enhancer for activation at a later time point (priming) (Spitz and Furlong, 2012), or directly trigger its activity. Enhancer activation requires both the deposition of active chromatin marks (Heintzman *et al.*, 2009; Creyghton *et al.*, 2010; Bonn, Zinzen, Girardot, *et al.*, 2012) and the binding of specific TFs (Rosenfeld, Lunyak and Glass, 2006; Kim *et al.*, 2010; Herz *et al.*, 2012). Active enhancers tend to be associated to condensates enriched in TFs and RNA polymerase II (Sabari *et al.*, 2018; Shrinivas *et al.*, 2019), which increase the local concentration of TFs and promoter/enhancer occupancy by locally changing the equilibrium between bound and unbound TFs.

Simultaneous TF co-binding at the same site could further stabilize or weaken the TF-DNA interaction, resulting in cooperativity, additivity or antagonism at a single enhancer (Sandmann *et al.*, 2007; Spitz and Furlong, 2012) (Figure 5). In addition, the transcriptional output of the enhancer's target gene may not fully mirror enhancer occupancy: in fact, the effect on gene expression may be cooperative, additive or sublinear depending on the TF combination at the enhancer (Ludwig, Patel and Kreitman, 1998;

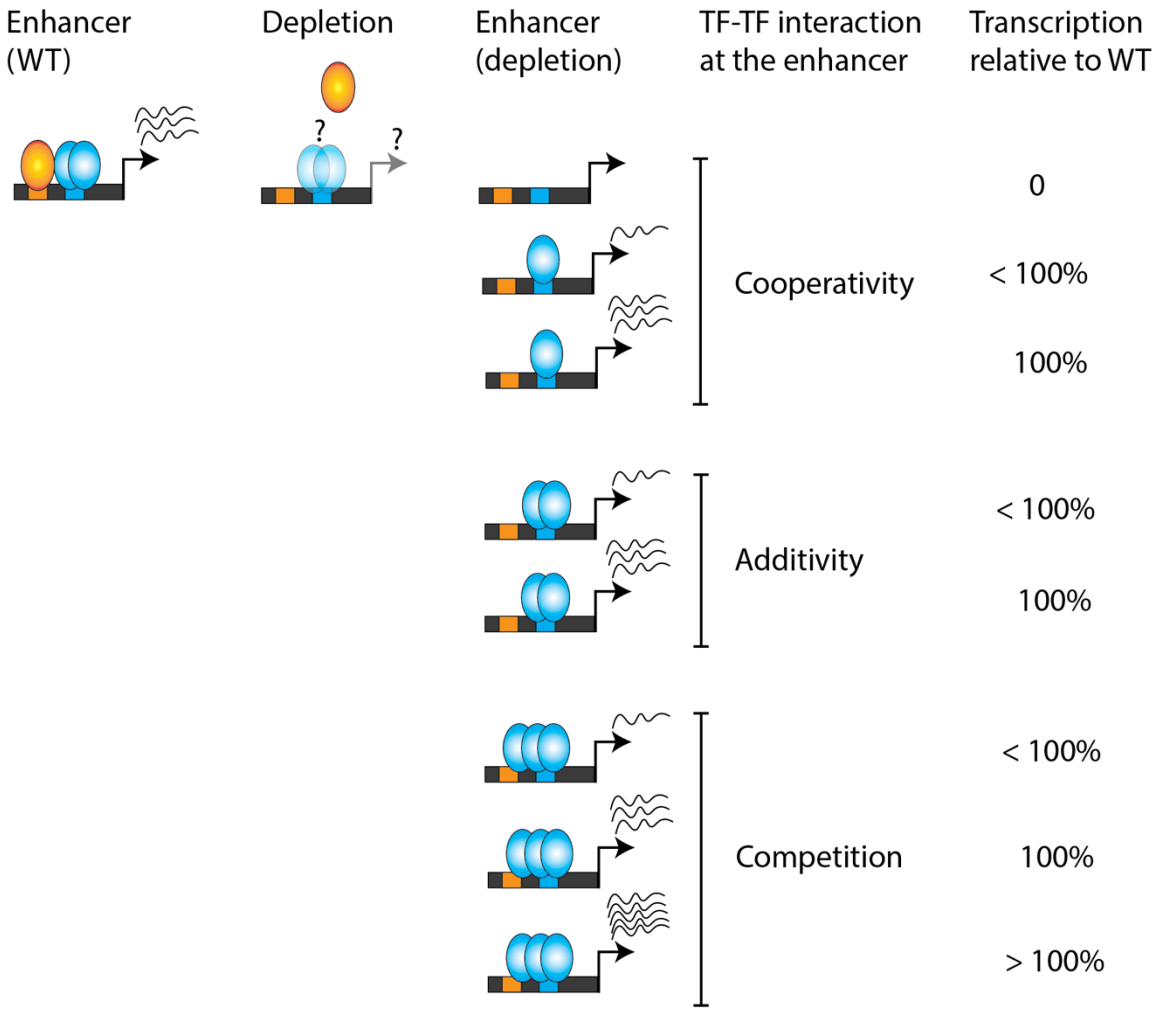


Figure 5: TF-TF interactions at co-bound enhancers and transcriptional output.
Scheme of the WT enhancer on the left, with binding of TF 1 (yellow) and TF 2 (blue) to the corresponding motifs. Upon depletion of TF 1, TF 2 binding can be (i) reduced, pointing to cooperative binding between TF 1 and TF 2 in the WT or to TF 1 being required for the binding of TF 2; (ii) invariant, indicating to independent binding (additivity); (iii) increased, in case there is competition between TF 1 and TF 2. Each of those situations may be associated with different effects on transcription.

Khoueiry *et al.*, 2017) (Figure 5). Moreover, enhancers mainly contribute to transcription initiation by modulating the frequency of transcriptional bursts, but their size depends on the core promoter (Haberle and Stark, 2018; Larsson *et al.*, 2019), which introduces an additional source of variation.

On top of these single elements, the GRN structure *per se* implements complex responses, including sequential or delayed activation, memory and epistasis (Alon, 2007). Since some TFs activate downstream TFs with different kinetics, there is a built-in hierarchical structure controlling the order and the starting time of transcription via an activation cascade (Rosenfeld and Alon, 2003). The requirement for multiple TFs to activate a downstream target delays its activation until all regulators are expressed at a sufficient level, which ensures precise and robust temporal control while minimizing the components in the network (Rosenfeld and Alon, 2003; Alon, 2007; Shoval and Alon, 2010). Moreover, previous binding of other TFs and the epigenetic state provide a history of local events, which can be accessed in following binding events (Srivastava and Mahony, 2020).

To deviate from linear responses and either dampen or enhance transcription, feed-back and feed-forward loops are leveraged in GRNs. In a feedback loop, transcription of a target inhibits that of its upstream activator, thus limiting target expression in time and extent, for instance, to preserve a specific state or achieve temporary identities (Alon, 2007; Shoval and Alon, 2010). On the contrary, in a developmental feed-forward loop, the target further activates its activator, leading to a delayed explosive response suitable to push cells towards a specific fate in developmental transitions (Alon, 2007).

Overall, a GRN integrates a large array of inputs in *cis*, in *trans* and from the architecture of the GRN itself in a reliable, robust and reproducible way, leading to stereotypical developmental patterns (Davidson and Erwin, 2006).

1.2.1 A historical perspective on cis-regulatory networks

The intuition of the existence of GRNs dates back to long before their experimental validation: their first theoretical model was presented in 1969 by Britten and Davidson (Britten and Davidson, 1969), who were the first to systematically describe and validate GRNs in sea urchin and over 40 years later in *Drosophila* (Levine and Davidson, 2005). Based on experimental evidence showing that: a) simple external clues can promote changes in differentiation state, b) differentiation requires the activation of many non-contiguous genes, and c) regulatory sequences are present and transcribed according to differentiation patterns, Britten and Davidson built a model of gene regulation in a differentiating organism considering four kinds of DNA sequences with an elementary function:

- sensor genes, which are bound by agents (TFs) in response to changes in cell state; they can be compared to modern enhancers
- producer genes, which are transcribed into RNA molecules; they correspond to the current definition of target genes. The set of producer genes activated in response to the same stimulus is known as a gene battery

- receptor genes, which trigger transcription of the producer gene via an activator RNA; they can be considered current promoters
- integrator genes, which synthetize the activator RNA and act as an intermediate between receptor and producer genes; they have similar features to enhancers, which combine the integrator and sensor gene classes.

In their model, these elements had a binary Boolean state (on/off) and were integrated into a circuit to achieve a complex output. The circuit leverages redundancy at different levels – receptor or integrator genes – to produce tightly coordinated responses at small and large scales, respectively. Moreover, an activation cascade where sensors (enhancers) respond to the product of producer genes (TFs) can lead to sequential activation or cell type stabilization. Albeit some of the basic elements have been updated – the products of producer genes with an effect on sensors are TFs, and sensors, receptors and integrators are CRMs –, the model could predict their existence and behaviour in the absence of experimental evidence; as such, it remains the theoretical foundation of GRNs.

Developmental GRNs were first validated in the endomesoderm of sea urchins (Davidson *et al.*, 2002). This research uncovered a network of connections between enhancers and target genes based on time and tissue specificity, the use of multiple types of perturbation (RNA knock-down [KD] by morpholino, change of activators into repressors, blocking signalling pathways) to disentangle direct and indirect effects, and rescue experiments. The use of perturbations was a key methodological intuition for assessing the contributions of single elements in the GRN. The network included all possible connections with a genome-wide perspective and allowed finding active connections at a certain time point and tissue. This experimental model, which refined the theoretical predictions by underlining the crucial role of TFs and signalling pathways, showcased the role of feed-forward connections in pushing development forward.

Overall, these GRN models capture many key properties of CRMs and their connections in development, and can predict cell fate based on the active elements in the network (Martín, Organista and de Celis, 2016). However, they are limited by the binary view of enhancer occupancy and activity (Peter, Faure and Davidson, 2012). They do not integrate the combined action of TFs binding to the same enhancers nor the relationship between the activity of enhancers and the transcriptional output, or the different levels of contributions to feed-forward loops. Despite current knowledge, implementing such information in a GRN model is still challenging: generally, causal relationships are modelled by co-expression (Fernandez-Valverde, Aguilera and Ramos-Dıaz, 2018), and models incorporating gene interactions do not account for deviations from linearity in TF-TF interactions and enhancer output (Shu *et al.*, 2021). In addition, intermediate stages in cell differentiation are difficult to address. Recent graph neural networks based on single-cell RNA-seq data like DeepSEM (Shu *et al.*, 2021) and CEFCON (P. Wang *et al.*, 2023) aim to elucidate intermediate cell states and model non-linear relationships and driver connections; however, they cannot disentangle TF contributions at single enhancers and do not leverage experimental perturbations to assess necessity.

1.3 A primer on *Drosophila* development

In my Ph.D. project, I focused on the *Drosophila* mesoderm GRN to understand the contribution of enhancers and their TF co-binding to enhancer activity and gene expression, to reconstruct the links from molecular to cellular phenotypes. Before delving into the role of the mesoderm GRN in muscle specification and differentiation, I will give a brief overview of the main developmental events taking place in *D. melanogaster*'s embryogenesis.

D. melanogaster embryogenesis lasts roughly 24 h from egg laying to hatching at room temperature and it is divided into 17 stages based on morphological landmarks (Hartenstein, 1993; Weigmann *et al.*, 2003) (Figure 6). Fertilization is followed by 13 rapid nuclei divisions within a syncytium, known as nuclear cycles. After the first seven divisions, most syncytial nuclei migrate to the surface of the egg, where they acquire either a somatic or germline identity. Once all nuclei (~6,000) are at the embryo's surface, they start to cellularize leading to the cellular blastoderm stage (stage 5, 2:10-2:50 hours post fertilization, hpf) (M., Zalokar and Erk, 1976). The anlage for the future three germ layers – ectoderm, mesoderm and endoderm – are already determined at this stage (Wakabayashi-Ito and Ip, 2006). The following stage only lasts 10 min (2:50-3 hpf) and corresponds to early gastrulation, when the mesoderm starts to invaginate internally on the ventral side of the embryo forming the ventral furrow (presumptive mesoderm), and the blastoderm cells close to germ cells at the posterior pole of the embryo form a dorsal plate to carry them into the next stages. Gastrulation completes within 20 min (stages 6–7) and is followed by germ band elongation. The germ band consists of invaginated mesodermal cells and of adjacent ectodermal cells (Bate and Arias, 1993). During stages 8–11, it extends up to the posterior pole of the embryo, retracting at stage 12–13. In parallel, important morphogenetic events take place in the mesoderm: at stage 8 (3:10-3:40 hpf), mesodermal cells flatten into a bilayer and undergo two mitoses (Bate, 1993; Hartenstein, 1993), and at stage 9 (3:40-4:20 hpf) they undergo a transient parasegmentation, further flattening into a monolayer before splitting up into different organ primordia at late stage 10. At stage 10 (4:20-5:20 hpf), the parasegmental pattern is also visible in the epidermis, and at stage 11 (5:20-7:20 hpf) the segmental furrows become visible (Weigmann *et al.*, 2003). Stage 11 and 12 are marked by extensive cell death (Weigmann *et al.*, 2003). The stomodeum, which will give rise to the foregut, appears at stage 10 and contacts the anterior midgut primordium, but will contact the posterior midgut primordium at the end of stage 12. The head mesoderm flanks the anterior midgut rudiment and moves to the anterior end of the head, giving rise to the head musculature and to macrophages (Campos-Ortega and Hartenstein, 2013). Early stage 12 marks the subdivision of the mesoderm into its four organ derivatives: somatic muscle (SM), visceral muscle (VM), heart (dorsal vessel) and fat body (Bate and Arias, 1993). The VM comprises the circular VM (CVM), which forms an inner circular ring surrounding the gut, and the caudal or longitudinal VM (LVM), which makes the external layer of longitudinal fibres along the digestive tract (Hartenstein and Campos-Ortega, 1985; Bate and Arias, 1993; Lee, Zaffran and Frasch, 2005). At the next stage (stage 13; 9:20-10:20 hpf), myoblasts fuse to form syncytia in the SM and VM (Bour *et al.*, 2000; Frasch and Leptin, 2000; Ruiz-

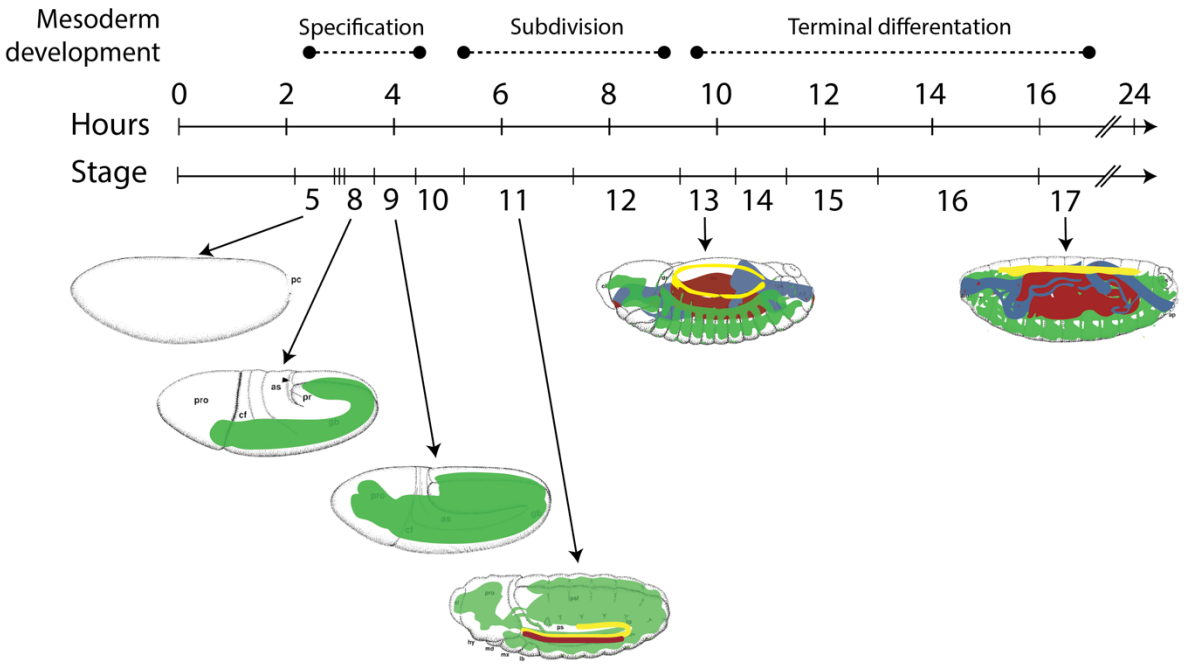


Figure 6: Mesoderm development in* D. melanogaster *embryos.
Timeline of the main phases in mesoderm development and corresponding hours post fertilization and stages. Schematic representation of embryos at selected stages, with the mesoderm (stages 8-9) and the somatic mesoderm (stages 11-17) in green, the heart in yellow, the visceral mesoderm surrounding the gut in red and the foregut and hindgut in blue. Embryo representation adapted from Hartenstein (1993).

Gómez *et al.*, 2000; San Martin *et al.*, 2001). As for the heart muscle, at this stage, the dorsal vessel primordium presents as two rows of cells on each side of the embryo; the cells in the dorsal row acquire a cuboidal shape and are specialized as cardioblasts, whereas those in the ventral rows maintain their rounded shape and become pericardial cells (Hartenstein, 1993). In stages 14–17, head and gut morphogenesis finalize, the embryo closes dorsally and the cardioblasts on each side meet on the dorsal side of the embryo, which is finally ready to hatch into a larva (Hartenstein, 1993).

1.4 The mesoderm GRN as a functional model

In my Ph.D. project, I focused on the mesoderm GRN of the *Drosophila* embryo for the conveniency of using a well-characterized system and the potential to drive generalizable conclusions. *Drosophila* mesoderm GRN has been extensively studied genetically, in terms of TFs and their target enhancers, from knock out (KO) mutants (Boomer, Jan and Jan, 1990; Leptin, 1991; Azpiazu and Frasch, 1993; Taylor *et al.*, 1995; Zaffran *et al.*, 2001) to genome-wide TF binding and enhancer accessibility (Furlong *et al.*, 2001; Sandmann *et al.*, 2006; Jakobsen *et al.*, 2007; Bonn and Furlong, 2008; Zinzen *et al.*, 2009). This valuable knowledge base allows us to identify the key events in mesoderm development, their timing, the TFs and enhancers that are likely involved, and to correlate such changes to specification and differentiation of different mesodermal tissues. Moreover, this network has been subject to positive selection in evolution, being vastly conserved across the animal kingdom, with TFs of the same family being in the same position in the GRN (Komuro and Izumo, 1993; Lints *et al.*, 1993; Ranganayakulu *et al.*, 1995; Schultheiss, Xydas and Lassar, 1995; Lee, Xu and Breitbart, 1996; Tanaka *et al.*, 1999;

Castanon and Baylies, 2002; Cripps and Olson, 2002; Wu *et al.*, 2011). This allows us to find general principles of interaction between GRN elements as well as GRN organization and their impact on the phenotype, and to translate them easily to other organisms.

1.4.1 Mesoderm TFs have spatio-temporally restricted expression

The first mesoderm TF expressed in *Drosophila* is the master regulator TF *twist* (*twi*), which is necessary and sufficient for mesoderm specification (Leptin, 1991; Baylies and Bate, 1996), in a manner similar to *MyoD* in vertebrates (Rudnicki *et al.*, 1993; Tajbakhsh and Cossu, 1997; Baylies, Bate and Gomez, 1998). *twi* is first expressed in the presumptive mesoderm before gastrulation, at stage 5 (Simpson, 1983; Nüsslein-Volhard, Wieschaus and Kluding, 1984; Thisse, Messal and Perrin-Schmitt, 1987; Leptin, 1991). *twi* expression is uniform throughout the trunk mesoderm until stage 10, when the mesoderm is divided into segmental stripes with high *twi* expression in the ventral posterior side of each segment, the future SM (Dunin-Borkowski and Brown, 1995; Baylies and Bate, 1996). The low Twi domain will give rise to the future VM, heart, fat body and mesodermal glia (Dunin-Borkowski and Brown, 1995; Baylies and Bate, 1996).

At blastoderm stage, shortly after the beginning of its expression, Twi activates its target TF *tinman* (*tin*) (Bodmer, 1993; Lee *et al.*, 1997; Yin, Xu and Frasch, 1997) (Figure 7A-B), which is necessary for the specification of the dorsal mesoderm into VM, cardiogenic mesoderm and dorsal somatic muscle (Azpiazu and Frasch, 1993; Bodmer, 1993). Together, Twi and Tin activate the expression of *Myocyte enhancer factor 2* (*Mef2*) (Gajewski *et al.*, 1997; Cripps *et al.*, 1998), which is required for the differentiation of all muscle types (SM, VM, heart) (Bour *et al.*, 1995). Both *tin* and *Mef2* are expressed uniformly in the trunk mesoderm up to stage 10, and at stage 11, during mesoderm diversification, *tin* gets restricted to the dorsal part of each segment, giving rise to the heart and the VM (Azpiazu and Frasch, 1993; Bodmer, 1993). In the presumptive VM, Tin activates the transient expression of the downstream TF *bagpipe* (*bap*), which in turn activates the FoxF TF *biniou* (*bin*) (Zaffran *et al.*, 2001).

The hierarchical activation of *twi*, *tin* and *Mef2*, and later *bap* and *bin* gives rise to the major mesodermal tissues: after mesoderm specification by Twi, Tin contributes to the dorsal mesoderm identity (dorsal SM, heart, VM) (Azpiazu and Frasch, 1993), Mef2 to myoblast fusion and terminal muscle differentiation (Bour *et al.*, 1995), whereas Bin and the transient expression of *bap* are essential for the specification and differentiation (Bin) of the VM (Zaffran *et al.*, 2001). In a simplified view of the GRN, the upmost three TFs – Twi, Tin and Mef2 – are active at stage 9; at stage 10–11, Twi, Tin and Mef2 are active in all tissues, Bap in the VM and heart and Bin in the VM; for the terminal differentiation (stage 16), Mef2 is active in the SM, VM and heart, Tin only in the heart and Bin only in the VM (Azpiazu and Frasch, 1993; Lilly *et al.*, 1995; Baylies, Bate and Gomez, 1998) (Figure 7B). However, mesoderm TFs have additional roles in the terminal differentiation of precise subsets of cells within these tissues.

Aside from being essential for gastrulation and mesoderm specification, Twi is necessary for the specification of the somatic mesoderm at stage 10 (Baylies and Bate, 1996; Baylies, Bate and Gomez, 1998), and could contribute to the maintenance of an undifferentiated state of the adult muscle precursor cells (Bate, Rushton and Currie, 1991; Anant, Roy and Vijayraghavan, 1998; Cripps and Olson, 1998).

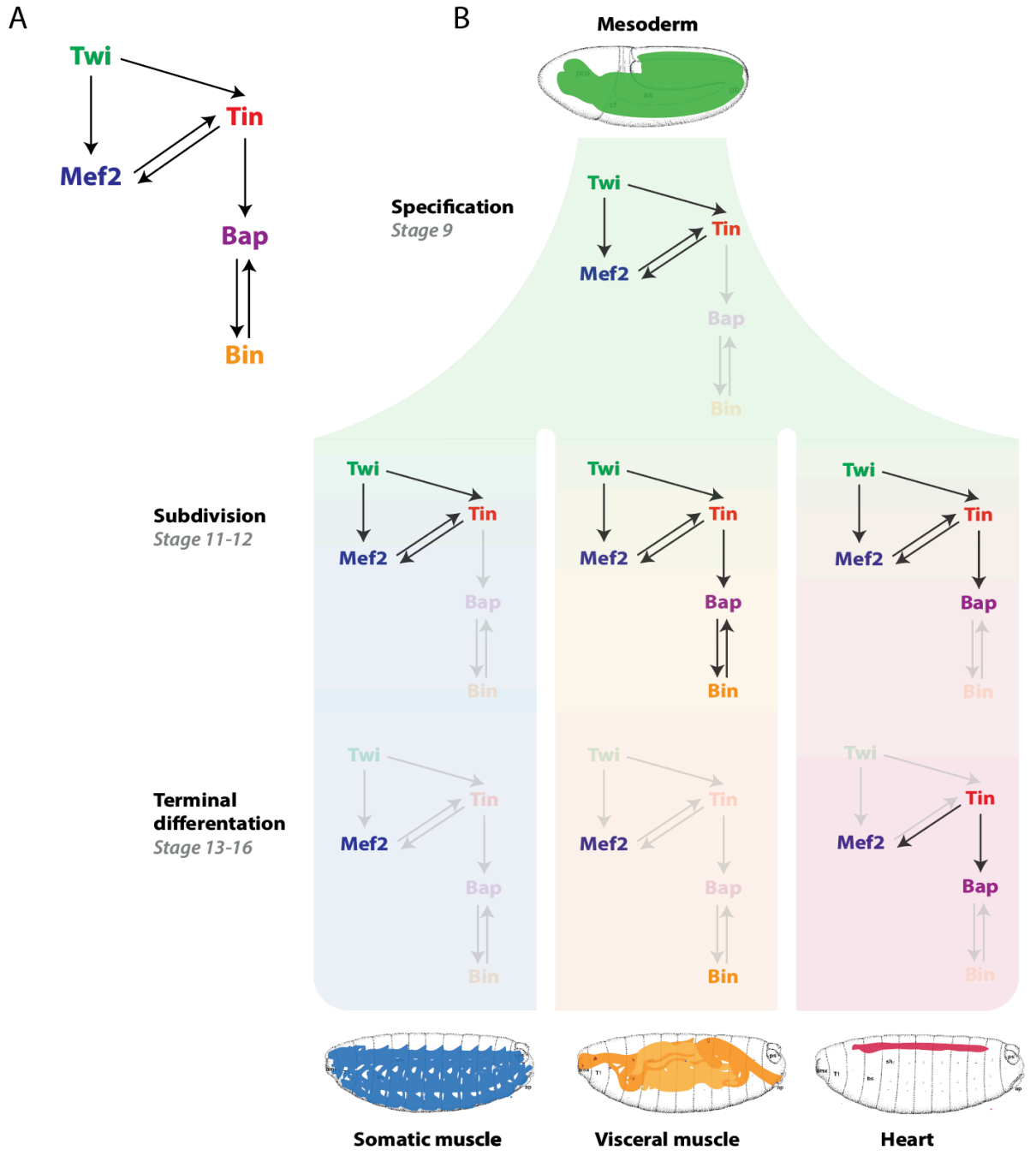


Figure 7: Activation of the mesoderm GRN in time and tissue.

(A) Scheme of the mesoderm GRN: Twi, at the top of the hierarchy, activates *tin* and *Mef2*. Tin and Mef2 activate each other's expression, and Tin is also necessary for *bap* expression. In turn, Bap activates *bin*, while Bin transiently sustains *bap* expression.

(B) Active TFs and connections within the mesoderm GRN across time and tissue. Expressed TFs and active connections are depicted in bright colours, inactive connections and not expressed TFs are faint. Embryo depictions adapted from Hartenstein (1993).

Tin not only is required for heart formation, but its activation of *bap* is fundamental for the emergence of the VM: indeed, *tin* homozygous mutants lack the VM (Azpiazu and Frasch, 1993; Bodmer, 1993). Multiple enhancers regulate *tin* expression at different times and tissues (Figure 8): *tin* B drives the expression in the entire trunk mesoderm at stage 9, *tin* D in the dorsal mesoderm at stage 11, *tin* C in cardioblasts and *tin* A in the pharynx (Yin, Xu and Frasch, 1997; Xu *et al.*, 1998; Lo and Frasch, 2001). A transgene with *tin* under the control of the enhancers *tin*-AB can restore all derivatives of the dorsal mesoderm in a deficiency that removes the *tin* locus, although with a decrease in cardioblast progenitors and mature cardioblasts, and defects in the dorsal vessel (heart) (Zaffran *et al.*, 2006). Similarly, a disruption in cardioblast differentiation and dorsal vessel morphology can be observed in the *tin*-ABD rescue (Zaffran *et al.*, 2006). Therefore, *tin* expression is required first for all dorsal mesoderm derivatives, i.e. heart and VM, and later for the correct terminal differentiation of cardioblasts and pericardial cells in the heart.

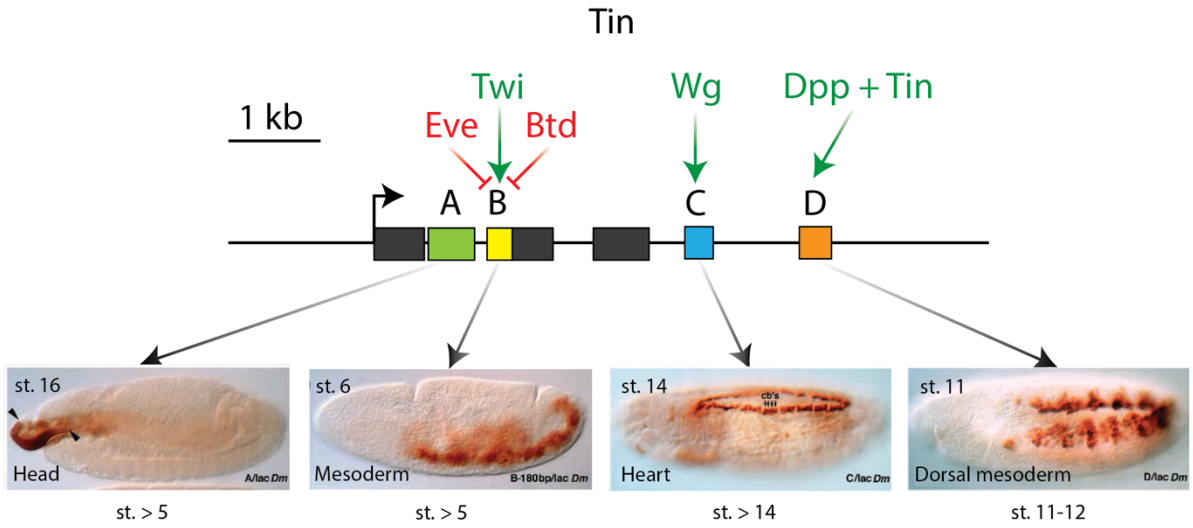


Figure 8:* tin *enhancers.
Schematic representation of *tin* locus with *tin* gene in black and the 4 enhancers *tin* A-D in colour. The TFs binding to each enhancer are indicated above: activators are in green, repressors in red. The domain where each enhancer promotes *tin* expression is shown by beta-galactosidase antibody staining on *D. melanogaster* embryos from reporter lines (Yin, X. L. Xu and Frasch, 1997). The stage of the embryo is indicated on the top left of the images, the tissue on the bottom left, and the activity window of the enhancer below each image.

In addition to its role in myoblast fusion in the somatic muscle, *Mef2* is also essential for the differentiation of heart and VM (Bour *et al.*, 1995; Lilly *et al.*, 1995; Taylor *et al.*, 1995). Its widespread expression pattern is under the control of numerous enhancers (Nguyen and Xu, 1998): a Twi-responsive enhancer (I-D_S) triggers *Mef2* transcription at postblastoderm stages, a second Tin-responsive element (I-D_L) is active at stage 9-11, and three enhancers dictate Mef2 expression in the cardioblast domain: the first enhancer (II-B_S) is active by stage 12 in *tin*-expressing dorsal mesoderm cells and cardiac precursors, and the second (I-E) is activated both in the somatic mesoderm and in the heart around stage 14, during terminal differentiation (Nguyen and Xu, 1998). The third enhancer is active in a set of cardioblasts which express *seven-up* instead of *tin* (Gajewski *et al.*, 2000). Additional enhancer elements contribute to the late expression in the VM (II-A, II-B_L, II-E) and in the SM (II-E,

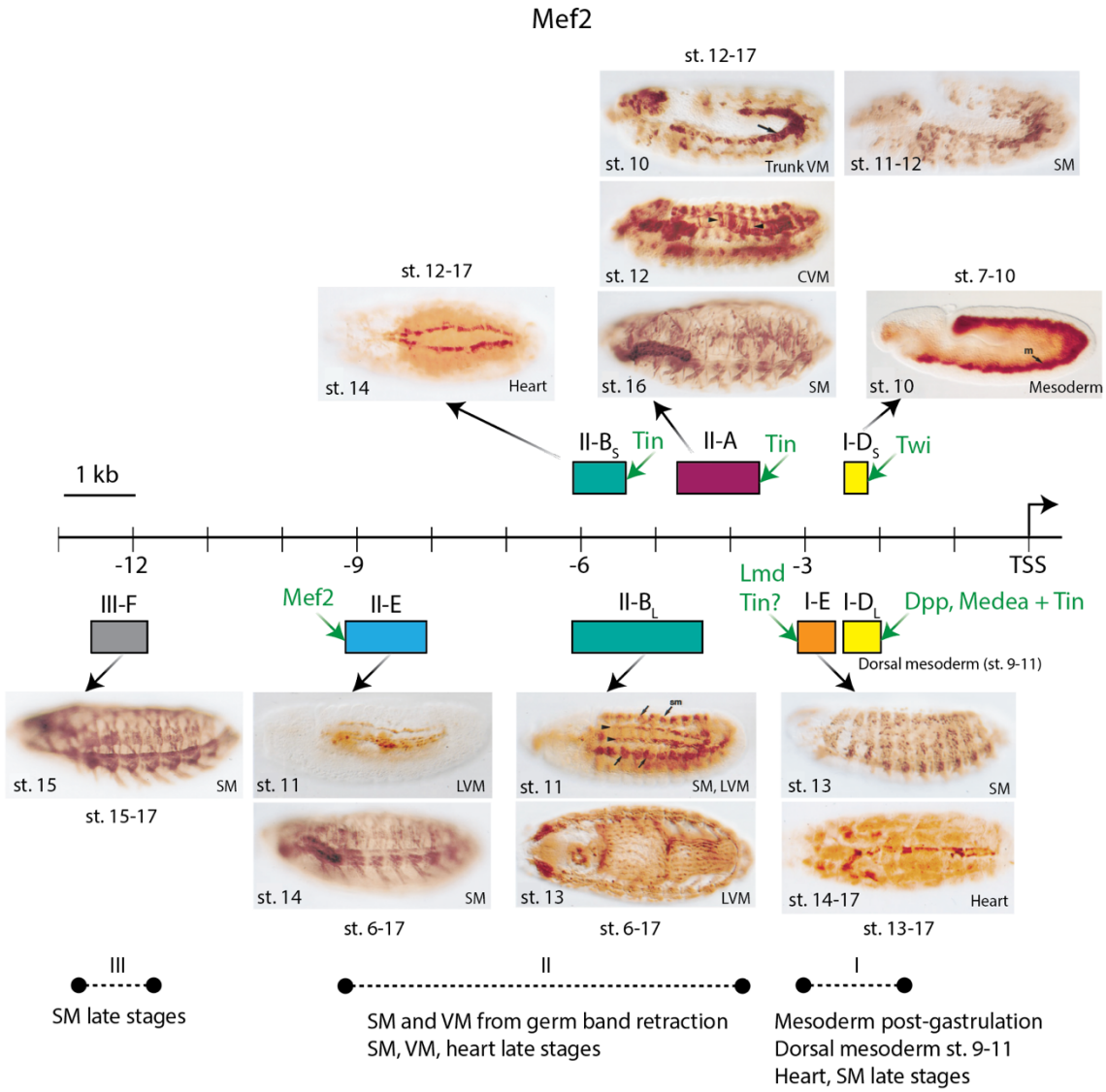


Figure 9:* Mef2 *enhancers.
Schematic representation of the *Mef2* genomic locus, with the TSS of *Mef2* (right) and the enhancer elements in different colours. The TFs activating each enhancer are indicated in green. The expression domain of the enhancers across time is shown by beta-galactosidase antibody staining in reporter lines (Nguyen and Xu, 1998); the stage of the embryo and the expression domain are reported on the bottom left and bottom right of the image, respectively (SM = somatic muscle, VM = visceral muscle, LVM = lateral visceral muscle, CVM = circular visceral muscle). The stages at which the enhancer is active are reported above or below the stainings. Below, a summary of the location, tissues and expression stages of the three main groups of enhancers from early (I) to late (III).

III-F). Multiple of those enhancers (II-A, I-E, II-E, II-B$_L$) are active in a combination of tissues (Nguyen and Xu, 1998; Cripps, Zhao and Olson, 1999; Cripps, Lovato and Olson, 2004; Lovato *et al.*, 2015).

bap is transiently expressed in segmental patches in the prospective visceral mesoderm, which correspond to the future CVM, and in the foregut and hindgut at stage 10–11 (Azpiazu and Frasch, 1993). Bap is essential for VM development, through its activation of *biniou* (Azpiazu and Frasch, 1993; Zaffran *et al.*, 2001). At stage 13, *bap* is expressed again in a subset of *tin*-expressing cardiomyocytes and pericardial cells, and from stage 14 up to adulthood in all *tin*-expressing cardiomyocytes (Seyres *et al.*, 2016). In the heart, Bap contributes to correct myofibrillary organization,

ensuring proper heart rate and cardiac output (Seyres 2016). Transgenic reporter assays have identified three *bap* enhancers: the early enhancer, bap3, drives metameric expression at stage 11 and includes multiple binding sites for mesodermal activators (one strong Tin binding site, two Bap binding sites for positive autoregulation, and one binding site for Bap product Bin, reflecting positive feedback regulation), as well as a long sequence for Sloppy (Slp) binding, partially overlapping with the Bin binding site and responsible for the segmental repression of Bap expression and its confinement to visceral but not cardiac mesoderm (Lee and Frasch, 2005). Additional enhancers include baplac4.5, which is partially redundant with bap3, and bapD3.5 for bap expression in the foregut and hindgut (Lee and Frasch, 2005).

bin is expressed in the entire VM from stage 10 as well as in the foregut and hindgut VM and the LVM (Zaffran *et al.*, 2001). Tin and Bap are required for *bin* expression in the trunk visceral mesoderm (Zaffran *et al.*, 2001). In the absence of Tin, Decapentaplegic (Dpp) can activate *bin*, albeit transiently and in smaller cell clusters. Moreover, positive autoregulation is required for *bin* expression at stage 11; at that stage, Bin is necessary to maintain *bap* expression through a cross-regulatory loop (Zaffran *et al.*, 2001).

A dissection of the roles of mesodermal TFs at later stages of embryogenesis, and their changes between a uniform mesoderm-wide expression (stage 9) and the subdivision of the mesoderm into different muscle primordia (stage 11), has not been possible so far due to the loss of all derived cell types in classic loss-of-function knock-out (KO) mutants, and to the high heterogeneity of TF abundance across cell types in each segment.

1.4.2 Mesoderm TFs bind to different enhancers with different combinations of TFs across time and tissues

In addition to the expression of mesoderm TFs being time- and tissue- specific, their targets are also dynamic across embryogenesis, often sharing multiple TFs engaging in cooperative enhancer interactions (Zinzen *et al.*, 2009; Junion *et al.*, 2012).

Twi binds to enhancers in the vicinity of a striking 25% of all *Drosophila* TFs; its targets include genes involved in anterior-posterior mesoderm patterning and segmentation, as well as mesoderm specification (Sandmann *et al.*, 2007). Half of Twi enhancers have continuous binding across multiple embryonic stages, the other half has dynamic occupancy: 23% of putative enhancers are bound in early development (stage 5–9) and 26% at later stages (stage 10–11), while ~50% are bound at all three time-points. Temporal specificity is achieved by the co-binding of other TFs: early enhancers are co-bound by Dorsal, and late enhancers by Tin. Moreover, Twi often co-binds with its targets Mef2 and Tin to the enhancers of their associated target genes, in a feed-forward loop; for instance, 42% of Mef2 bound enhancers are co-bound by Twi.

The same temporal dynamics of enhancer binding applies to Tin, Mef2 and Bin: each TF occupies about 50% of its targeted enhancers either at earlier or later stages of development (Zinzen *et al.*, 2009), and

transient occupancy is correlated to the presence of motifs of other mesoderm TFs, hinting at the necessity of TF cooperativity in the binding of stage-specific enhancers. Transient occupancy is correlated to concomitant expression of the genes involved in the developmental events specific of that embryonic stage.

Even TFs with very transient expression such as *bap*, which is expressed in the visceral mesoderm at stage 10–11, can substantially contribute to enhancer co-binding and cell fate: Bap occupies only a small number of enhancers on its own, in the majority of cases it co-binds to enhancers with Bin at early bound visceral mesoderm enhancers, together with Tin (Jakobsen *et al.*, 2007). More than half of those enhancers are co-bound only at stage 10–11, and they can be specifically associated with the specification, but not the maintenance, of the VM.

Moreover, multiple simultaneous programs controlling general muscle properties and tissue-specific features can be driven by different sets on enhancers bound by different combinations of TFs: for instance, in the visceral mesoderm a set of Mef2-bound enhancers may be responsible for muscle differentiation and contractility, whereas a different Bin-bound enhancer set maintains tissue identity (Jakobsen *et al.*, 2007).

In conclusion, extensive TF cooperativity at shared enhancers is the major determinant for time- and tissue-specific activation of developmental programs. However, the contribution of each TF to cooperative binding and enhancer activity remains unclear.

1.5 Limitations of current approaches to understanding the mesoderm GRN

The numerous inputs and the complex connections between regulatory elements make it challenging to unravel all contributions to a GRN.

Initial studies either addressed the role of single regulatory elements in *cis* (enhancers and promoters) by reporter assays, or in *trans* (TFs) using KO mutants or knock down (KD) of TFs (Azpiazu and Frasch, 1993; Bodmer, 1993; Bour *et al.*, 1995; Lilly *et al.*, 1995). More recently, with the advent of genome-wide chromatin profiling, correlative studies explored TF binding at enhancers at a genome-wide level throughout development (Sandmann *et al.*, 2007; Liu *et al.*, 2009; Zinzen *et al.*, 2009).

1.5.1 Insights from KO or KD mutants

The first insights into the mesoderm GRN arose from phenotypic screening of *Drosophila* spontaneous and induced mutants. Such a large-scale forward genetic screen, using mutagens such as ethyl methane-sulfonate (EMS) (Wieschaus and Nüsslein-Volhard, 2016), combines scale with an easy setup and readout, but did not specifically aim to address the temporal evolution of active GRNs and their hierarchical structure. The main limitation stems from the fact that in null mutants the gene of interest is ablated since the beginning of embryogenesis. If the gene product is required for multiple essential developmental events, its requirement can only be assessed for the first event, and subsequent ones will

be masked either by early lethality or the complete absence of the parent cell type. Twi, for example, is essential for mesoderm gastrulation and the subsequent subdivision of the mesoderm – however as *twi* mutant embryos cannot gastrulate, all mesoderm development is blocked (Simpson, 1983). Further, the time estimation of the first necessity window might be itself imprecise: even in the absence of the TFs necessary for a specific embryogenetic transition, embryos tend to continue their (aberrant) development, especially at early stages, and therefore the effect of TF removal (i.e. lethality) could manifest much later than their activity window. In fact, in the Heidelberg screen (Nüsslein-Volhard, Wieschaus and Kluding, 1984), embryonic lethal mutant phenotypes were examined after cuticle deposition at stage 16; as for the mesoderm TFs, *twi*, *tin* and *Mef2* KO embryos develop until stage 16, despite the absence of the full mesoderm or of some crucial parts (Bour *et al.*, 1995; Lilly *et al.*, 1995; Wieschaus and Nüsslein-Volhard, 2016). Moreover, if the gene encodes a TF at the top of the GRN, its KO will hinder transcription of all its direct targets as well as their own, leading to direct and indirect effects. Additional indirect effects may be due to long-term compensation from other TFs (Tao, Christiansen and Schulz, 2007; Drechsler *et al.*, 2013; Hollfelder, Frasch and Reim, 2014; Frasch, 2016). Moreover, the absence, reduction or disfunction of a particular cell type may hinder differentiation of other cells which rely on cell-cell interactions and signalling.

Conditional RNA interference (RNAi) KO/KD have been used to reduce indirect effects and achieve a higher temporal resolution (Kim *et al.*, 2004; Neely *et al.*, 2010). However, both still have significant limitations in their temporal resolution: a conditional KO will have a time lag effect dependent on the mRNA and protein half-life of the KO gene, and RNAi lines, in spite of the sharper temporal control at the RNA level, still have kinetics in the order of hours. Moreover, confounder and specific effects could still arise in RNAi lines due to the high RNA dosage, and the signal might be masked by the often-incomplete depletion.

A complementary strategy to address TF contribution in time and tissue while minimizing indirect effects is the combination of reporter assays with deletions and targeted mutagenesis in *cis*, whereby a specific enhancer bound by a known TF is ablated or mutated. The reporter assay provides basic information on the time and tissue where a putative enhancer is sufficient to drive gene expression (Stanojevic, Small and Levine, 1991; Small, Blair and Levine, 1992; Harafuji, Keys and Levine, 2002; Kvon *et al.*, 2014; Kvon, 2015), although it might fail to capture additional regulation layers in its endogenous context. Putative enhancers from the reporter assay are then assessed for necessity by mutagenesis, which allows to finely map enhancers to phenotype with temporal and spatial precision, down to the single cell level (Stanojevic, Small and Levine, 1991; Wenick and Hobert, 2004; Canver *et al.*, 2015). Despite being extremely informative, this approach is limited to a single locus and not scalable to the high number of enhancers within a GRN, especially *in vivo*, in an embryonic context. Furthermore, testing of an enhancer's necessity overlooks the regulatory interactions between TFs at the enhancer.

Partial rescue under the control of specific enhancers in a null mutant allows more precise temporal and spatial control via the time- and tissue-specificity of the enhancer, but it is highly dependent on previous

knowledge of the locus and on low enhancer redundancy, and not scalable to the whole GRN (Zaffran *et al.*, 2006).

1.5.2 Insights from bulk and single-cell genome-wide assays

Genome-wide techniques such as RNA-seq, ATAC-seq and ChIP-seq provide a scalable, time-resolved view of embryonic development at the molecular level through a series of static snapshots of embryos at specific time points (Sandmann *et al.*, 2007; Zinzen *et al.*, 2009). For RNA-seq and ATAC-seq, single-cell protocols enable a higher spatial resolution down to subpopulations of cells within a tissue, provided that they are large enough, overcoming the need for markers for FAC sorting or precise dissection to isolate the cells of interest (Islam *et al.*, 2011; Cusanovich *et al.*, 2015, 2018; Hwang, Lee and Bang, 2018). The massive advantage of a time- and cell- specific readout, however, could not be coupled with a perturbative approach matching the same temporal resolution until recently. Therefore, the large body of evidence from genome-wide studies, spanning single TFs and their enhancers over time, as well as the bound enhancer and targets of multiple TFs, relied mainly on correlative, not causal, evidence.

Although computational modelling has been leveraged on this data to predict the final state of the GRN, it could not provide full mechanistic insights. Together, enhancer occupancy and chromatin state could accurately predict the temporal patterns of gene expression in *Drosophila* mesoderm (95% success), but had only a 50% success rate for spatial patterns (Wilczynski *et al.*, 2012); therefore, the role of tissue-specific enhancers may be underestimated in such models. Independent studies have confirmed that physical information from TF binding at enhancers and motifs, and functional information from chromatin state and coordinated expression, are complementary and both required to accurately predict GRN structure and refine time- and tissue- specific connections (Marbach *et al.*, 2012). Notably, the expression level of target genes is more accurately predicted by models based on a GRN structure than by TF binding and motif presence. This is in agreement with previous research in which a regulatory graph with logical rules based on current literature could simulate and predict mutant phenotypes (Mbodj *et al.*, 2016). Overall, computational modelling underscores the importance of combining enhancer occupancy and GRN architecture, particularly the connections between enhancers and targets, for the prediction of gene expression. However, the contribution of single TFs at enhancers is not clarified by these models, which simplify enhancer occupancy and activity to a binary enhancer-to-gene link. Moreover, such models can only capture spatial resolution across tissues with known differential regulatory connections.

To enable a finer dissection of the mesoderm GRN and more direct measurement of the functional contribution of TFs, and their effect on cell differentiation at each developmental stage, without spatial resolution limitations, in my Ph.D. thesis I combined the potential to uncover causal inference from a time-resolved perturbative approach with the power of genome-wide methods.

1.6 Fast perturbations and single-cell readouts could overcome current limitations

Investigating stage-specific roles of TFs and their contribution to the differentiation of various cell types from a common progenitor poses two major challenges.

First, TF perturbation must occur on a faster timescale than stage transitions and allow to observe its effect on enhancer binding and activity, and on gene expression, within the same stage. Since enhancer binding is almost simultaneously affected by changes in TF concentration, and enhancer activity can lead to transcriptional changes on a timescale of minutes (Shamir *et al.*, 2016; Dufourt *et al.*, 2021), the system needs to respond quickly to changes in TF availability, provided that its magnitude is sufficient to affect binding at the enhancer. Considering the timescale of changes within chromatin and in the transcriptional activation of new genes, the limiting factor is the duration of the shortest embryonic stage of interest, which can range from 10 min to 2 h (Weigmann *et al.*, 2003). Linking time-specific TF depletion to changes in total mRNA, protein and phenotypes introduces more time restrictions, due to mRNA and protein half-life and abundance, which may mask smaller, more recent transcriptional changes and extend beyond the selected stage. Fast perturbation must thus be combined with an up-to-date readout at the chromatin level.

Second, intermediate developmental states cannot be preventively isolated by examining previously characterized markers of differentiated cell types, and they may only be present in a rare and highly heterogeneous cell population. Single-cell approaches are a powerful instrument to investigate such a system: they are sensitive enough to identify rare cell populations and capture granularity, and they enable *de novo* identification of cell types and their markers (Park *et al.*, 2018; Aizarani *et al.*, 2019).

1.6.1 The iLEXY system enables fast stage-specific depletion of nuclear TFs

For a perturbation with effects at a timescale of minutes, it is crucial to bypass the lag due to mRNA and protein half-life. Conditional KO, KD or ubiquitination-mediated degradation of the TF of interest are too slow to study TF necessity at very rapid embryonic stages; similarly, conditional degradation with the auxin-induced degron (Nishimura *et al.*, 2009; Holland *et al.*, 2012) requires hours to be effective, and even more fundamental, it is very difficult if not impossible to deliver auxin to all cells within an embryo in a uniform manner. A faster way to alter TF availability is to sequester the TF in the cytoplasm by unmasking a nuclear export signal upon external stimulation. This stimulation should be orthogonal, cover the whole embryo, scale to a large number of embryos and be minimally invasive, to avoid introducing unwanted perturbations. A huge limitation of inducible perturbations in *Drosophila* embryos is their inability to uptake external substances through the vitelline membrane (Rand *et al.*, 2010); therefore, exogenous molecules need to be injected into each embryo, limiting the number of embryos which can be processed at a given time and introducing additional stress. Optogenetic stimulation is a valid and well-established alternative to overcome this limit (Guglielmi *et al.*, 2015; Johnson *et al.*, 2017; Kaur, Saunders and Tolwinski, 2017; Kögler *et al.*, 2021). In my Ph.D. project, I

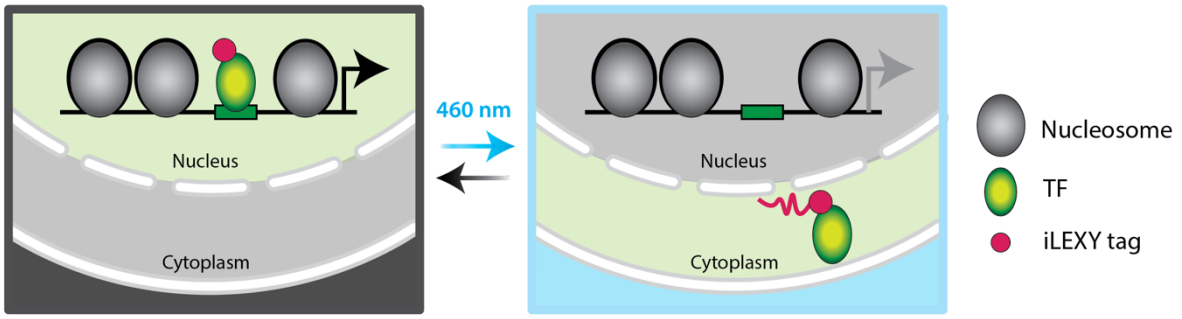


Figure 10: The iLEXY system enables rapid and reversible TF nuclear export.
In the dark (left), the TF (in green) with the iLEXY tag (in red) localizes to the nucleus as in the WT. Upon exposure to blue light (right), the modified Jα helix exposes 3 nuclear export signals, leading to TF relocalization to the cytoplasm. The perturbation is reversible in the dark. The green hue in the cell indicates the expected localization of the iLEXY-tagged TF in immunostaining.

used the improved light-inducible nuclear export system (iLEXY) developed by a former Ph.D. student in the lab, Anna Kögler, to trigger TF nuclear export.

In the iLEXY system, the endogenous TF of interest is fused to the plant-derived AsLOV2 domain, which changes its conformation in response to blue light by exposing the Jα helix from the core (Figure 10). In the engineered AsLov2, three copies of the nuclear export signal (NES) are incorporated in the Jα helix to counteract the endogenous nuclear localization signal (NLS) of the TF (Kögler *et al.*, 2021). Thus, blue light exposes the NES on the Jα helix promoting nuclear export of the TF, which relocates to the cytoplasm and cannot exert its original function on chromatin. Blue light induction leads to a rapid depletion, with a half-life < 30 s, which can disrupt the developmental processes triggered by a TF, as demonstrated for Twi (Kögler *et al.*, 2021). Moreover, this perturbation is reversible in the dark, and its strength and recovery time can be tuned by using different iLEXY variants.

Despite its huge potential, this system has some caveats. First, shuttling of the TF between the nucleus and cytoplasm is an equilibrium process where the equilibrium concentration in the nucleus under physiological conditions is significantly higher than in the cytoplasm, and delocalising a significant proportion of the nuclear TF will bring the cytoplasmic concentration far from its basal levels. Under such conditions, the rate of TF relocalisation to the nucleus will increase, likely causing leakage and incomplete depletion (Kögler *et al.*, 2021). On the contrary, a process involving TF degradation might lead to a more complete depletion by decreasing the total TF level.

For the same reason, cells with higher TF concentration may have a lower response to optogenetic perturbation. Within each cell, enhancers may show heterogeneous depletion levels: high-affinity enhancers will be depleted less efficiently than low-affinity ones. Despite this limitation, this technique can help classify enhancers based on actual affinity instead of motif strength, and analyse the components contributing to enhancer affinity and the role of TF-TF interactions. Moreover, the heterogeneity in depletion levels across cells can signal different cell identities, which may be differently perturbed upon depletion. Finally, less extreme effects might preserve the intermediate cell states which would be short-lived in the null mutant.

1.6.2 Molecular changes in the GRN over development can be captured by bulk and single-cell readouts

To capture heterogeneity in the chromatin state at single-cell level, single-cell measurements of TF binding would be ideal. However, state-of-the-art techniques are not sensitive enough due to high sparsity and dropout rates combined with limited, small binding sites (Gopalan and Fazzio, 2022). An alternative approach is to combine quantitative bulk TF binding data from ChIP or CUT&Tag with single-cell accessibility data from single-cell ATAC-seq: at a single enhancer, accessibility can be used as proxy for the binding of any TF, while bulk TF binding data indicate which TFs can bind at those positions (Calderon *et al.*, 2022; Secchia *et al.*, 2022). At enhancers bound by a single TF, accessibility will be uniquely associated to binding of the given TF. Combinatorically bound sites are harder to interpret, since the accessibility signal will result from the signal from each bound TF, and may have different outcomes according to the underlying TF interactions (Spitz and Furlong, 2012; Park *et al.*, 2019; Waymack *et al.*, 2020; Kim and Wysocka, 2023) (Figure 5). In case of compensation, after removal of a bound TF, the enhancer is still active and accessible to a similar extent due to the action of other TFs; instead, in case of high cooperativity, loss of one TF could impair binding of all others, leading to loss of enhancer accessibility. Enhancers with high affinity motifs for a given TF will likely not show accessibility changes, resembling highly compensated enhancers – in both cases, adding towards the robustness of the transcriptional programme. On the other hand, highly cooperative sites may be involved in mutually exclusive developmental fates. Overall, despite not providing a full picture of all TF-TF interactions, accessibility data in combination with occupancy data or information on TF motifs can be used to identify single TF binding, cooperativity, additivity and competition at enhancers.

Regardless of binding, accessibility can reflect enhancer activity: the intensity of the accessibility signal can often be used to approximate the activity of enhancers in multiple cell types (Bozek *et al.*, 2019; Shashikant and Ettensohn, 2019; Reddington *et al.*, 2020), and cell-type specific accessibility can predict *in vivo* enhancer activity in > 80% of tested cases (Cusanovich *et al.*, 2018; González-Blas *et al.*, 2020). This principle holds true globally, although some enhancers may be occupied but not active due to priming (Liber *et al.*, 2010; Mercer *et al.*, 2011) or binding of ubiquitous factors or repressors (Uyehara *et al.*, 2017). For enhancers with known targets, the relationship between accessibility and activity can be confirmed by the accessibility at the gene body of the target gene due to RNA polymerase II elongation, which is a good proxy for gene expression.

Another advantage of single-cell accessibility is the possibility to identity novel tissue-specific enhancers which may be active only in a cell subpopulation (Cusanovich *et al.*, 2018; González-Blas *et al.*, 2020), and be driven by TFs beyond the classic set known for the mesoderm GRN, overcoming technical and practical limitations in the number of tested TFs and the availability of appropriate antibodies. This is particularly informative in case of mutant cells which may deviate from the wild-type (WT) developmental trajectory.

Secchia *et al.* (2022) used single-cell ATAC-seq in loss-of-function mutant embryos to investigate mesoderm TFs at the end of mesoderm diversification, at 10–12 h. In *tin* and *bin* KO embryos, the

mutant cells still aligned on the WT developmental trajectory but had significantly lower numbers in the target tissues, i.e. in the cardiogenic and in the visceral mesoderm, respectively. Only the *Mef2* KO showed two mutant cell clusters detaching from the WT developmental trajectory, which were characterized by loss of accessibility at Mef2 enhancers, pointing to the role of Mef2 as an activator for the muscle cell state. This result uncovered the molecular phenotype of KO mutants when tissue differentiation is close to completion; however, it could still include epistatic and indirect effects, and it does not provide any information on the developmental trajectory of the mutant cells or on the roles of those TFs in previous developmental events. Thus, a time-specific perturbation followed by an immediate readout is required to explore intermediate cell fates and address the roles of mesoderm TFs at crucial stages in mesoderm differentiation and diversification.

1.7 The genome is hierarchically organized in three dimensions

In the nucleus, the chromatin is organized at many scales, leading to a hierarchy of compacted structures from large chromosomal territories to small looping interactions (Figure 11). The first evidence of nuclear organization was the microscopy observation of the differential chromatin density between active euchromatin and silenced heterochromatin (Heitz, 1928), later integrated with a more precise localization: heterochromatin is enriched mainly at the nuclear periphery close to the lamina, in the nucleolus and in the interior part of the chromosomes, whereas euchromatin is present at the chromosome surface (Cremer and Cremer, 2001; Guelen *et al.*, 2008; van Bemmel *et al.*, 2010; van Steensel and Belmont, 2017). Initial studies also focused on the association between chromatin state and transcription: the euchromatic chromosome surface was associated to highly transcribed genes, with subdivisions based on the chromatin state within single chromosomes (Cremer and Cremer, 2001). Later, Pombo and Dillon (Pombo and Dillon, 2015) identified supra-chromosomal clusters of active genes enriched in RNA polymerase II, also called transcription factories. This observation raised interest for subnuclear membrane-less compartments such as nuclear bodies, which engulf proteins associated to specific roles in DNA and RNA metabolism, from gene regulation to DNA repair and post-transcriptional processing (Chen and Belmont, 2019). Nowadays, condensates and phase separation are considered to be generally involved in organizing chromosome separation (Boija *et al.*, 2018; Ahn *et al.*, 2021; Demmerle, Hao and Cai, 2023).

On a smaller scale, looping domains in the tens of kb to Mb range further contribute to genome architecture, chromatin states and gene regulation. Known as topologically associating domains (TADs), they were first identified by microscopy as regions with coordinated replication, by pulse labelling of DNA with fluorescent nucleotide analogues (Jackson *et al.*, 1993), and rediscovered as domains with increased internal interactions within chromosomes over 25 years later (Lieberman-Aiden *et al.*, 2009).

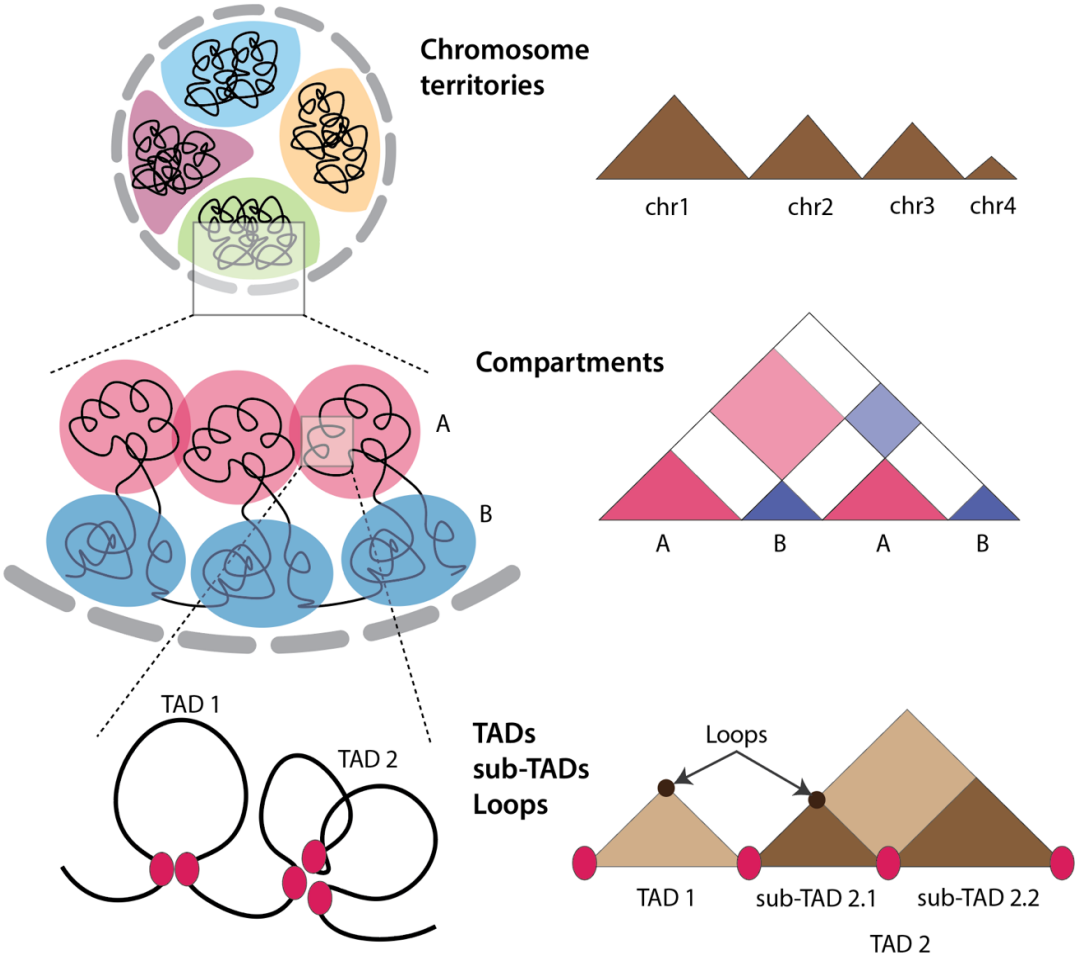


Figure 11: 3D genome organization.
Schematic representation of chromatin arrangement in the nucleus from large to small scale (left) and of the corresponding structures in Hi-C contact maps (right). Chromosomes occupy distinct territories in the nucleus, and within each chromosome there is separation between transcriptionally active (A) and transcriptionally inactive (B) compartments. Zooming in to the kb-Mb scale, the chromatin is organized into TADs, which can contain smaller domains (sub-TADs). Focal interactions between two sequences (loops) can also be detected. Adapted from Boltsis *et al.* (2021).

1.8 TADs are functionally important structures in 3D genome organization

TADs are genomic segments that self-interact with a higher frequency than with neighbouring regions (Dixon *et al.*, 2012; Nora *et al.*, 2012; Sexton *et al.*, 2012). TADs can be visualized by imaging-based techniques in single cells, and by genomic techniques like Hi-C in a cell population. TAD detection by imaging relies on *in situ* hybridization (ISH) of a tiling array of fluorophore-conjugated probes to the DNA in the fixed sample (DNA FISH, ORCA (Mateo *et al.*, 2019)); these techniques enable the reconstruction of the 3D structure of the genome in each cell. Here, TADs appear as looping structures sampling an ensemble of configurations across different cells with different probability. Therefore, slightly different loops will be detected in different cells or in the same cell across time, indicating that TADs are dynamic structures; nonetheless, the most represented configurations are the most frequent, which have a longer half-life. The optical reconstruction of TADs allows investigating TAD structures in single cells, but it is very low-throughput and restricted to a specific locus (Mateo *et al.*, 2019).

On the contrary, genomic techniques offer a genome-wide readout which is typically an average over hundreds of thousands or millions of cells, and therefore gives the most frequency configurations at the expense of single-cell resolution. Among conformation capture techniques, Hi-C is the most suitable to visualize 3D architecture at the TAD level (Lieberman-Aiden *et al.*, 2009). In Hi-C, interactions between genomic regions are detected by crosslinking the chromatin to preserve proximity between regions close to each other and by ligating the proximal regions to produce a chimeric fragment. The occurrences of each chimeric fragment correspond to the interactions between the two DNA sequences ligated. The interactions across the whole genome are represented in a frequency matrix where the axes indicate the positions along a chromosome, and the *(i, j)* value in the matrix corresponds to the frequency of the interaction between positions *i* and *j* in the genome. In a frequency (or contact) matrix, TADs appear as squares with higher contact frequency deviating from the diagonal with linear interactions.

The operative definition of TADs is based on the contact matrix from Hi-C experiments, which understates some conceptual simplifications. The Hi-C signal is the average chromatin conformation across millions of cells (Lieberman-Aiden *et al.*, 2009), and therefore a high interaction frequency may either represent a very frequent interaction in the cell population, or an extremely stable one in a smaller cell subset. In single cells, TADs are heterogenous and span an ensemble of conformations (Bintu *et al.*, 2018); thus, Hi-C results reflect the most abundant conformation across the population in time. This masks dynamics across the cell cycle, differentiation and tissues: indeed, TADs appear largely invariable between cell types and during differentiation (Phillips-Cremins *et al.*, 2013; Dixon *et al.*, 2015). In addition, the identification of TADs from Hi-C matrices is somewhat arbitrary. TAD boundary calling algorithms are based on differences in the interactions within a given radius from a sliding genome location: if the interaction frequency at a specific genomic location is a local minimum, the sequence will be more insulated than the neighbouring DNA and get classified as a boundary (Crane *et al.*, 2015). Algorithms generally agree on the position of most boundaries, but there is divergence in some calls (Zufferey *et al.*, 2018). Moreover, the resolution of the contact matrix used for TAD boundary calls has a significant impact on the definition of boundaries and TADs: at higher resolution, both TADs and boundaries will be smaller and more numerous (Dali and Blanchette, 2017).

Moreover, a huge limitation of TAD callers is the assumption that only consecutive boundaries form a TAD. Nonetheless, consecutive boundaries differ across resolutions, and the TADs found at each resolution likely reflect hierarchical levels of organization present at the same time. Indeed, TADs may have nested substructures with a similar conformation known as nested TADs or sub-TADs (Phillips-Cremins *et al.*, 2013), and boundaries in different TADs or sub-TADs could engage in looping interactions. Further, there are reports on domains driven by the transcription of a single gene (Rowley *et al.*, 2017; Hsieh *et al.*, 2020).

1.8.1 TADs are associated with gene regulation and chromatin states

Similar to higher organization levels such as lamina-associated domains and compartments, which are linked to repressed and active transcriptional states, TADs have functional transcriptional relevance:

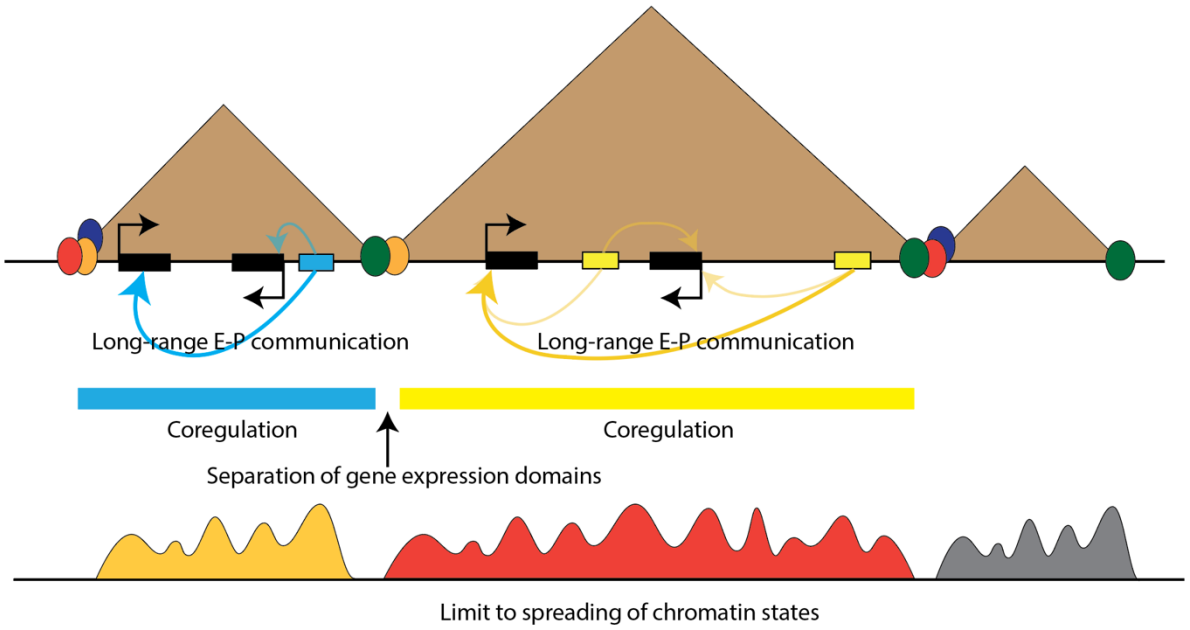


Figure 12: TADs contribute to gene regulation by modulating E-P contacts, and they are associated to coregulation and chromatin states.
Scheme of the functions proposed for TADs in gene regulation. TADs with genes (black rectangles), enhancers (blue/yellow rectangles) and insulator proteins (circles), with arrows pointing from the enhancer to the regulated genes (top); coregulation domains of the blue/yellow enhancer set (middle); distribution of chromatin marks (bottom).

they can constrain regulatory activity (Symmons *et al.*, 2014; Lupiáñez *et al.*, 2015), and are thought to enable distal enhancers to contact promoters within the same TAD, and form higher-order structures where E-P interactions can span multiple TADs (Figure 12).

TADs can constrain gene regulatory programs by modulating E-P contacts (Dehingia *et al.*, 2022 for a review) and by potentially limiting the spreading of chromatin states (Dixon *et al.*, 2012; Nora *et al.*, 2012; Le Dily *et al.*, 2014). Closer proximity within a TAD facilitates E-P interactions; often developmental genes within the same TAD are spatially and temporally co-regulated, pointing to the existence of a TAD-wide expression program (Nora *et al.*, 2012; Shen *et al.*, 2012; Laat and Duboule, 2013; Symmons *et al.*, 2014). Even reporters inserted into the same TAD often share a reproducible pattern (Symmons *et al.*, 2014). The search space restriction for E-P interaction is coherent with the presence of insulator proteins at the TAD boundaries, which by definition block an enhancer or promoter from communicating with a regulatory element on the other side of the boundary. For example, in mouse, insulators shield promoters from activation by spurious enhancers in the *α/δ T-cell receptor* locus (Zhong and Krangel, 1997) and in the *globin* locus (Hanssen *et al.*, 2017), and insulator deletion may lead to gene misexpression and morphological defects (Hagstrom, Muller and Schedl, 1996; Zhou *et al.*, 1996, 1999). Similarly, in *Drosophila*, deletion of some TAD boundaries leads to misregulation of genes in the neighbouring TAD, as seen for the *wnt4*/*wg* and *sog*/*CG12708* loci (Cavalheiro *et al.*, 2023).

The effects of TAD boundary perturbations may range from major gene dysregulation to very minor or no changes. Inversions typically have the most dramatic effects, as they change the relative position of

the enhancer and the boundary, in addition to increasing the distance between the enhancer and its endogenous target gene, as well as typically also causing deletions at the breakpoints. In the mouse *WNT6-EPHA4-PAX3* locus, deletions and inversions affect the number of digits (Lupiáñez *et al.*, 2015), and in the *HoxA* and *HoxC* loci boundary deletion leads to homeotic transformations (Narendra *et al.*, 2016). Changes in sub-domain organization within a TAD may also have important effects: in the *alpha globin* locus, removal of sub-TAD boundaries causes aberrant gene upregulation (Hanssen *et al.*, 2017). Similarly, aberrant gene expression in cancer (Flavahan *et al.*, 2016; Vian *et al.*, 2018) and neurological disorders, such as schizophrenia (Gallagher *et al.*, 2017) and frontotemporal lobal degeneration (Rajarajan *et al.*, 2018), have been associated to genetic variants at insulator binding sites in TAD boundaries. Similarly, at the *Notch* locus in *Drosophila*, TAD fusion and long-range topological alterations caused transcriptional defects, as a result of the loss of insulator proteins and RNA polymerase II binding (Arzate-Mejía *et al.*, 2020).

However, at other loci, fusion of TADs or inversions have little impact on gene expression, including the mouse *Sox9-Kcnj2* locus (Despang *et al.*, 2019), *Shh* (Williamson *et al.*, 2019), *TFAP2A* (Laugsch *et al.*, 2019), and *HoxD* (Rodríguez-Carballo *et al.*, 2019) loci. More large-scale studies, which were unbiased to the loci or genes involved, also showed that in many cases the disruption of TADs in both *Drosophila*, in the context of balancer chromosomes (Ghavi-Helm *et al.*, 2019), and in humans, in the context of cancer (Akdemir *et al.*, 2020), only impacts the expression of a small subset of genes, in the range of 10-14%. Similarly, depletion of CTCF protein in mESCs, which abolished > 80% of TAD boundaries, resulted in only a few hundred genes with changed expression (Nora *et al.*, 2017). This suggests that additional mechanisms likely ensure a tighter control of E-P interactions. This effect is more apparent at single loci.

TADs often coincide with domains with specific epigenetic modifications: their boundaries tend to mark the transition between active and repressive chromatin domains in mammals (Nora *et al.*, 2012; Le Dily *et al.*, 2014; Dixon *et al.*, 2015) and in *Drosophila* (Hou *et al.*, 2012; Ulianov *et al.*, 2016). This may prevent aberrant gene activation or silencing in specific loci: for instance, removal of insulator binding sites in the *Hox* locus leads to the spreading of active chromatin and the aberrant transcriptional activation of *HoxA7*, causing a homeotic transformation (Narendra *et al.*, 2016). In *Drosophila*, one study subdivided TADs into four classes based on their chromatin state: active, null (not enriched in any known marks), heterochromatic (enriched in the heterochromatin protein HP1 and linked to centromeres) and polycomb (PcG) (Sexton *et al.*, 2012). However, TADs are unlikely to have exclusive chromatin states, as they contain many genes which are often not co-expressed.

1.8.2 The functional importance of TADs is linked to the properties of the insulator proteins at their boundaries

Insulators (described in section 1.1.3) were first discovered in *Drosophila* as DNA sequences capable of shielding transcription from the influence of neighbouring elements and of blocking E-P activation when inserted in between.

Locus-specific perturbations in the *Drosophila Hox* gene locus confirmed that insulators contribute to inhibiting cross-talk between enhancers (Gyurkovics *et al.*, 1990) and thereby suppress the ectopic expression of the Hox gene *Abdominal-A*. Notably, other insulators in the same locus are required for long-distance E-P communication (Postika *et al.*, 2018; Kyrchanova *et al.*, 2019)

The grammar of insulator sites – i.e. their number, type, orientation and interactions – have been proposed to substantially contribute to their function in E-P communication (Kyrchanova, Chetverina, *et al.*, 2008; Nègre *et al.*, 2010; Ramírez *et al.*, 2018). A single insulator placed between an enhancer and promoter, within a transgenic reporter, obliterates E-P interactions, but when two copies of an insulator are placed between the enhancer and promoter, they cancel each other out, leading to insulator bypass (Muravyova *et al.*, 2001) (Figure 13). While the mechanism was not understood at the time, this is due to the creation of an insulator-insulator interactions, which loops the intervening DNA out of the way, such that the enhancer and promoter are even closer together. Similarly, stacked insulators in the right configuration can bring enhancers closer to their target promoters. Loop stacking by insulator proteins and their interactors are involved in many levels of 3D organization even in mouse, contributing to the robustness of gene expression (Hafner *et al.*, 2023).

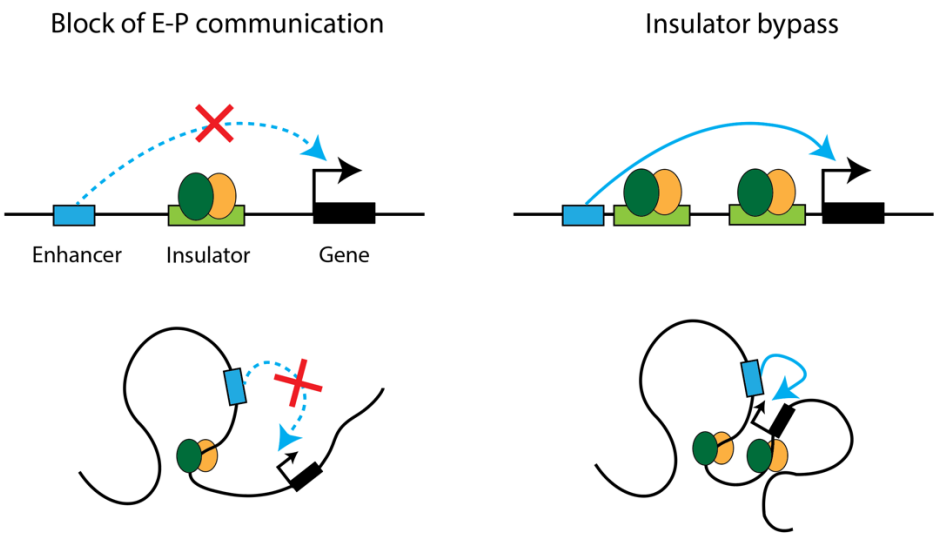


Figure 13: Insulator bypass.
A single insulator blocks E-P communication (left), but two copies of the same insulator can promote E-P communication by insulator bypass (right). Representation of the linear DNA sequence (top) and of the 3D arrangement of the loops (bottom).

Insulator composition can also impact insulator function. Insulators are bound by different insulator proteins: *scs* has Zw5 binding (Gaszner, Vazquez and Schedl, 1999), *scs'* has multiple BEAF-32 binding sites (Gilbert, Tan and Hart, 2006), *gypsy* contains 12 Su(Hw) sites (Gdula, Gerasimova and Corces, 1996) and the *H19* locus is bound by CCCTC-binding factor CTCF (Kurukuti *et al.*, 2006). The identity and the arrangement of the bound insulator proteins are crucial for insulator function, especially in *Drosophila* where many insulator proteins exist: interchanged insulators may not have the same enhancer-blocking strength (Hogga *et al.*, 2001; Iampietro *et al.*, 2008; Kyrchanova *et al.*, 2016) and insulator function may be compromised upon loss of ≥ 1 insulator protein (Gerasimova and Corces,

1996). Accordingly, insulators rely on specific insulator protein combinations for their function. The insulator proteins were subsequently found to be enriched at TAD boundaries, which lead to the hypothesis that TAD boundaries might act as insulator to restrict regulatory element activity across the boundary. In *Drosophila*, these proteins are often present in combination at boundaries, potentially ensuring robustness and effectiveness of insulation.

Insulators not only impinge on E-P communication, but a subset also contributes to enhancer binding by blocking the spread of heterochromatin (Gaszner and Felsenfeld, 2006). This was described as "position effect variegation" in the *white* gene locus in *Drosophila*, where X-ray induced mutations could lead to a combination of WT red and white mutant facets in the same eye, pointing to a functional gene which could not be expressed in some cells due to a heterochromatic context (Muller, 1930). This is consistent with the correlation between TADs and chromatin states (Sexton *et al.*, 2012).

In summary, the insulator proteins at TAD boundaries explain many TAD properties and may harbour the necessary information to make a functional TAD.

1.8.3 Mammalian TADs are marked by convergent CTCF sites and formed by loop extrusion

Although insulators were first discovered in *Drosophila*, most of our current knowledge on TADs comes from vertebrates, particularly mammals. In mammals, TAD boundaries are marked by convergent CTCF motifs. CTCF depletion in *trans* causes loss of > 80% of the boundaries (Nora 2017), and removal of typically multiple CTCF sites in *cis* disrupts topology at multiple genomic loci (Splinter *et al.*, 2006; Sanborn *et al.*, 2015). Conversely, addition of a CTCF site in convergent orientation can trigger formation of a new TAD boundary. Therefore, CTCF is necessary and sufficient for most boundaries in mammals. The convergence of CTCF motifs was first noticed by Rao *et al.* (2014), confirmed by Nanni, Ceri and Logie (2020), and proven by CTCF inversion (de Wit *et al.*, 2015) and genome editing of CTCF motifs, which disrupts loops containing the edited motif and extends the TAD to the next convergent CTCF motif (Sanborn *et al.*, 2015).

The functional impact of those mutants is consistent with predictions from the loop extrusion model, which integrated polymer modelling with data on CTCF and cohesin localization (Sanborn *et al.*, 2015; Fudenberg *et al.*, 2016). Whereas CTCF is recruited at its motifs in convergent orientation at TAD boundaries, the cohesin complex is specifically loaded and unloaded on the genome by accessory proteins (Peters, Tedeschi and Schmitz, 2008), but both colocalize at TAD boundaries and loops (Wendt *et al.*, 2008; Rao *et al.*, 2014). Once loaded on the genome, cohesin starts to extrude a chromatin loop until stalling at convergent CTCF sites. This stalling increases the residency time of cohesin and constrains the two cohesin anchors in tight proximity, resulting in persistent loops, visible as focal interactions at TAD corners in mammalian Hi-C. Single-molecule interactions between CTCF and cohesin on DNA have been observed *in vitro*, where CTCF sites could sequester freely diffusing cohesin, probably causing its accumulation at TAD boundaries, and to block extruding cohesin, to stabilize the TAD-forming loop (Davidson *et al.*, 2023). DNA tension contributes to CTCF action by

establishing the direction of loop extrusion (Sanborn *et al.*, 2015; Davidson *et al.*, 2023). The need for cohesin in loop extrusion was confirmed by removing some components of the cohesin complex. In fact, auxin-induced degradation of the core component Rad21 in a human cell line caused rapid loss of loop domains, which was partly compensated by strengthening of compartments with chromatin marks (Rao *et al.*, 2017); a similar phenotype was observed with depletion of the cohesin loader NIPBL in mice (Schwarzer *et al.*, 2017). Conversely, TAD borders were strengthened upon removal of the cohesin release factor Wapl (Haarhuis *et al.*, 2017).

Currently, the loop extrusion model with CTCF-cohesin interactions is the prevailing model for TAD formation in mammals; however, it does not explain the ~15% of boundaries which do not show CTCF binding (Nora *et al.*, 2017). Those might be associated with active transcription, which could stall cohesin in the absence of CTCF. Indeed, the R-loops formed at the beginning of transcription by hybridization of the nascent RNA with the template DNA can stall cohesin *in vitro*, albeit secondary to CTCF (Zhang *et al.*, 2023). Moreover, TAD boundaries are generally accessible and transcriptionally active, which may ensure their recognition and binding by insulator proteins; thus, transcription and CTCF binding could reinforce each other via boundary-associated RNAs (Islam *et al.*, 2023). Several boundaries, in both mammals and flies, overlap with the transcription of mRNA and tRNA-coding genes and the presence of RNA polymerase II or III, respectively (Dixon *et al.*, 2012; Rao *et al.*, 2014; Yuen, Slaughter and Gerton, 2017). The establishment of boundaries in developing mammalian embryos occurs at transposable retroelements, which are transiently active and could facilitate boundary positioning (Zhang *et al.*, 2019; Choudhary *et al.*, 2023). In addition, the cell-specific boundaries formed during differentiation often correlate to transcription and TF binding (Bonev *et al.*, 2017; Stadhouders *et al.*, 2018). Similarly, the formation of a transcription-dependent TAD boundary is critical for antibody class switch recombination (Costea *et al.*, 2023). However, this may depend on the context, as in one instance Cas9-induced transcriptional activation was not sufficient to establish a TAD boundary during neuronal differentiation (Bonev *et al.*, 2017). Alternatively, RNA instead of transcription could stabilize insulator proteins at TAD boundaries. Indeed, boundary-associated enhancer RNAs are necessary to induce boundary-associated transcription, facilitate CTCF recruitment at TAD borders and increase TAD insulation (Islam *et al.*, 2023). Additional interactors are also suggested by the observation that some conserved TAD boundaries with CTCF binding are maintained even after deletion of the CTCF binding site (Barutcu *et al.*, 2018), and by the discovery that the TFs of the ZBTB family facilitate cohesin binding in mice (R. Wang *et al.*, 2023).

Despite the different actors involved in stalling cohesin or stabilising insulator protein binding, the loop-extrusion model is compatible with both *in vitro* and i*n vivo* evidence. Thus, it can explain the behaviour of boundaries with multiple CTCF motifs in different orientations and more complex loop stacking configurations (Anania *et al.*, 2022).

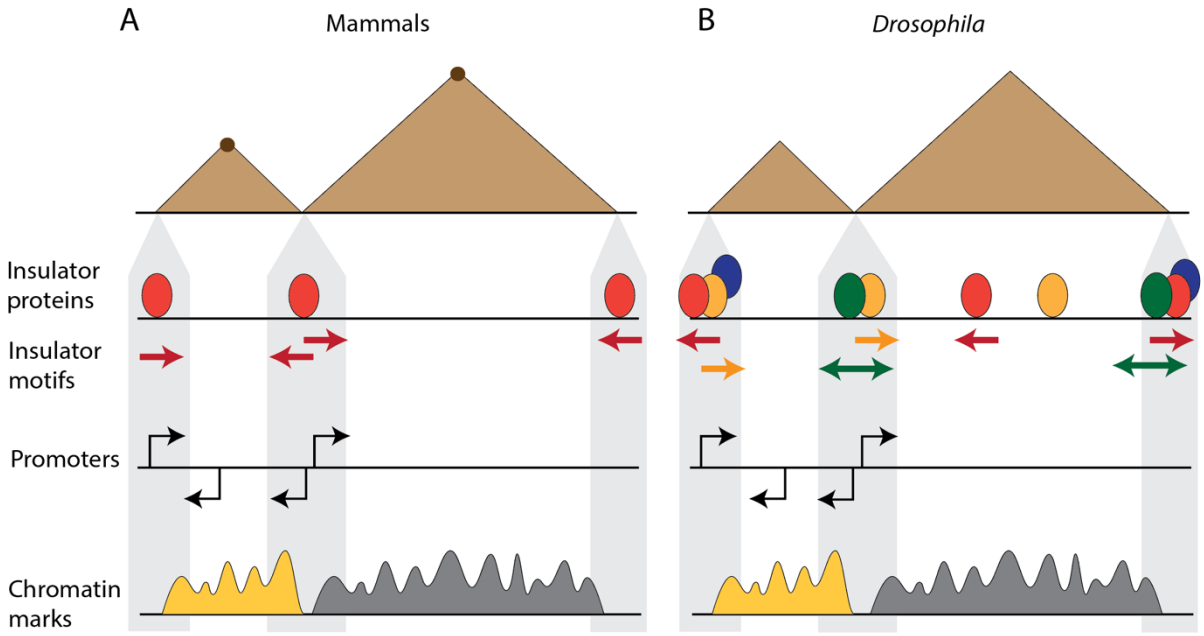


Figure 14: Features of TAD boundaries in mammals and in Drosophila.

(A) In mammals, TADs are marked by high-frequency looping interaction involving convergent CTCF sites (in red), and also present promoters and a separation of chromatin marks.

(B) In Drosophila, TAD boundaries have multiple insulator proteins, both direct (BEAF-32, CTCF, Su(Hw)) and indirect binders (CP190). The presence of promoters and the separation of chromatin marks are in common with mammals.

1.8.4 Drosophila TAD boundaries show a combination of insulator proteins

In *Drosophila*, TAD boundaries are enriched in CTCF and multiple other insulator proteins, including both direct and indirect binders.

The insulator proteins such as the historical insulator proteins BEAF-32, CTCF, Su(Hw), and more recently described insulator proteins like GAF, M1BP, Ibf1, Ibf2, Pita, Zw5 and ZIPIC directly bind DNA and recognize their own motif. Other insulator proteins, like CP190 and Mod(Mdg4), can only bind to TAD boundaries indirectly, and colocalize with a direct binder that targets them to the correct DNA region (Özdemir *et al.*, 2019).

The direct binding insulator proteins recognize their motif through C2H2 zinc-finger domains (CTCF, Su(Hw), Pita, ZIPIC) (Kim *et al.*, 1996; Gaszner, Vazquez and Schedl, 1999; Moon *et al.*, 2005), a single zinc-finger (Ibf1, Ibf2) (Cuartero *et al.*, 2014) or a zinc-finger BED domain (BEAF-32) (Avva and Hart, 2016). They also have another domain for protein-protein interaction: for instance, the ZAD domain (in Pita and ZIPIC) (Zolotarev *et al.*, 2016), hypothesized to mediate dimerization and necessary for nuclear localization, the BTB/POZ homodimerization domain (in CP190 and Mod(Mdg4)) (Melnikova *et al.*, 2017, 2018), and the BESS domain (BEAF-32) (Avva and Hart, 2016).

Multimerization and co-binding are important features of *Drosophila* insulators proteins: CTCF, Zw5, ZIPIC, Pita, CP190 and Mod(Mdg4) can homodimerize *in vitro*, and CP190 and Mod(Mdg4) can form

multimers (Kyrchanova and Georgiev, 2014). Multimerization could facilitate and stabilize co-binding at TAD boundaries.

The co-binding of insulators at TAD boundaries is not random, it actually depends on the boundary type. Boundaries with a promoter are enriched in M1BP, BEAF-32 and ZIPIC motifs, whereas non-promoter boundaries are associated to CTCF, Su(Hw) and Ibf (Ramírez *et al.*, 2018). Among indirect binders, the cohesin subunit Rad21 is mostly enriched at M1BP promoter boundaries, and CP190 at non-promoter boundaries. Notably, some motif combinations are associated with stronger boundaries, but the convergent CTCF sites from mammals have not been observed to date in *Drosophila*. The mechanism of TAD formation in *Drosophila* is therefore not clear.

To simplify insulator combinatorics, multiple studies addressed the need for insulator proteins through single and combined depletion. Single maternal depletions of BEAF-32, CTCF and CP190 in *Drosophila* embryos have limited global effects on TAD establishment, but locally they each disrupted about 10% of the boundaries, representing different subsets (Cavalheiro *et al.*, 2023). Therefore, at most boundaries, single insulator proteins are not necessary for TAD establishment or maintenance. Similar results have been obtained using BEAF-32 RNAi in S2 cells (Ramírez *et al.*, 2018), by *CP190* or *CTCF* KO in Ras3 cells (Kahn *et al.*, 2023), and by *CTCF* KO in larvae (Kaushal *et al.*, 2021). This points either to a highly redundant, robust system, or to unknown additional regulators. However, the fact that some boundaries (6–10%) with multiple insulators have different sensitivity to TF depletion suggests that they are not simply redundant, there is some specificity (Cavalheiro *et al.*, 2023).

1.8.5 Mechanisms other than loop extrusion could establish Drosophila TAD boundaries

To date, there are currently no studies that have tested cohesin mediated loop extrusion in *Drosophila*. Multiple alternative mechanisms for compartmentalization have been proposed so far.

First, given the co-occurrence of promoters at TAD boundaries observed in *Drosophila* (Dixon *et al.*, 2012; Sexton *et al.*, 2012) and in many other species, including mammals, transcription may contribute to TAD boundaries. In fact, 77% of *Drosophila* TAD boundaries colocalize with promoters (Ramírez *et al.*, 2018), and most are constitutively expressed, since they belong to housekeeping genes; moreover, some TFs recruited by promoter motifs (M1BP, Dref) are enriched at TAD boundaries (Hou *et al.*, 2012; Rao *et al.*, 2014; Ramírez *et al.*, 2018). However, the subtle effect of inhibition of transcriptional elongation by RNA polymerase II in *Drosophila* embryos and cell lines on genome-wide insulation at TAD boundaries implies that transcription is not a crucial factor in boundary maintenance (Hug *et al.*, 2017; Jiang *et al.*, 2020). Nonetheless, this does not exclude the contribution from the presence of the polymerase alone, which could help stabilizing a local protein hub.

A second candidate for compartmentalization is the chromatin state. In *Drosophila*, compartments appear more accentuated than in mammals, and TAD boundaries can coincide with compartment boundaries. The observation that *Drosophila* TADs are often associated to a chromatin state (Sexton *et*

al., 2012) could be due to the fact that chromatin modification within the TAD, and not the TAD boundaries *per se*, define the TAD extension. This would be in agreement with the ability to predict TADs from histone marks and transcriptional state (Ulianov *et al.*, 2016; Rozenwald *et al.*, 2020). However, the biological mechanism whereby a chromatin mark would partition a TAD is unclear. Given the evidence that TAD boundaries can prevent the spread of chromatin states to a neighbouring region (Chathoth *et al.*, 2022), the segregation of chromatin marks would result from TAD boundaries, rather than the other way around.

To date, the contribution of insulator proteins, polymerase and chromatin state to TAD boundaries in *Drosophila* is under debate, and no definite single mechanism has been proven to drive or be sufficient for TAD formation. Understanding TAD formation in flies could shed light on processes contributing to 3D organization in other species, while the genome topology of other species could inform a fly-centric approach.

1.8.6 TAD formation may rely on orthogonal evolutionary mechanisms

The multiscale 3D organization of chromatin is conserved throughout evolution and linked to different features in different species (Szabo, Bantignies and Cavalli, 2019). Plants typically show compartments marked by transcriptional activity and epigenetic signatures (Feng *et al.*, 2014; C. Wang *et al.*, 2015; Dong *et al.*, 2018), whereas animals have TADs marked by different insulator proteins at their boundaries.

Whereas the cohesin-mediated looping interaction between convergent CTCF sites is a strongly conserved TAD feature in mammals (Nasmyth and Haering, 2009; Vietri Rudan *et al.*, 2015; Kabirova *et al.*, 2023), other coexisting mechanisms could be observed in the absence of cohesin: condensin II (an SMC complex like cohesin) and the general transcription factor TFIIIC are enriched at TAD boundaries (Yuen, Slaughter and Gerton, 2017), and condensin can extrude naked DNA *in vitro* (Ganji *et al.*, 2018), while transcription-generated R-loops cause cohesin stalling and might have the same impact on condensin (Zhang *et al.*, 2023). Moreover, global loss of TADs upon cohesin depletion results in an average of highly randomized positions of TAD borders in single cells (Bintu *et al.*, 2018), pointing to the role of cohesin in stabilizing the location of TAD borders but not in creating TAD-like domains. Therefore, even in mammals, additional mechanisms could contribute to TAD formation.

In the nematode *Caenorhabditis elegans*, TAD-like self-interacting domains of 1 Mb can be observed mainly on the X chromosome, and only sporadically in autosomes (Crane *et al.*, 2015). In hermaphrodites, the X chromosome is bound by the dosage compensation complex (DCC), a condensin complex fundamental to reinforce weak TAD boundaries and form new ones by increasing insulation. The DCC is also involved in long-range contacts. This points to a potentially more ancient role of condensin, although its interactors remain unknown.

Plants do not uniformly show TAD structures. *Arabidopsis* has compartmental domains linked to the open or close chromatin state, whereas other plants, including rice, cotton, maize and tomato, have

TAD-like domains linked to the same four chromatin types as *Drosophila* (active, null, PcG and heterochromatin domains) (Dong *et al.*, 2017, 2018; Liu *et al.*, 2017; Wang *et al.*, 2018). So far, no insulators have been reported in plants; thus, this process seems to rely on chromatin state (C. Wang *et al.*, 2015) and, at least in some cases, on active transcription at boundaries (Dong *et al.*, 2017), or on yet unknown factors.

TAD domains have also been described for simpler organisms such as bacteria and yeast. In the yeast *Schizosaccharomyces pombe*, they are known as globules, span 50–100 kb in size and are associated to cohesin binding and active transcription (Duan *et al.*, 2010; Mizuguchi *et al.*, 2014; Hsieh *et al.*, 2015; Benedetti *et al.*, 2017). Among bacteria, *Caulobacter crescentus* (Le *et al.*, 2013) and *Bacillus subtilis* (Marbouty *et al.*, 2015) have chromosomal interaction domains (CIDs) similar to TADs in the 30–420 kb size range, delimited by highly expressed genes which are sufficient to create a new boundary. SMC complexes form processive loops in those species, which however does not seem necessary for CID formation in *C. crescentus* (Le *et al.*, 2013); thus, the interplay between loop extrusion and transcription-dependent boundaries in bacteria remains to be elucidated.

Overall, the presence of an SMC complex able to extrude DNA loops in *trans* and of transcription or insulators in *cis* is a common paradigm across evolution, with the exception of some plants. However, the mechanistic interactions between the two differ among species, and could even leverage multiple actors in the same species.

1.8.7 The CTCF-cohesin loop: far from a universal model

Cohesin stalling at CTCF convergent sites in TAD boundaries has been consistently described in mammals. However, about 20% of TAD boundaries are CTCF-independent (Nora *et al.*, 2012), and some species lack CTCF altogether (Heger *et al.*, 2012), whereas in others, convergent CTCF motifs do not stop loop extrusion (Matthews and White, 2019; for a review, Kabirova *et al.*, 2023). In fact, the CTCF-cohesin interaction could be recent. While cohesin could date back to bacteria, CTCF itself appeared only about 500 million years ago (Mya) in Bilateria, 30 My after the first aquatic vertebrates and close to the first colonization of the earth by nematodes (Heger *et al.*, 2012).

Differential sequence conservation across the domains of CTCF proteins points to potentially different interactomes across species, which could explain the different contributions of CTCF to genome topology: the C2H2 DNA-binding domain is conserved as is its motif, whereas the N-terminal domain, which harbours the YXF motif critical for CTCF-cohesin-mediated loop extrusion (Nora *et al.*, 2012), is highly variable. Notably, the YXF motif is not conserved in species where there is no evidence of CTCF-cohesin loop extrusion, including *Drosophila*. Therefore, there could be additional requirements for extrusion of a CTCF-cohesin loop, like the structure of the cohesin complex or further interactors: for instance, the motif for CTCF interaction with the PDS5 cohesin-associated factor A is conserved in jawed vertebrates but absent in *Drosophila* (Kaushal *et al.*, 2021). Moreover, Platyhelminthes and some nematode species, including *C. elegans*, underwent secondary CTCF loss (Heger *et al.*, 2012), pointing to alternative mechanisms.

1.8.8 TAD prediction: learning the cause or the effect of 3D organization?

TADs can be the product of different actors with an additive or a synergistic effect. Whereas the main drivers for TAD formation may be captured by experimental approaches like depletions of proteins in *trans* or DNA feature deletion in *cis*, additional contributions are more elusive. Moreover, evaluating the extent of each contribution and of their combination at genome-wide scale may be challenging for *trans* factors whose depletion is lethal or causes side effects.

To weight individual contributions and potentially find new roles of known factors in genome topology, machine-learning approaches have been applied to the prediction of chromatin interactions. While TAD boundary sequence can be reliably predicted, predicting interacting regions (i.e. the pairs of boundaries that will form a TAD) is more challenging, due to the difficulty in defining domains and their strong dependence on resolution. This leads to different choices of interacting DNA sequences in the training and test sets, which may impact the type of interactions best predicted by each model, and the top features associated to interacting sequences.

Even in the well-consolidated model of CTCF-cohesin loop extrusion in mammalian TADs, predictions from different groups emphasize different aspects of the interactions. Kai *et al.* (2018) predicted CTCF-mediated long-range interactions in mammalian cell lines using genomic and epigenomic features, mainly focusing on CTCF and cohesin binding, the loop length and motif pattern. This model outperformed others based only on CTCF motif orientation, showing that more features could significantly contribute to the identification of interactions. Abbas *et al.* (2024) obtained a similar result using a different set of input features, namely ChIP-seq for the cohesin subunit Rad21 and for epigenetic marks H3K27ac and H3K27me3. Strikingly, the presence of convergent CTCF motifs under CTCF peaks was associated with stronger interactions than in any other orientation, but could also be found at weaker sites in 50% of cases, versus 70% in the absence of a convergent motif. Therefore, although CTCF convergent sites improve boundary strength predictions, they are not necessary to predict a strong boundary nor are they critical for predicted interactions. This agrees with the experimental observation that the characteristics of individual CTCF binding sites can shadow motif orientation and number, pointing to redundancy and other nuances of CTCF motifs as a mechanism to ensure robustness (Anania *et al.*, 2022).

Although the computational models above were validated on the full WT genome and have some independent experimental correlates (Li *et al.*, 2020; Nanni, Ceri and Logie, 2020; Chang *et al.*, 2023), further advances have been made in predicting the effect of mutations. A mathematical model of CTCF-cohesin mediated loop competition and extrusion could predict looping interaction in the WT and upon mutation of CTCF sites in *cis* (Xi and Beer, 2021); further, a regression framework could detect and predict differential TAD borders after chromosomal rearrangement in humans and mice (Mourad, 2022).

In *Drosophila*, the high multiplicity of insulators and their combinations poses additional challenges in the prediction of interactions between TAD boundaries. Moreover, new insulators at TAD boundaries

are still being discovered, suggesting that our knowledge may still be partial. Various results point to different features related to TADs and TAD boundaries. For single boundaries, Ulianov *et al.* (2016) reported that active chromatin and transcriptional are better predictors than CTCF and Su(Hw), proposing a model of transcription-driven TADs as in *C. crescentus*. On the other hand, the sequence-based deep learning model by Henderson *et al.* (2019) found BEAF-32, Trl, some promoter motifs and other unknown or non-insulator motifs among the top features.

For TAD prediction, epigenetic marks were used as input for multiple model architectures (Rozenwald *et al.*, 2020). Since TADs are highly correlated with chromatin states, the model had a high predictive power; however, it does not provide insights into the underlying cause of TAD formation.

In all these methods, the choice of the boundary set, the separation of the training and test sets, feature annotation and the choice of background have a very significant impact on the outcome. Boundary predictions could provide insights on genome-wide data but should be validated experimentally. Another limitation is that they mostly extract a set of single features with little to no information on feature combinations within a boundary, or on the two elements of a boundary pair.

1.9 Aims

During my Ph.D., I worked on two projects, each presented in a separate results chapter:

- *Functional characterization of the mesoderm GRN*. In this project, I aimed to assess the causal relationship between TF binding at enhancers, co-binding of other TFs, activation of enhancers and promoters and how they link to phenotype within a GRN. I focused on the mesoderm GRN in *Drosophila* embryos. Experimentally, I combined TF depletion at precise timepoints with single-cell measurements, to investigate TF roles at pivotal stages for mesoderm development. In particular, I addressed the contribution of each TF to both enhancer occupancy and activity, the TF-TF interactions at single enhancers and the effect on the transcription of putative direct target genes. Further, I characterized the new cell states arising upon TF depletion, their molecular features and their relationship to the KO phenotype (Chapter 3).
- *Finding rules for TAD boundary function in* Drosophila. Here, I aimed to clarify which intrinsic features contribute to the function of a TAD boundary, and how two boundaries interact to form a TAD. To this end, I tested boundary function experimentally by inserting endogenous boundaries into an ectopic location in the middle of a TAD and checked if they still formed a boundary. Then, I examined the features of the working boundaries to find elements correlated to boundary function and pairing, and I evaluated if they could be generalized to TAD boundaries genome-wide (Chapter 5).

2 Materials and methods

2.1 Plasmid design and cloning

2.1.1 Generation of TF-iLEXY mutants by CRISPR cloning

The *tin-mCherry-iLEXY*, *Mef2-mCherry-iLEXY, bin-mCherry-iLEXY* lines were generated in collaboration with Dr. Katharina Bender as in Kögler *et al.* (2021). In this thesis, their names are abbreviated as *tin-iLEXY*, *Mef2-iLEXY* and *bin-iLEXY.*

The mCherry-iLEXY cassette was inserted at the C-terminus of the TF of interest by CRISPR/Cas9-mediated homology-directed repair: a CRISPR donor and a gDNA plasmid were co-injected into the *vasa-cas9 (III)* (w[1118]; PBac{y[+mDint2] GFP[E.3xP3]=vas-Cas9}VK00027 - BDSC #51324) line, which has *cas9* expression in the germline, to derive F1 flies with a heterozygous *TF-mCherry-iLEXY* allele.

The donor plasmids were derived from the pHD-Scarless-dsRed vector (DGRC #1364) with insertion of the mCherry-LEXY(V416L)-dsRed cassette from the plasmid pAct-MCS-XhoI-mCherry-LEXY-BstBI. The cassette was flanked by two homology arms for the correct genomic integration at the double-strand break introduced by Cas9 at the canonical PAM sequence (5'-NGG-3') targeted by the gRNA. For C-terminal fusion, the mCherry-iLEXY cassette is in frame with the TF and inserted before the stop codon. The *dsRed*, downstream of the stop codon and flanked by two TTAA sites for piggyBac transposase-mediated excision, is used as a marker to track successful insertion.

For scarless integration, I targeted the NGG site closest to the stop codon of the TF, and designed and PCR-amplified:

- a 2-kb left homology arm directly upstream of the stop codon, with a point mutation in the NGG to prevent further cutting by Cas9 after the first integration
- a genomic linker from the stop codon up to the first downstream TTAA sequence
- a 2-kb right homology arm downstream of the TTAA sequence

These elements were PCR-amplified from the genomic DNA of the vasa-cas9 (III) line and inserted individually into the pHD-Scarless-mCherry-LEXY(V416L)-DsRed vector by InFusion cloning (Takara). The iLEXY constructs are reported in Supplementary figure 1.

gRNA sequences were generated by oligo annealing and inserted into the BbsI site of the pU6-BbsI-gRNA vector. All oligos used for the cloning are listed in Supplementary table 1.

2.1.2 DNA FISH probe cloning

DNA fluorescent *in situ* hybridization (DNA FISH) probes were designed on TAD boundaries and anchors of high-frequency interaction, as visualized in the Hi-C browser Juicebox (Durand *et al.*, 2016).

Primers for amplification of the selected regions from genomic DNA were designed using Primer3 in the Primer Blast tool from NCBI (https://www.ncbi.nlm.nih.gov/tools/primer-blast/). The primers for DNA FISH probe cloning are reported in Supplementary table 8. After PCR and gel band extraction, the A-tailed purified DNA was ligated into the linearized pGEMT-Easy vector (#A1360, Promega), according to manufacturer's instructions. The vector with the inserted probe was used for fluorescent labelling to generate the final DNA FISH probes, as described in section 2.6.162. The insertion was confirmed by gel electrophoresis and sequence-verified with T7 and Sp6 standard primers.

Alternatively, BACs from the BACPAC Resource Center (Supplementary table 9) were directly used as probes after fluorescent labelling as I did for PCR-derived probes.

2.1.3 TAD boundary cloning

The selected TAD boundaries were amplified from genomic DNA and inserted into the pBS-KS-attB1-2-DsRed-loxP vector via InFusion cloning in the PstI site. The insertion site lays within a cassette that contains attB sites in inverted orientation for recombination-mediated cassette exchange with the attP sites in the acceptor fly line, and a reporter gene (*dsRed*) flanked by loxP sites (Supplementary figure 2). The oligonucleotides used for TAD boundary cloning are reported in Supplementary table 4, and the boundary coordinates in Supplementary table 6.

2.2 Drosophila genetics and husbandry

All fly lines were maintained on standard food at a temperature between 18 and 25°C. All fly lines used in my project are listed in Supplementary table 3 (mesoderm GRN) and Supplementary table 7 (TAD boundaries).

2.2.1 CRISPR/Cas-mediated homology-directed repair

The mCherry-iLEXY cassette was fused in-frame to the coding sequence of the endogenous *Mef2*, *tin*, and *bin* as described above (section 2.1.1), by CRISPR/Cas9-mediated homology-directed repair (Gratz *et al.*, 2013).

For injection into *Drosophila* embryos, I prepared a 20 µL mix of 250 ng/µL CRISPR donor plasmid and 50 ng/µL of the gRNA plasmid for *tin-iLEXY* and *bin-iLEXY*, and 225 ng/µL donor and 75 ng/µL gRNA plasmid for *Mef2-iLEXY*, as to keep a 2:1 donor:gRNA molar ratio. All vectors were prepared via a Midi Prep with the QIAfilter Plasmid Midi Kit (#12243, Quiagen) and further purified using the QIAquick PCR Purification kit (#28104, Quiagen). The mix was injected into *vasa-cas9 (III)* embryos by Dr. Alessandra Reversi of the EMBL *Drosophila* injection service. Positive transformants were screened by DsRed expression in the eyes of adult flies, which is visible under a fluorescent lamp. Positive transformants were used to establish a stable homozygous line. The correct incorporation of the *mCherry-iLEXY* tag was checked by PCR amplification and by sequencing of three overlapping amplicons from upstream of the left homology arm up to downstream of the right homology arm. Genotyping primers are reported in Supplementary table 2.

To prevent recombination between *mCherry* and *dsRed*, *dsRed* was removed by crossing homozygous *mCherry-iLEXY-dsRed* flies to a line expressing the PiggyBac transposase. *dsRed* excision and integrity of the remaining sequence were confirmed by PCR amplification and sequencing of a single or two overlapping amplicons from upstream of the left homology arm up to downstream of the right homology arm. For some lines, *dsRed* excision was not successful due to the low number of transformants and minor changes in the sequence, or to the reduced viability of the resulting fly line. Lines where *dsRed* excision was not successful were regularly monitored by checking fluorescence intensity in the eye, to ensure that no recombination happened.

2.2.2 TAD boundary insertions

TAD boundaries were inserted into TAD A or TAD B by recombination-mediated cassette exchange using the MiMIC system (Venken *et al.*, 2011).

The donor plasmid was injected into BDGP lines with the MiMIC cassette, the phiC31 integrase and Flp recombinase (BDSC #33194, BDSC #60957) by Dr. Alessandra Reversi from the EMBL *Drosophila* injection service. All male adults derived from the injected embryos were crossed to *yw* virgins. In the F1, the positive male transformants were screened for fluorescent red eyes from DsRed expression, and were crossed to virgin females with a GFP balancer chromosome. In the F2, positive siblings were crossed, and the F3 screened for homozygous mutants, marked by the absence of the GFP balancer. Homozygous viable stocks were maintained in homozygosity, whereas the homozygous lethal TAD B lines were maintained over the GFP balancer.

2.2.3 Handling of photosensitive embryos

The maintenance and handling of homozygous iLEXY lines were always performed in safelight conditions, avoiding exposure to blue or white (ambient) light, which also includes short (blue) wavelengths. To this end, room lights, stereoscopes and other light sources were covered with amber light filter foil (Lee Filter 135 Deep Golden Amber; #LF050135, Huss Licht & Ton), and red LED lamps were used during fixation of dark-reared samples. The fly stocks were either maintained in safelight rooms or in fibreboard boxes (#KA1100BU1, National Lab).

2.2.4 Sensitized background

To enhance the phenotypic effect of Tin depletion, *tin-mCherry-iLEXY* homozygous adult male flies were crossed to virgin females of the classic *tin* KO mutant tin^{EC40}*/TM3 Ser GFP*, to obtain 50% *tin-mCherry-iLEXY*/tin^{EC40} sensitized background embryos in the next generation. Recently eclosed flies from both lines were collected separately for up to 5 days and mixed in the same cage with a 2:1 female-to-male ratio just before the beginning of the embryo collection. Homozygous *tin-mCherry-iLEXY* flies and *tin-mCherry-iLEXY*/tin^{EC40} embryos were kept and handled under safelight conditions.

2.3 Genotyping

2.3.1 Dirty gDNA extraction for single-fly genotyping

Single flies were collected in a 0.5 mL Eppendorf tube, mashed for 10 s with a pipette and incubated at 37°C for 30 min in squishing buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 25 mM NaCl, 200 µg/µL proteinase K). At the end of the incubation time, proteinase K was inactivated at 95°C for 5 min. The extracted DNA was stored at 4°C.

2.3.2 Clean gDNA extraction for genotyping

A total of 50 flies were collected in an Eppendorf tube and snap-frozen in liquid nitrogen. Flies were crushed with a pestle in 250 µL of solution A (0.1 M Tris-HCl pH 9.0, 0.1 M EDTA pH 8.0, 1% SDS) and rinsed in 250 µL of the same solution. Crushed flies were then incubated at 70°C for 30 min and cooled down for 10 min at room temperature (RT). After addition of 70 µL of 8M K-Acetate, flies were incubated on ice for 30 min and centrifuged at 14000 rpm for 15 min at 4°C. Half a volume of isopropanol was added to the flies at RT and flies were spun at maximum speed for 5 min at RT. Then, the pellet was washed with 300 µL 70% ethanol, centrifuged, dried and resuspended in 100 µI of 5 mM Tris-HCI pH 8.5. After RNA digestion for 30 min at 37°C with RNAse 1 µg/µL, DNA was extracted by phenol-chloroform extraction and precipitated overnight. Following precipitation and centrifugation at 14000 rpm for 30 min at 4°C, the pellet was washed in ethanol, dried and resuspended in 50 µL TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0).

2.3.3 Genotyping the orientation of TAD boundary insertions

The PCR to detect the orientation of the boundary insertion was performed as in Venken *et al.* (2011).

Following a successful F1 cross, insertion orientation was determined in single DsRed-positive flies. DNA was extracted as described in section 2.3.1, and two PCR reactions using the same primer internal to the donor fragment, and a different primer on the genomic region outside of the donor cassette, on the left side for reaction 1 or on the right side for reaction 2, were performed. Flies with the insertion in the forward orientation (orientation 1, Or 1) had a 1.4 kb amplicon in reaction 1 and no product in reaction 2, and flies with the insertion in the reverse orientation (orientation 2, Or 2) had a 1.4 kb PCR product from reaction 2 but not from reaction 1. Primers are reported in Supplementary table 5.

2.3.4 Genotyping the identity and location of TAD boundary insertions

To confirm that the insertion occurred in the correct landing site and to sequence-verify the inserted boundary, I used primers in the genomic region flanking the MiMIC insertion site to amplify the full cassette. Primers are indicated in Supplementary table 5. The amplicons range from 3.4 to 4.4 kb in size depending on the size of the inserted boundary. Next, the PCR product was sent for Sanger sequencing to confirm the identity and correctness of the inserted boundary.

2.4 iLEXY depletion

2.4.1 Photoactivation in the LED illumination box

To achieve uniform blue light illumination on the 10-cm agar plates used for embryo collection, I used custom-made blue light-emitting diode (LED) illumination boxes built by Christian Kieser from the EMBL electronic workshop. The boxes have two boards with 36 LEDs each on the top and bottom to allow illumination from both directions. LED intensity, pulse frequency and pulse-width-modulation (pwm) can be set via a microcontroller. Different conditions were employed for short and long incubations: continuous illumination at 100% intensity for short incubations (up to 2 h), and 100% intensity and pulsed light (2 s on / 1 s off) for long incubations (2–14 h). Pulsed light reduced phototoxicity and heating over long time periods. All blue light incubations were performed at 25°C to maintain the same temperature throughout collection, carried out at 25°C. To ensure uniform illumination on all embryos, the number of flies in the cage was adjusted to avoid crowding and prevent overlapping embryos, and collection plates were cleared of opaque elements such as yeast paste and dead flies before blue light exposure.

2.5 Embryo collection

For all embryo collections, adult male and female flies were placed in a cage in an apple juice agar plate and yeast paste, at 25°C. After three consecutive 1 h pre-lays to remove the old eggs retained by the females, embryos were collected in a new plate. At the end of collection, plates were maintained at 25°C up to the desired time point. Embryos were then washed with dH2O into a container with a mesh, dechorionated in 8% NaOCl (50% household bleach in H_2O) for 2 min, rinsed in dH_2O and PBTr 0.1% (PBS 1X + Triton 0.1%) and dried on blotting paper.

2.5.1 Embryo collection for immunostaining and DNA FISH

Dechorionated embryos were fixed in a 1:1 mixture of fixative solution (4% methanol-free paraformaldehyde diluted in PBS 1X) and heptane on a shaker at 470 rpm for 20 min. Then, the lower fixative phase was removed. Devitellinization was performed by adding methanol to the remaining heptane phase and vigorously shaking for 1 min, until most embryos settled at the bottom of the tube. Embryos were finally washed three times in pure methanol and stored in methanol at -20°C for DNA FISH. For immunostaining, after the three methanol washes, embryos were rehydrated with two 10-min washes in 50% methanol in PBTr 0.2% (PBS 1X + Triton 0.2%) and three 5-min washes in PBTr 0.2%, and stored at 4°C in PBTr 0.2% for up to 1 week. This minimized alteration to methanol-sensitive epitopes, especially in membrane proteins.

2.5.2 Embryo collection for CUT&Tag, sci-ATAC-seq and capture Hi-C

After dechorionation, embryos were fixed in a 1:3 mixture of fixative solution (1.8% formaldehyde in PBS 1X) and heptane, by shaking at 470 rpm for 15 min. After a brief spin at 500 g for 1 min, fixative and heptane were removed, and fixation was quenched in a stopping solution of 125 mM glycine in

PBTr 0.1% by vigorously agitating for 1 min. Then, embryos were precipitated at 500 g for 1 min, washed twice in PBTr 0.1% and dried on a 125 µm Nitex membrane (#03-125/45, Sefar Nitex Switzerland) over blotting paper. Dried embryos were then transferred to a clean tube, weighted, snap-frozen and stored at -80°C.

A small aliquot of the embryo sample from the same collection was fixed for immunostaining to check the embryonic stage. For the iLEXY lines, both the embryonic stage and the depletion of the iLEXY-tagged TF were verified.

2.5.3 Embryo collection of iLEXY lines

Embryos of iLEXY lines were fixed as previously described, but under LED lamps to ensure the appropriate light condition: red light for dark-reared embryos, and 458 nm blue light for depleted embryos. Embryos were kept in the appropriate light condition up to the end of fixation.

2.5.4 Hand sorting of iLEXY embryos with sensitized background

A sensitized background was used to enhance the depletion and phenotypic alterations of the *tin-mCherry-iLEXY* line. The homozygous *tin-mCherry-iLEXY* line was crossed to a classic *tin* KO mutant over a balancer with GFP expression under the control of pan-mesodermal *twi* enhancer. From stage 8 onwards, GFP expression could be detected in *tin-mCherry-iLEXY*/*TM3 Ser GFP* embryos but not in *tin-mCherry-iLEXY*/tin^{EC40} embryos. The delay with respect to *twi* expression may be due to GFP maturation time (~1.5 h) (Sniegowski *et al.*, 2005; Clark *et al.*, 2011; Balleza, Kim and Cluzel, 2018). Since collections occur at stage 9 and 11, GFP was clearly detectable on an epifluorescence microscope.

Embryos were collected, fixed in 1.8% formaldehyde as described above (section 2.5.2), and stored in PBTr 0.1%. For hand-sorting, embryos were transferred into a Petri dish with PBTr 0.1% on ice, and selected using an embryo needle under the epifluorescence microscope using the 488 nm filter. GFP-negative embryos were collected into a new tube and stored at 4°C in PBTr 0.1%. Then, nuclei were extracted within 24 h from collection to minimize reversal from formaldehyde crosslinking (Kennedy-Darling and Smith, 2014).

2.6 DNA Fluorescence *In Situ* Hybridization (DNA FISH)

2.6.1 Probe preparation

DNA FISH was performed with 8 kb PCR-derived probes for TAD A, and with 80 kb BACs for TAD B. Probes were fluorescently labelled using the Nick Translation Kit (#7J0001, Abbott Bioscience) and fluorescently labelled dUTP nucleotides (Aminoallyl-dUTP-ATTO-550 #NU-803-550-S, Jena Bioscience; Alexa Fluor 647-aha-dUTP #A32763 Thermo Scientific), according to manufacturer's instructions. After nick translation, probes were stored at -20°C for up to 1 month.

For each embryo sample, 5 µL of fluorescently labelled probe were mixed with 2 µL salmon sperm DNA (#D7656, Sigma Aldrich), 0.1 volume 3 M NaOAc (pH 5.2) and 3 volumes of ethanol, vortexed and precipitated by centrifugation at 17000 g and 4°C for 30 min. After washing in 70% ethanol, the pellet was dried and x µL of formamide were added according to the following formula: x = N × 6.25 µL ÷ n (N = number of embryo samples, n = number of probes per sample). The resuspended probes were left on a shaker at 37°C until prior to hybridization. At that point, probes were denatured at 75°C for 10 min, briefly cooled down on ice, and mixed with 2x hybridization buffer (80% Dextran 50, 4x SSC pH 7.5, 0.055 µg/µL BSA, and 2 µg/µL RNAse A).

2.6.2 DNA FISH

Embryos fixed in 4% paraformaldehyde and stored in 100% methanol at -20°C were rehydrated by sequential washes in 75%, 50% and 25% methanol in 2x SSCT (2x SSC pH 7.0 + 0.1% Tween-20) for 5 min, followed by three 5-min washes in 2x SSCT. All washes are performed at RT on a nutator, unless otherwise specified. Embryos were then washed for 10 min in 20% FA/2x SSCT pH 7.4 (20% formamide, 2x SSC pH 7.0, 0.1% Tween-20) and for 10 min in 50% FA/2x SSCT pH 7.4 (50% formamide, 2x SSC pH 7.0, 0.1% Tween-20). Embryos were washed twice for 1 h each in 50% FA/2x SSCT pH 7.4 at 37°C on a nutator; at the end of the 2 h, the buffer was removed and embryos were denatured at 80°C for 15 min in a water bath, placed on ice and mixed with the probes denatured at 75°C for 10 min in the hybridization mix. Samples were overlaid with a drop of mineral oil to prevent evaporation and incubated overnight at 37°C for hybridization. The next day, after removing the mineral oil, samples were washed twice for 30 min in 50% FA/2x SSCT pH 7.4 at 37°C in a thermomixer, and once in 20% FA/2x SSCT pH 7.4 at RT on a nutator for 30 min, followed by three 10-min washes in 2x SSCT.

Finally, the supernatant was removed and embryos were mounted in ProLong Gold Antifade Mountant with DAPI (#P36935, Thermo Fisher) on a glass slide, covered with a coverslip and sealed with nail polish.

2.6.3 Combined DNA/RNA FISH

The combined DNA/RNA FISH was performed in all TAD B insertion lines. Since some lines were homozygous lethal and had to be maintained over a GFP balancer, RNA FISH was incorporated in the DNA FISH protocol to detect the GFP signal. This way, during imaging I could select embryos homozygous for the insertion by the absence of GFP. For consistency, DNA/RNA FISH was performed on all TAD B lines, including the homozygous viable lines.

The first part of the combined DNA/RNA FISH consists of the DNA FISH protocol described above, except that the hybridization mix contained a RNase inhibitor VRC (2 µM) (#R3380, Sigma Aldrich) instead of the RNAse. After the last SSCT wash, RNA detection was started using the HCR v3.0 kit (Molecular Instruments), following manufacturer's instructions. After three 5-min washes in PBTw 0.1% (PBS 1X + Tween 0.1%), samples were postfixed in 4% formaldehyde in PBTw 0.1% for 20 min

on a nutator, and washed five times for 5 min each in PBTw 0.1%. Samples were then pre-hybridized in 30% probe hybridization buffer for 30 min at 37°C, and incubated overnight at the same temperature with the probes diluted in 30% probe hybridization buffer. The next day, excess probes were removed by four 15-min washes in 30% probe wash buffer at 37°C, and samples were washed three times for 5 min each in 5x SSCT (5x SSC + 0.1% Tween-20). The probe signal was amplified using fluorescently labelled hairpins: following a pre-amplification step in amplification buffer for 30 min on a nutator at RT, samples were incubated with activated fluorescently-labelled hairpins overnight. The next day, excess hairpins were removed by five washes in 5x SSCT at RT on a nutator, to a total of 1 h washing time. Finally, the supernatant was removed and embryos were mounted in ProLong Gold Antifade Mountant with DAPI (#P36935, Thermo Fisher) on a glass slide.

2.6.4 Image acquisition

DNA FISH and DNA/RNA FISH samples were imaged on a Leica SP8 confocal microscope equipped with the HC PL Apo CS2 64x/1.4NA Oil and the HC PL Apo CS2 100x/1.4NA Oil objectives. Dr. Gabriel Rodrigues Cavalheiro and I selected ventrally oriented stage 11 embryos, and for 3–4 embryos per line we acquired a Z-stack with 200 nm step size spanning the entire outermost nuclei layer in the centre of the ventral epidermis. For TAD B, only GFP-negative homozygous embryos were imaged. Leica HyD detectors allowed efficient detection of DNA FISH spots, especially from short probes.

2.6.5 Quantification of 3D distance

Images were deconvolved using the Huygens Professional software (SVI) with default parameters. To measure 3D distances between DNA FISH probes, I used a custom Fiji (Schindelin *et al.*, 2012) plugin (*Analyze FISH spots*) developed by Dr. Christian Tischer in EMBL the Advanced Light Microscopy Facility (ALMF). The plugin first performs spot detection, and then computes 3D distances between within each cluster of close spots. The coordinates of each spot were determined starting from a manual spot selection using the multi-point tool, where I required each spot to be present in ≥ 3 consecutive stacks, to have highest signal in the middle stack and be in an interphase nucleus. At the end of the manual selection of spot clusters, the plugin reports the x, y and z coordinates with the maximum intensity in each of the DNA FISH probe channels for each spot cluster, and computes the pairwise distance between the spots in the cluster. Specifically, the plugin examined the Gaussian distribution of the fluorescence intensity of the spot along the x, y and z axis, finding the centre of the spots as the maximum of the Gaussians, and uses the difference of Gaussians (DoG) or the Center of Mass (CoM) to calculate 3D distance between spots. The DoG was used to quantify distance in this study, as it is more reproducible between experimenters than the CoM, even though the measurements generally agree. Multiple lines were quantified independently by Dr. Gabriel Rodrigues Cavalheiro and myself to ensure reproducibility, and the DoG measurements were highly similar. The distribution of distances between each given condition and the WT was compared by a Kolmogorov-Smirnov test.

2.7 RNA FISH

Embryos fixed in 4% paraformaldehyde and stored in 100% methanol at -20°C were rehydrated by sequential washes in 70%, 50% and 30% methanol in PBTw 0.1% for 5 min, followed by three 5-min washes in PBTw 0.1%. All washes are performed at RT on a nutator, unless otherwise specified. Embryos were postfixed in 4% formaldehyde in PBTw 0.1% on a nutator, followed by 5 5-min washes in PBTw 0.1% and a 5-min wash in 100% Hyb-B (50% formamide, 5x SSC pH 5.0). Then the embryos were prehybridized for 2 h in Hyb-A (50% formamide, 5x SSC, 100 mg/ml salmon sperm, 1% Tween-20, 50 mg/ml heparin) at 65°C, followed by the addition of RNA FISH probes (100 µg/µL) denatured at 80°C for 5 min and diluted 1:50 in 250 µL Hyb-A. For double and triple ISHs, all the probes were added simultaneously. The samples were incubated with the probes overnight at 65°C. The next day, embryos were washed 3 times for 30 min and 3 times for 1 h in Hyb-B at 65°C, followed by 4 15-min washes is Hyb-B:PBTw 0.1% with decreasing Hyb-B concentration (80%, 50%, 20%) and 4 5-min washes in PBTw 0.1%. Samples were incubated twice for 30 min in 2x Western Blocking Reagent (WBR) (#11921673001, Roche) in PBTw 0.1%, and then incubated with horseradish peroxidase (HRP)-conjugated primary antibodies diluted 1:2000 in WBR + PBTw 0.1% overnight at 4°C on a nutator (Table 1).

The excess antibody was removed with 6 10-min washes in PBTw 0.1%. Tyramide Signal Amplification (TSA) was performed by adding the fluorophore diluted 1:50 in amplification buffer (# NEL760, Perkin Elmer) to the samples and incubating for 3 min in the dark, followed by 2 quick rinses and 3 10-min washes in PBTw 0.1%.

For double and triple ISHs, the peroxidase was inactivated by incubating the embryos in stop solution (10 mM HCl in PBT) for 5 min followed by 2 5-min washes in PBTw 0.1%, repeated twice, and then adding fresh 3% H2O2 in PBTw 0.1% for 1 h, followed by 6 10-min washes in PBTw 0.1%. Blocking, incubation with the primary antibody, TSA and peroxidase inactivation were performed sequentially for each target.

Finally, embryos were washed overnight in PBTw 0.1% and mounted in ProLong Gold Antifade Mountant with DAPI (#P36935, Thermo Fisher) on a glass slide.

For combined RNA FISH and immunostaining, RNA FISH was performed first, followed by the immunostaining (section 2.8.1) starting from the blocking step.

Table 1: HRP-conjugated antibodies used for RNA FISH

Primary Antibody	*Source/Vendor*
Anti-Fluorescein-POD, Fab fragments	Roche (# 11426346910)
Anti-Digoxigenin-POD (poly), Fab fragments	Roche (# 11633716001)
HRP Anti-Biotin antibody	abcam (# ab19221)

2.8 Immunostaining

2.8.1 Immunohistochemistry

Rehydrated embryos stored in PBTr 0.1% at 4°C were briefly washed three times in PBTr 0.1% and blocked for 1 h in 10% BSA in PBTr 0.2%. All incubations and washes were performed on a nutator at room temperature unless otherwise specified. After blocking, embryos were incubated on a nutator overnight at 4°C with the primary antibodies diluted in 300 µL of 10% BSA in PBTr 0.2%. Antibodies, dilution and source are indicated in Table 2. On the next day, embryos were washed in PBTr 0.2% six times (10 min each) to remove excess unbound antibody, and incubated with secondary antibodies (Table 3) for 2 h in the dark. Embryos were then rinsed and washed in PBTr 0.2% for six times (20 min each or overnight), mounted in ProLong Gold Antifade Mountant with DAPI (#P36935, Thermo Fisher) on a glass slide and covered with a No. 1.5 coverslip.

Table 2: Primary antibodies for immunohistochemistry

Primary Antibody	***Working Dilution***	***Source/Vendor***	***Use***
Mouse anti-Lamin IgG	1: 200	DSHB (ADL101)	Nuclear marker for test of depletion in all iLEXY lines
Rabbit anti-DsRed IgG	1:300	Takara	Test of depletion in *bin-iLEXY* and *tin-iLEXY* lines
Rabbit anti-Mef2 IgG	1:300	Furlong lab	Test of depletion in the *Mef2-iLEXY* line
Rabbit anti-Tinman IgG	1:1000	Furlong lab	Heart marker for phenotype after depletion in the *tin-iLEXY* line
Rat anti-Tropomyosin IgG1	1:4000	Babraham (P6694)	Pan-muscle marker for phenotype after depletion in all iLEXY lines
Rabbit anti-β3Tubulin IgG	1:2500	Buttgereit lab	Pan-muscle marker for phenotype after depletion in all iLEXY lines
Chicken anti-GFP IgY	1:500	Abcam (ab13970)	Detection of GFP balancer
Chicken anti-β-Galactosidase IgG	1:500	Abcam (ab9361)	Detection of LacZ balancer

Table 3: Secondary antibodies for immunohistochemistry

Secondary Antibody	***Working Dilution***	***Source/Vendor***
Wheat Germ Agglutinin (WGA), Alexa Fluor 488, 1 mg/mL	1:250	Thermo Fisher Scientific (# W11261)
Goat anti-chicken IgY, Alexa Fluor 488	1:500	Thermo Fisher Scientific (# A-11039)
Goat anti-mouse IgG, Alexa Fluor 647	1:500	Thermo Fisher Scientific (# A-21235)
Goat anti-rat IgG, Alexa Fluor 555	1:500	Thermo Fisher Scientific (# A-21434)
Goat anti-rabbit IgG, Alexa Fluor 568	1:500	Thermo Fisher Scientific (# A-11011)
Donkey anti-rabbit IgG, Alexa Fluor 647	1:500	Thermo Fisher Scientific (# A-31573)

2.8.2 Image acquisition

For the quantification of depletion with the iLEXY system, images from immunostained embryos were acquired on a Leica SP8 confocal laser scanning microscope with HC Pl APO CS2 20x/0.75 IMM for the whole-embryo view, or HC PL Apo CS2 100x/1.4NA Oil objective for the regions of interest (ROIs) within the expression domain of the depleted TF.

Embryos under dark and blue light exposure of the same line were stained together and imaged on the same day. For each stage, the imaged embryos were in matching orientation, and the ROIs approximately from the same area. The laser power in the channel of the depleted TF was constant across dark and blue light, and across stages. This ensured consistency and comparability across conditions.

For the phenotypic readout, images of immunostained embryos were acquired on a Zeiss LSM780 NLO microscope using the Plan-Apochromat 20x/0.8 M27 air (FWD=0.55mm) or the Plan-Apochromat 63x/1.4 Oil DIC M27 objective. Laser settings were maintained constant between WT, dark-reared and blue light-exposed embryos.

Image processing and analysis were performed in Fiji (Schindelin *et al.*, 2012).

2.9 Nuclear isolation

2.9.1 Small scale nuclear isolation for CUT&Tag and sci-ATAC-seq

Nuclei were isolated from up to 0.5 g of 1.8% formaldehyde-fixed embryos stored at -80°C. Frozen embryos were dounced 20 times with a loose pestle and 20 times with a tight pestle in a 5 mL tissue homogenizer (Wheaton) in 2 mL cold HB buffer (15 mM Tris-HCl pH 7.4, 15 mM NaCl, 60 mM KCl, 340 mM sucrose, 0.2 mM EDTA pH 8.0, 0.2 mM EGTA pH 8.0, 1x Roche cOmplete Protease inhibitors). The suspension was then filtered through two perpendicular layers of Miracloth (#475855-1R, Calbiochem) and centrifuged for 10 min at 2000 g at 4°C. After removing the supernatant, nuclei were resuspended in 1 mL HB buffer and centrifuged for 10 min at 2000 g at 4°C. Then, the supernatant was removed and nuclei resuspended in 1 mL PBTr 0.1% with 1x Roche cOmplete Protease inhibitors and dissociated by passing them 10 times through a 20G needle and 10 times through a 22G needle using a 5 mL syringe.

Nuclei were counted using CountBright Absolute Counting Beads (# C36950, Invitrogen) on BD LSRFortessaTM X-20 Flow Cytometer at the EMBL Flow Cytometry Facility, aliquoted in freezing buffer (50 mM Tris-HCl pH 8.0, 25% glycerol, 5 mM Mg-Acetate, 0.1 mM EDTA pH 8.0, 5 mM DTT, 1x Roche cOmplete Protease inhibitors), snap-frozen and stored at -80°C.

2.9.2 Large scale nuclear isolation for Hi-C

Nuclei were isolated from 0.5–1.2 g of 1.8% formaldehyde-fixed embryos stored at -80°C. Frozen embryos were thawed in 10 mL cold HB buffer (15 mM Tris-HCl pH 7.4, 15 mM NaCl, 60 mM KCl, 340 mM sucrose, 0.2 mM EDTA pH 8.0, 0.2 mM EGTA pH 8.0, 1x Roche cOmplete Protease inhibitors), and dounced 20 times with a loose pestle and 20 times with a tight pestle in a 15 mL tissue homogenizer (Wheaton). The homogenized suspension was filtered through two perpendicular layers of Miracloth (#475855-1R, Calbiochem) and centrifuged for 10 min at 3200 g at 4°C. After removing the supernatant, the nuclei pellet was resuspended in 10 mL HB buffer and centrifuged for 10 min at 3200 g at 4°C. The supernatant was removed, and the nuclei in the pellet resuspended in 3 mL PBTr 0.1% with 1x Roche cOmplete Protease inhibitors and dissociated by passing them 10 times through a 20G needle and 10 times through a 22G needle using a 5 mL syringe. The suspension was then filtered through a 20 µm Nitex membrane (#03-20/14, Sefar Nitex Switzerland).

Nuclei were counted using CountBright Absolute Counting Beads on BD LSRFortessaTM X-20 Flow Cytometer, aliquoted in freezing buffer (50 mM Tris-HCl pH 8.0, 25% glycerol, 5 mM Mg-Acetate, 0.1 mM EDTA pH 8.0, 5 mM DTT, 1x Roche cOmplete Protease inhibitors), snap-frozen and stored at -80°C.

2.10 Spike-in CUT&Tag

2.10.1 Generation of spike-in CUT&Tag libraries

A mixture of *D. melanogaster* and *D. virilis* nuclei in a precise ratio was used as input material. A total of 50,000 *D.virilis* nuclei were employed for all conditions; the amount of *D. melanogaster* for each antibody differed based on its relative affinity for the two species: 50,000 nuclei of *D. melanogaster* were used for Bin, Twi and Mef2 antibodies, and 200,000 for Tin. To ensure that the ratio was correct, nuclei were FACS-sorted in the EMBL Flow Cytometry Core Facility using the BD FACSAria Fusion Flow Cytometer with the 100 µm nozzle. Dark-reared and depleted embryos were processed together in duplicate or triplicate. A Concanavalin A (ConA) bead slurry (#86057-3, Polysciences) was washed twice and resuspended in Binding Buffer (20 mM Hepes pH 7.5, 10 mM KCl, 1 mM $CaCl_2$, 1mM $MnCl_2$), and then added to the nuclei in Dig-wash buffer (20 mM Hepes pH 7.5, 150 mM NaCl, 0.5 mM spermidine, 0.01% digitonin, 1x Roche cOmplete Protease inhibitors) (10 µL ConA bead slurry/100,000 nuclei). Nuclei were absorbed to the ConA beads on a rotating wheel for 10 min at 4°C, washed once in Dig-wash buffer and permeabilized for 30 min in Permeabilization Buffer (PBTr 0.1%, 5% BSA, 0.2% Igepal CA-630, 1x Roche cOmplete Protease inhibitors) on a nutator at 4°C. After a wash in Dig-wash buffer, the primary antibody diluted in 50 µL Antibody buffer (2 mM EDTA pH 8.0, 0.05% digitonin, 0.1% BSA, 20 mM Hepes pH 7.5, 150 mM NaCl, 0.5 mM spermidine, 1x Roche cOmplete Protease inhibitors) was added to each sample and incubated overnight at 4°C on a rotating wheel at 10 rpm. The antibodies and their corresponding amounts are reported in Table 4.

Table 4: Primary antibodies for CUT&Tag

Primary Antibody	*Volume*	*Source/Vendor*
Rabbit anti-Mef2 IgG	3 µl	Furlong lab
Rabbit anti-Tinman IgG	1 µl	Furlong lab
Guinea pig anti-Twist	1 µl	Furlong lab
Rat anti-Biniou	3 µl	Furlong lab

The next day, samples were washed in Dig-wash buffer and incubated with the secondary antibody at 1:100 (guinea pig anti-rabbit immunoglobulin G (IgG; H + L) - #ABIN6923140, Antibodies-online) for 1 h at room temperature on a nutator. Sample were washed three times in Dig-wash buffer and processed according to Kaya-Okur *et al.* (2019). Briefly, samples were incubated with an adaptor-loaded transposase conjugated to protein A (pA-Tn5) for 1 h for Tn5 targeting to sites bound by the secondary antibody, washed three times in Dig-300 buffer (20 mM Hepes pH 7.5, 300 mM NaCl, 0.5 mM spermidine, 0.01% digitonin, 1x Roche cOmplete Protease inhibitors) to remove specifically bound Tn5 and tagmented for 1 h at 37°C in V2 tagmentation buffer (36.3 mM Tris-Acetate pH 7.8, 72.6 mM K-Acetate, 11 mM Mg-Acetate, 17.6% DMF). Tagmentation was stopped in stop buffer (50 mM EDTA pH 8.0, 50 mM NaCl, 25 mM Tris pH 8.0, 0.1% SDS, 0.5 mg/mL proteinase K) and samples were de-crosslinked overnight at 37°C for 6 h and at 65°C for 10 h. The next day, DNA was purified using the DNA Clean & Concentrator-5 Kit (# D4014, Zymo) according to manufacturer's instructions, and the number of PCR cycles for library amplification was estimated by qPCR. Finally, samples were amplified for the appropriate number of cycles, followed by post-PCR cleanup with 1.3 x AMPure XP beads (#A63881, Beckman Coulter) to remove primer dimers.

2.10.2 Spike-in CUT&Tag data analysis

Spike-in CUT&Tag reads were trimmed using TrimGalore (v. 0.4.3.1) (Martin, 2011) and aligned to a joint reference genome containing *D. melanogaster* (dm6 assembly) and *D. virilis* (r1.07 assembly) using bwa mem (v. 0.7.17) (Li and Durbin, 2009). After annotating the duplicates with picard (v. 3.0.0) (*Picard Toolkit*, 2019) Markduplicates, the bam files were filtered (samtools v. 1.16.1) (Danecek *et al.*, 2021) to include only reads mapping in proper pairs and exclude non-primary alignments, reads not passing the quality checks, PCR and optical duplicates and supplementary alignments. The filtered reads were then separated in two bam files, one for *D. melanogaster* and another for *D. virilis* genome. *D. melanogaster* reads were normalized using the number of uniquely mapped exogenous reads as spike-in scale factor. When a condition included multiple samples, the number of reads was divided by the geometric mean of the reads of the exogenous species (*D. virilis*) in the condition, so that scale factors would add to 1. Spike-in normalized coverage tracks were generated using bamCoverage (deeptools v. 3.5.1) (Ramírez *et al.*, 2016) with options "--normalizeUsing RPKM --exactScaling --extendReads --ignoreDuplicates --effectiveGenomeSize 125464728 --scaleFactor scale_factor". Replicates were merged by averaging the signal with bigwigCompare.

Reads with a fragment size < 600 bp were used for a lenient peak calling via macs2 (v. 2.2.7.1) (Zhang *et al.*, 2008) with options "-g 125464728 -p 0.6 -f BAMPE --call-summits". To check peak reproducibility, idr (v. 2.0.4.2) (Li *et al.*, 2011) with options "-i 0.05 --use-best-multisummit-IDR" was performed on peak summits expanded by 500 bp upstream and downstream. The resulting high-confidence peaks were centred and expanded by 500 bp upstream and downstream or up to the next peak.

The signal under the peaks called on dark-reared samples was quantified by htseq-count (v. 2.0.2) (Anders, Pyl and Huber, 2015) and the counts within the peaks were used to quantify reads per kilobase of peak length (RPKM by millions of total mapped CUT&Tag reads). RPKM were averaged across replicates. Heatmaps were generated with computeMatrix and plotHeatmap (deeptools v. 0.1.9) (Ramírez *et al.*, 2016) based on RPKM normalized CUT&Tag signals. Differential binding analysis was performed using DESeq2 on the read counts normalized by the above scale factors. A differentially accessible peak was defined based on an absolute log2 fold change difference > 0.5 and FDR-adjusted p value < 0.05. Genome-wide peak distribution was calculated on all the peaks using the annotatePeak function from the R package ChIPseeker (v. 1.8.6) (Yu, Wang and He, 2015). To find peaks at muscle enhancers, CUT&Tag peaks were overlapped with tissue-specific DNase I Hypersensitive Sites (DHSs) from Reddington *et al.* (2020) using the GenomicRanges R package (v. 3.19) (Lawrence *et al.*, 2013).

2.11 sci-ATAC-seq

I used a homemade single-cell combinatorial indexing ATAC-seq (sci-ATAC-seq) protocol developed in the lab by Dr. Stefano Secchia (Cusanovich *et al.*, 2018; Calderon *et al.*, 2022).

2.11.1 Generation of Tn5 transposomes

Hyperactive quad-mutant Tn5 transposase E54K E110K P242A L372P was purified by the EMBL Protein Expression and Purification facility as previously described in Rossi, Lai and Pugh (2018) and stored at −20°C in storage buffer (25 mM Tris pH 7.5, 800 mM NaCl, 0.1 mM EDTA pH 8.0, 1 mM DTT, 50% glycerol). Oligonucleotides with unique indexes (Cusanovich *et al.*, 2018) were each annealed to common pMENTS oligos in a PCR machine, by heating up to 95°C for 5 min, slowly cooling down to 65°C (0.1 °C/s), incubating at 65°C for 5 min and slowly cooling down to 4°C (0.1 °C/s). The annealed oligonucleotides were stored at -20°C until further use.

On the day of the experiment, the Tn5 was loaded with the annealed oligonucleotides by incubation at 23°C with shaking at 350 rpm for 50 min, and the loaded Tn5 transposomes were diluted 1:10 (0.02 mg/mL final concentration) immediately before tagmentation. The combination of annealed oligonucleotides produced 96 unique indexes, one for each well of the 96-well plate.

2.11.2 Generation of sci-ATAC-seq libraries

Sci-ATAC-seq libraries were generated as in Cusanovich *et al.* (2018), with minor modifications. Nuclei stored at -80°C were quickly thawed in ice-cold PBTr 0.1%, washed twice and resuspended in 300 µL PBTr 0.1%. After the addition of DAPI 1:100, 2500 nuclei were FACS-sorted into each well of a 96-well plate containing 5 µL of Omni-ATAC buffer (Corces *et al.*, 2017) with 1x Roche cOmplete Protease inhibitors and 12 µL of 2x TD buffer (#C01019043, Diagenode), using the BD FACSAria Fusion Flow Cytometer with a 100 µm nozzle. Tagmentation was performed by adding 2 µL of the corresponding custom uniquely indexed transposome into each well and incubating at 55°C for 1 h. Tagmentation was stopped by adding 1:1 40 mM EDTA pH 8.0 (20 mM final concentration) and incubating at 37°C for 15 min. Nuclei were then pooled in a Falcon tube and stained with DAPI 1:100; 25 nuclei were FACS-sorted into each well of 12–16 plates with 12 µL reverse crosslinking buffer (10 mM Tris-HCl pH 8.5, 0.0004% SDS, 0.0004% Quiagen Proteinase K #19131). Sorted nuclei were decrosslinked at 65°C for 16 h.

Next, libraries were amplified according to the index type. For the original sci-ATAC-seq protocol, used for the *twi-iLEXY* line, the DNA was PCR-amplified by adding 5 µL of 5 µM forward and reverse indexed primers, 8 µL NEBNext Ultra II PCR master mix (#M0544, NEB) and 1 µL Triton 10% with the following cycling conditions: 72°C 5 min; 98°C 30 s; 98°C 10 s, 63°C 30 s, 72°C 1 min for 19 cycles, and hold at 10°C.

For the S3-ATAC-seq protocol (used for the *Mef2-iLEXY* line) (Mulqueen *et al.*, 2021), which introduces a dUTP on Tn5 adaptors, gap filling and adaptor switching was performed by adding 4 µL of the uracyl-insensitive polymerase NEBNext Ultra II Q5 Master Mix (#M0544, NEB), incubating at 72°C for 10 min, adding 1.5 µL of 1 µM A14_ME_LNA adaptor (5'-TCGTCGGCAGCGTCAGATGTGTA+TA+AG+AG+ AC+AG/3InvdT-3') and performing multiple rounds of linear extension (98°C 30 s; 98°C 10 s, 59°C 30 s, 72°C 30 s for 10 cycles before holding at 10°C). Library amplification was performed by adding 5 µL of 5 µM forward and reverse indexed primers and 9 µL NEB Q5U 2x master mix (# M0597, NEB), and running the following PCR program: 98°C 30 s; 98°C 10 s, 63°C 20 s, 72°C 40 s for 19 cycles; 72°C 2 min, and hold at 10 °C.

The uniquely indexed oligos used for library amplification were the same for both protocols (Cusanovich *et al.*, 2018).

2.11.3 Library clean-up and sequencing

After library amplification, clean-up was performed using the DNA Clean & Concentrator-5 Kit (Zymo #D4014), following manufacturer's instructions. Primer dimer removal and size selection were performed with AMPure XP beads following manufacturer's instructions: left side selection was achieved with 1x AMPure XP beads, followed by double side size selection with 0.9-0.5x AMPure XP beads.

DNA concentration was quantified with Qubit dsDNA HS Assay Kit and the library size distribution was visualized on a Bioanalyzer.

Sci-ATAC-seq libraries were sequenced on a NextSeq2000 sequencer 50 PE (Illumina) using custom primers (Table 5) and a custom sequencing protocol in Amini 2014. S3-ATAC-seq libraries were sequenced on a NextSeq2000 sequencer 50 PE (Illumina) with standard primers. Sequencing was performed by the EMBL GeneCore Facility.

Table 5: Sequencing primers for sci-ATAC-seq

Oligo name	***Oligo sequence (5'-3')***
Read 1 Sequencing Primer	GCGATCGAGGACGGCAGATGTGTATAAGAGACAG
Read 2 Sequencing Primer	CACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
Index 1 Sequencing Primer	CTGTCTCTTATACACATCTGAGGCGGAGACGGTG
Index 2 Sequencing Primer	CTGTCTCTTATACACATCTGCCGTCCTCGATCGC

2.12 sci-ATAC-seq data analysis

2.12.1 Raw sequencing data processing, mapping and duplicate removal

Raw sequencing data were processed using the scripts and pipeline from Cusanovich *et al.* (2018). BCL files were converted into fastq files using bcl2fastq v. 2.16. Barcode fixation was performed by assigning the barcodes from the experiment to their best match among all potential barcodes, using Levenshtein distance < 3 and distance from the best match > 2. This way, barcodes with sequencing or PCR amplification error can be preserved and linked to the correct barcode. Not matched barcodes were flagged as ambiguous or unknown. Reads were trimmed with Trimmomatic v. 0.32 (Bolger, Lohse and Usadel, 2014) and mapped to the dm6 reference genome using Bowtie2 v. 2.3.4.1 (Langmead and Salzberg, 2012) with options -X 2000 -3 1. Reads not mapping to autosomes or sex chromosomes, with a mapping quality < 10 or with ambiguous or unknown barcodes were discarded. Duplicates were removed with the python script *sc_atac_library_deconvoluter.py* (Cusanovich *et al.*, 2018).

2.12.2 Count matrix generation, LDA, clustering and UMAP

For each experiment, accessibility matrices of peaks by cells were generated using the script *sc_atac_window_counter.py* (Cusanovich *et al.*, 2018). The peak set from Cusanovich *et al.* (2018) was used for preliminary whole-embryo clustering; then, clustering was refined by calling peaks per cluster on each dataset with macs2 v. 2.2.7.1 (Zhang *et al.*, 2008) and parameters "'--nomodel --keep-dup all --extsize 200 --shift -100". To avoid sex bias in the clustering, dimensionality reduction was performed by Latent Dirichlet Allocation (LDA) with the cisTopic function 'runModels' (cisTopic v. 0.2.2 (Bravo González-Blas *et al.*, 2019), R package) on autosomal peaks. The resulting topic by cell matrix was used to compute UMAP plots and define clusters in Seurat (v. 3.2.2, R package) (Hao *et al.*, 2021) and

Signac (v. 1.13.0, R package) (Stuart *et al.*, 2021), with 'RunUMAP' and 'findClusters' functions. Clusters were checked for the presence of technical biases, like barcode collisions (doublets) or aberrant coverage, and all cells in biased clusters were excluded from the analysis.

2.12.3 Gene activity matrix generation

Gene activity scores were calculated by recording the accessibility across the whole gene body and 500 bp upstream of the TSS, for all genes in the 'TxDb.Dmelanogaster.UCSC.dm6.ensGene_3.4.4' R Bionconductor package, using the script 'sc_atac_window_counter.py' (Cusanovich *et al.*, 2018). The resulting gene per cell matrix was imported in Seurat for the analysis of differentially accessible genes.

2.12.4 Cell type annotation

The cell identity of each cluster was assigned starting from cluster-specific accessible peaks, as identified by Seurat v. 3.2.2 function 'FindAllMarkers', with the logistic regression method and using the total counts per cell as a latent variable. Peaks with log2 fold change > 0.5 and Bonferroni-adjusted p value < 0.05 were considered differentially accessible (DA) cluster-specific peaks, and non-significant peaks were assigned to the background. Cluster markers and background peaks were matched to peaks annotated in the CAD4 developmental enhancer collection (Cusanovich *et al.*, 2018) or in the BDGP gene expression database (Wilczynski *et al.*, 2012) with a characterized tissue-specific enhancer activity *in vivo* (Bonn, Zinzen, Girardot, *et al.*, 2012; Kvon *et al.*, 2014; Rivera *et al.*, 2019). The terms describing the activity of the enhancers were tested for overrepresentation in the cluster markers against the background via one-tailed Fisher test, and the most consistent categories within the enriched terms were used to assign the cell type.

2.12.5 Mesoderm reclustering

To increase resolution, refine cell type annotation and a stronger focus on the effect of depletion in the tissue of interest, cells annotated as mesoderm were isolated from the whole-embryo data and reclustered. Clustering was further refined by peak calling on the new mesoderm clusters using macs2 v. 2.2.7.1 (Zhang *et al.*, 2008) with the parameters in section 2.12.2.

2.12.6 Calculation of TF deviation scores

To understand which TFs are mostly affected in their binding by depletion of the iLEXY-tagged TF, I examined the binding of mesoderm and other TFs and the presence of motifs under non-ubiquitous TF peaks. For the binding of mesoderm TFs, I used 16 ChIP-on-chip and ChIP-seq datasets from our lab (Jakobsen *et al.*, 2007; Zinzen *et al.*, 2009; Cunha *et al.*, 2010; Junion *et al.*, 2012), and for other TFs I used 280 ChIP datasets from the modERN database (Kudron *et al.*, 2018). For motifs, I used the JASPAR Core motif collection (Castro-Mondragon *et al.*, 2022). Deviations in accessibility were calculated using chromVAR v. 1.10.0 (Schep *et al.*, 2017). I used the deviations in accessibility to track

developmental time and to find potential determinants of the phenotypic changes between mutants and WT by focusing on TFs and motifs with changes in accessibility in the mutant clusters.

2.12.7 Identification of differential ATAC peaks and genes

Differentially accessible ATAC peaks between clusters were identified using the FindMarkers function in Seurat (Hao *et al.*, 2021), using the logistic regression method and considering the total reads per cell as a latent variable. Sci-ATAC-seq is prone to background noise, which could lead to a higher number of peaks called by macs2 (Zhang *et al.*, 2008) with the parameters described in section 2.12.2, as well as with the FindMarkers function. To remove background noise and obtain a high-confidence set of peaks, for each pairwise comparison between clusters, the union of cluster-specific peaks was filtered by requiring peaks to have a coverage 10 times higher than the average background signal in ≥ 1 cluster. Regions with no ATAC peaks and no known genes were used to estimate the background signal, and the coverage threshold was chosen by inspecting the distribution of reads per base in peaks and background regions. Among the high-confidence peaks, peaks with absolute log2 fold change > 0.5 and Bonferroni-adjusted p value < 0.05 were deemed differentially accessible (DA).

For the genes, a similar approach was used to exclude genes with low coverage. Due to the generally lower coverage on the gene body of ATAC peaks, peaks were distinguished from the background by requiring 5x more coverage and being accessible in ≥ 10% of cells in one cluster. Similarly to DA peaks, DA genes were defined by an absolute log2 fold change > 0.5 and Bonferroni-adjusted p value < 0.05.

2.12.8 Integration with WT mesoderm time course

The Mef2-iLEXY and Twi-iLEXY sci-ATAC-seq datasets were separately aligned to the WT timecourse datasets from Cusanovich *et al.* (2018), Secchia *et al.* (2022) and Calderon *et al.* (2022). First, a single peak per cell count matrix was created for the two datasets to be aligned, using the merged peaks for integration with the Cusanovich *et al.* (2018) and Secchia *et al.* (2022) dataset, and the Calderon *et al.* (2022) peak set for integration with Calderon *et al.* (2022). The count matrix was used as input to find the topics by LDA with the cisTopic function 'runModels' (cisTopic v. 0.2.2 (Bravo González-Blas *et al.*, 2019), R package), and then batch correction was performed using the Mutual Nearest Neighbour (MNN) algorithm (reducedMNN function in the R package LTLA/batchelor v. 1.19.1) (Haghverdi *et al.*, 2018). Since both datasets include WT embryos, the WT genotype was used as anchor for batch correction. As well as the principal components, the superposition of the WT embryos from the two datasets in the UMAP, the annotation and the correct time progression were checked to evaluate the batch correction efficacy. Cluster similarity was assessed by Pearson correlation on pseudobulk cluster RPKM-corrected bigwigs using multiBigwigSummary and plotCorrelation from the deeptools library (Ramírez *et al.*, 2016), and by hierarchical clustering based on accessibility for known TFs, as calculated by chromVAR (Schep *et al.*, 2017).

2.12.9 Analysis of DA peaks

To analyse the genomic distribution of ATAC peaks, peaks were intersected with several genomic features using BEDTools v. 2.27.1 (Quinlan and Hall, 2010). Sites within 1 kb (+/- 500 bp) from the TSS of a gene were classified as proximal, and more distant sites as distal. As putative enhancers reside in the distal sites, they were further inspected by overlapping the peaks with a collection of tissue-specific DHSs of FACS-purified mesodermal cells (Reddington *et al.*, 2020), a dataset of CRMs active specifically in the mesoderm (Zinzen *et al.*, 2009), the CAD4 collection of tissue-specific enhancers active at different timepoints (Bonn, Zinzen, Girardot, *et al.*, 2012; Kvon *et al.*, 2014; Rivera *et al.*, 2019) and the BDGP gene expression data from ISH (Tomancak *et al.*, 2002). Moreover, the time-resolved occupancy of mesoderm TFs was assessed by their overlap with ChIP peaks from our lab (Jakobsen *et al.*, 2007; Zinzen *et al.*, 2009; Cunha *et al.*, 2010; Junion *et al.*, 2012); additional TFs were investigated using whole-embryogenesis ChIP-seq data from the modERN collection (Kudron *et al.*, 2018). For the Mef2-iLEXY dataset, which includes embryos collected at both stage 9 and 11, sites overlapping Mef2 ChIP peaks at 4–6 h or 6–8 h were deemed Mef2-bound. For Twist-iLEXY, since embryo collection is centred on stage 9, ATAC peaks were considered to be bound by Twist if they overlapped with Twist ChIP-on-chip peaks at 4–6 h.

For the heatmaps, accessibility was averaged per cluster using the 'AverageExpression' function in Seurat (Hao *et al.*, 2021). Motif enrichment was calculated on the motifs from JASPAR Core *D. melanogaster* 2020 (Castro-Mondragon *et al.*, 2022), from CISBP *D. melanogaster* (Weirauch *et al.*, 2014) and from a collection of mesoderm motifs from the lab. Motif enrichment under differential peaks was scored with the 'AddMotifs' function from the R package Signac v. 1.1.0 (Stuart *et al.*, 2021). The ChIP signal under ATAC peaks was calculated with bamCoverage (Ramírez *et al.*, 2016), and deciles were computed using the function 'ntile' from the R package dplyr v. 1.0.2 (Wickham H, François R, Henry L, Müller K, 2023).

2.12.10 Analysis of DA genes

The global identity of DA genes was assessed by gene ontology (GO) on Biological Process by hypergeometric test, using the hyperGTest function in the GOstats R package v. 2.38.1 (Falcon and Gentleman, 2007). The gene universe included all genes with a coverage 5x above the background in ≥ 1 of the compared clusters, and a p value < 0.01 was used as cut-off for differential enrichment. The enriched pathways among down and up DA genes were then assigned to macro-categories by manual annotation. To facilitate the assignment, I used REVIGO (http://revigo.irb.hr/) (Supek *et al.*, 2011) on the list of GO terms sorted by p value, using the *D. melanogaster* database with medium list size and SimRel (default) semantic similarity.

2.13 Capture Hi-C

Capture Hi-C libraries were generated by Dr. Rebecca Rodriguez Viales using a Bridge-Adaptor *in situ* Hi-C protocol (Rao *et al.*, 2014), which leverages biotinylated bridge oligo adaptors for multiplexing between samples (Ramani *et al.*, 2017), combined to the capture of the genomic regions of interest (ROIs).

2.13.1 Design of capture Hi-C probes

Capture Hi-C probes were designed by Dr. Adam Rabinowitz. The captured regions include the two acceptor TADs, extended by ≥ 150 kb on both sides, and three control regions. The genomic coordinates of the captured regions are shown in Table 6.

Table 6: Genomic coordinates of captured regions in capture Hi-C

Captured region	*Coordinates of captured region*	*TAD coordinates*
Acceptor TAD A	chr3L: 10950000-11900000	chr3L: 11213000-11525000
Acceptor TAD B	chr2L: 14850000-15650000	chr2L: 15271500-15500000
Control region 1	chr3L: 6400000-7300000	
Control region 2	chr2R: 5600000-6100000	
Control region 3	chr3R: 14550000-15100000	

Since the restriction enzyme DpnII was used for Hi-C, the probes were designed on either end of a DpnII fragment in the ROIs. A set of 13784 75-bp probes, covering 90% of each region, was employed for a total panel size of 3.7 Mb. Digestion fragments shorter than the probe length (75 bp) were excluded.

2.13.2 Generation of capture Hi-C libraries

A total of 25 million nuclei stored at -80°C were thawed in 1 mL ice-cold lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal CA-630, 1x Roche cOmplete Protease Inhibitors), incubated on ice for 30 min and centrifuged at 1000 g 4°C for 5 min. After removing the supernatant, nuclei were resuspended in 200 µL 0.5% SDS and permeabilized at 65°C for 10 min. SDS was quenched by adding 100 µL of 10% Triton and incubating at 37°C for 15 min. Chromatin was digested overnight at 37°C by adding 8 µL of the restriction enzyme DpnII (#R0543, NEB) in 50 µL of 10x DpnII buffer and 130 µL nuclease-free water. Then, another 5 µL DpnII were added in the morning. After 1 h of incubation, the sample was centrifuged for 5 min at 1000 g at RT and the pellet resuspended in 30 µL water. This step was followed by ligation of biotinylated bridge oligos to the fragmented DNA by adding 1 µL T4 DNA ligase HC (#EL0013, Thermo Fisher Scientific) and 5 µL 90 µM annealed biotinylated oligos (MKVII_F_22: 5'-GATCGGCGCTCAGAA/iBiodT/T-3' and MKVII_R_22: 5'-CTGAGCGCC-3') to the resuspended pellet, followed by the addition of 4 µL 10x BSA, 5 µL ligation buffer and 5 µL PEG

4000. The reaction was incubated at 16°C overnight and stopped by adding 2.5 µL 0.5 M EDTA. Nuclei were centrifuged and washed four times in a solution with 1x BSA and decreasing SDS concentration from 0.2% to 0%. The nuclei were then resuspended in 250 µL water containing 1x BSA and 0.2% Triton X-100.

At this point, the biotinylated bridge adaptor oligos were ligated to the DNA fragments but not to each other, due to the lack of a phosphate group at their ends. To enable adaptor ligation, nuclei were treated with PNK (#M0201L, NEB) in 1x T4 ligase buffer for 1 h at 37°C, followed by ligation at RT for 4 h. The nuclei were then centrifuged at 1000 g, 4°C for 5 min, and resuspended in 500 µL of extraction buffer (50 mM Tris-Cl pH 8.0, 50 mM NaCl, 1 mM EDTA pH 8.0, 1% SDS). Proteins were digested by incubation with 20 µL of 20 mg/mL Proteinase K (Invitrogen) at 55°C, 1000 rpm shaking for 30 min, followed by the addition of 130 µL of 5 M NaCl and decrosslinking on a shaker at 68°C and at 1000 rpm overnight.

The next day, 63 µL of 3 M NaOAc pH 5.2, 2 µL of 15 mg/mL GlycoBlue (Life Technologies) and 1 mL of absolute ethanol were added, and the samples incubated at -80°C for 15 min. The DNA was precipitated by centrifugation for 1 h at 20000 g at 4°C. The pellet was washed, dried and resuspended in 50 µL of 10 mM Tris-HCl pH 8.0. RNA was digested with 0.2 µg/µL RNase A (#10109142001, Roche) at 37°C for 15 min. Then, DNA was sheared ~200-400 bp with Diagenode Bioruptor Pico (4°C, 8 cycles, 30"/90" ON/OFF). The resulting library was size-selected with AMPure XP beads 1–0.6x according to manufacturer's instructions, eluted in 100 µL of 10 mM Tris-HCl pH 8.0, and quantified with Qubit. Biotinylated DNA fragments were captured with 30 µL of Dynabeads Streptavidin M-280 beads (#11205D, Thermo Fisher Scientific), followed by washing in bead wash buffer (5mM Tris-HCl pH 7.4, 0.5 mM EDTA pH 8.0, 1M NaCl, 0.1% Tween-20) and resuspension in 40 µL of 10 mM Tris-HCl pH 8.0.

The sequencing library was prepared using the Swift Biosciences Accel-NGS 2S Plus DNA Library Kit (Swift biosciences, 21024), with slight modifications to allow library preparation from DNA bound to Streptavidin beads (see library preparation published by Arima https://arimagenomics.com/wp-content/files/User-Guide-Library-Preparation-using-Swift-Biosciences-Accel-NGS-2S-Plus-DNA-Library-Kit.pdf). The library was amplified with eight PCR cycles, quantified by Qubit and checked on a Bioanalyzer.

Samples with compatible indexes were pooled and processed according to the Twist Target Enrichment protocol, with minor modifications. Briefly, the indexed libraries were dried using a vacuum concentrator and resuspended in a blocker solution containing 5 µL SeqCapEZ from NibleGen (#06684335001, Roche) and 7 µL of universal blockers. Then, 1 µL of 10 µM block bridge oligo (block_bridge_22: 5'-GATCGGCGCTCAGAATT[SpcC3]-3') was added as an additional blocker to block the bridge adapter sequence. Meanwhile, a probe solution was prepared by adding 4 µL of capture Hi-C probes to the Hybridization Mix, heated at 65°C for 10 min and slowly cooled down for 5 min at RT. For hybridization, the probe solution was heated to 95°C for 2 min and cooled on ice for 5 min, the library pool was heated to 95°C for 5 min, and both were equilibrated at RT for 5 min before mixing.

Next, 30 µL Hybridization Enhancer was added and the reaction incubated at 70°C for 16 h. The hybridized library was then bound to streptavidin beads, PCR-amplified for 8–9 cycles and purified using DNA Purification beads. After quantification with Qubit and quality control on a Bioanalyzer (displaying a single peak between 350 and 500 bp), the libraries were sent for sequencing on an Illumina NextSeq 500 platform 150 PE at the EMBL Genomics Core Facility, using standard Illumina primers.

2.14 Capture Hi-C data analysis

Capture-Hi-C data analysis, from preprocessing of raw sequencing data to data normalization, was performed by Dr. Charles Girardot using a local installation of a Galaxy platform maintained by GBCS, and the HiCExplorer suite (Ramírez *et al.*, 2018). Read separation on homozygous lethal TAD B lines was performed by Dr. Mattia Forneris. I performed the subsequent analysis steps starting from TAD visualization.

2.14.1 Read separation

As homozygous lethal TAD B lines were maintained over a GFP balancer, there were 25% homozygous embryos with the insertion, 25% homozygous balancer embryos and 50% heterozygous embryos with the insertion over the balancer. To examine the effect of the insertion without confounders from the balancer signal, reads were separated by genotype based on the GATK Best Practice Workflow in the MiMIC and in the balancer chromosome (inferred from a virginizer|Balancer F1 cross and Virginizer line, as in Ghavi-Helm *et al.*, 2019), using a custom script.

The parental genomic DNA from three parental lines (MiMIC, virginizer and virginizer|Balancer cross) were trimmed using trim_galore (v. 0.5.0) (Martin, 2011) and options "-q 30 -phred33 --fastqc --illumina --length 25 --paired" and mapped to the joint reference genome of *D. melanogaster* (dm6) and *Wolbachia* (AE017196) with bwa mem (v. 0.7.17) (Li and Durbin, 2009) and option "-T 20". Duplicates were annotated with picard (v. 2.16.0) (*Picard Toolkit*, 2019) MarkDuplicates and reads were filtered with samtools (v. 1.9) and options "-F 1548 -f3". The resulting bam files were then annotated with picard AddOrReplaceReadGroups to create one group per parental line. Before proceeding with variant call, we performed a base recalibration with GATK (v. 4.1.0.0) (McKenna *et al.*, 2010) BaseRecalibrator and ApplyBQSR using the DGRP vcf as reference (MacKay *et al.*, 2012) for common variants in *Drosophila* populations.

Variants were then called with GATK HaplotypeCaller (McKenna *et al.*, 2010) using a joint genotyping strategy and options "-G StandardAnnotation --min-base-quality-score=20". Variants on balancer chromosomes were inferred from the virginizer|Balancer F1 cross and Virginizer line as follows: (1) biallelic heterozygous variants in the virginizer|Balancer F1 cross genotyped as homozygous in the Virginizer line were genotyped as homozygous for the oppositive allele in the Balancer line; (2) biallelic variant homozygous for the same allele in both virginizer|Balancer F1 cross and Virginizer line were annotated for the same genotype in the Balancer chromosome; (3) all other variants were annotated as unknown genotype for the Balancer chromosome. Next, variants were filtered applying a set of hard

cut-offs using bcftools (v. 1.9) (Li *et al.*, 2009): "MQ > 58 & MQRankSum > -2.5 & MQRankSum < 2.5 & QD > 15 & SOR < 1.5 & FS < 10 & ReadPosRankSum > -6 & ReadPosRankSum < 6".

Capture-Hi-C reads from F1 crosses were then processed as other reads down to mapped and sorted reads. The resulting bam files were then split in three .bam files corresponding to MiMIC, balancer and unknown genotypes with a custom script. Reads were piled up on variants obtained from the previous step. Reads unambiguously harbouring alleles from MiMIC or balancer genotype were assigned to the corresponding genotype, while reads that did not overlap known variants or harboured ambiguous alleles were assigned to the unknown .bam file. The original fastq files were then filtered for reads belonging to the split bam files with seqtk (v. 1.3) to produce a MiMiC and a Balancer fastq file for each sample. These fastq files were then processed using the common capture-Hi-C pipeline.

Since trans interactions were not used for subsequent analysis, all data from the allele with the insertion (i.e. from the homozygous insertion and heterozygous embryos) was retrieved. Reads from the insertion allele were processed as described in the next section.

2.14.2 Capture Hi-C data processing

The Illumina sequencing files were demultiplexed, converted into FASTQ format and quality-checked using FastQC (Andrews *et al.*, 2015) in Galaxy (Afgan *et al.*, 2016) The reads were then mapped on the *D. melanogaster* dm6 genome assembly using BWA-MEM (Li and Durbin, 2009) with default parameters (Galaxy Tool version 0.7.17.1). Hi-C matrices were generated using the HiCExplorer (Ramírez *et al.*, 2018). BAM files were sorted by name using the Galaxy tool "Sort reads by name using Samtools sort" (v. 2.0.2) (Li *et al.*, 2009) and processed using hicBuildMatrix (Galaxy tool version 3.6+galaxy0) with the list of DpnII restriction sites as produced by hicFindRestSite (Galaxy version 3.6+galaxy0, with --searchPattern GATC) ran on the dm6 FASTA genome and options –restrictionSequence GATC --danglingSequence GATC --binSize 1000 –skipDuplicationCheck to yield a matrix with 1 kb resolution; 2 and 5 kb resolution matrices were also computed.

The subsequent operations were performed using custom scripts and the HiCExplorer v. 3.7.2. Trans reads were removed by hicAdjustMatrix, keeping only chromosomes 2L, 2R, 3L, 3R, 4 and X (options -a keep --interIntraHandling inter --chromosomes chr2L chr2R chr3L chr3R chr4 chrX). Technical replicates were merged by hicSumMatrices and converted from hic to cool format using hicConvertFormat. To avoid bias from the interactions with only one end in the captured region during the normalization step, the matrix was subset using hicAdjustMatrix to include only the contacts with both ends within the captured region. Each line with an insertion was compared to the WT of the same TAD separately: the total reads in each condition were computed, and the matrix of the sample with the highest coverage was downsampled to the reads of the sample with the lowest read number using cooltools random-sample (v. 0.4.0). Both the WT and the insertion line capture Hi-C matrices were then Knight-Ruiz(KR)-normalized using hicCorrectMatrix with the –correctionMethod KR parameter (Knight and Ruiz, 2013).

2.14.3 Quality control

Quality control was performed on the initial complete raw data using the quality reports produced by hicBuildMatrix and the MultiQC tool (Ewels *et al.*, 2016), to obtain the number of duplicated read pairs, dangling end, self-ligations, the distribution by read orientation and the proportion of short-range, long-range and trans-chromosomal interactions.

2.14.4 TAD visualization

Acceptor TADs were visualized using hicPlotTADs (in HiCExplorer v. 3.7.2) (Ramírez *et al.*, 2018) on subsampled and KR-normalized capture Hi-C matrices at 5 kb resolution, including only the captured ROI.

2.15 Identification and annotation of TAD boundary pairs

2.15.1 Manual annotation of TAD boundaries and boundary pairs from WT Hi-C data

In the WT 6–8 h Hi-C data used for genome-wide boundary annotation, TADs and boundaries called at different resolutions (1, 2, 5, and 10 kb) by hicFindTADs were visualized using the HiGlass tool (Kerpedjiev *et al.*, 2018) in a Docker container. TAD calls by hicFindTADs define consecutive boundaries as a pair, and do not consider multi-way interactions or sub-TADs; this may be partially overcome by using the union of all the TADs called at different resolution, but, in case of a shift in the position of the same boundary across resolutions, this approach could lead to inaccurate TAD calls. For reliable annotation of TAD boundary pairs, the boundary calls at different resolutions were inspected alongside Hi-C maps, and boundary pairs were manually annotated. Each boundary was resized to 2 kb, consistent with the length of the experimentally inserted boundaries (2–3 kb). TAD nesting and multi-way interactions were allowed, and each region between two manually annotated boundary pairs was considered a TAD.

2.15.2 Identification of boundary pairs by virtual 4C from capture Hi-C data

The boundary pairs formed by TAD boundary insertion were deduced by visualizing the capture Hi-C matrices in hicPlotTADs and by examining the virtual 4C profile. First, the effect of the insertion was evaluated based on the capture Hi-C matrix, and on the subtraction of the contact frequency in the insertion versus the WT using hicCompareMatrices (in HiCExplorer v. 3.7.2) (Ramírez *et al.*, 2018). The result was confirmed by computing the insulation score at 5 kb resolution at the MiMIC site with hicFindTADs and FDR-adjusted p value < 0.01.

The effect of the insertion is linked to its ability of forming new boundary pairs with elements in the acceptor TAD. In TAD B, the inserted boundaries can only interact with one acceptor TAD boundary on each side; on the contrary, in TAD A they could interact with two different boundaries on the left side and three on the right, namely the left TAD boundary, the left, middle, and right anchor, and the right boundary. To understand the interactions occurring upon TAD boundary insertion, virtual 4C plots

were generated from capture Hi-C data with hicPlotViewpoint (in HiCExplorer v. 3.7.2) (Ramírez *et al.*, 2018), using the MiMIC insertion site or the location of DNA FISH probes (left and right anchor in TAD A, left and right boundary in TAD B) as viewpoint. Interactions were computed in the pairwise normalized WT and insertion lines, and the contacts made by the insertion were detected as a gain of interactions between the acceptor TAD boundaries and the insertion site.

2.15.3 Annotation of features of TAD boundary pairs

Genomic and experimental TAD boundary pairs were annotated based on known boundary features from the literature (Ramírez *et al.*, 2018; Szabo, Bantignies and Cavalli, 2019). For time-resolved data, I chose the 6–8 h window, in accordance to the time window of the WT Hi-C. Unless otherwise specified, to calculate the presence and number of a specific genomic feature within the boundary, the summit of the feature was overlapped with the 2 kb boundary region using the countOverlaps function in the GenomicRanges R package v. 1.54.1 (Lawrence *et al.*, 2013). There were three main classes of examined features:

i. Binding of known insulators

    a. Number, presence, coverage under the summit and on the whole boundary region for the insulator proteins BEAF-32, CTCF, CP190, Su(Hw), Trl, M1BP from 6–8 h ChIP-seq data from the lab (BioProject accession number PRJEB38618). Publicly available ChIP-seq data were used for Ibf1 (SRA accession number SRR837792), Ibf2 (SRA accession number SRR837793) (Cuartero *et al.*, 2014), Pita (SRA accession number SRR3109310), ZIPIC (SRA accession number SRR3109311), Zw5 (SRA accession number SRR3109312) (Zolotarev *et al.*, 2016)

    b. Number, presence and orientation of motifs for the insulator proteins BEAF-32, CTCF, CP190, Su(Hw), Trl, M1BP from CISBP 2.00 *D. melanogaster* (Weirauch *et al.*, 2014); motif scanning was performed separately for each TF under the respective ChIP-seq peaks using FIMO (MEME suite v. 5.0.2) (Grant, Bailey and Noble, 2011) with a DHS background from Reddington *et al.* (2020) and q value < 0.8

ii. Presence and activity of a promoter

    a. For each orientation, the number, presence, type and signal of promoters in general from TxDb.Dmelanogaster.UCSC.dm6.ensGene (*TxDb.Dmelanogaster.UCSC.dm6.ensGene: Annotation package for TxDb object(s)*, 2019) or promoters active at 6–8 h, based on CAGE data from the lab (Mikhaylichenko *et al.*, 2018)

    b. For each strand, the RNA expression level from modENCODE (Roy *et al.*, 2010)

iii. Chromatin accessibility and chromatin marks

a. Number and presence of any DHS, and of DHS under insulator peaks (DHS summit within 500 bp of the insulator summit), from ≥ 2 tissues (Reddington *et al.*, 2020)

b. DHS signal from the whole-embryo and in FACS-sorted mesoderm, nervous system and non-mesodermal non-neuronal (double negative) tissue at 2–4, 4–6, 6–8 and 10–12 h, from Reddington *et al.* (2020);

c. Presence and signal of chromatin marks H3K27ac, H3K27me3, H3K4me3, H3K4me1 in the FACS-sorted mesoderm, nervous system and both, from our lab (Bonn, Zinzen, Girardot, *et al.*, 2012).

Moreover, I recorded the insulation score from the Hi-C 6–8 h data at 2 kb resolution (Pollex, Marco-Ferreres, *et al.*, 2024) and the distance between boundaries.

2.16 Evolutionary conservation of motifs in TAD boundaries

2.16.1 Alignment of inserted and acceptor TAD boundaries across the D. melanogaster clade

The evolutionary conservation of inserted and acceptor TAD boundaries was evaluated by aligning their sequence in *D. melanogaster* to the genome of 18 other *Drosophila* species in the same clade. Dr. Aaron Comeault (Bangor University) performed the alignments and evaluated the sequence conservation of each base using a conservation score. The conservation score was defined as

$$CS = \frac{1}{2}\, p\left\{base_{other\ species} = base_{D.mel}\right\} + \frac{1}{2}\, p\left\{any\ base\right\}$$

Where the first component is the frequency at which the other species have the same base as in *D. melanogaster*, and the second component is the frequency of any other base (i.e. not having a gap in the alignment).

From the alignment, the sequences were mapped back to the *D. melanogaster* genome.

To find conserved regulatory regions, I considered the non-coding sequences in DHS regions (Reddington *et al.*, 2020) extended by +/- 100 bp, and for each boundary I defined as conserved bases with a conservation score in the top quartile. Consecutive conserved bases were considered a conserved sequence; up to three consecutive nucleotides with lower conservation were allowed within a conserved sequence. The allowed gap in highly conserved bases was calculated based on the motif composition for known TFs. Conserved sequences with length < 6 bp were discarded.

2.16.2 Motif search and annotation in aligned boundaries

The non-coding DHS and conserved sequences were scanned for known motifs using FIMO (MEME suite v. 5.5.5) (Grant, Bailey and Noble, 2011), with a background model based on nucleotide frequency

in the DHSs from Reddington *et al.* 2020 and q value < 0.8. A conserved sequence was assigned to a known motif if it was found in the DHSs and overlapped with the sequence. Since most motifs found in DHSs are also called in the conserved sequences, but short conserved sequences may be penalized in matching to a motif, I used DHSs for the motif search. For the motif search, three different motif sources were used: CISBP 2.00 *D. melanogaster* (Weirauch *et al.*, 2014) for TFs in general, a collection of insulator protein motifs obtained by MEME-ChIP on ChIP-seq data from our lab (for BEAF-32, CTCF, su(Hw), Trl, M1BP) and other labs (Ibf1, Ibf2, Pita, Zw5, ZIPIC) (Cuartero *et al.*, 2014; Zolotarev *et al.*, 2016) (as reported in section 2.15.3), and core promoter motifs (Ohler *et al.*, 2002; Ohler, 2006).

In addition, conserved sequences were compared to known motifs using the motif comparison tool Tomtom in the MEME suite v. 5.5.5 (Gupta *et al.*, 2007) with E < 0.05, but fewer matched could be retrieved due to the lower statistical power. All the matches from Tomtom were recapitulated by FIMO, which was thus used as the only reference.

Conserved sequences were classified based on the contained motifs into insulators, other TF and promoter motifs, and sequences with multiple motifs were assigned to mixed categories (e.g. insulator + promoter). Sequences not assigned to any known motif were defined as new motifs. The known motifs found in conserved sequences and the sequences not assigned to known motifs ("new motifs") were used for downstream analysis.

2.16.3 Finding motifs linked to working TAD boundaries in the experimental and genome-wide dataset

To simplify motif annotation of experimental boundaries, FIMO (MEME suite v. 5.5.5) (Grant, Bailey and Noble, 2011) was run on the DHSs extended by +/- 100 bp within the boundaries, using the DHS background from Reddington *et al.* (2020), a q value < 0.8 and the known and new motifs in the conserved sequences as motif dataset. The position-weight matrices (PWMs) of known motifs were retrieved from the original motif database, and those for new motifs were generated from the evolutionary conservation, using the frequency of each base in the species alignment from section 2.16.1, and rescaled to 1. FIMO (MEME suite v. 5.5.5) (Grant, Bailey and Noble, 2011) was run on genome-wide boundaries using the same parameters.

Motifs were tested for enrichment in TAD boundaries genome-wide using AME from MEME suite v. 5.5.5 (McLeay and Bailey, 2010), with parameters --scoring max and --method fisher, and a DHS background from Reddington *et al.* (2020). Only motifs enriched in TAD boundaries with E < 0.05 were used for further analysis.

To find the motifs associated to working TAD boundaries in the experiments, the percentage of working boundaries among those with each motif was computed separately for each TAD. Working boundaries were defined as boundaries with an effect in DNA FISH and capture Hi-C; in case of discrepancy (3/36 lines), capture Hi-C was used as reference, as it is more sensitive to subtle changes.

Since the presence of a motif in an inserted boundary is not sufficient to describe the behaviour of orientation- and context- specific boundaries, to investigate rules for boundary interactions, I focused on pairs of two motifs, each on one element of an experimentally created boundary pair. For each motif pair, all possible reciprocal motif orientations were considered (convergent, divergent, tandem plus or tandem minus). The frequency of interacting boundary pairs among those containing the motif pair in a specific orientation was separately computed for each TAD, using a similar approach as the one for single boundaries. For genome-wide boundary pairs, the frequency of boundaries with the motif pair in a specific orientation over all possible orientations was computed to derive the preference of a motif pair for a specific reciprocal orientation. One-tailed Fisher test was used to assess orientation preference. The results were visualized with the ComplexHeatmap R package v 2.14.0 (Gu, Eils and Schlesner, 2016; Gu, 2022). Due to the difference in baseline proportion of working boundary pairs in TAD A (32%) and TAD B (80%), oriented motif pairs always associated to working boundaries and TAD B, and in 80% of the cases in TAD A, were considered as motifs potentially driving successful pairing.

For a preliminary evaluation of the sensitivity and specificity of this set of motif pairs, I computed the percentage of working and non-working experimental pairs with the motif pair set. To evaluate if these motif pairs could explain boundaries in the whole genome as well, I examined their frequency in genome-wide boundary pairs.

Combinations of more than 2 motifs were evaluated in experimental boundaries by the frequency of co-occurrence of two motif pairs in the same working boundary pair. Motif pairs co-occurring > 70% of the times and not composed of highly similar sequences were considered co-occurring, and pairs with 50–70% co-occurrence were considered to be potentially facilitating each other's function.

3 Results I

Functional characterization of the mesoderm GRN

Understanding the molecular basis of development and tissue and cell differentiation has been a long-standing question, tackled in increasing detail, by considering multiple sources of complexity. As well as contributing to our knowledge of organism development, such information sheds light on the working mode of TFs, on the elements driving enhancer-TF specificity, and on the differential usage of CRMs in time and tissue.

3.1 Motivation

3.1.1 Single-cell approaches overcome sensitivity limitations in the detection of aberrant developmental states

The first efforts to understand the roles of TFs in development mostly relied on bulk techniques and morphological readouts, which showcased large-scale effects but were limited in tackling heterogeneity and effects in small cell populations. For example, bulk genome-wide readouts, which output a signal average across all cells with the feature of interest (a TF for ChIP-seq, or any open chromatin region for ATAC-seq), often lack sensitivity to detect changes in small populations. This issue may be partially overcome by sorting cells with tissue-specific antibodies to enrich for the tissue of interest (i.e. using the BiTS-ChIP protocol from our lab (Bonn, Zinzen, Girardot, *et al.*, 2012; Bonn, Zinzen, Perez-Gonzalez, *et al.*, 2012)), but the recovery and separation of small populations may still be suboptimal; moreover, cells which lost the epitope as a result of a perturbation cannot be recovered, and even more problematic some cell states do not yet have discovered markers that can be used for sorting.
Single-cell techniques like sc-ATAC-seq and sc-RNA-seq relinquished the need for a previous knowledge of the markers of the population of interest, and the requirement for a large population size, enabling the identification of new cell types with little to no previous knowledge. Further, it allows finding the regulatory elements driving the expression of the genes which define cell identity, linking the molecular and cellular phenotype.

3.1.2 Time-resolved perturbations increase time resolution and prevent indirect effects

Most studies in the development of a variety of organisms and tissues focused on wild-type developmental context, making correlation-based conclusions based on the TF expression and binding over development (Suo *et al.*, 2018; Joung *et al.*, 2023). More recently, a combination of sc-RNA-seq and sc-ATAC-seq served to infer gene-regulatory networks and predict the effect of TF knockout (KO) mutants (Kamimoto *et al.*, 2023). In *Drosophila*, our lab has assessed the role of mesodermal TFs Mef, Tin, Bap and Bin by sci-ATAC-seq in KO mutant embryos after the major differentiation events in

muscle development (10–12 h) (Secchia *et al.*, 2022). This showed the cumulative effects of TF absence from the beginning of embryogenesis up to the 10–12 h window. However, despite providing important information on the general targets and the activation of alternative cell fate programs, it did not resolve temporal changes, and transient events were masked as they had already disappeared by the detection time. For instance, *bap*, *bin* and *tin* KO mutants did not lead to the emergence of new cell types, but rather showed loss of cell numbers as they are essential for specific cell fate decisions (cardiogenic and early visceral mesoderm for *tin*, visceral mesoderm for *bap* and *bin*). Conversely, the *Mef2* KO mutant led to two distinct mutant cell populations (Secchia *et al.*, 2022): one with a clear reduction in accessibility at Mef2 sites, as a direct effect of Mef2 absence, and another showing more indirect effects. For the second cluster, it is unclear whether it is caused by indirect effects over time or technical issues.

A second limitation is the inability to capture dynamics at shared enhancers. As mentioned in the introduction (section 1.4.2), *Drosophila* mesoderm enhancers have extensive TF co-binding, and the modality of their interaction – cooperativity, additivity, competition – could impact accessibility and transcription of the target gene(s). Moreover, the set of TF-bound enhancers is dynamic and changes across development (Sandmann *et al.*, 2006, 2007; Jakobsen *et al.*, 2007; Liu *et al.*, 2009; Zinzen *et al.*, 2009). By examining stage 12–13 (i.e. 10–12 h), only the connections active at that stage were detected; in addition, co-binding is less prominent at this stage than at earlier stages (stage 9–11). When considering the contributions towards a specific fate, the choice of stages critical for the development of the tissue of interest and a higher time resolution are better suited to address time specificity in TF-enhancer binding and TF-TF interactions. Moreover, a short perturbation can limit the indirect effects arising from long-term compensation.

To assess the role and interaction of mesoderm TFs in time with higher confidence, limiting indirect effects and preserving intermediate cell states, I leveraged a short time-specific perturbation using the iLEXY optogenetic system followed by single-cell readout by sci-ATAC-seq.

3.2 Using iLEXY to uncover the regulatory network of mesoderm development

To deplete the endogenous TFs from the nucleus, I cloned the mCherry-AsLov2-3xNES cassette downstream of the coding sequence of the endogenous *tin* and *Mef2* genes, before the stop codon, by CRISPR/Cas9-mediated homology-directed repair. The *twi-iLEXY* line was already generated and characterized by Anna Kögler (Kögler *et al.*, 2021). *twi* and *tin* have only one isoform; thus, the stop codon is unique; for *Mef2*, which has multiple isoforms with stop codon readthrough, I chose the first stop codon, which corresponds to the most expressed form during embryonic development based on lab RNA-seq data (unpublished). Readthrough mostly occurs at later stages and leads to a spliced mRNA of the same length as the shorter isoform; therefore, by tagging the first codon, all Mef2 populations will have the iLEXY tag. Moreover, the integration at the C-terminus maximized the chance of exposing the AsLov2 domain and minimized interference with the DNA-binding domain at the N-terminus. To prevent confounders due to the genetic background, all lines were derived from injections into the *vasa-*

cas9 (III) (BDSC#51324) line. After generating the lines, I confirmed the correct integration of the iLEXY cassette by sequencing.

3.3 Rationale for the choice of depletion and collection time points

As mentioned in the introduction (section 1.3), the development of *Drosophila* mesoderm occurs in 3 main phases: first, mesoderm specification at stage 5-9 by Twist and then by Mef2 and Tin, all in the same domain; then, at late stage 10 and stage 11 the mesoderm is subdivided into its 4 main derivatives somatic muscle (SM), visceral muscle (VM), heart (HM) and fat body, each with the expression of a specific but partially overlapping set of TFs (Bate and Arias, 1993; Baylies, Bate and Gomez, 1998), and this stage marks the specification of founder cells (FCs) and fusion-competent myoblasts (FCMs); finally, at stage 13-16 terminal differentiation takes place, supported by a smaller set of TFs. At stage 9 and early stage 10, expression patterns within the mesoderm are very homogeneous; at stage 11 the TFs have separate expression domains. I focused on these two stages to capture TF roles and the accessibility profile before and after the subdivision of the mesoderm into different muscle primordia, which marks the transition between a uniform and a tissue-specific program. Additionally, these steps involve larger cell populations than terminal differentiation, which allowed capturing intra-tissue heterogeneity more easily.

As for timing and duration of the depletion, it must centre on the correct stage and be long enough to observe changes in the chromatin state and in TF binding. Since the magnitude of the effect may be affected by depletion duration, I used 1 h depletion at all examined stages (Figure 15). For the genomics experiments, I collected embryos for 2 h. For stage 9 depletion, I exposed the embryos to blue light from 2.5 to 3.5 h after finishing collection; for stage 11, the exposure was from 4.5 to 5.5 h after collection (Figure 15). Embryos were fixed immediately after collection to collect 3.5–5.5 h embryos for stage 9 and early stage 10, and 5.5–7.5 h embryos for stage 11. The relatively large collection window enabled me to examine the effect of TF depletion throughout the whole stage, and increase the yield, especially for lines with reduced egg laying. For Twi, I used samples fixed as described in Kögler *et al.* (2021), depleted after gastrulation.

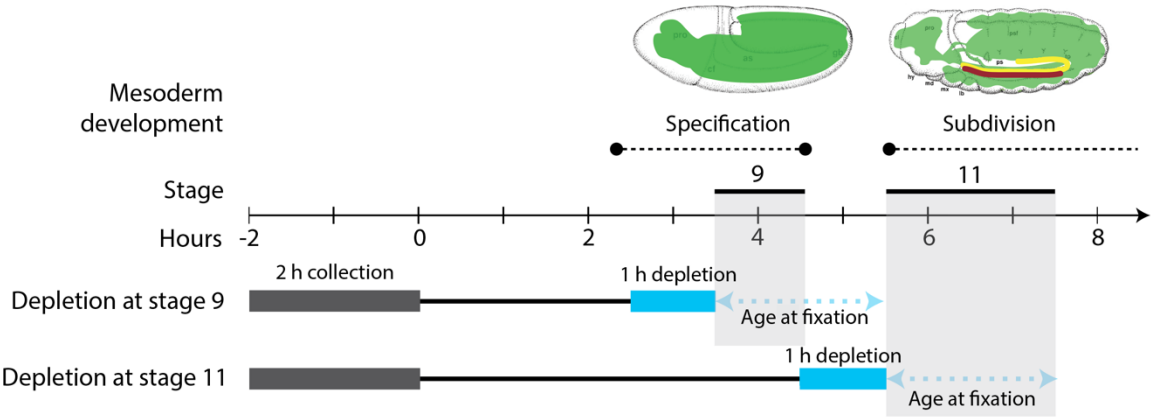


Figure 15: Timeline of the iLEXY depletion.
Embryos were collected for 2 h and aged in safelight conditions. They were exposed to blue light 2.5 h after the end of the collection for stage 9, and 4.5 h for stage 11, and fixed immediately after the end of the depletion.

3.4 iLEXY-tagged mesoderm TFs are depleted from the nucleus and from chromatin upon optogenetic stimulation

To assess if the iLEXY system can deplete these factors, I first verified if the nuclear depletion occurred efficiency and comparably across stages, and if depletion for the whole of embryogenesis recapitulated the well characterised loss-of-function mutant phenotypes.

TF-iLEXY lines recapitulated WT development in the dark: under safelight conditions, all lines were homozygous viable and fertile, and the expression domain of the TF was the same as in WT embryos, indicating that the presence of the tag has little or no impact on the TFs' function. Upon blue light activation, all four TFs were exported from the nucleus in a cell culture assay (not shown) and in *Drosophila* embryos at all developmental stages examined (Figure 16 B-D).

In the *bin-iLEXY* line, in the dark, Bin was localized exclusively in the nucleus of the visceral muscle cells as in WT embryos (Figure 16B). Similarly, Twist had mainly nuclear localization (Kögler *et al.*, 2021). On the other hand, in the *tin-iLEXY* and *Mef2-iLEXY* lines, the tagged TF was partly cytoplasmic in the dark, as opposed to the full nuclear localization in WT (Figure 16 C-D). This was also observed in the *twi-iLEXY* line (Kögler *et al.*, 2021). However, as these lines are homozygous viable and fertile, this low level of cytoplasmic protein has no deleterious function. In the ideal scenario, iLEXY-tagged TFs would be fully nuclear in the dark, and fully cytoplasmic under blue light; however, the system is based on an equilibrium where the baseline is set by the balance of the strength of the endogenous nuclear localization signal (NLS) and the NES array in the iLEXY tag. Therefore, it could be leaky in the dark, or cause incomplete depletion in blue light, depending on the relative strength of the localization signals.

In addition, I checked the nuclear depletion across stages by immunostaining for the iLEXY-tagged TF and nuclear lamina on broad stage embryo collections exposed to blue light for 1 h before fixation (Figure 16A). Upon optogenetic stimulation, the TF was exported to different extents in different lines and at different timepoints in the same line (Figure 16 B-D): the percentage of cytoplasmic TF was about 50% in the *bin-iLEXY* and in the *tin-iLEXY* lines, and higher in the *Mef2-iLEXY*. Depletion efficiency varied at different stages: for instance, at stage 9 depletion efficiency was generally lower. Within each stage, most cells were depleted with comparable efficiency regardless of how internal the tissue is, suggesting that blue light can penetrate the whole embryo and that its settings are in the saturation range. However, some cells may have a higher retention due to their identity or the molecular complex that the TF is in at that stage. For instance, at stage 11 Bin can be depleted from the nucleus of FCMs (*sns*+ cells), but not from most of the adjacent muscle FCs (*duf*+) (Figure 17), pointing to different TF roles and requirements in the two cell types. A difference in TF retention between FC and FCMs is also apparent for the other TFs (not shown).

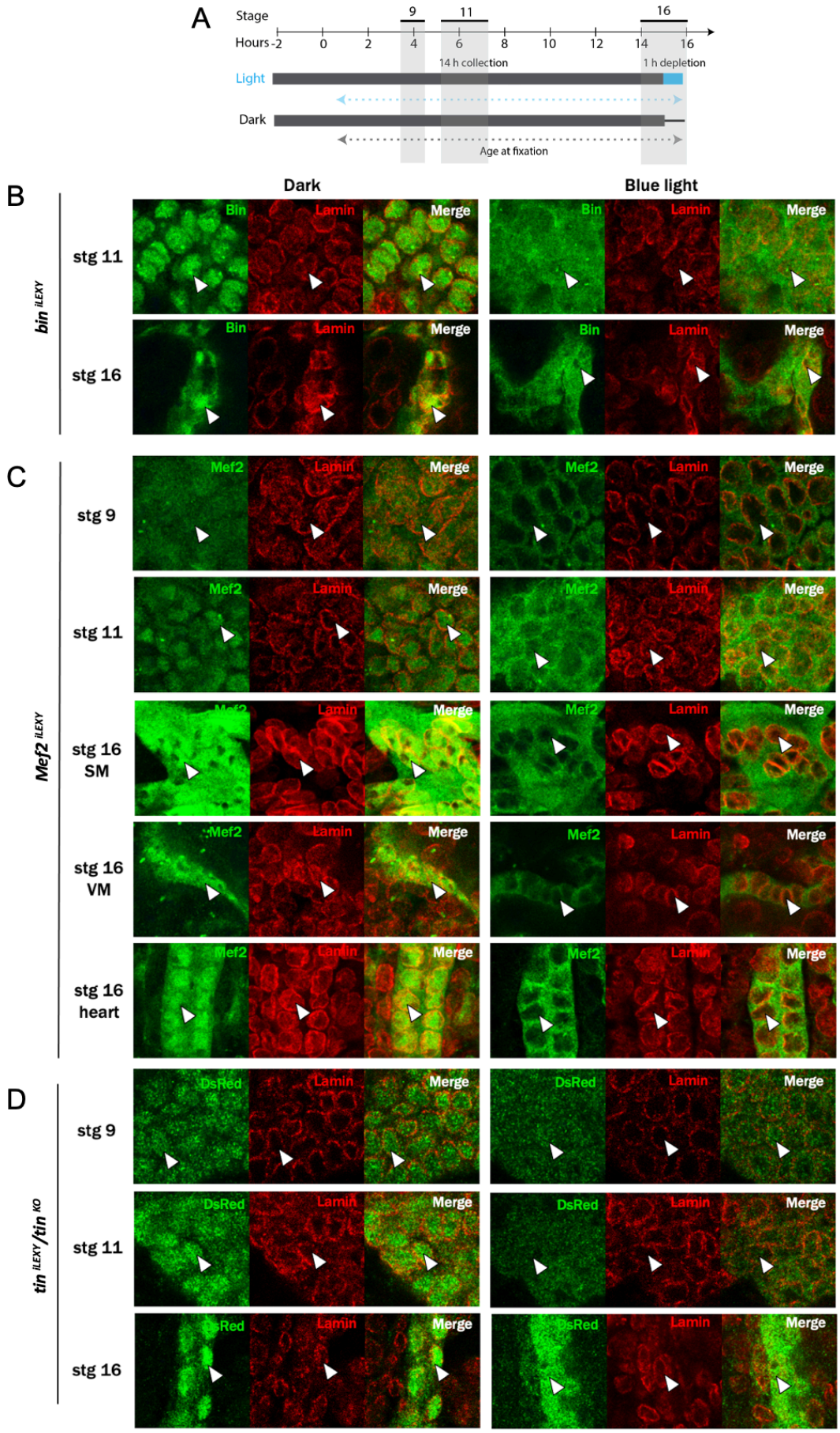

A
Stage
9
11
16
Hours -2 0 2 4 6 8 10 12 14 16
14 h collection
1 h depletion
Light
Dark
Age at fixation
B
Dark
Blue light
bin iLEXY
stg 11
stg 16
Bin
Lamin
Merge
C
Mef2 iLEXY
stg 9
stg 11
stg 16 SM
stg 16 VM
stg 16 heart
Mef2
Lamin
Merge
D
tin iLEXY/tin KO
stg 9
stg 11
stg 16
DsRed
Lamin
Merge

Figure 16: iLEXY-tagged TFs are depleted from the nucleus upon optogenetic stimulation.

(A) Scheme of the depletion. Embryos were collected overnight in safelight conditions, and exposed to blue light or dark-reared for 1 h before fixation.

(B-D) Immunostaining for the iLEXY-tagged TFs (with the Bin and Mef2 antibody for the respective lines, and for the mCherry tag for Tin; in green) and for lamin (in red). White arrows point to a single nucleus in dark vs blue light. Intensity scaling may vary between different TFs and stages.

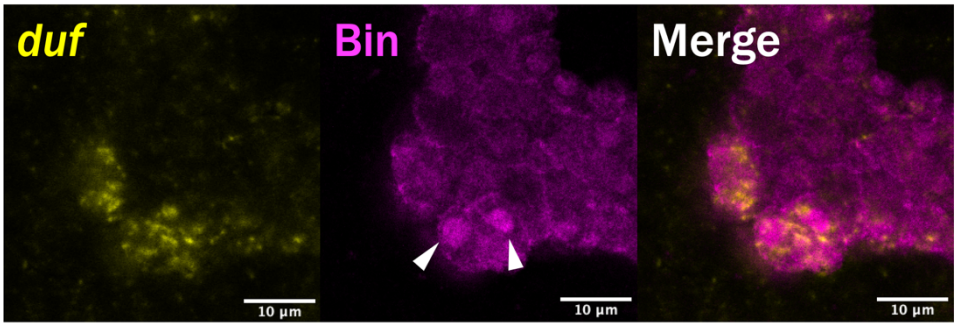
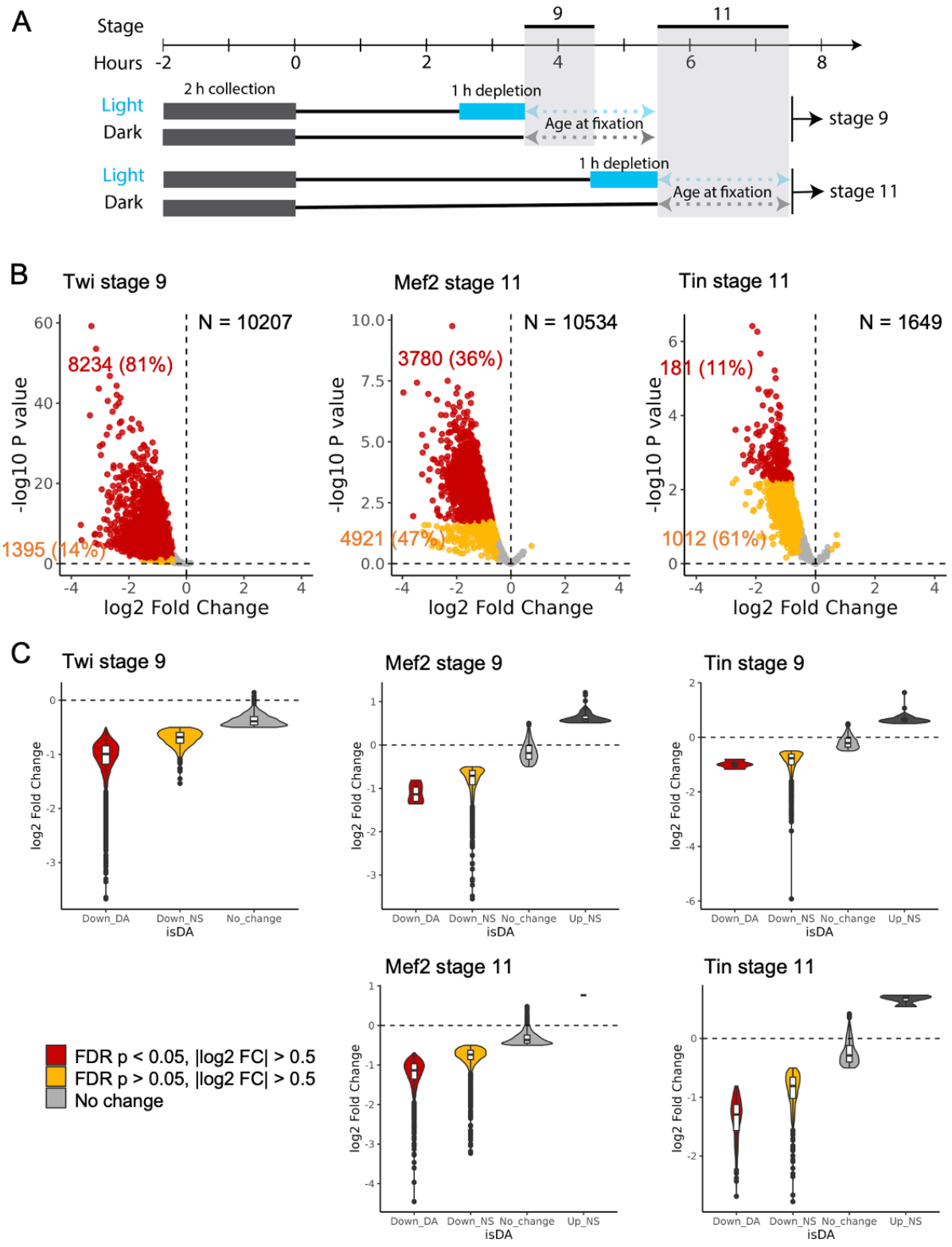


Figure 17: Bin has higher retention in FCs.

RNA FISH for the FC marker *duf* (left), immunostaining for Bin in the *bin-iLEXY* line exposed to blue light (centre), and merge (right) on stage 11 *bin-iLEXY* embryos exposed to blue light. The white arrows indicate cells with higher Bin signal and retention in the nucleus, which overlap the *duf*+ domain.

From these experiments, Bin depletion appears to be partial (Figure 16B). Mef2 was strongly depleted from the nucleus in blue light in all three muscle types (somatic, visceral and heart) (Figure 16C). To improve the efficiency of Tin depletion, I used a sensitized background by crossing the homozygous *tin-iLEXY* to the *tin* KO line tin^{EC40}*/TM3 Ser GFP*. In the sensitized background, Tin was very efficiently depleted, although not completely (Figure 16D). Twi was almost fully nuclear in the dark, and relocated to the cytoplasm in blue light with with high export efficiency (31-fold reduction in the nucleocytoplamic ratio) (Kögler *et al.*, 2021).

To quantitative the extent of depletion of each factor off chromatin, across all TF binding sites, I performed spike-in CUT&Tag (Kaya-Okur *et al.*, 2019) in bulk for each iLEXY-tagged TFs after depletion at stage 9 or 11 (Figure 18A). The spike-in enables a correct estimation of the depletion in the presence of global changes. The spike-in provides us with a reference independent of the difference in binding in the sample of interest, leading to a more accurate estimation. The reference must be constant in all CUT&Tag experiments for the same TF, separable from *D. melanogaster* reads in the mapping, and must express the TF of interest; therefore, I chose *D. virilis* nuclei at 4–12 h for the Twi spike-in, and at 4–20 h for the spike-in of the other TFs. These time windows correspond to the examined phases of mesoderm development. Indeed, *D. virilis* development is slower the *D. melanogaster*'s, and the stages have about 1.5 h delay: 4 h correspond to stage 5, with gastrulation and beginning of germ band elongation, 12 h to the start of germ band retraction (stage 12), 20 h to the completion of embryogenesis (stage 16) (Ninova, Ronshaugen and Griffiths-Jones, 2014). The large evolutionary distance leading to SNP accumulation and different genome size between the two species allows easy read separation. Importantly, the ratio of reads between the spike-in (*D. virilis*) and the sample (*D. melanogaster*) should be the same under dark and blue light.


A
Stage
Hours
9
11
-2
0
2
4
6
8
2 h collection
1 h depletion
Light
Dark
Age at fixation
stage 9
1 h depletion
Light
Dark
Age at fixation
stage 11
B
Twi stage 9
N = 10207
8234 (81%)
1395 (14%)
Mef2 stage 11
N = 10534
3780 (36%)
4921 (47%)
Tin stage 11
N = 1649
181 (11%)
1012 (61%)
-log10 P value
log2 Fold Change
C
Twi stage 9
Mef2 stage 9
Tin stage 9
Mef2 stage 11
Tin stage 11
log2 Fold Change
Down_DA
Down_NS
No_change
Up_NS
isDA
FDR p < 0.05, |log2 FC| > 0.5
FDR p > 0.05, |log2 FC| > 0.5
No change

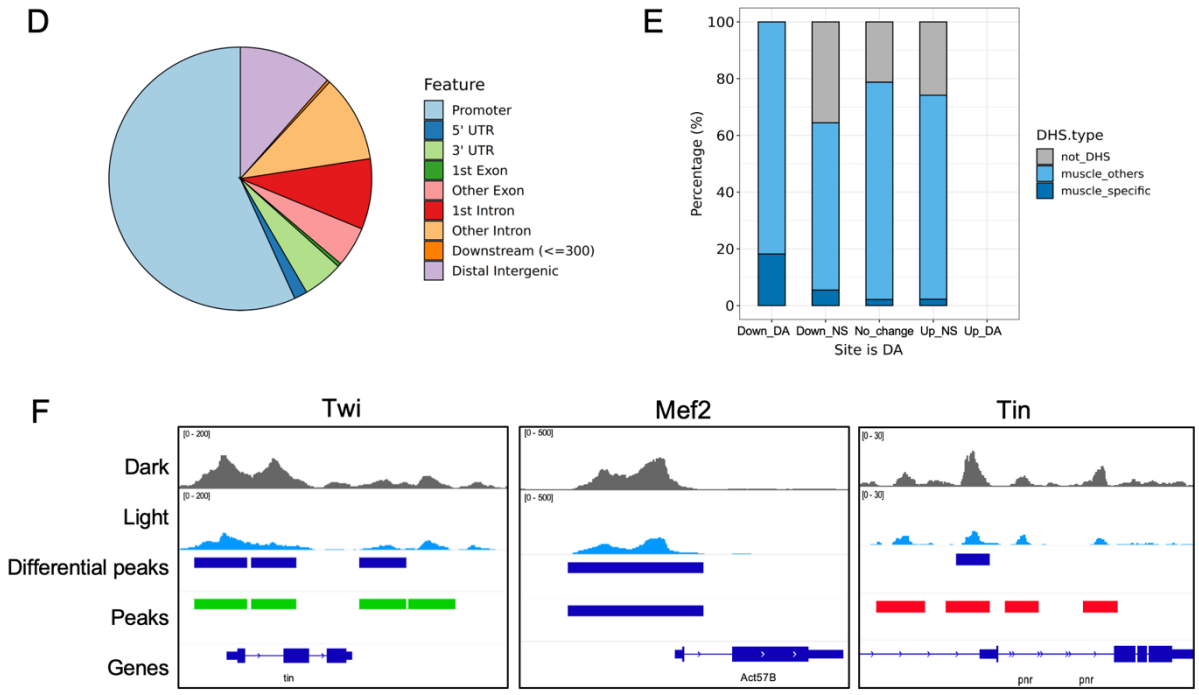


Figure 18: iLEXY-tagged TFs are depleted from the chromatin upon optogenetic stimulation.

(A) Scheme of the depletion. Embryos were collected for 2 h in safelight conditions, aged in the dark for 2.5 h for stage 9 and for 4.5 h for stage 11, exposed to blue light or dark-reared for 1 h and fixed immediately after.

(B) Volcano plots of the spike-in CUT&Tag for depletion of Twi, Mef2 and Tin in the respective iLEXY lines at stage 11, light VS dark. In red, the significant differential peaks, with FDR-adjusted p value < 0.05 and |log2FC| > 0.5; in orange, the peaks with |log2FC| > 0.5 but not significant by FDR. There is a trend for TF reduction in all the lines, albeit they have different depletion efficiency.

(C) Violin plot with the distribution of the log2FC in significantly reduced (FDR-adjusted p value < 0.05 and log2FC < -0.5, in red), not significantly reduced (FDR-adjusted p value > 0.05 and log2FC < -0.5; in orange), not changing (in light grey), and not significantly increased peaks (FDR-adjusted p value > 0.05 and log2FC > 0.5; in dark grey). The log2FC distribution confirms the trend for TF reduction upon depletion.

(D) Genome-wide distribution of CUT&Tag peaks for Twi, Mef2 and Tin. The distribution is very similar for all the TFs. Most of the peaks are at promoters, followed by intronic and distal intergenic sites.

(E) Percentage of muscle and other DHSs at significantly decreasing, not significantly decreasing, not changed and increasing peaks. The distribution is very similar for all the tested TFs. Overall, peaks are strongly enriched in DHSs present in the muscle and in other tissues. Peaks that significantly decrease upon depletion have the highest percentage of DHS exclusive to muscles and shared with other tissues.

(F) Genome browser tracks with average CUT&Tag signal in the dark and blue light from 3 replicates for Twi, Mef2 and Tin at known target genes. The known targets have reduced TF signal.

The CUT&Tag experiments showed good correlation (> 90%) between replicates. I called peaks on the dark-reared embryos, and I used a cutoff FDR < 0.05 and absolute log2FC > 0.5 to identify differentially bound peaks between the dark and blue light. Upstream TFs tended to have a higher number of differential peaks: Twi had the most differential peaks (81%, 8234/10207), followed by Mef2 at stage 11 (36%, 3780/10534) (Figure 18B). At stage 11, Tin had fewer significantly depleted peaks (11%, 181/1649), but an overall tendency to decrease, suggesting an issue with reproducibility (Figure 18B). Depletion of Mef2 and Tin at stage 9 caused a global decrease in TF signal but few significant changes (Figure 18C), which could be due to the lower expression levels of those TFs at stage 9, or stabilization at enhancers co-bound by other non-depleted TFs.

All TFs had a similar genomic distribution of the peaks: promoter peaks were the most abundant (55–60%), followed by intronic peaks (20%, equally distributed between the first and all other introns) and distal intergenic peaks (10%) (Figure 18D). The peaks distribution is compatible with TF binding at enhancers: indeed, intronic enhancers are common in the mesoderm, along with classic distal enhancers. Notably, these two classes were slightly enriched in peaks with decreased accessibility upon depletion, pointing to preferential depletion at enhancers instead of promoters.

To further evaluate if the peaks occurred at muscle CRMs, I examined the overlap of depleted and non-depleted peaks with DHSs specific for the muscle, common to the muscle and other tissues, or exclusively not muscle but specific for other tissues (from Reddington *et al.* , 2020). All peak classes had on average 80% overlap with DHSs common to the muscle and other tissues (65–100%) (Figure 18E). Sites with reduced accessibility, both significant and non-significant, had a higher percentage of muscle-specific DHSs (Figure 18E).

Moreover, I evaluated the quality of CUT&Tag peaks by comparing them to other stage-matched TF binding data from ChIP-on-chip. CUT&Tag peaks recapitulated 77–85% of ChIP-chip peaks from Sandmann 2006 (Mef2), Sandmann 2007 (Twi) and Liu 2009 (Tin), with the exception of Tin at stage 11, which had lower recovery (43%). The poor performance of Tin could be due to its smaller expression domain and lower expression levels at stage 11, leading to a strong reduction in the overall bulk signal, especially in a low-input protocol such as CUT&Tag.

Notably, the depleted peaks included regions proximal to validated direct target genes. For Twist, 10/10 validated targets have a depleted peak: *tin* (Yin, Xu and Frasch, 1997) (Figure 18F), *Mef2* (Cripps *et al.*, 1998), *Zinc finger homeodomain 1 (zfh1)* and *2 (zfh2)* (Fortini, Lai and Rubin, 1991), *heartless (htl)* (Shishido *et al.*, 1993), *breathless (btl)* (Shishido *et al.*, 1993), *snail (sna)* (Ip *et al.*, 1992), *rhomboid (rho)* (Leptin, 1991), *even-skipped (eve)* (Halfon *et al.*, 2000) and *twi* itself (Leptin, 1991). For Mef2, the depleted peaks include 11/14 known targets, as confirmed in the Mef2 ChIP-on-chip from Sandmann *et al.* (2006): the general muscle genes *Actin 57B (Act57B)* (Kelly, Meadows and Cripps, 2002) (Figure 18F), *Muscle LIM protein at 60A (Mlp60A)* (Stronach *et al.*, 1999), *inflated (if)* (Ranganayakulu *et al.*, 1995), *β-tubulin60D (βTub60D)* (Damm *et al.*, 1998), *Tropomyosin I* (*Tm1*) and *II (Tm2)* (Lin and Storti, 1997), *Myosin heavy chain (Mhc)* (Bour *et al.*, 1995; Lilly *et al.*, 1995), *meso18E* (Taylor, 2000) and *Mef2* itself (Cripps, Lovato and Olson, 2004), the muscle identity gene *nautilus (nau)* (Lin *et al.*, 1997), and the apoptotic gene *muscleblind (mbl)* (Artero *et al.*, 1998). Tin peaks recapitulated a lower number of known targets, and the depletion was apparent at *pannier* (*pnr*) (Gajewski *et al.*, 2001) (Figure 18F) and *tin* itself (Xu *et al.*, 1998).

Overall, CUT&Tag confirmed that TFs are depleted from the chromatin, albeit not completely, showing that different sites have different susceptibility to depletion. Since the depletion is partial, it was crucial to test if it was sufficient to elicit a mutant phenotype.

3.5 Depletion over the whole of embryogenesis partially recapitulates the loss-of-function mutant phenotypes

To assess if iLEXY depletion of each TF is sufficient to phenocopy characterized KO mutants, I depleted the embryos from before the beginning of TF expression (0–2 h) up to the end of embryogenesis (stage 16; 14–16 h) (Figure 19A) and stained them for the pan-muscle markers *Tm1* and *βTub60D*. iLEXY dark-reared, WT and KO mutant embryos were used as controls.

All dark-reared embryos had a WT phenotype, showing nicely patterning muscles, confirming that the system has no effect when inactive, and that any cytoplasm leakage of TFs is not sufficient to elicit a phenotypic change (Figure 19B-E, left column).

Upon blue light exposure, the phenotype of iLEXY embryos recapitulated to a large extent the KO. *Mef2-iLEXY* embryos had muscle fusion defects as the KO mutant (Figure 19B), albeit with less severe morphological defects: part of the somatic muscle (SM) is still present, but FCMs fail to fuse with FCs, leading to slender single fibres as opposed to the normal interwoven muscle syncytia (Figure 19B). Fibres spatial organization was similar to that of segmental muscles. Some residual unfused SM fibres were also present in homozygous *Mef2* KO mutants (Lilly *et al.*, 1995) and in a *Mef2* loss-of-function (LOF) allele over a deletion (Bour *et al.*, 1995), but in smaller numbers than in iLEXY lines, and with a more disordered spatial pattern. KO mutants also have defects in the differentiation of the VM, which was severely bloated and did not form the typical constrictions (Lilly *et al.*, 1995); *Mef2-iLEXY* embryos failed to recapitulate this defect, showing a WT organization of the VM into four chambers at stage 16 (not shown). In an attempt to fully mimic the KO mutant phenotype, I crossed the *Mef2-iLEXY* line to the *Mef2* KO, but the sensitized background did not exacerbate the phenotype: residual fibres were still present in the SM, and the VM was not affected (not shown).

bin-iLEXY embryos have no gut constriction under blue light (Figure 19C). However, VM cells were still specified, which does not occur in the *bin* LOF mutant (Zaffran *et al.*, 2001). The VM around the gut comprises a layer of circular visceral muscle (CVM) and a layer of longitudinal muscle (LVM), also known as caudal VM (Hartenstein and Campos-Ortega, 1985; Bate and Arias, 1993; Lee, Zaffran and Frasch, 2005). According to the Frasch lab (Ismat *et al.*, 2010), formation of gut constrictions may be hindered either by muscle fusion defects in the CVM, or by lack of the LVM. *In situ* hybridization for the LVM marker *beaten path IIa (beat-IIa)* (Ismat *et al.*, 2010) in *bin-iLEXY* embryos exposed to blue light showed that the LVM fails to form, and that the gut was lined only by CVM, which may correspond to cells with higher Bin retention at late stage 11 (Figure 20). This result suggests that Bin binds to, and regulates, enhancers differently in the LVM and CVM, and within the CVM differently between FCs and FCMs; however, it is difficult to distinguish those populations already at stage 11. To boost the phenotype, I used a sensitized background *bin-iLEXY/bin KO* with two different *bin* KO lines, but the embryos underwent developmental arrest at stage 10–11 in both cases; thus, I could not pursue it further. Due to the specification of VM cells and the inability to use a sensitized background to enhance the depletion, the *bin-iLEXY* line was not used for further analysis.

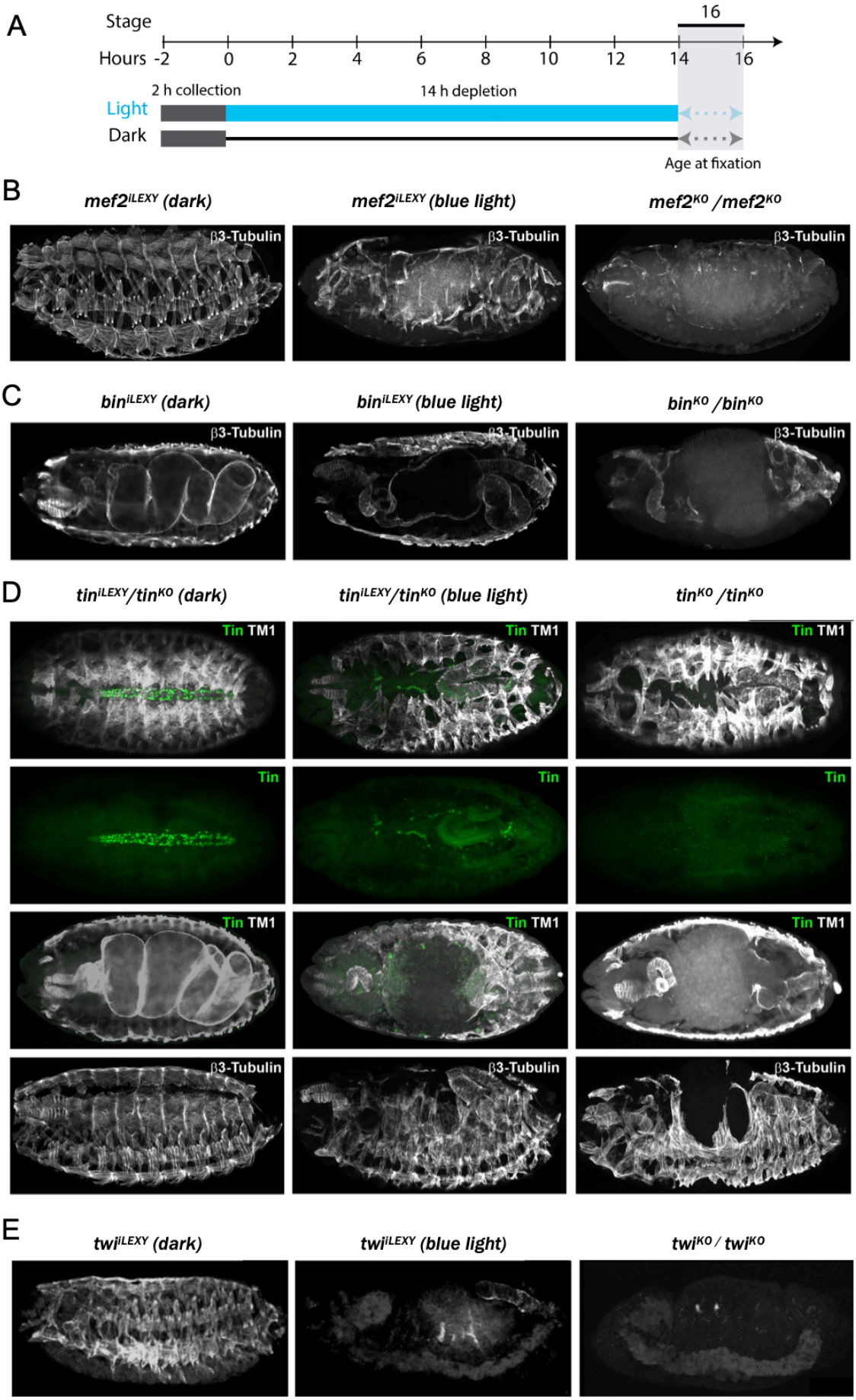

A
Stage
16
Hours -2 0 2 4 6 8 10 12 14 16
2 h collection
14 h depletion
Light
Dark
Age at fixation
B
mef2iLEXY (dark)
mef2iLEXY (blue light)
mef2KO /mef2KO
β3-Tubulin
C
biniLEXY (dark)
biniLEXY (blue light)
binKO /binKO
β3-Tubulin
D
tiniLEXY/tinKO (dark)
tiniLEXY/tinKO (blue light)
tinKO /tinKO
Tin TM1
Tin
β3-Tubulin
E
twiiLEXY (dark)
twiiLEXY (blue light)
twiKO / twiKO

Figure 19: The iLEXY system partially recapitulates the phenotype of the LOF mutants upon optogenetic depletion for the whole embryogenesis.

(A) Scheme of the depletion. Embryos were collected for 2 h in safelight conditions and immediately exposed to blue light or dark-reared for 14 h, until stage 16, as to achieve continuous depletion for the whole embryogenesis.

(B-E) Immunostaining of iLEXY embryos in the dark (left column), exposed to blue light (middle column) and of KO mutants (right column), stained for Tin, Tropomyosin (TM1) and β3-Tubulin: (B) *Mef2-iLEXY*, (C) *bin-iLEXY*, (D) *tin-iLEXY*, (D) *twi-iLEXY* (from Kögler *et al.*, 2021). Dark-reared iLEXY embryos have a WT phenotype. Upon depletion, *Mef2-iLEXY* has thin muscle fibres pointing to fusion defects; *bin-iLEXY* has no constrictions in the VM; and *tin-iLEXY/tinKO* loses the heart and the dorsal SM almost completely, and it has no constrictions in the VM. These phenotypes mimic the KO mutants but they are milder.

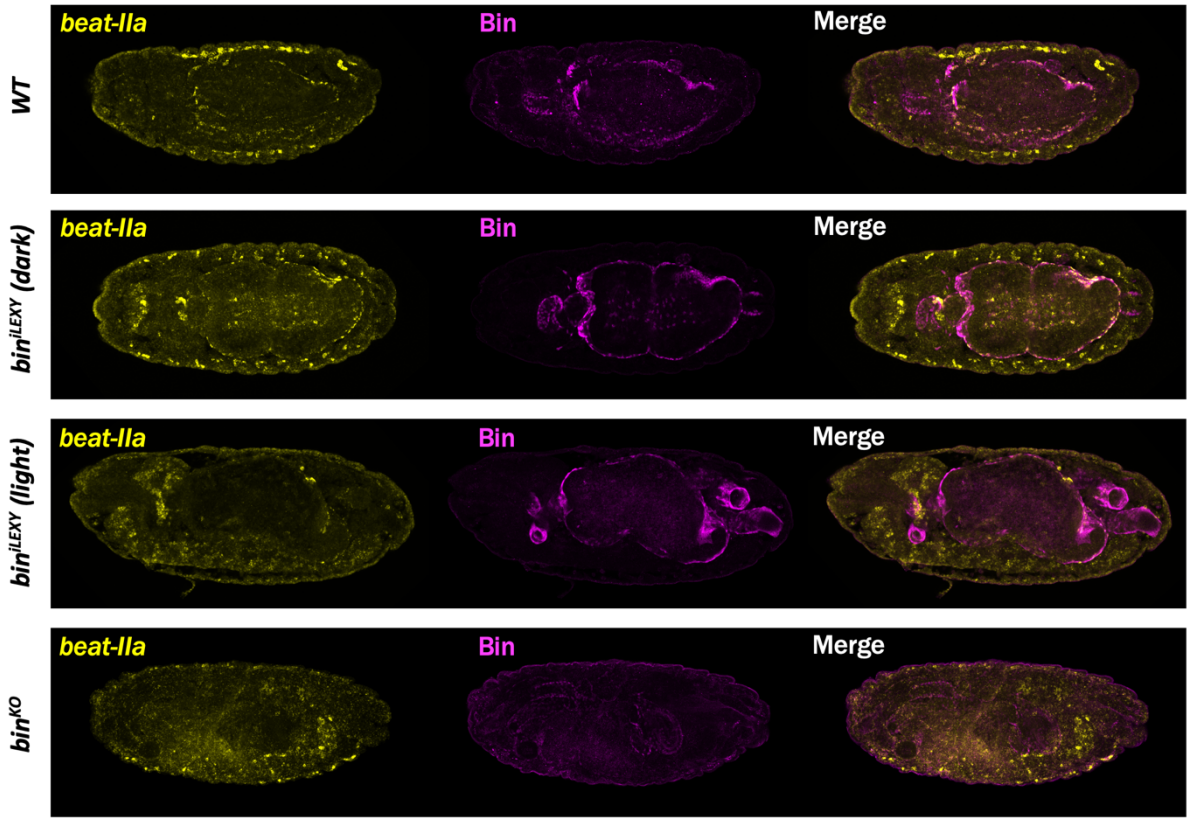


Figure 20: The LVM fails to form upon Bin depletion.
RNA FISH for the LVM marker *beat-IIa* (yellow, left), immunostaining for Bin (magenta, centre) and merge (right) in terminally differentiated embryos. In the WT and in *bin-iLEXY* in the dark, *beat-IIa*+ cells line the gut. In *bin-iLEXY* with continuous light exposure, there are very few residual *beat-IIa*+ cells, as in the bin KO mutant.

Despite successful nuclear export, the homozygous *tin-iLEXY* line showed mild defects, with only small areas of the heart missing and minor effects on other muscle types (Figure 21). Therefore, I used a sensitized background by crossing the homozygous *tin-iLEXY* to the *tin* KO line tin^{EC40}*/TM3 Ser GFP*. Transheterozygous *tin-iLEXY/*tin^{EC40} embryos recapitulated the *tin* KO phenotype under blue light, with very few residual fibroblast-like heart cells, complete loss of the dorsal SM and no constrictions in the VM (Figure 19D). On the contrary, balancer embryos *tin-iLEXY/TM3 Ser GFP* and dark-reared homozygous *tin-iLEXY/*tin^{EC40} embryos were fully normal (not shown). This indicates that *tin-iLEXY* in *trans* to a LOF mutant can recapitulate most of the *tin* phenotypes, although some Tin function might remain.

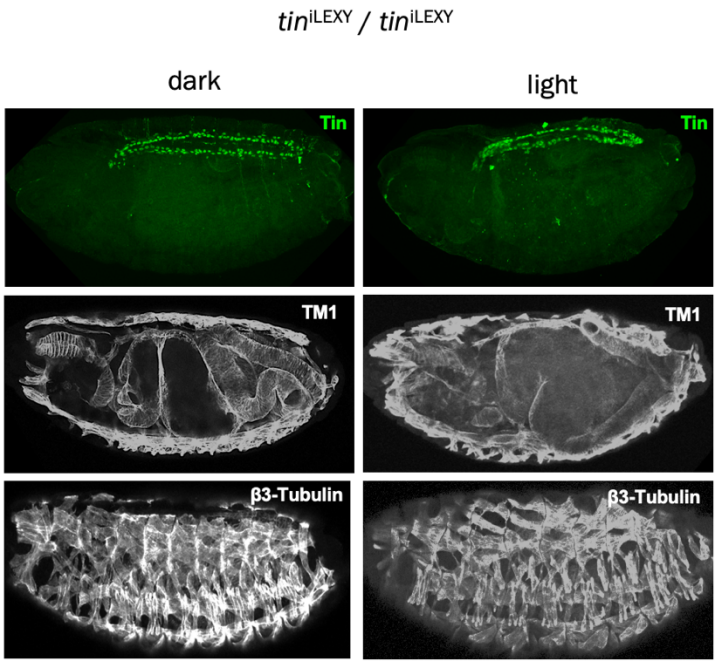


Figure 21: **tin-iLEXY** ***homozygous mutants have mild defects***.
Immunostaining for Tin, Tropomyosin 1 (TM1, middle), and β3-Tubulin in stage 16 homozygous *tin-iLEXY* embryos kept in the dark or in blue light for the whole embryogenesis. Tin depletion causes minor defects, mainly in the VM.

As for the *twi-iLEXY* line, Kögler *et al.* (2021) showed that Twi depletion could mimic the LOF mutant phenotype, with gastrulation failure and almost complete absence of mesodermal derivatives (Figure 19E).

The Twi, Mef2 and Tin (in transheterozygous embryos) iLEXY depletions therefore recapitulate the features of classic LOF mutants, albeit with a slightly milder effect, and I therefore focused on these three TFs for the remainder of this study. For *bin-iLEXY*, although tissue morphogenesis was impaired, a large number of VM cells were still specified, indicating that the iLEXY depletion is to weak and not informative enough to investigate mutant phenotypes. I describe below the single cell results for Mef2 and Twist depletions. The Tin experiments are still ongoing (the experiments take longer due to the requirement for a transheterozygous cross), and will be added to this study for publication.

3.6 Investigating the contributions of TFs to cell fate over time by sci-ATAC-seq

To understand the contribution of each TF to regulatory trajectories at key developmental transitions in the mesoderm across all muscle lineages, I performed single-cell ATAC-seq by combinatorial indexing (sci-ATAC-seq) (Cusanovich *et al.*, 2015, 2018) on *twi-iLEXY*, *Mef2-iLEXY* and *tin-iLEXY/tin*^{EC40} embryos after depletion at stage 9 and 11. Following a 2 h collection, I aged the embryos for 2.5 h for stage 9 depletion and for 4.5 h for stage 11 depletion, I exposed them to blue light for 1 h and formaldehyde-fixed the embryos immediately after depletion (Figure 22A). I performed sci-ATAC-seq on whole embryo nuclei, following a modified protocol as Secchia *et al.* (2022). I included WT embryos from stage 9 and 11 and combined different timepoints from the same line in the same experiment to

separate batch effects and biological variation. These experiments were done on nuclei isolated from whole embryos as any staining with mesodermal markers for FACS sorting would be sensitive to the depletion of the TFs Mef2, Tin and Twi. After pre-processing the data (Methods section 2.12.1, 2.12.2), I computationally isolated the mesoderm cells by clustering. The WT nuclei were used as an anchor population in the batch correction before integrating data from multiple lines and datasets.

3.6.1 Mef2

3.6.1.1 Mef2 depletion drives more than half of mesoderm cells towards two mutant cell states

To investigate the role of Mef2, I performed two replicates of symmetrical strand sci-ATAC-seq (S3-ATAC-seq) (Mulqueen *et al.*, 2021) using the same setup, each starting from 240000 nuclei, of which 40% *Mef2-iLEXY* were depleted at stage 9, 40% *Mef2-iLEXY* at stage 11 and 10% WT stage 9 and 10% WT stage 11. From the two replicates combined, a total of 19998 single nuclei passed standard quality filters based on nucleosomal patterning, median reads per cell and doublet exclusion. I obtained 3671 median reads per cell, in line with the expected average yield.

The analysis was performed as for previously generated sci-ATAC-seq time course datasets in our lab (Cusanovich *et al.*, 2018; Secchia *et al.*, 2022), to ensure consistency and facilitate direct comparisons. For each experiment, I created a count matrix based on consensus peaks from the Cusanovich *et al.* (2018) whole-embryo dataset, and further refined it by peak calling on my data. To overcome data sparsity, a common limitation in single-cell techniques, and to simplify the large number of dimensions in the peak space, I performed dimensionality reduction on the cell-by-peak count matrix using the Latent Dirichlet Allocation (LDA) reduction in cisTopic (Bravo González-Blas *et al.*, 2019). I identified clusters with Seurat (Hao *et al.*, 2021), I visualized them on a Uniform Manifold Approximation and Projection (UMAP) and assigned cluster identity based on term overrepresentation from our database of previously characterised developmental enhancers (Cusanovich *et al.*, 2018) and from proximal genes in the BDGP gene expression database (Wilczynski *et al.*, 2012), on all peaks and on differentially accessible (DA) peaks. The annotations from all approaches were highly concordant. To confirm the annotation and further increase the resolution, especially in mutant clusters, I inferred TF activity by calculating TF deviation scores based on ChIP peaks of mesoderm TFs and on known motifs from cisBP and JASPAR, and on modENCODE, using chromVAR (Schep *et al.*, 2017). To increase the resolution within the mesoderm, I subset cells annotated as mesoderm and mesodermal derivatives and repeated the peak calling (Methods section 2.12.5).

3.6.1.2 Half of the mesoderm cells deviate from the wild-type trajectory upon Mef2 depletion

Using the starting peak set from Cusanovich *et al.* (2018), a cluster of *Mef2-iLEXY* mutant cells emerged from the mesoderm in the whole embryo Mef2-iLEXY dataset. After clustering refinement on the dataset-specific cluster and genotype peaks, yielding 32131 unique peaks, a large cluster of mutant cells

emerged in a central position in the UMAP, branching towards all cell identities (Figure 22 B-C). Cells in this cluster were not assigned to any known cell type and may regress instead to an intermediate and possibly less specified state, as suggested by their central position and branching.

Apart from the emergence of a mutant cluster, there were no major differences in the distribution of WT and *Mef2-iLEXY* cells for non-muscle lineages, demonstrating the specificity of the effect to the mesoderm: the proportion of cells per tissue reflected the embryo composition, and, in all non-mutant clusters, the WT:mutant ratio was close to the expected value (75% *Mef2-iLEXY*, 25% WT cells) (Figure 22D). Remarkably, this was also true for the mesodermal cluster, which included part of the mutant cells; this reflects Mef2 role in muscle differentiation, and not specification. The muscle cells are still specified in *Mef2* mutant embryos (Bour *et al.*, 1995; Lilly *et al.*, 1995), but fail to differentiate into mature contractile muscles.

All conventional cell types showed a clear temporal progression from stage 9 to 11; conversely, Mef2-depleted cells from both stages were mixed in the centre of the mutant cluster, hinting to a high similarity between mutants with depletion at stages 9 and 11 (Figure 22E). Moreover, stage 11 was more represented than stage 9 in the mutant cluster; this could be explained by less efficient Mef2 depletion at stage 9, possibly linked to Mef2 stabilization.

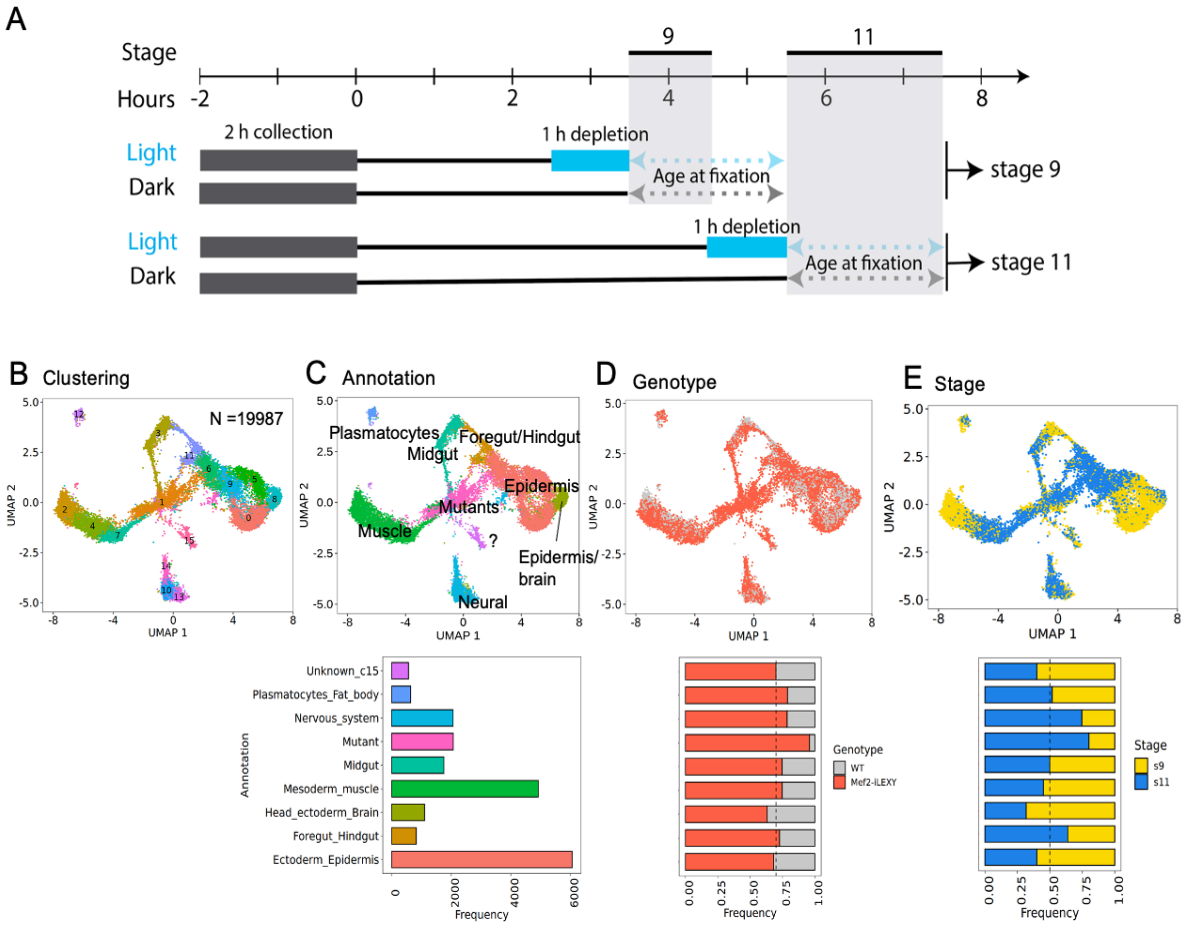


Figure 22:* Mef2-iLEXY *mutants segregate in a central cluster in the whole-embryo dataset.

(A) Scheme of the depletion. Depletion was performed as for CUT&Tag: embryos were collected for 2 h in safelight conditions, aged in the dark for 2.5 h (stage 9) or for 4.5 h (stage 11), exposed to blue light or dark-reared for 1 h and fixed.
(B) UMAP visualization of the whole embryo Mef2-iLEXY dataset on cluster and genotype peaks, with cells coloured by cluster.
(C) Annotation based on BDGP terms enriched among the DA peaks. UMAP and histogram with cell number by annotation.
(D) UMAP as in B, coloured by genotype, and bar chart with genotype frequency by annotation.
(E) UMAP as in B, coloured by stage, and bar chart with stage frequency by annotation.

To understand the relationship between mutant cells and their embryonic origin (i.e. the mesoderm), I reclustered the mutant cluster and the cells annotated as mesoderm and its derivatives (mesoderm, muscle, plasmatocytes, fat body), for a total of 7626 cells (Figure 23A). The main cluster included the mesoderm and muscle primordium (clusters 1,5,0) and a tip with VM identity (cluster 4) (Figure 23B). There was a clear temporal progression, cluster 1 being the youngest and cluster 4 the oldest, as displayed by the enrichment for binding sites of mesoderm TFs at different timepoints (Figure 23 E-F). Clusters 1, 5 and 0 were enriched for accessible binding sites of Twi, Mef2 and Tin, according to their mesoderm identity, whereas cluster 4 had the highest enrichment for Bap and was the only cluster enriched for Bin, confirming its VM identity (Figure 23 E-F). Plasmatocytes segregated in cluster 6. Interestingly, *Mef2-iLEXY* mutants divided into two clusters close to each other and completely separated from the main mesodermal cluster (Figure 23 A-C). The two clusters did not show obvious differences in coverage, suggesting that the separation is biological rather than technical. Since the mutant clusters did not map to any known cell identity, by mapping the peaks to characterised enhancers (Cusanovich *et al.*, 2018) and BDGP genes (Wilczynski *et al.*, 2012), and had a striking reduction in the accessibility of the sites bound by mesodermal TFs Twist, Mef2 and Tin (Figure 23 E-F), I used the accessibility of binding sites and motifs of a broad panel of mesodermal and non-mesodermal TFs to further investigate their identity.

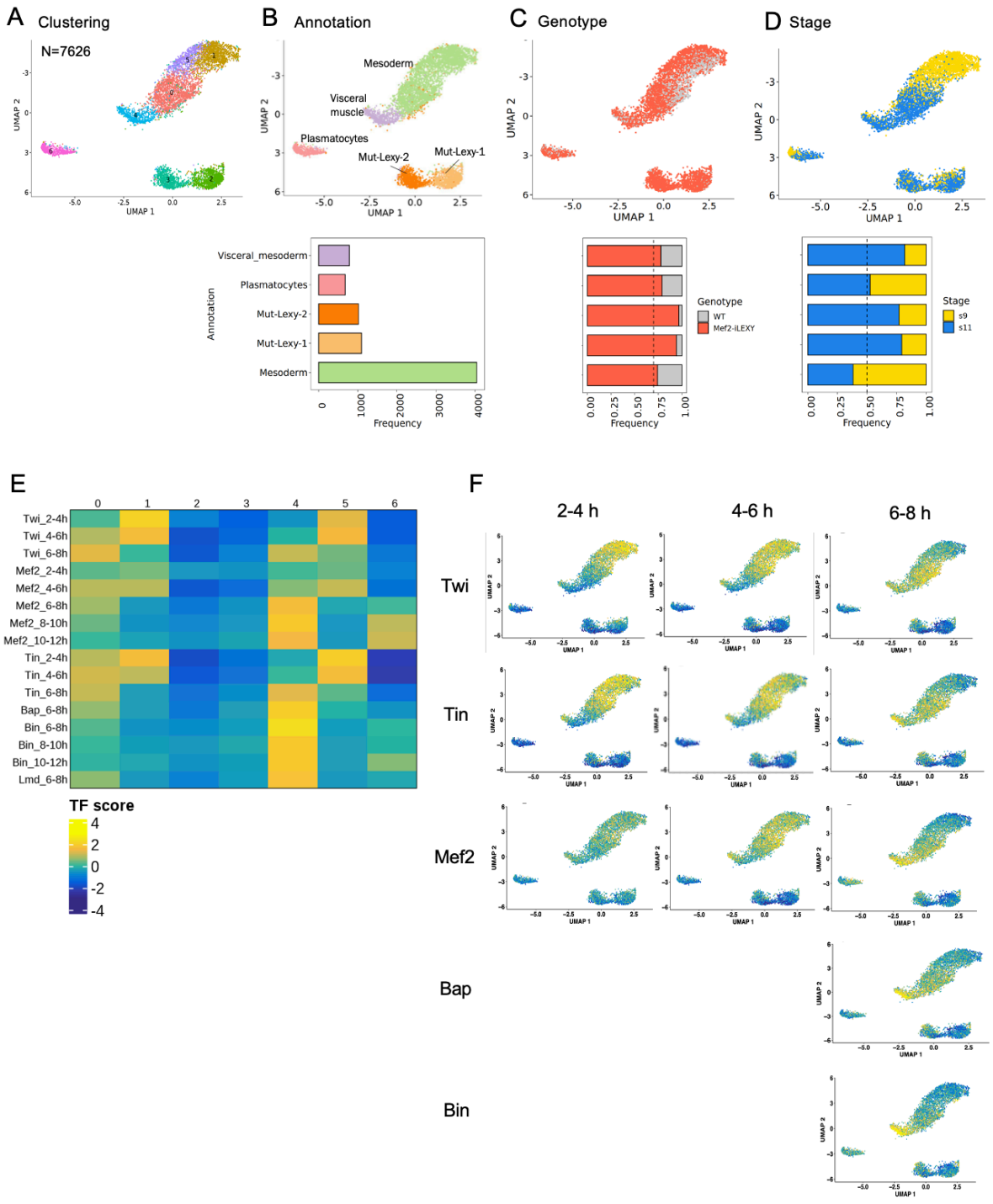


Figure 23: After mesoderm reclustering,* Mef2-iLEXY *mutants are separated into two clusters deviating from the normal mesoderm trajectory.

(A) UMAP visualization of the mesoderm from the Mef2-iLEXY dataset on mesoderm cluster and genotype peaks, with cells coloured by cluster.

(B) Annotation based on BDGP terms enriched among the DA peaks. UMAP and histogram with cell number by annotation.

(C) UMAP as in A, coloured by genotype, and bar chart with genotype frequency by annotation.

(D) UMAP as in A, coloured by stage, and bar chart with stage frequency by annotation.

(E) Heatmap of accessibility deviation scores for TF-bound regions at different timepoints from bulk ChIP data.

(F) UMAPs with cells coloured by accessibility deviation score (TF score) for the indicated mesoderm TF using bulk ChIP data for specific timepoints.

3.6.1.3 Mef2 depletion leads to two distinct mutant states

3.6.1.3.1 Both mutant clusters are compatible with failure in muscle differentiation at the equivalence cluster stage

The two *Mef2-iLEXY* mutant clusters had numerous common markers.

Motifs for the gastrulation gene *sna* and for *escargot* (*esg*) and *worniu* (*wor*) were highly enriched in both mutant clusters (Figure 24A). Since these three genes have similar binding motifs, only one or two could in fact be enriched. Sna defines the mesodermal domain in opposition with Twi, by preventing expression of genes destined to be active in more lateral or dorsal mesectodermal regions (Leptin, 1991). Esg could also play a role in defining the mutant cell identity, since in the adult midgut it is necessary and sufficient to maintain the stem cell population by repressing differentiation genes in progenitor cells (Korzelius *et al.*, 2014).

Mutant clusters were also enriched for *daughterless* (*da*) (Figure 24A), which homodimerizes with Twi to repress genes for somatic myogenesis, allowing the development of other mesodermal tissues. Remarkably, overexpression of the Twi-Da dimer in the mesoderm caused a phenotype similar to iLEXY depletion of Mef2 for the whole embryogenesis (Figure 19), with few SM cells and patterning defects (Castanon *et al.*, 2001; Wong, Castanon and Baylies, 2008).

Overall, the mutants appeared to deviate from the SM trajectory, and get stuck or regress to earlier broader identities, possibly similar to those at gastrulation.

The enrichment for members of the *achaete-scute* complex like *asense* (*ase*) and *lethal of scute* (*l(1)sc*) (Figure 24A) suggests a more precise timeframe for the mutant identity, which is similar to the mesoderm cells during the formation of equivalence clusters. Equivalence clusters are groups of cells from which the central one will become a FC, and the surrounding ones will turn into FCMs (Dobi, Schulman and Baylies, 2015). A similar process occurs during neuronal specification (Simpson and Carteret, 1990; Cubas *et al.*, 1991). *l(1)sc* is first expressed in the whole equivalence cluster at stage 10, and restricted to FCs at stage 11 (Carmena, Bate and Jiménez, 1995). *ase* is better known for its contribution to the emergence of progenitors in neuronal equivalence clusters (Dominguez and Campuzano, 1993; Williams *et al.*, 1993), but could have a similar role in the muscle.

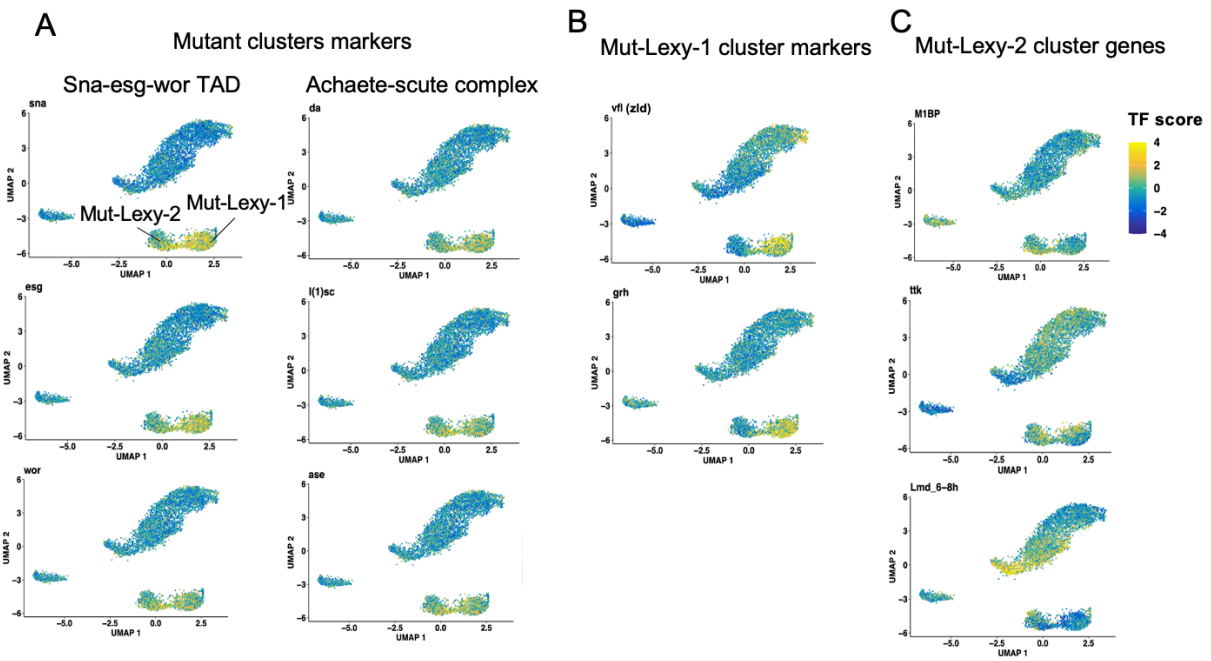


Figure 24: Both the* Mef2 *mutant clusters have increased accessibility at target sites of TF linked to gastrulation and specification of equivalence clusters, and they have specific enrichment for other regions.

(A) UMAPs with cells coloured by the deviation score for specific TFs and motifs enriched in the accessible regions of both the iLEXY mutant clusters.

(B) UMAPs with cells coloured by the deviation score for specific TFs and motifs enriched in the accessible regions of the Mut-Lexy-1 cluster (cluster 2).

(C) UMAPs with cells coloured by the deviation score for specific TFs and motifs enriched in the accessible regions of the Mut-Lexy-2 cluster (cluster 3).

3.6.1.3.2 Mutant cluster 1 has more stemness features, whereas mutant cluster 2 could undergo developmental arrest during FCM specification

Some markers were enriched in a single mutant cluster, showing that cells in the two clusters might have slight differences in their identity (Figure 24 B-C).

Cluster 2 (Mut-Lexy-1) was enriched for *zelda* (*zld*), a pioneer TF responsible for opening the chromatin during zygotic genome activation (ZGA) (Liang *et al.*, 2008; Harrison *et al.*, 2011; Schulz *et al.*, 2015). Zld had many interactors enriched in the mutant clusters: it binds both the primary and shadow enhancer of *sna* to regulate transcription coordination (Dufourt *et al.*, 2018), promotes Twi and Bicoid (Bcd) binding (Yáñez-Cuna *et al.*, 2012; Xu *et al.*, 2014),and interacts with both Ultrabithorax (Ubx) and Tin to ensure gene expression in the mesoderm (Carnesecchi *et al.*, 2020). Constitutively open Zld-bound loci contain GAGA motifs and show Trl extensive binding (Schulz *et al.*, 2015), also enriched in cluster Mut-Lexy-1 (Figure 24B). Moreover, the Zld LOF phenotype shows some analogy to Mef2 iLEXY depletion: it causes segmentation defects and strong abnormalities in the muscle pattern (Staudt *et al.*, 2006). Mut-Lexy-1 was also enriched in a second pioneer TF, *grainyhead* (*grh*) (Figure 24B), which can act both as an activator and repressor, being involved in epithelial cell specification (Nevil *et al.*, 2017). Zld and Grh compete for occupancy of the TAGteam element for transcription of many early

genes (Harrison, Botchan and Cline, 2010). The enrichment for both *zld* and *grh* suggests that cluster Mut-Lexy-1 could regress to an earlier cell identity, with higher stemness, although maintaining some mesodermal features.

Cluster 3 (Mut-Lexy-2) did not have exclusive markers but higher enrichment for some genes like *tramtrack* (*ttk*), *DNA replication-related element factor* (*Dref*) and *Motif 1 Binding Protein* (*M1BP*) (Figure 24C). Along with Lmd, Ttk is involved in FCM specification in equivalence clusters (Ciglar *et al.*, 2014). Ttk mediates transcriptional repression of FC genes, and Lmd – enriched only in a small section of the cluster – mediates the activation of FCM genes (Ciglar *et al.*, 2014). Prospective FCMs in Mut-Lexy-2 could thus be stuck in an intermediate state after losing the FC identity, since they cannot activate the transcriptional program of FCMs without Lmd.

Overall, the *Mef2-iLEXY* mutants appear to suffer developmental arrest or regress to earlier cell states at the time of the establishment of equivalence clusters and the specification of FCs and FCMs. Some cells (Mut-Lexy-1 cluster) could reactivate an earlier and more stem transcriptional program by increased chromatin accessibility at earlier targets through pioneering factors and TFs expressed in gastrulation, whereas other cells (Mut-Lexy-2 cluster) seem to be arrested at FCM specification. This may explain the lack of fused muscle fibres and the survival of FCs, but not FCMs, in the terminally differentiated mutant (stage 16). Such different responses of sub-populations of cells to mutants are masked in normal bulk experiments, and highlight the advantage of analysing mutants with single cell genomics.

3.6.1.3.3 Both mutant clusters are present at stages 9 and 11 but show temporal progression

To investigate if any mutant cluster identities was more associated to Mef2 depletion at a specific stage, I computationally separated stage 9 and stage 11 mesoderm, and reclustered the cells based on stage-specific peaks. At stage 9, I retrieved 3426 cells, 2723 from the *Mef2-iLEXY* line and 703 from the WT; at stage 11, I had a total of 4200 cells, 3445 from *Mef2-iLEXY* and 755 from the WT. Clusters highly overlapping Mut-Lexy-1 and Mut-Lexy-2 clusters were retrieved at both stages (Figure 25 A-B), indicating that they are not specifically associated to a single stage and not due to the use of a common peak set. The proportion of cells in the two new clusters corresponding to Mut-Lexy-1 and Mut-Lexy-2 changed with time: at stage 9, Mut-Lexy-1 had less cells (244/2723, 8.96%) than Mut-Lexy-2 (341/2723, 12.52%) (Z-score $p = 2.655e-05$); at stage 11, the trend was inverted, with Mut-Lexy-1 accounting for 23.60% (813/3445) and Mut-Lexy-2 for 14.34% (494/3445) of the mutant cells (Z-score $p < 2.2e-16$). The overall proportion of *Mef2-iLEXY* cells in the mutant clusters was also higher at stage 11 (31.12% vs 16.34%, Z-score $p < 2.2e-16$).

The difference in the proportion of *Mef2-iLEXY* cells in mutant clusters may be due to more incomplete depletion at stage 9, or to a slower accumulation of changes in accessibility. Depletion at stage 9 seems to mostly impinge on FCM specification, and to lead to more extreme aberrations in a slightly lower number of cells; on the contrary, the cell fate associated to stemness markers is predominant at stage

11. This would be compatible with a two-phase model, where cells first fail to follow the standard FCM differentiation pathway, transforming then into an aberrant or more primordial state.

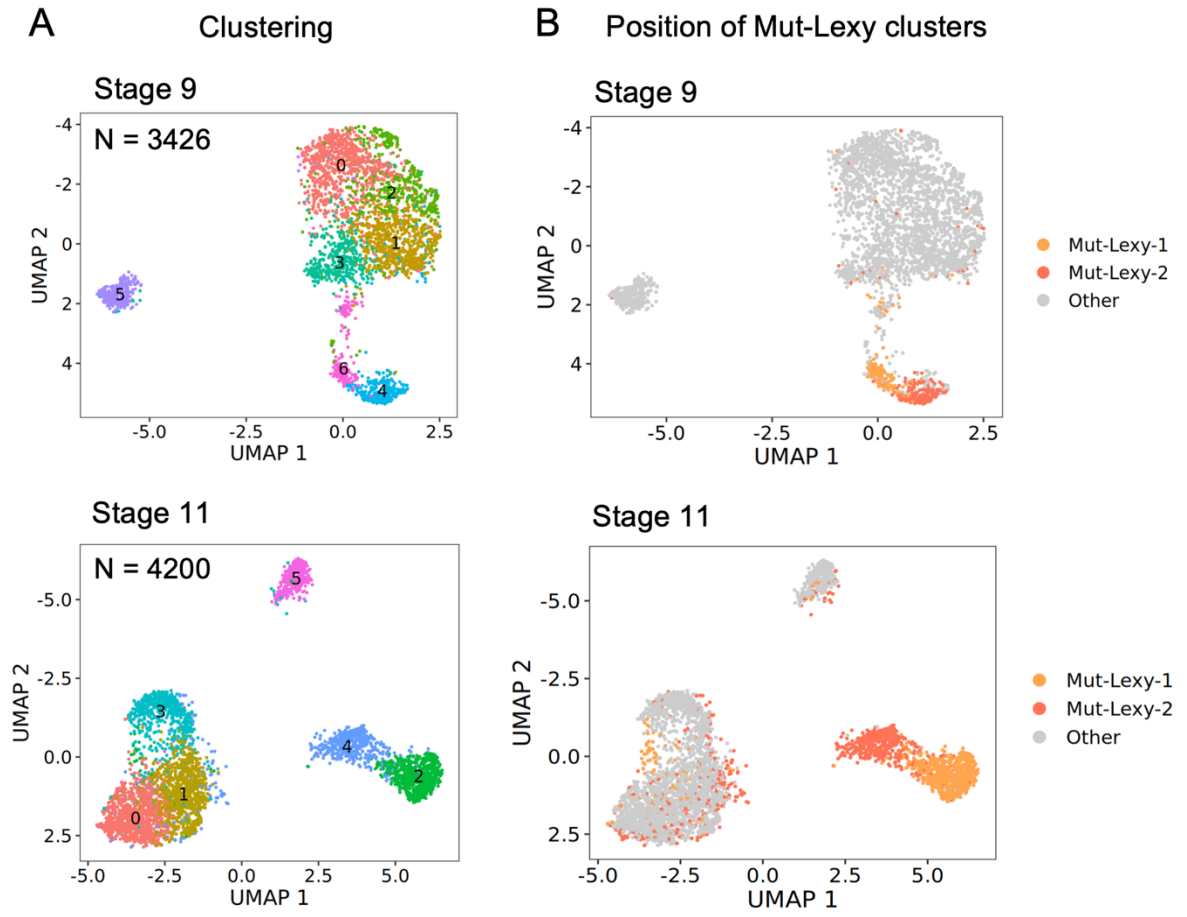


Figure 25: Stage-specific clustering of* Mef2 *mutants shows that the mutant clusters Mut-Lexy-1 and Mut-Lexy2 are present both at stage 9 and stage 11, and more abundant at stage 11.

(A) UMAP coloured by *de novo* clusters on stage 9 (top) and stage 11 (bottom) mesoderm.

(B) UMAP showing the position of cells in the *Mef2-iLEXY* mutant clusters Mut-Lexy-1 (light orange) and Mut-Lexy-2 (dark orange).

3.6.1.4 Mutant states do not regress to previous developmental state, but branch off the WT developmental trajectory at the right time

Since the Mut-Lexy-1 had markers compatible with higher stemness, I investigated whether the trajectory of *Mef2-iLEXY* mutants overlapped earlier timepoints in mesoderm development, as perhaps would be expected for a 'simple' block in differentiation. To increase the resolution within the mesoderm, I compared the Mef2-iLEXY dataset to two previous sci-ATAC-seq datasets from the lab (Figure 26, Figure 27): a dataset of 3–12 h FAC sorted mesodermal nuclei from WT embryos and 10–12 h *Mef2* KO embryos (Secchia et al., 2022), and an extensive whole-embryo sci-ATAC-seq time course, from 0 to 20 h (DEAP dataset) (Calderon et al., 2022).

I generated a new cell-per-peak count matrix using the merged peaks from all integrated datasets, for a total of 74,946 peaks for the integration with Secchia *et al.* (2022) and 95,446 peaks for the integration

with Calderon *et al.* (2022). Using the WT from each dataset as a reference, I performed LDA dimensionality reduction in cisTopic and batch correction with MNN (Haghverdi et al., 2018) (Methods section 2.12.8) (Figure 26A).

For integration with Secchia *et al.* (2022), the integrated dataset after batch correction recapitulates mesoderm development in time. The WT aligned on a trajectory from left to right, according to their stage: from the 3–5 h unspecified mesoderm cells on the far left, there was a progression towards mesoderm specification halfway through the time course, roughly in the centre of the UMAP, to the final cell identities at 10–12 h on the right (Figure 26 C-D). At the time of mesoderm specification (5–8 h), the early mesoderm was divided into the SM (clusters 3, 4, 5 and 7) and a first branch of VM (clusters 13, 18). This branch was later joined by a second branch of VM emerging from the SM at 10–12 h (clusters 8, 11) (Figure 26 D, F). The heart mesoderm also stemmed from the muscle at 10–12 h (cluster 11): a population of dorsal/cardiac cells could already be identified at 5–8 h (Figure 26 D, F). Plasmatocytes were in a separate cluster far from the WT mesoderm trajectory at 4–8 h (Figure 26D). Mesoderm TF deviation scores confirmed temporal progression and cluster annotation (Figure 26 E, G). The SM was divided into two branches, which appear as two parallel lines in the UMAP (Figure 26D). Since this was not observed in the initial dataset, I re-clustered after removing the 10,693 unique peaks from the Mef2-iLEXY dataset from the common set (leading to 64,253 peaks); this did not affect branching (not shown). Thus, SM splitting was not due to technical issues.

Both *Mef2-iLEXY* and *Mef2* KO mutant clusters were clearly separated from the WT trajectory (Figure 26B). They branched out in the same direction from the WT of the corresponding timepoint. The two *Mef2* KO clusters corresponded to the two mutant clusters in Secchia *et al.* (2022), whereas *Mef2-iLEXY* clusters were now separated based on proximity to the WT trajectory (Figure 26B): cluster Mut-Lexy-1, which had a more stem-like identity, was the farthest from the WT, while cluster Mut-Lexy-2 was closer and divided into two sub-clusters (Figure 26B). Therefore, iLEXY mutants were still isolated from the WT trajectory; their degree of separation from the WT was consistent with previously observed markers. These clusters were also separated from the *Mef2* KO mutants (which were also off the WT trajectory), which may be due to the later embryonic stage of the *Mef2* KO experiment (stage 13 versus 9–11), and/or to the accumulation of changes due to absence of Mef2 over all previous stages of embryogenesis.

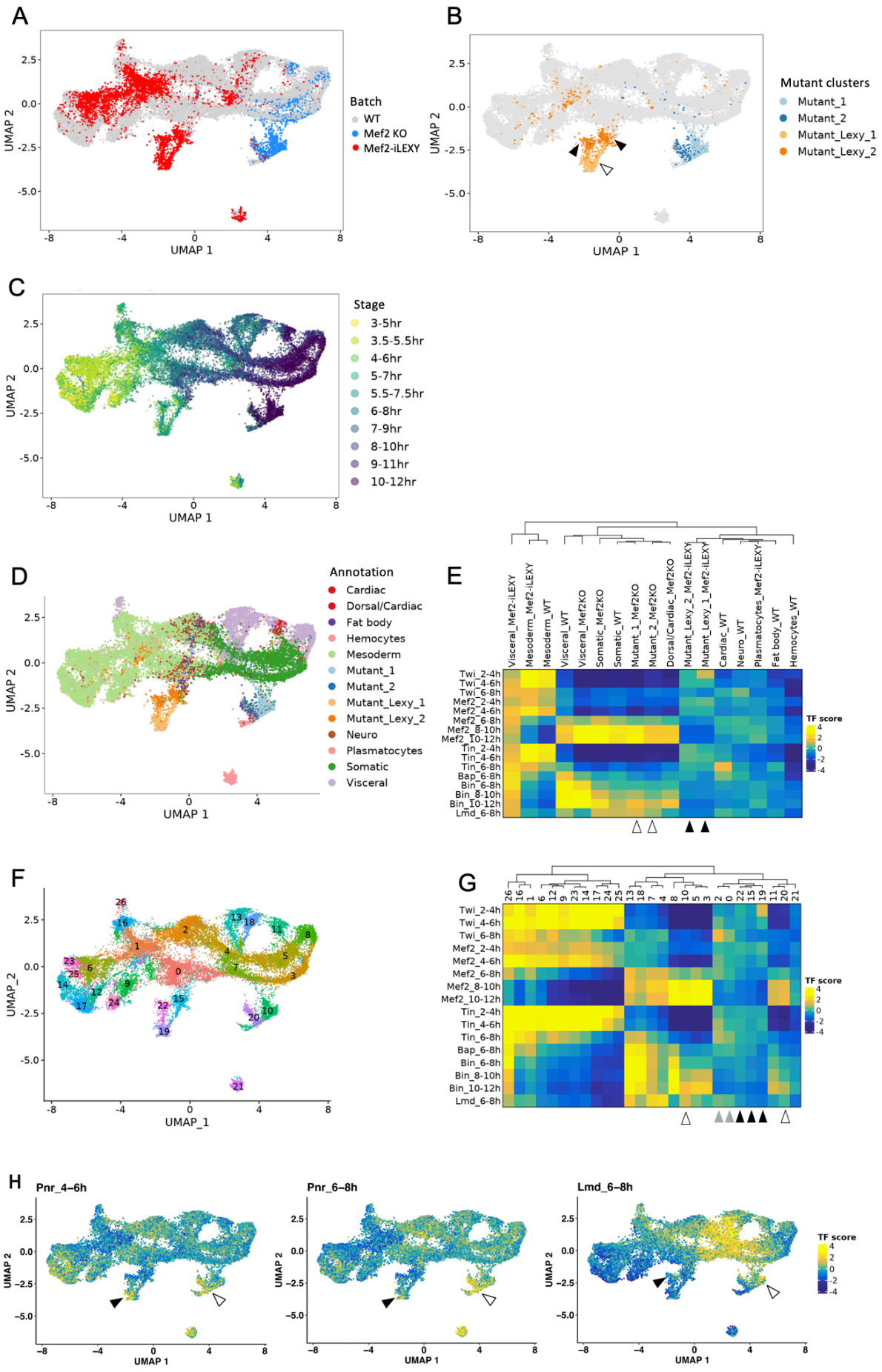

A
UMAP 2
UMAP 1
Batch
WT
Mef2 KO
Mef2-iLEXY
B
Mutant clusters
Mutant_1
Mutant_2
Mutant_Lexy_1
Mutant_Lexy_2
C
Stage
3-5hr
3.5-5.5hr
4-6hr
5-7hr
5.5-7.5hr
6-8hr
7-9hr
8-10hr
9-11hr
10-12hr
D
Annotation
Cardiac
Dorsal/Cardiac
Fat body
Hemocytes
Mesoderm
Mutant_1
Mutant_2
Mutant_Lexy_1
Mutant_Lexy_2
Neuro
Plasmatocytes
Somatic
Visceral
E
Visceral_Mef2-iLEXY
Mesoderm_Mef2-iLEXY
Mesoderm_WT
Visceral_WT
Visceral_Mef2KO
Somatic_Mef2KO
Somatic_WT
Mutant_1_Mef2KO
Mutant_2_Mef2KO
Dorsal/Cardiac_Mef2KO
Mutant_Lexy_2_Mef2-iLEXY
Mutant_Lexy_1_Mef2-iLEXY
Cardiac_WT
Neuro_WT
Plasmatocytes_Mef2-iLEXY
Fat body_WT
Hemocytes_WT
Twi_2-4h
Twi_4-6h
Twi_6-8h
Mef2_2-4h
Mef2_4-6h
Mef2_6-8h
Mef2_8-10h
Mef2_10-12h
Tin_2-4h
Tin_4-6h
Tin_6-8h
Bap_6-8h
Bin_6-8h
Bin_8-10h
Bin_10-12h
Lmd_6-8h
TF score
F
UMAP_2
UMAP_1
G
TF score
H
Pnr_4–6h
Pnr_6–8h
Lmd_6–8h
TF score

Figure 26: Comparison of the Mef2-iLEXY dataset to the WT and* Mef2 *KO dataset from Secchia* et al. *(2022).

(A) UMAP coloured by dataset: WT dataset in grey, Mef2-iLEXY in red and *Mef2* KO dataset in blue.
(B) UMAP showing the position of Mef2-iLEXY mutant clusters Mut-Lexy-1 (white arrow) and Mut-Lexy-2 (split in two, black arrows), and of *Mef2* KO mutant clusters Mutant_1 and Mutant_2.
(C) UMAP coloured by stage.
(D) UMAP coloured by the original annotation of the datasets.
(E) Heatmap of accessibility deviation scores for selected mesoderm TFs based on ChIP data at different timepoints, for each annotation and dataset. Mef2-iLEXY mutant clusters are indicated by black arrows, and Mef2 KO mutant clusters by white arrows.
(F) UMAP coloured by clusters called on the merged datasets.
(G) Same as E, with clusters instead of annotations. *Mef2-iLEXY* mutant clusters are indicated by black arrows, *Mef2* KO mutant clusters by white arrows, and the WT clusters closest to the *Mef2-iLEXY* mutants by grey arrows.
(H) UMAPs with cells coloured by the deviation score for the mesoderm TFs Pnr and Lmd. *Mef2-iLEXY* mutant clusters enriched in those TFs are indicated by a black arrow, *Mef2* KO clusters by a white arrow.

To further investigate the identity of mutant iLEXY clusters and their relationship to WT and KO mutant clusters, I compared accessible regions by TF deviation scores. Using the original annotation, mesoderm clusters from the Mef2-iLEXY dataset clustered with the WT mesoderm from Secchia *et al.*, 2022, the plasmatocytes were close to the fat body and hemocytes, and the VM from both datasets clustered together (Figure 26E). Overall, the same tissues at the same timepoints from different datasets clustered together (Figure 26D), confirming the lack of residual batch effects after the batch correction, and the consistency among the two datasets. As for mutant clusters, the *Mef2* KO clusters from 10–12 h embryos were closest to the SM, in accordance with Secchia *et al.* (2022). On the contrary, stage 9 and 11 Mef2-iLEXY clusters were grouped together and isolated from the rest of the mesoderm, closer to non-SM identities (Figure 26E).

After calling *de novo* clusters on the integrated dataset, the old two *Mef2-iLEXY* mutant clusters were split into three, of which cluster 19 corresponds to the core of former Mut-Lexy-1, and clusters 22 and 15 to cluster Mut-Lexy-2 and some branches from Mut-Lexy-1 (Figure 26F). In the matrix of TF deviation scores based on the new clustering, the iLEXY mutant clusters were grouped together, closer to *Mef2* KO cluster 20 and to WT mesoderm clusters at the same stage (Figure 26G).

The two *Mef2* KO clusters had some similarity with iLEXY mutant clusters in the accessibility at specific TF sites (Figure 26H): one cluster was enriched in *pnr*, similarly to the iLEXY mutant cluster Mut-Lexy-1, and the other in *lmd*, also enriched in part of the iLEXY mutant cluster Mut-Lexy-2.

Overall, iLEXY mutant clusters have some similarity to *Mef2* KO clusters, but developmental time was the main source of variability within the integrated dataset. Moreover, mutant cells, from both iLEXY and the KO, did not regress on the WT developmental trajectory, rather acquiring a new identity with some features typical of earlier cells.

Using the 4–8 h WT dataset as a reference, the integration of the Mef2-iLEXY dataset with the mesoderm data from Calderon *et al.* (2022) confirmed that the Mef2-iLEXY trajectory is separate from the WT (Figure 27B). Here, the two Mef2-iLEXY original clusters were still grouped together based on accessibility at mesoderm TF sites; the closest WT identities were the 4–6 h muscle primordium and,

secondarily, the 2–4 h mesoderm anlage, again indicating via an independent dataset, that after the depletion of Mef2 the cells are in a more immature mesoderm-like state (Figure 27 D, E).

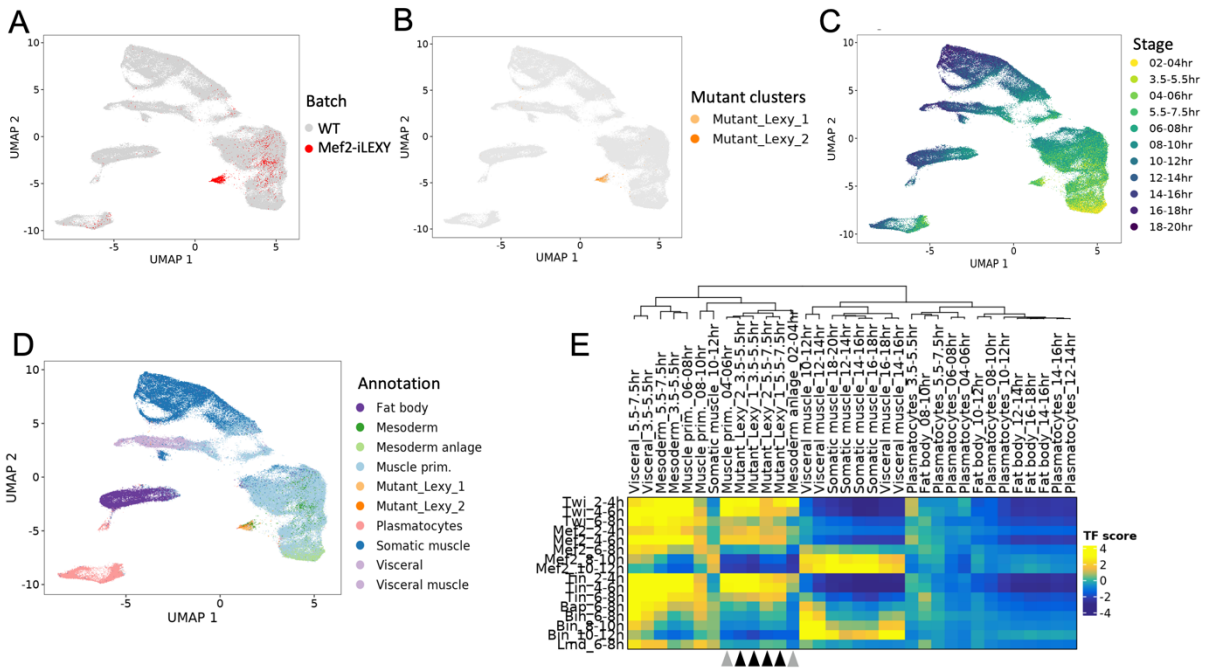


Figure 27: Comparison of the Mef2-iLEXY dataset to the WT Calderon* et al. *(2022) dataset.

(A) UMAP showing the position of the Mef2-iLEXY dataset (red) on the WT Calderon *et al.* (2022) dataset (grey).
(B) UMAP showing the position of *Mef2-iLEXY* mutant clusters.
(C) UMAP coloured by stage.
(D) UMAP coloured by cell annotation in the original datasets.
(E) Heatmap of the accessibility deviation scores of TFs by ChIP peaks at different timepoints, by age and annotation. Black arrows point to the Mef2-iLEXY mutant clusters, grey arrow to the closest WT clusters.

3.6.1.5 Accessibility is mostly reduced at Mef2-bound regions

To identify the underlying molecular changes leading to these changes in cell behaviour, I inspected the changes in accessibility in *Mef2-iLEXY* mutant clusters. I compared each mutant cluster to the closest WT mesoderm cluster (cluster 0) by calling DA peaks with absolute logFC > 0.5 and FDR-corrected p value < 0.05. Cluster Mut-Lexy-1 had a prevalence of peaks with loss of accessibility (858 peaks out of 3257 total peaks), and a moderate number of peaks with increased accessibility (252 peaks) (Figure 28A). On the other hand, cluster Mut-Lexy-2 had almost only peaks with reduced accessibility among DA peaks (951 out of 3027 total peaks), and only 16 up DA peaks (Figure 28A), hinting to a more direct effect of Mef2 depletion.

To assess directly if the differentially accessible peaks are Mef2-depleted peaks, I examined the overlap between ChIP peaks of Mef2 at 4–6 and 6–8 h and ATAC peaks. In the Mut-Lexy-1 cluster, 63.3% of down DA peaks were bound by Mef2, versus only 33% of non-significant and 6% of up peaks (Figure 28B), showing that accessibility loss is correlated to Mef2 depletion. The Mut-Lexy-2 cluster had a similar percentage of Mef2 sites as Mut-Lexy-1 among the down DA (64.9%) and non-significant peaks (30.9%), and no binding at up DA peaks (not shown). Accordingly, Mef2 depletion as measured by

CUT&Tag was significantly higher at down DA than at non-significant or up DA sites for Mut-Lexy-1 at stage 9 and 11, and for Mut-Lexy-2 at stage 9 (Figure 28C).

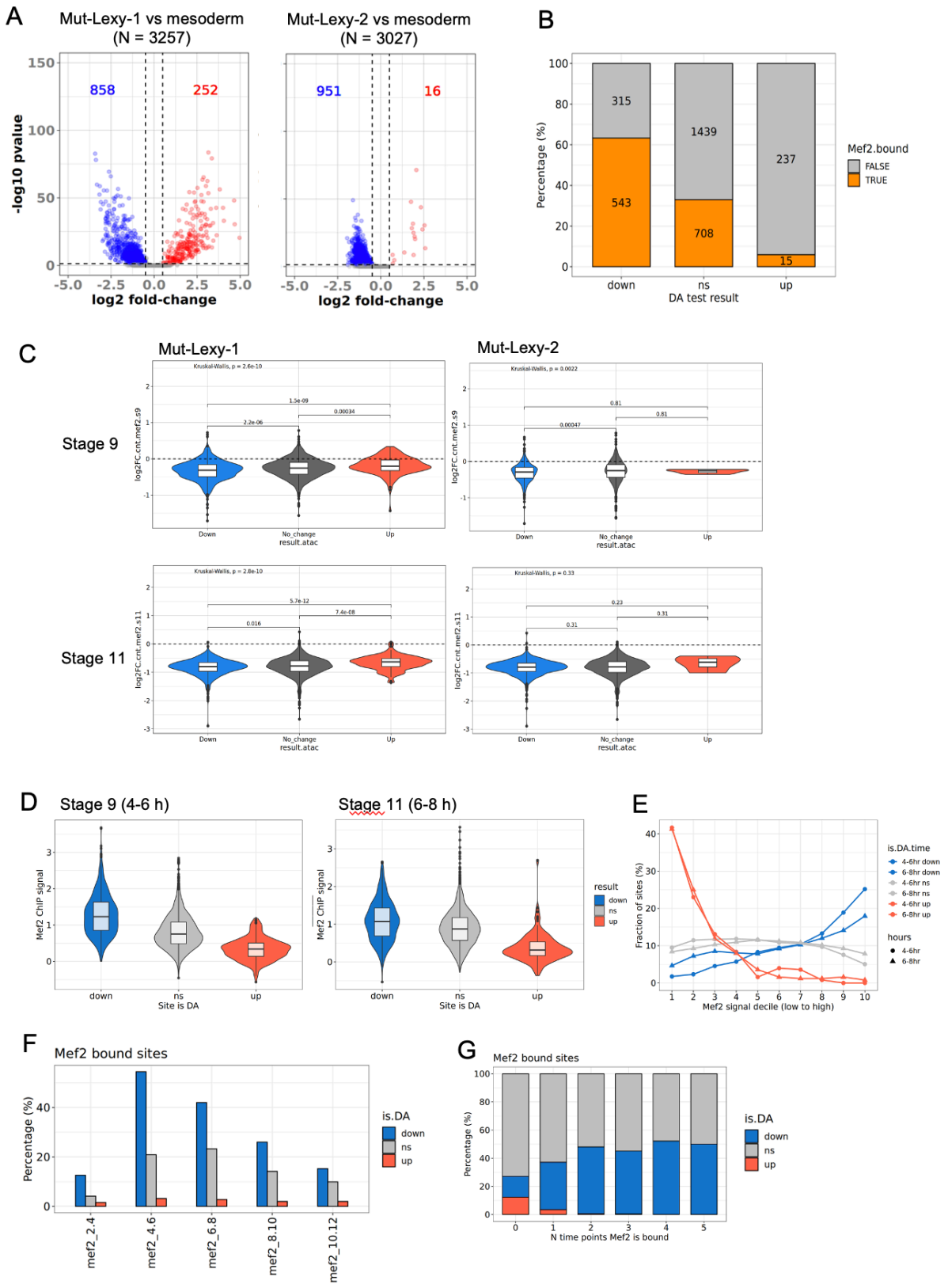


Figure 28: The* Mef2-iLEXY *mutant clusters mostly have reduced accessibility, especially at Mef2-bound sites.

(A) Volcano plot of ATAC peaks in clusters Mut-Lexy-1 (left) and Mut-Lexy-2 (right) with respect to the closest WT mesoderm cluster (cluster 0). Differential peaks have FDR-adjusted p value < 0.05, |log2FC| > 0.5 and signal 10x above background. Mut-Lexy-2 has almost exclusively reduced peaks, and Mut-Lexy-1 has 3.4x more reduced than increased peaks.
(B) Percentage of Mef2-bound peaks in reduced, not changed and increased peaks in Mut-Lexy-1. The percentages are similar for Mut-Lexy-2 (not shown).
(C) Log2FC from CUT&Tag in reduced (blue), not changed (grey) and increased (red) ATAC peaks in the two *Mef2-iLEXY* mutant clusters. The stage refers to the CUT&Tag stage; ATAC peaks include both stage 9 and 11. Reduced ATAC peaks have the highest reduction upon light exposure in CUT&Tag.
(D) Mef2 ChIP signal at stage 9 and 11 in reduced, not changed and increased ATAC peaks in Mut-Lexy-1. Reduced ATAC peaks have the highest ChIP signal.
(E) Fraction of total reduced, not changed and increased ATAC peaks by Mef2 ChIP signal decile. Data from ChIP binding at stage 9 is indicated with triangles, from stage 11 with dots. Increased ATAC peaks have the lowest Mef2 ChIP signal, and decreased peaks the highest.
(F) Percentage of total reduced, not changed and increased ATAC peaks bound by Mef2 at different timepoints (2-4, 4-6, 6-8, 8-10, 10-12 h). Most of the down peaks are bound at 4-6 and 6-8 h, according to the chosen depletion stage.
(G) Percentage of total reduced, not changed and increased ATAC peaks bound by Mef2 for N timepoints. Down sites are mostly bound in at least 2 timepoint, and up sites never or at a single timepoint.

In both mutant clusters, the down DA peaks had higher Mef2 ChIP signal than non-significant or up DA peaks, especially at stage 9 (4–6 h); up peaks had the lowest ChIP signal (Figure 28D). The fraction of DA sites increased steadily together with the Mef2 signal, albeit with a steeper slope at stage 9 than at stage 11 peaks (Figure 28E). This suggests a possible higher compensation of Mef2 loss at stage 11. In addition, 4–6 h Mef2 ChIP peaks showed the highest overlap with down DA peaks, matching the stage 9 time point of the experiments, followed by 6–8 and 10–12 h peaks (Figure 28F). Sites with Mef2 binding at > 1 timepoint had a higher chance of being less accessible upon Mef2 depletion (Figure 28G). However, the fraction of DA sites was similar from two to five timepoints (continuous binding). This suggests that Mef2 presence impacts accessibility, but the occupancy across developmental time is not a major contribution. On the contrary, all (Mut-Lexy-2) or most (Mut-Lexy-1) up DA peaks were never bound by Mef2, or only at one timepoint, pointing to indirect effects.

3.6.1.6 Accessibility is mostly affected at co-bound sites

Since accessibility is not strongly correlated to the magnitude of Mef2 ChIP signal or its persistence in time, there could be additional mechanisms contributing to an open chromatin state. Mesoderm TFs have high co-occupancy at enhancers (Sandmann *et al.*, 2006, 2007; Jakobsen *et al.*, 2007), and regulated them in a combinatorial manner (Nguyen and Xu, 1998; Jakobsen *et al.*, 2007; Zinzen *et al.*, 2009; Jin *et al.*, 2013). To assess if co-occupancy correlates to maintenance or loss of accessibility, I examined the percentage of DA sites among those bound by a number of mesoderm TFs other than Mef2. Sites bound by Mef2 and 1 or 2 other TFs had 20–30% DA peaks; this percentage increased steadily in sites with an increasing number of TFs, especially $\geq$ 4 TFs, reaching 80% of sites (Figure 29A).

Most down DA sites exhibited co-binding of Twi, Tin, Bap, and Lmd (Figure 29B). This suggests that they may engage in cooperative binding with Mef2, which itself could promote or stabilize their binding. Upon Mef2 removal, the other TFs would bind to the enhancer at a significantly lower level,

if at all. Alternatively, there could be additive binding, with the highest contribution from Mef2; however, the correlation between the number of co-bound TFs and the percentage of DA peaks strongly supports the cooperativity hypothesis.

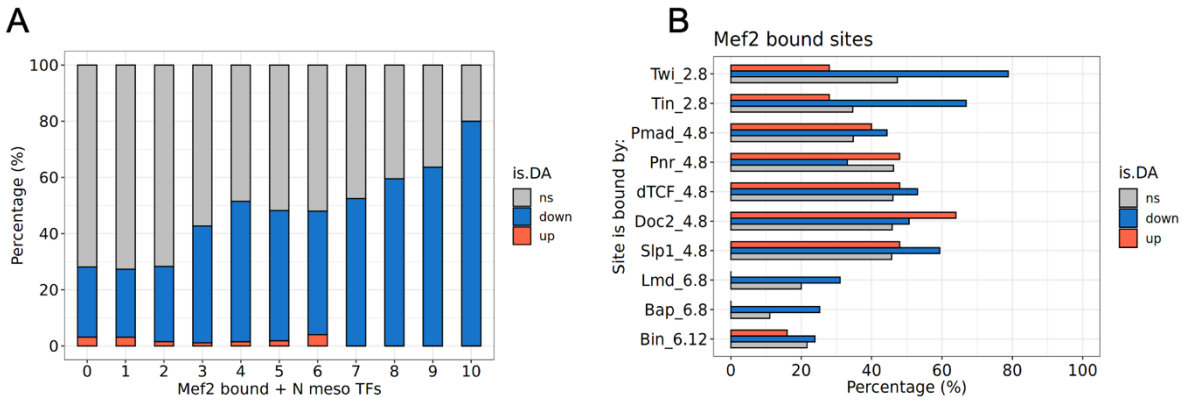


Figure 29: Sites with reduced accessibility are mostly co-bound by multiple mesoderm TFs.

(A) Percentage of reduced, not changing and increased ATAC peaks bound by Mef2 and N other mesoderm TFs. Sites bound by ≥ 3 mesoderm TFs have a higher percentage of reduced peaks.

(B) Percentage of reduced, not changing and increased ATAC peaks co-bound by Mef2 and another mesoderm TF at a specific time point, indicated on the y axis. Reduced peaks are mostly co-bound by Mef2 and Twi or Tin.

3.6.1.7 Affected enhancers and target genes mark the loss of mesodermal identity

To understand the biological role of the altered peaks after Mef2 depletion, I examined regulatory regions and their annotation within DA peaks. The percentage of distal sites, putative enhancers, was high in all peaks, especially in down and up DA peaks: distal peaks account for 64.6% of the down DA peaks in Mut-Lexy-1 and 59.5% in Mut-Lexy-2, for 68.3% of up DA peaks in Mut-Lexy-1 and all 16 up peaks in Mut-Lexy-2, and for ~40% of the non-significant peaks in both (Figure 30A). For Mut-Lexy-1, some peaks represent characterised muscle-specific enhancers and genes and were mostly within the down DA peaks (~50%); for Mut-Lexy-2, they were about 40% in both down DA and non-significant peaks (Figure 30 B, C). Muscle-specific and general muscle putative enhancers from DHS data (Reddington 2020) were enriched in down DA sites and depleted in up DA sites: up DA sites included 50% of enhancers of other tissues in Mut-Lexy-1, and 50% of non-enhancer sites and 30% of other tissue enhancers in Mut-Lexy-2 (Figure 30D). This suggests that sites with increased accessibility are mostly secondary effects in Mut-Lexy-2, and could be linked to alternative cell states in Mut-Lexy-1.

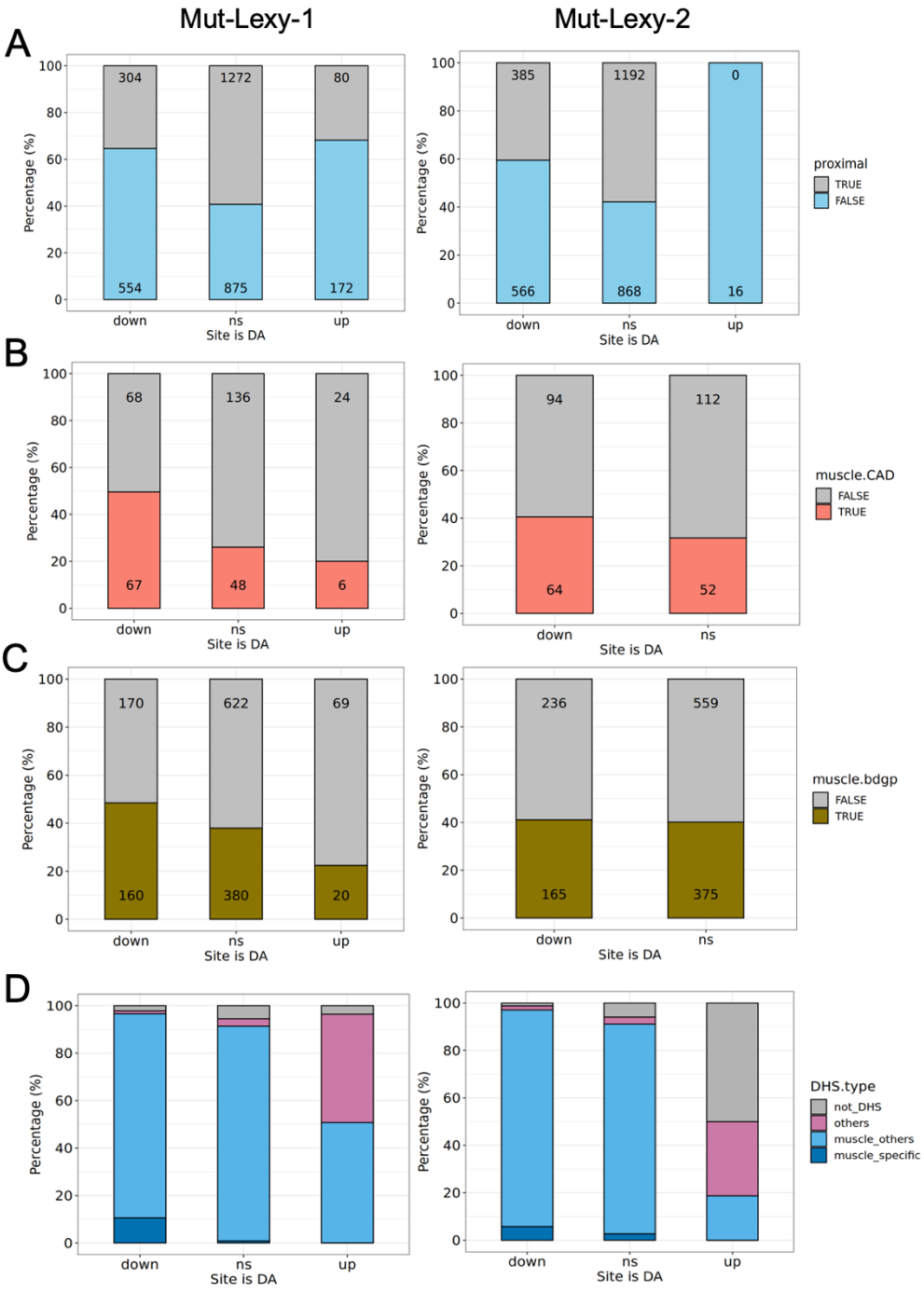


Figure 30: Sites with reduced accessibility are enriched in putative muscle enhancers.

(A) Percentage of proximal (grey) and distal (light blue) sites among reduced, not changing and increased ATAC peaks in Mut-Lexy-1 (left) and Mut-Lexy-2 (right) vs WT. Distal peaks are > 40% in all sets, and more enriched in changing ATAC peaks.

(B) Percentage of sites overlapping with characterized enhancers among reduced, not changing and increased ATAC peaks in Mut-Lexy-1 (left) and Mut-Lexy-2 (right). Reduced ATAC peaks are the most enriched in enhancers.

(C) Percentage of sites overlapping with muscle genes from BDGP enhancers among reduced, not changing and increased ATAC peaks in Mut-Lexy-1 (left) and Mut-Lexy-2 (right). Reduced ATAC peaks are the most enriched in muscle genes.

(D) Percentage of sites overlapping with DHSs among reduced, not changing and increased ATAC peaks in Mut-Lexy-1 (left) and Mut-Lexy-2 (right). Reduced ATAC peaks are the most enriched in DHSs specific for the muscle and shared between the muscle and other tissues, whereas increased peaks have more non-muscle DHSs or non-DHSs.

Changes in cell state are not only reflected in enhancers, but also in the transcriptional output, which has a more direct connection to the phenotype. To identify genes correlated to changes in cell state, I used accessibility on the TSS and gene body as a proxy for transcription, as previously demonstrated in Secchia *et al.* (2022). DA genes (absolute log2FC > 0.5 and FDR-corrected p value < 0.05) were a small proportion of all genes with accessibility above the background: 132 genes out of 2660 (4.96%) were DA for Mut-Lexy-1, 124 down and 8 up; 163 out of 2626 genes (6.21%) were DA for Mut-Lexy-2, 161 down and 2 up (Figure 31A). Overall, almost all DA genes had reduced accessibility, suggesting a blockage of one gene regulation program for muscle differentiation, without active establishment of a new identity. This is confirmed by the GO enrichments, which showed that down DA genes are associated with developmental processes, organ development, gene expression and negative regulation of signalling (Figure 31 B-C). Specifically, Mut-Lexy-2 was marked by the negative regulation of Notch signalling, which is consistent with developmental arrest in the specification of FCs and FCMs in equivalence clusters (Giebel, 1999) (Figure 31C). Instead, Mut-Lexy-1 had a strong component of cell migration, pattern specification and nucleic acid metabolism (Figure 31B), which may hint to a more general imbalance in cell motility and metabolic state. Among down DA genes, we recapitulated known Mef2 targets like *Act57B*, *Act87E* and *zfh1*. The fact that the genes only reflect loss of a mesodermal state, but the enhancers also capture the emergence of a new state in Mut-Lexy-1, shows how chromatin accessibility can be more informative than expression for early identification of cell fate changes, which is especially important for intermediate and short-lived states.

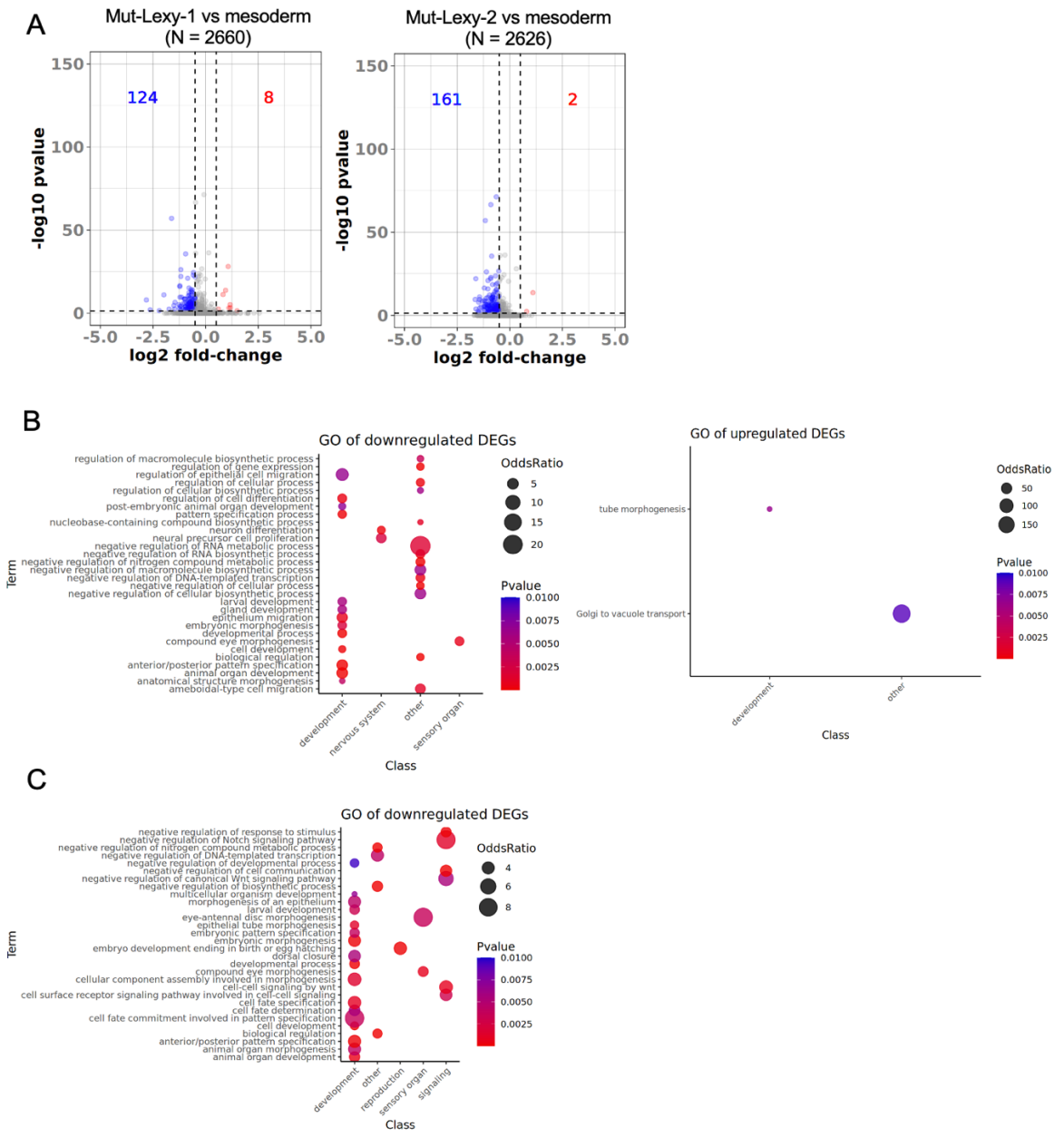


Figure 31: DA genes are mostly downregulated and involved in development.

(A) Volcano plot of DA genes in Mut-Lexy-1 (left) and Mut-Lexy-2 (right) with respect to the closest WT mesoderm cluster (cluster 0). Differential genes have FDR-adjusted p value < 0.05, | log2FC | > 0.5 and signal 5x above background. Both mutant clusters have almost exclusively downregulated genes.

(B) GO of downregulated (left) and upregulated (right) genes in Mut-Lexy-1.

(C) GO of downregulated (left) genes in Mut-Lexy-2.

To further investigate the difference between the two mutant clusters, I compared them directly to extract DA peaks and genes (Figure 32). Among 1375 ATAC peaks, 390 were up DA in cluster Mut-Lexy-1, and 15 in cluster Mut-Lexy-2 (Figure 32A). The up DA peaks in cluster Mut-Lexy-1 only included 24.1% of Mef2-bound peaks, versus 50% in non-significant and down DA peaks (Figure 32C). They had 60% of distal sites (Figure 32D), the lowest Mef2 ChIP signal and they were the least enriched in muscle enhancers and genes annotated in characterised enhancers (Figure 32F), BDGP (Wilczynski *et al.*, 2012) (Figure 32G) and DHSs (Reddington *et al.*, 2020) (Figure 32E). This supports that the

more accessible peaks in cluster Mut-Lexy-1 were mainly indirect targets, in line with the comparison between this cluster and the WT mesoderm. As for DA genes, 13 were more accessible in Mut-Lexy-1 and only 3 in Mut-Lexy-2: among others, the former included two genes which also gained accessibility in the 10–12 h *Mef2* KO mutant in Secchia *et al.* (2022) (*brother of odd* [*bowl*], *Cadherin 99 C* [*Cad99C*]). On the contrary, genes less accessible in Mut-Lexy-1 compared to Mut-Lexy-2 were the Mef2 antagonist *Holes in muscle* (*Him*), the mesoderm-associated markers *cousin of atonal* (*cato*) and *Ventrally-expressed protein D* (*VepD*) (Secchia *et al.*, 2022), which indicate that Mut-Lexy-1 is more distant from the original mesoderm identity.

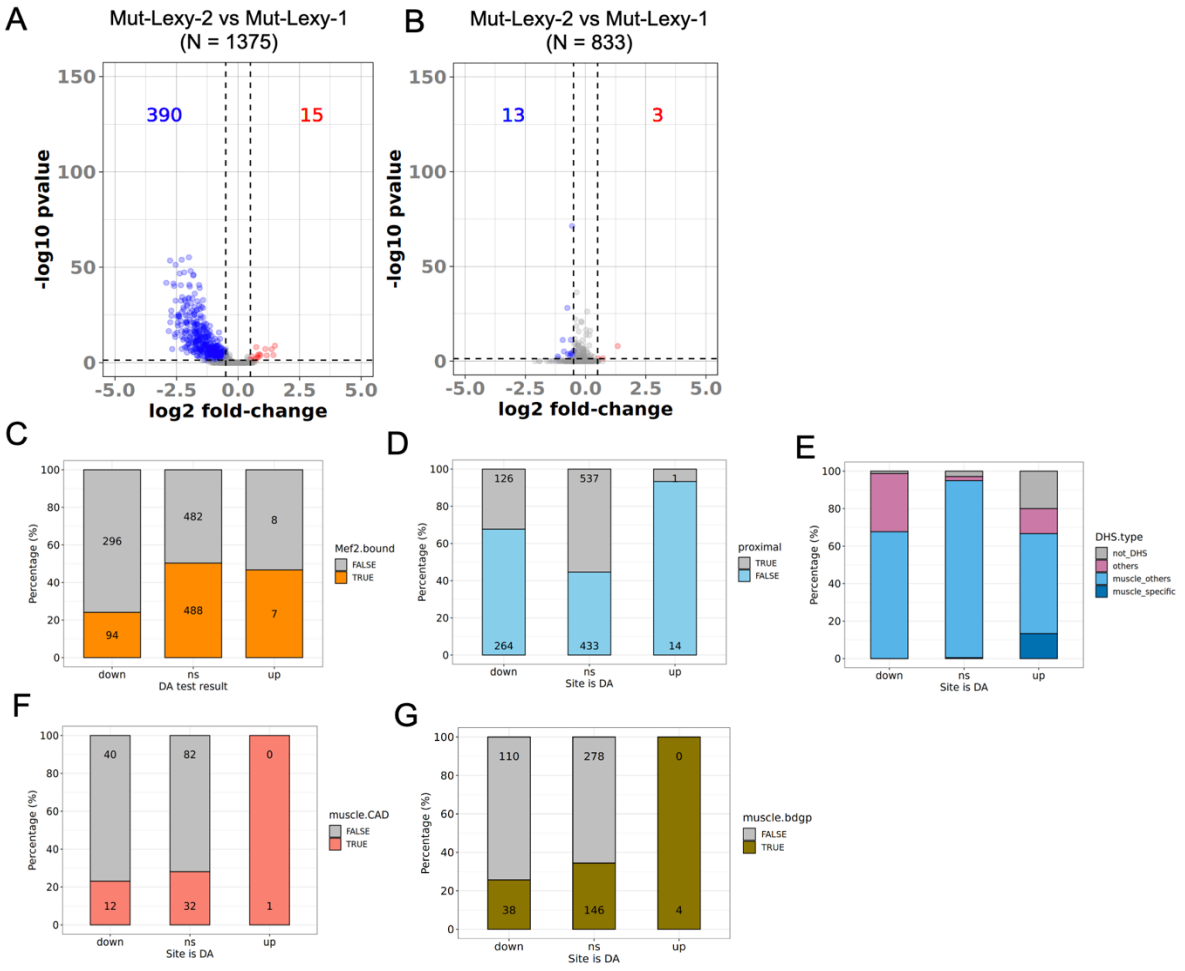


Figure 32: Mut-Lexy-1 has more indirect effects than Mut-Lexy-2.

(A) Volcano plot of DA peaks in Mut-Lexy-2 vs Mut-Lexy-1. Close to 400 peaks have increased accessibility in Mut-Lexy-1.

(B) Volcano plot of DA genes in Mut-Lexy-2 vs Mut-Lexy-1.

(C) Percentage of Mef2-bound peaks in reduced, not changed and increased peaks in Mut-Lexy-2 vs Mut-Lexy-1. Peaks increased in Mut-Lexy-1 (down) have less Mef2-bound sites, pointing to more indirect effects.

(D) Percentage of proximal (grey) and distal (light blue) sites among reduced, not changing and increased ATAC peaks in Mut-Lexy-2 vs Mut-Lexy-1.

(E) Percentage of sites overlapping with DHSs among reduced, not changing and increased ATAC peaks in Mut-Lexy-2 vs Mut-Lexy-1.

(F) Percentage of sites overlapping with annotated enhancers among reduced, not changing and increased ATAC peaks in Mut-Lexy-2 vs Mut-Lexy-1.

(G) Percentage of sites overlapping with muscle genes from BDGP enhancers among reduced, not changing and increased ATAC peaks in Mut-Lexy-2 vs Mut-Lexy-1.

3.6.2 Twi

3.6.2.1 Twi depletion obliterates the mesoderm, redirecting cells to an aberrant epidermis-like state, even after gastrulation

I examined the changes upon Twist depletion after gastrulation by sci-ATAC-seq starting from 240000 nuclei, 70% from *twi-iLEXY* embryos depleted from stage 8 onwards and collected at stage 9–10, and 30% from WT stage 9 embryos. After filtering for read number and doublet removal, I obtained 19559 nuclei, 6238 WT and 13321 *twi-iLEXY*, in line with the input ratio. Whole-embryo clustering on the 13071 peaks called on this dataset showed a marked reduction of *twi-iLEXY* cells in the mesoderm (23.75% of mesoderm cells, Z-score p < 2.2e-16), and a large cluster of cells strongly enriched for the *twi-iLEXY* mutant between the mesodermal and epidermal identities (Figure 33 B-C). This suggests that these cells could shift from a mesodermal to an epidermal-like fate. The other tissues tended to have a similar composition, except for the nervous system and plasmatocytes, which had an increased number of cells in *twi-iLEXY* mutants (89.4% of the total cells in the nervous system cluster with Z-score p < 2.2e-16, and 81.41% with p < 4.086e-16 in the plasmatocyte cluster) (Figure 33C).

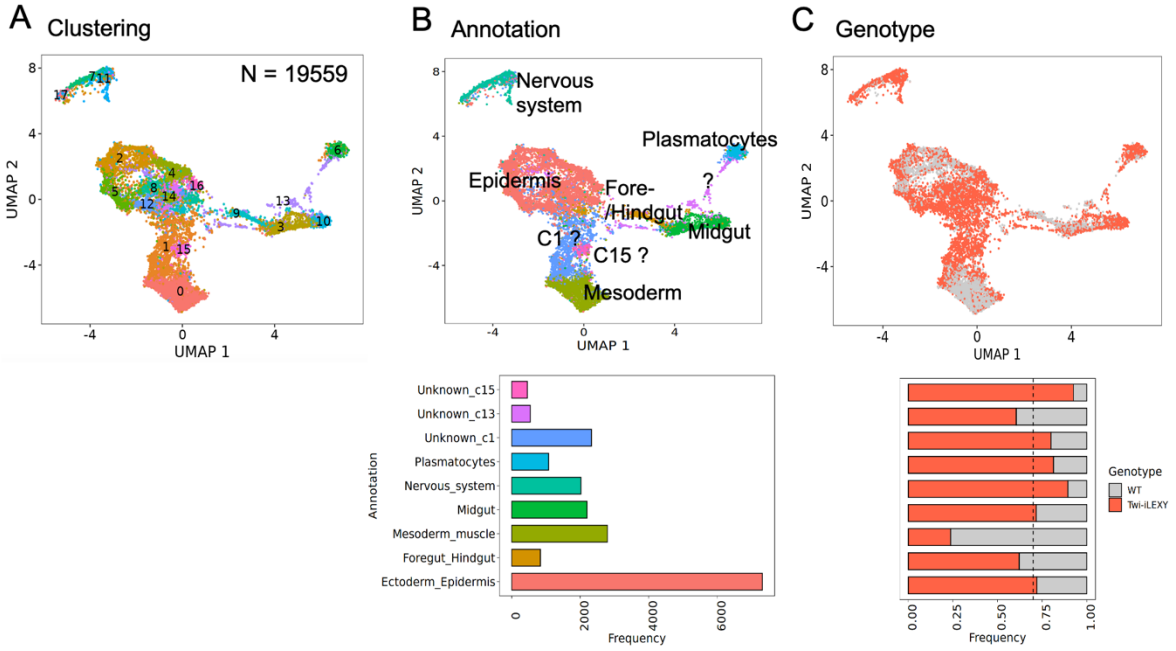


Figure 33:* twi-iLEXY *mutants are depleted in the mesoderm and move towards an epidermal-like identity; they are also enriched in the nervous system and plasmatocytes.

(A) UMAP visualization of the whole embryo Twi-iLEXY dataset on cluster and genotype peaks, with cells coloured by cluster.

(B) Annotation based on BDGP terms enriched among the DA peaks. UMAP and histogram with cell number by annotation.

(C) UMAP as in B, coloured by genotype, and bar chart with genotype frequency by annotation.

To improve resolution, I reclustered the *twi* mutant cluster and the mesoderm, for a total of 7003 cells with 10122 peaks (Figure 34A). In the 'whole embryo context' the main mutant cluster was a continuum between the epidermal and ectodermal identity (Figure 33C). Isolating and reclustering the mesoderm

created a clear gradient between WT and mutants but did not visibly separate the two populations (Figure 34C). As the experiment was done at a single early timepoint (stage 9) when the mesoderm is quite uniform, there may be decreased internal differences, leading to a less complex UMAP representation, where the primordia of many tissues were not yet fully separated from their embryonic origin (Figure 34B). The main mesodermal identity was the trunk mesoderm (2205 cells) where *twi* is expressed, in which a small subpopulation leading to the future VM could already be identified, but not isolated (Figure 34B). The trunk mesoderm is in contiguity with the *twi-iLEXY* mutant cluster, which was fragmented into smaller units assigned to ectodermal and unknown identities (Figure 34B). The only cluster with a clear separation was the plasmatocyte cluster (Figure 34B). Mutant cells were strongly depleted in the trunk mesoderm (22.29% of the total cells in the cluster, Z-score $p < 2.2e-16$) and in the VM primordium (11.14%, Z-score $p < 2.2e-16$), and enriched in the ectoderm-like (82.14%, Z-score $p = 3.631e-09$) and plasmatocyte (80.41%, Z-score $p = 1.674e-13$) clusters (Figure 34C). Accordingly, the accessibility of sites bound by mesoderm TFs at early stages was lower in *twi-iLEXY* mutants than in WT trunk mesoderm (Figure 35E), pointing to Twi depletion from the nucleus impacting the binding of the downstream TFs, Tin and Mef2. The decrease in accessibility was especially prominent at distal enhancer sites.

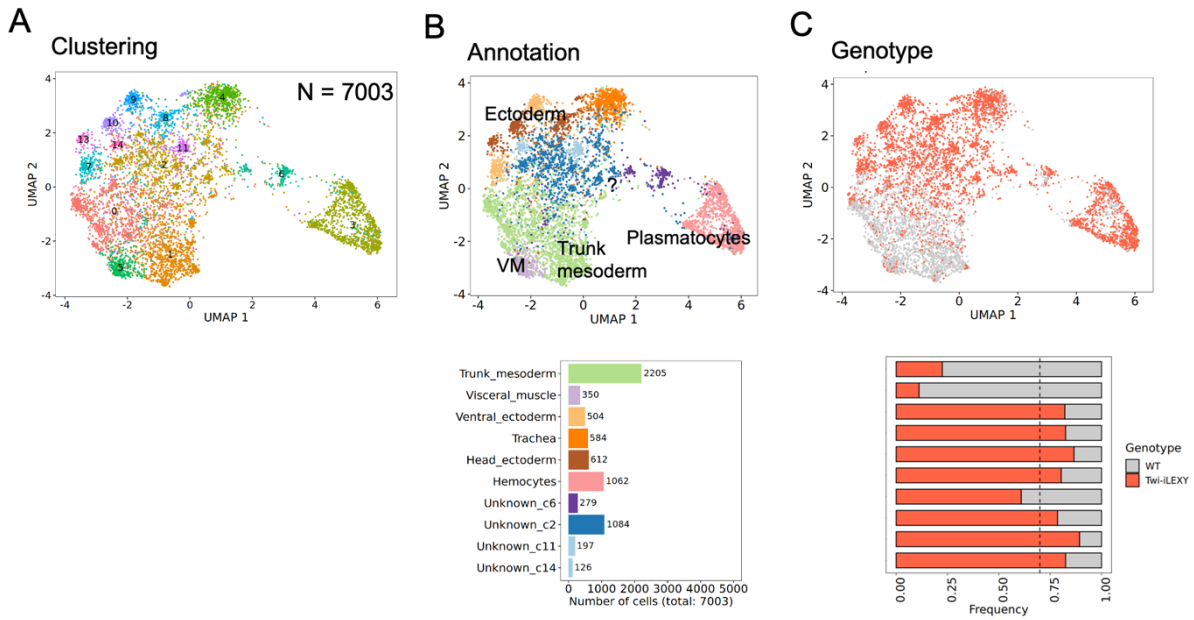


Figure 34: After mesoderm reclustering,* twi-iLEXY *mutants cannot be fully separated from the mesoderm.

(A) UMAP visualization of the mesoderm from the Twi-iLEXY dataset on cluster and genotype peaks, with cells coloured by cluster.

(B) Annotation based on BDGP terms enriched among the DA peaks. UMAP and histogram with cell number by annotation.

(C) UMAP as in B, coloured by genotype, and bar chart with genotype frequency by annotation.

3.6.2.2 Twist mutants form a single cluster separated from the WT mesoderm trajectory

To further examine the differences between *twi-iLEXY* mutants and the WT mesoderm, I isolated the *twi-iLEXY* mutants between an epidermal and a mesodermal fate, and the WT mesoderm, and I integrated them with the Secchia *et al.* (2022) WT mesoderm time course (Figure 35A). In this way, I could increase the number of data points for the WT and have a potentially clearer separation of the two populations. For integration, I generated a new cell-per-peak count matrix with 28499 cells and 50489 merged peaks from the two datasets, and performed LDA dimensionality reduction in cisTopic (Bravo González-Blas *et al.*, 2019) and batch correction with MNN (Haghverdi et al., 2018), using the WT from each dataset as reference (Figure 35A).

Despite the absence of information on WT stage, stage 9 WT from my dataset were correctly aligned on stage 9 WT from the Secchia *et al.* (2022) dataset (Figure 35 C-D). *twi-iLEXY* mutants were clearly separated from the WT trajectory and mostly grouped into a single cluster closest to the WT of the same timepoint (clusters 5, 6, 8) (Figure 35 B-C). The fact that mutant cells are reminiscent of their age agrees with the observations for *Mef2-iLEXY* mutants. A minority of *twi-iLEXY* mutants clustered with the plasmatocytes located close to the fat body, at a later timepoint (Figure 35 C-D). The main *twi-iLEXY* cluster included all non-plasmatocyte mutant cells from the mesoderm (Figure 35). The accessibility profile at TF motifs and binding sites confirmed the cell identity assignment; the main mutant cluster was closest to the mesoderm identity of the same age, but had higher accessibility for Zld and Trl and lower accessibility for mesoderm TFs Twi and Tin; the plasmatocyte cluster was most similar to the fat body (Figure 35E). The similarity of mutant cluster 1 to the trunk mesoderm anlage was confirmed by Pearson correlation between annotations (Figure 35F), further indicating that all mutant cells were more similar to each other than to any other identity.

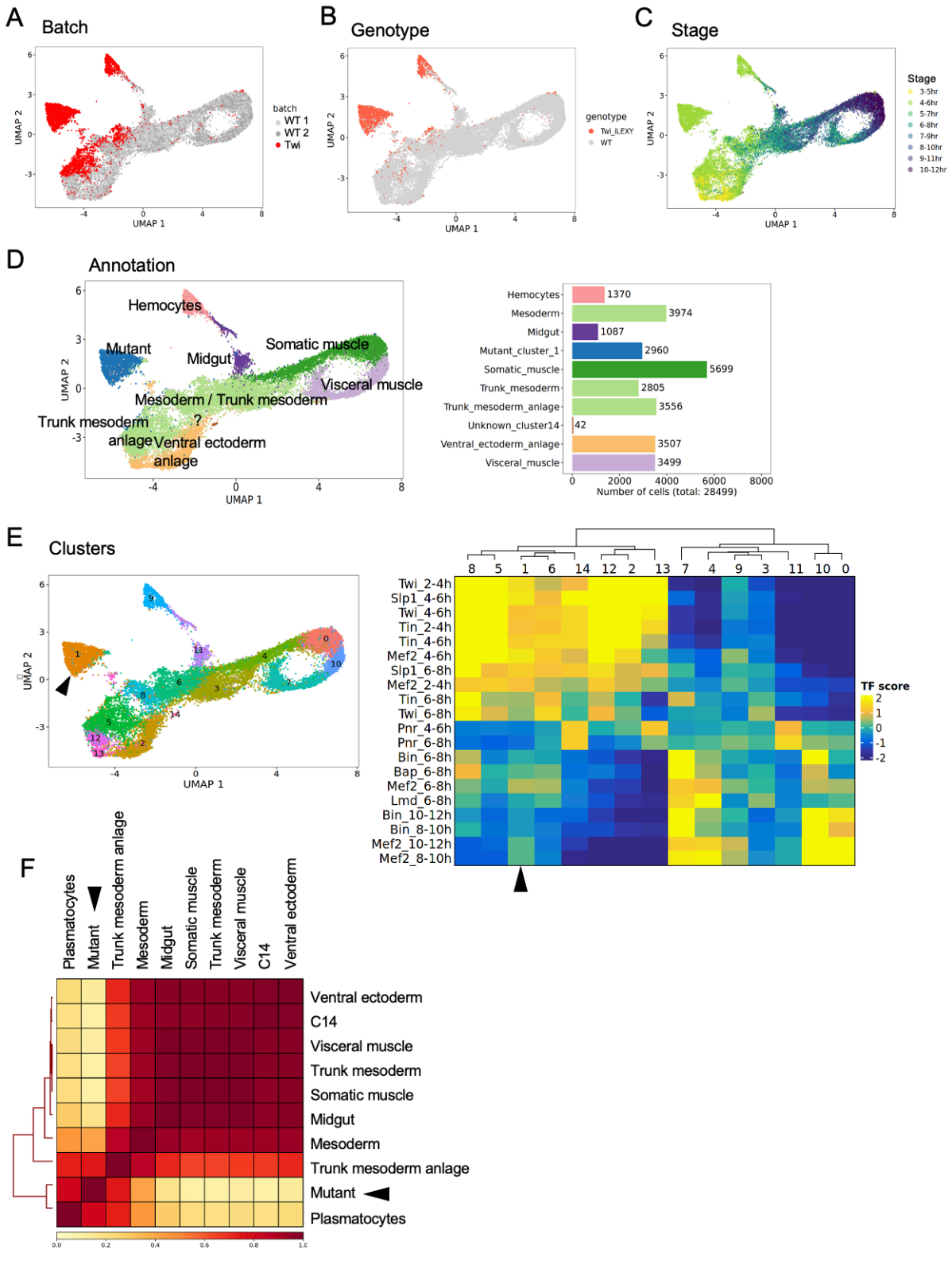


Figure 35: Comparison of the Twi-iLEXY dataset to the WT dataset from Secchia* et al. *(2022).

(A) UMAP coloured by dataset: WT dataset in grey, twi-iLEXY dataset in red.

(B) UMAP showing the position of *twi-iLEXY* mutant cells (in red).

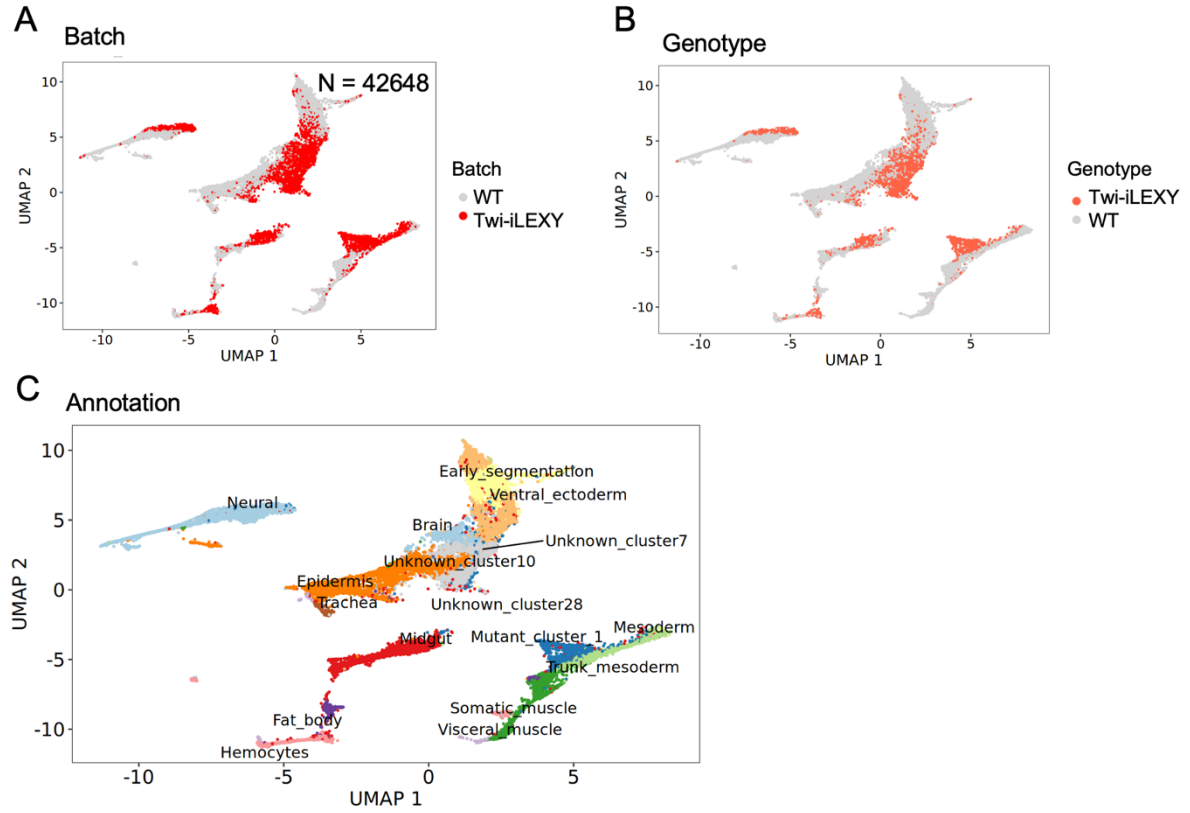
(C) UMAP coloured by stage.

(D) UMAP coloured by the original annotation of the datasets, and histogram of the number of cells per annotation.

(E) UMAP coloured by cluster. Heatmap of accessibility deviation scores for selected mesoderm TFs based on ChIP data at different timepoints, for each cluster. Cluster 1, which contains most of the *twi-iLEXY* mutants, is indicated by a black arrow; it is closest to the WT clusters of the same timepoint (cluster 6, 5, 8).

(F) Heatmap with Pearson correlation by annotation. The mutant cluster, indicated by a black arrow, is closest to the WT trunk mesoderm anlage and to the plasmatocytes, which are enriched in *twi-iLEXY* mutants.

The integration with the WT dataset from Secchia *et al.* (2022) showed that *twi-iLEXY* cells are clearly separated from the mesoderm, but closest to the trunk mesoderm identity of the same stage (Figure 35 C-D). However, in the whole-embryo clustering, those cells were intermediate between ectoderm and mesoderm (Figure 33B). To resolve the closest identity, I integrated the whole-embryo data from the Twi-iLEXY dataset with the Cusanovich *et al.* (2018) WT whole-embryo time-course at 2–4, 4–6, 6–8 and 10–12 h by LDA-based clustering on the 47191 merged peaks and MNN batch correction using the WT as reference (Figure 36A). The WT nuclei were assigned to the correct matching tissue and stage (Figure 36 A, C), while *twi-iLEXY* mutants occupied aberrant trajectories split between the mesoderm (cluster 5) and ectoderm (clusters 7, 10, 28) (Figure 36 B-D). The accessibility profile of mesoderm TF binding sites showed that the mesoderm-associated cluster 5 was more similar to epidermis-associated clusters 7 and 10 than to the WT trunk mesoderm (clusters 6, 11, 14); this was especially evident for distal sites (Figure 36 E). This result emphasizes the continuum in mutant identity, where most cells were rerouted from the mesodermal fate to an aberrant ectoderm-like state.



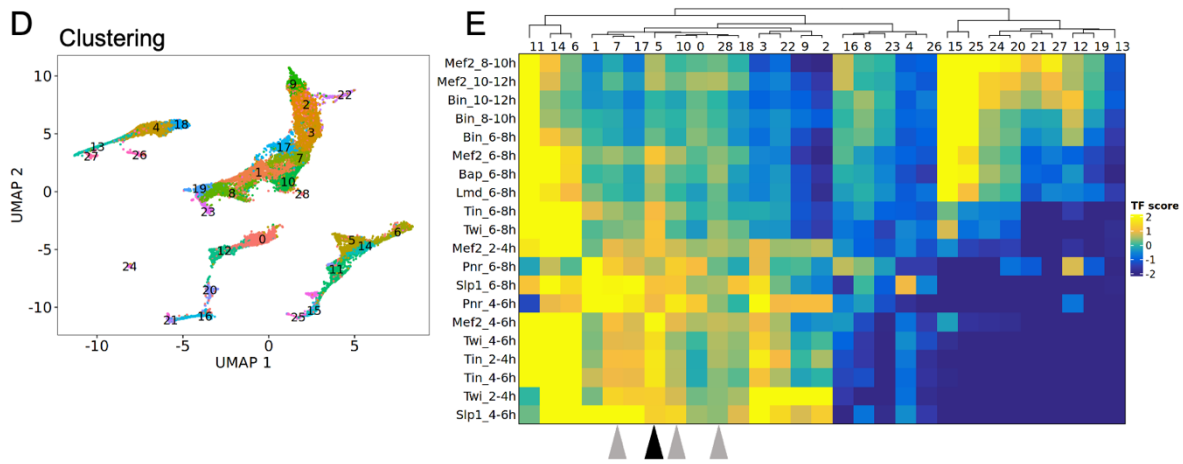


Figure 36: Comparison of the Twi-iLEXY dataset to the WT dataset from Cusanovich* et al. *(2018).

(A) UMAP coloured by batch: in grey, the WT embryos from Cusanovich *et al.* (2018); in red, the embryos (mutant and WT) from the Twi-iLEXY dataset.

(B) UMAP as in A, coloured by genotype.

(C) UMAP as in A, coloured by the original annotation of the datasets. *twi-iLEXY* mutants are mostly close to the mesoderm (Mutant cluster 1, in blue), but some epidermal-like clusters are associated to the epidermis and brain (Unknown clusters 7, 10, 28).

(D) UMAP coloured by cluster on the integrated datasets.

(E) Heatmap of accessibility deviation scores for selected mesoderm TFs based on ChIP data at different timepoints, for each cluster. The *twi-iLEXY* mutant clusters 7, 10 and 28, which are closest to the WT epidermis, are indicated by grey arrows; cluster 5, which is close to the WT mesoderm, is indicated by a black arrow. Cluster 5 is more similar to epidermal than to mesodermal clusters.

3.6.2.3 Loss of accessibility is prominent at Twi-bound sites

To understand which sites are correlated to the change in cell state in the mutant, I performed DA analysis on ATAC peaks. Since the *twi-iLEXY* mutants and the WT trunk mesoderm were not clearly separated in mesoderm reclustering, but there was splitting in the integrated datasets, I based the analysis of DA peaks on the genotype rather than on the clusters, by considering all *twi-iLEXY* mutants in the mesoderm-ectoderm cluster and the WT trunk mesoderm.

Of the 5911 merged genotype-specific peaks, most (5517) were not affected; 205 had significantly lower accessibility (log2FC < -0.5, FDR-corrected p value < 0.05), and 189 significantly higher accessibility (log2FC > 0.5, FDR-corrected p value < 0.05) (Figure 37A). The small difference in numbers between DA sites with a loss and a gain in accessibility may be due to indirect effects of Twi depletion; this would be consistent with the acquisition of a more ectoderm-like identity. To test whether the gain of accessibility occurs at non-Twi sites, I examined Twi ChIP binding at 4–6 h under DA peaks. Twi was bound at 50% of down DA peaks, but only at 16% of non-significant peaks, and at even less (12.7%) of up DA peaks (Figure 37B), supporting that the accessibility gain is not directly mediated by Twi. This is consistent with the expectation that Twi depletion from chromatin causes a loss in accessibility. Indeed, the decrease in Twi signal in CUT&Tag upon depletion is very prominent at down DA peaks, but not at non-significant or up DA peaks (Figure 37C).

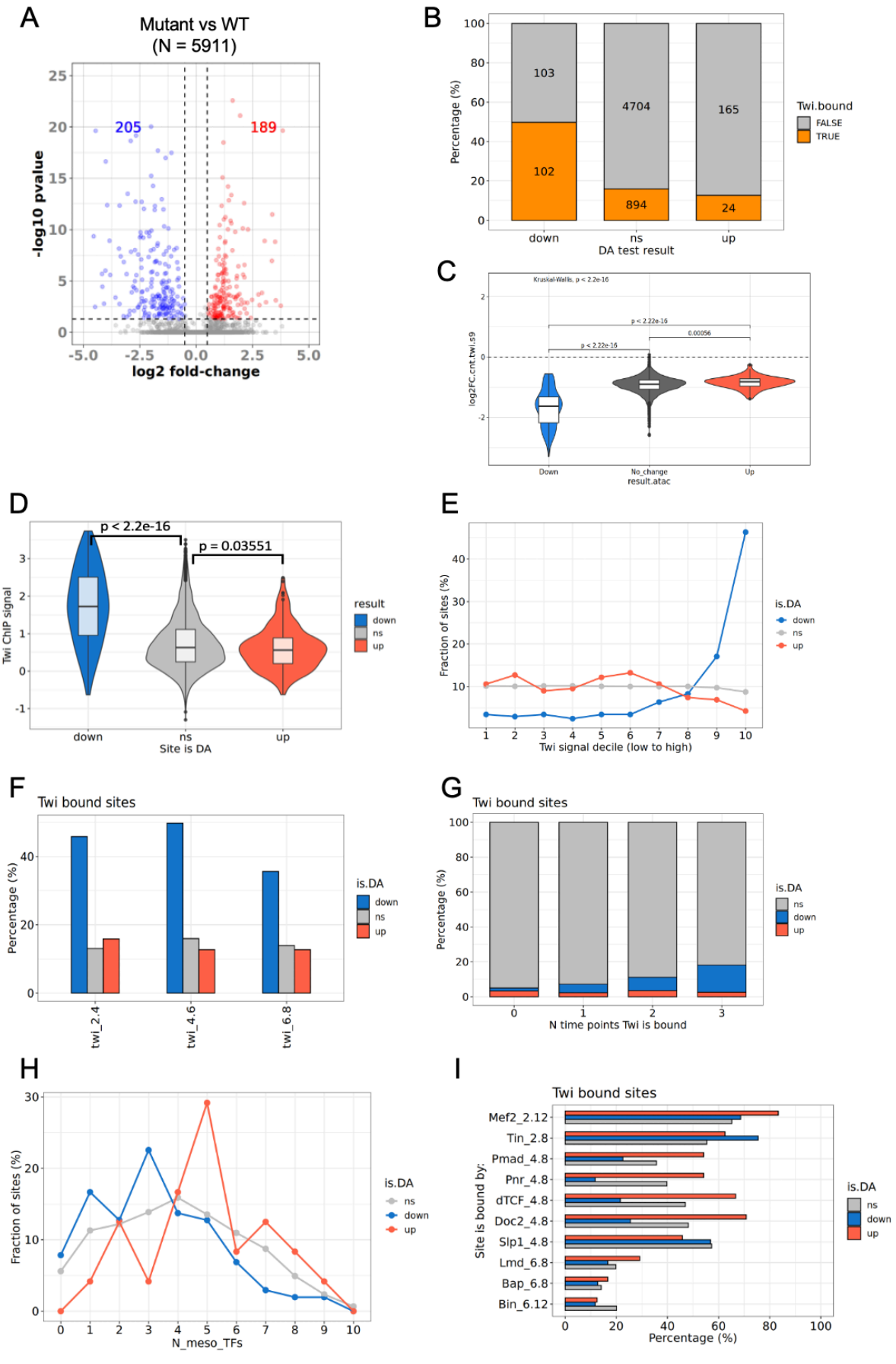

A
Mutant vs WT
(N = 5911)
205
189
-log10 pvalue
log2 fold-change
B
103
4704
165
102
894
24
Percentage (%)
down
ns
up
DA test result
Twi.bound
FALSE
TRUE
C
Kruskal-Wallis, p < 2.2e-16
p < 2.22e-16
0.00056
p < 2.22e-16
log2FC_cnt.twi.s9
Down
No_change
Up
result.atac
D
p < 2.2e-16
p = 0.03551
Twi ChIP signal
down
ns
up
Site is DA
result
E
Fraction of sites (%)
Twi signal decile (low to high)
is.DA
F
Twi bound sites
Percentage (%)
twi_2.4
twi_4.6
twi_6.8
is.DA
G
Twi bound sites
Percentage (%)
N time points Twi is bound
is.DA
H
Fraction of sites (%)
N_meso_TFs
is.DA
I
Twi bound sites
Site is bound by:
Mef2_2.12
Tin_2.8
Pmad_4.8
Pnr_4.8
dTCF_4.8
Doc2_4.8
Slp1_4.8
Lmd_6.8
Bap_6.8
Bin_6.12
Percentage (%)
is.DA

Figure 37:* twi-iLEXY *mutants have similar number of peaks with increased and decreased accessibility, and the down DA peaks are associated with strong Twi binding sites.

(A) Volcano plot of ATAC peaks in the *twi-iLEXY* mutant vs the WT. Differential peaks have FDR-adjusted p value < 0.05, |log2FC| > 0.5 and signal 10x above background.
(B) Percentage of Twi-bound peaks in reduced, not changed and increased peaks.
(C) Log2FC from CUT&Tag in reduced (blue), not changed (grey) and increased (red) ATAC peaks in the *twi-iLEXY* mutants. Reduced ATAC peaks have the highest Twi reduction upon light exposure in CUT&Tag.
(D) Twi ChIP signal at stage 9 in reduced, not changed and increased ATAC peaks. Reduced ATAC peaks have the highest ChIP signal.
(E) Fraction of total reduced, not changed and increased ATAC peaks by Twi ChIP signal decile. Close to 50% of the decreased ATAC peaks is in the top 10% Twi ChIP peaks.
(F) Percentage of total reduced, not changed and increased ATAC peaks bound by Twi at different timepoints (2-4, 4-6, 6-8 h). Most of the down peaks are bound at 4-6 h, according to the chosen depletion stage.
(G) Percentage of total reduced, not changed and increased ATAC peaks bound by Twi for N timepoints. The number of down sites increases with the number of timepoints, suggesting that continuously bound peaks are the most affected.
(H) Fraction of DA sites bound by 0 to 10 mesoderm TFs. Most down sites are bound by 3 TFs on average, and most up sites by 5.
(I) Percentage of reduced, not changed or increased ATAC peaks co-bound by Twi and a mesoderm TF indicated on the y axis. Co-bound peaks mostly have increased accessibility.

Down DA sites also corresponded to regions strongly bound by Twi, in contrast to non-significant and up DA sites (Figure 37D); Twi ChIP signal in the top 10% accounted for > 40% of the down DA sites, and sites below the 7th signal decile were very scarce within down DA sites (Figure 37E). Non-significant sites had a uniform distribution regardless of the Twi signal, and up DA sites tended to decrease in the highest Twi signal decile (Figure 37E). The percentage of down DA sites bound by Twi was highest when compared to ChIP peaks from the matching time point (4–6 h) (Figure 37F); on the other hand, peaks with Twi occupancy at multiple time points had a higher chance of being down DA (Figure 37G). The peaks present at multiple stages could have higher Twi affinity to maintain binding in a broader Twi concentration range, and may not require additional partners to further stabilize Twi binding. To test if the down DA peaks were indeed mostly Twi-only peaks, I examined the fraction of down, up and non-DA sites with binding of a number of other mesoderm TFs beside Twi: most down DA sites had up to three mesoderm TFs co-bound with Twi, and Twi-exclusive sites never showed increased accessibility (Figure 37H). On the contrary, up DA sites had on average five additional mesoderm TFs binding (Figure 37H). With respect to the identity of co-bound TFs, co-binding of Tin or Sloppy 1 (Slp1) was associated to down DA sites, whereas Dorsocross 2 (Doc2), Pnr, Mothers against dpp (Pmad), pangolin (dTCF, pan) and Lame duck (Lmd) were mostly present at up DA sites (Figure 37I). This was further confirmed by motif enrichment at DA sites: the Tin motif was present among those enriched in down DA sites alongside Twi, and the dTCF motif was enriched at up DA sites (not shown). Zld motif was also enriched at up DA sites, suggesting reactivation of a less specified cell state.

Co-binding at down DA sites was consistent with the known roles of mesoderm TFs. Slp, enriched at down DA sites, controls the segmental allocation of mesodermal fate, being necessary for the establishment of high-Twi domains which will differentiate into SM, and to prevent expansion of low-Twist domains, which will give rise to the heart, VM and fat body (Riechmann *et al.*, 1997). The reduced accessibility at Twi and Slp co-bound sites (Figure 37I) is consistent with the loss of SM cells. The

additional loss of Tin, expressed in heart and early VM, suggests that cells could be mostly directed towards fat body and plasmatocyte fate.

Up DA sites may be bound by repressors not assessed here, or have increased binding of other TFs but insufficient transcriptional activation of target genes. This could be the case for Lmd, which requires Twi co-binding to activate individual FCM enhancers (Busser *et al.*, 2012), and has extensive co-binding with Twi target Mef2 at muscle genes (Cunha *et al.*, 2010). Other cases involve TFs which act as activators or repressors in different contexts. In the absence of Armadillo and in the presence of Groucho, dTCF represses *wingless* (*wg*) (Cavallo *et al.*, 1998), which activates *slp* (Lee and Frasch, 2000); this results in SM loss. On the contrary, it is a transcriptional activator in the presence of Armadillo, but not Groucho (Song *et al.*, 2019). dTCF-mediated repression would be compatible with the SM loss in *twi* KO and *twi-iLEXY* mutants.

Overall, Twi is necessary for the maintenance of accessibility in a subset of its binding sites, even after gastrulation, facilitating binding of its downstream TF, Tin. The pivotal role of Twi in mesoderm development may involve actively counteracting alternative fates which could emerge upon Twi depletion through the action of other TFs.

To support the potential importance of DA sites in cell fate decisions, I assessed their enrichment in regulatory regions. Regulation generally occurs at distal sites, and 75.1% of *twi* down DA were distal, compared to 43.6% of non-significant and 37.6% of up DA sites (Figure 38A). Down sites also had the highest overlap with muscle enhancers from our database of characterised enhancers (Figure 38B) and muscle-specific gene expression in the BDGP database (Wilczynski *et al.*, 2012) (Figure 38C) (83.7% against < 30% in the enhancers, and 60.6% against < 40% in BDGP). Muscle-specific DHSs from Reddington *et al.* (2020) were also enriched in down DA but not in non-significant or up DA sites (Figure 38D). This confirms that accessibility is lost mainly at muscle-specific enhancers.

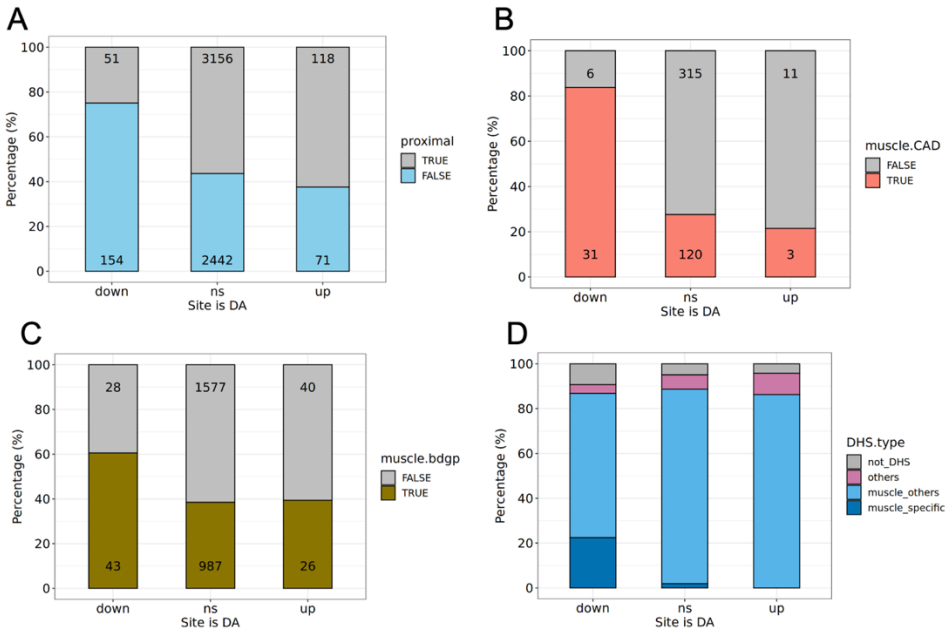


Figure 38: Sites with lower accessibility are enriched in putative muscle enhancers.

(A) Percentage of proximal (grey) and distal (light blue) sites among reduced, not changing and increased ATAC peaks in *twi-iLEXY* mutants vs WT. Distal peaks are ~40% in non-changing and up peaks, and more enriched (~80%) in down ATAC peaks.
(B) Percentage of sites overlapping with enhancers among reduced, not changing and increased ATAC peaks in *twi-iLEXY* mutants. Reduced ATAC peaks are by far the most enriched in characterized enhancers.
(C) Percentage of sites overlapping with muscle genes from BDGP enhancers among reduced, not changing and increased ATAC peaks in *twi-iLEXY* mutants. Reduced ATAC peaks are the most enriched in muscle genes.
(D) Percentage of sites overlapping with DHSs among reduced, not changing and increased ATAC peaks in *twi-iLEXY* mutants. All sets have > 80% DHSs for muscle and other tissues, and reduced ATAC peaks are the most enriched in muscle-specific DHSs.

3.6.2.4 Twist depletion has many indirect effects, especially at the level of differential genes

To investigate the transcriptional changes occurring in the *twi-iLEXY* mutant, I examined accessibility on the TSS and gene body. Among the 1584 genes with accessibility levels above background, 125 had a significant gain in accessibility in *twi-iLEXY* mutants compared to the WT (log2FC > 0.5, FDR-corrected p value < 0.05), and 32 genes had reduced accessibility (log2FC > 0.5, FDR-corrected p value < 0.05) (Figure 39A). This points to an overall gain of transcription, potentially by activation of non-mesodermal transcription programs, as suggested by the nature of the DA enhancers. To infer the pathways involving down and up regulated genes, I performed a GO analysis for biological processes. Down DA genes were mostly associated to mesoderm and muscle development and to nucleosome organization and protein-DNA interaction, whereas up DA genes were mostly related to signal transduction, especially Notch signalling and neurogenesis (Figure 39B). The enrichment for neurogenesis terms is in accordance with the increase in *twi-iLEXY* mutants in the nervous system (Figure 33 B-C).

Down DA genes included 13 genes in the vicinity of Twi ChIP peaks from Sandmann *et al.* (2007), together with the classic Twi targets *tin* and *Mef2*, and the direct targets *retained* (*rtn*), *CG12177* and *CG8788*, which require Twi for their expression (Sandmann *et al.*, 2007).

Up DA genes showed a moderate percentage of Twi targets (31.2%, 39/125), in addition to many neural genes: *cactus* (*cac*), *SoxNeuro* (*SoxN*), *glial cells missing* (*gcm*), *deadpan* (*dpn*), *empty spiracle* (*ems*), *naked cuticle* (*nkd*) and others. Remarkably, *twi-iLEXY* mutants were enriched in the nervous system, suggesting that in the absence of Twi, cells could acquire an identity closer to the neuronal fate. Different cells could be fully reprogrammed towards a physiological neuronal fate, ending up in the neuronal cluster, or be only partially altered, leading to a mutant identity. A similar phenomenon may occur for the plasmatocyte cluster, as suggested by the upregulation of *serpent* (*ser*) and *gcm*, which contribute to the development of fat body and plasmatocytes (Rehorn *et al.*, 1996; Sam, Leise and Hoshizaki, 1996; Bernardoni, Vivancos and Giangrande, 1997). Ser is sufficient to transform the trunk mesoderm into plasmatocytes (Spahn *et al.*, 2014), and if expressed in the early mesoderm it prevents formation of other mesoderm derivatives besides the fat body and plasmatocytes. *gcm* is downstream of Ser (Lebestky *et al.*, 2000). *ser* expression is activated by Sna and Buttonhead (Btd), further enhanced by Ems and repressed by Eve (Spahn *et al.*, 2014); notably, it is independent of Twi, since Ser is not affected in *twi* KO mutants (Spahn *et al.*, 2014). Since in the *twi-iLEXY* line Twi could still promote

early transcription of *sna* (Leptin, 1991), I hypothesize that Sna might be active in a larger domain during Twi depletion, inducing more widespread *ser* expression.

Notably, up DA genes included genes involved in myoblast fusion (*myoblast city* [*mbc*] and *schizo* [*siz*]) or expressed in FCMs in the SM and VM (*goliath* [*gol*]). High Twist levels define FCs, promote SM fate and reprogram the VM to a SM-like syncytium, while disrupting the heart (Baylies and Bate, 1996); low Twi levels, on the contrary, are associated to VM specification. Based on my results, Twi depletion could activate genes involved in FCM fate, potentially by releasing Twi-mediated inhibition, and it could reduce the SM in favour of the VM and plasmatocytes.

Finally, the apoptotic genes *scylla* (*scyl*) and *charybde* (*chrb*) were also upregulated. This may be reminiscent of Twi's anti-apoptotic role in mouse (Maestro *et al.*, 1999; Jen *et al.*, 2020). In *Drosophila*, *twi* ectopic expression promotes apoptosis by activating the expression of the *reaper* (*rpr*) gene (Gullaud, Delanoue and Silber, 2003). However, in my data *reaper* is not DA in the *twi-iLEXY* mutant, suggesting that the upregulation of apoptotic gene is an indirect response to cell-wide dysregulation.

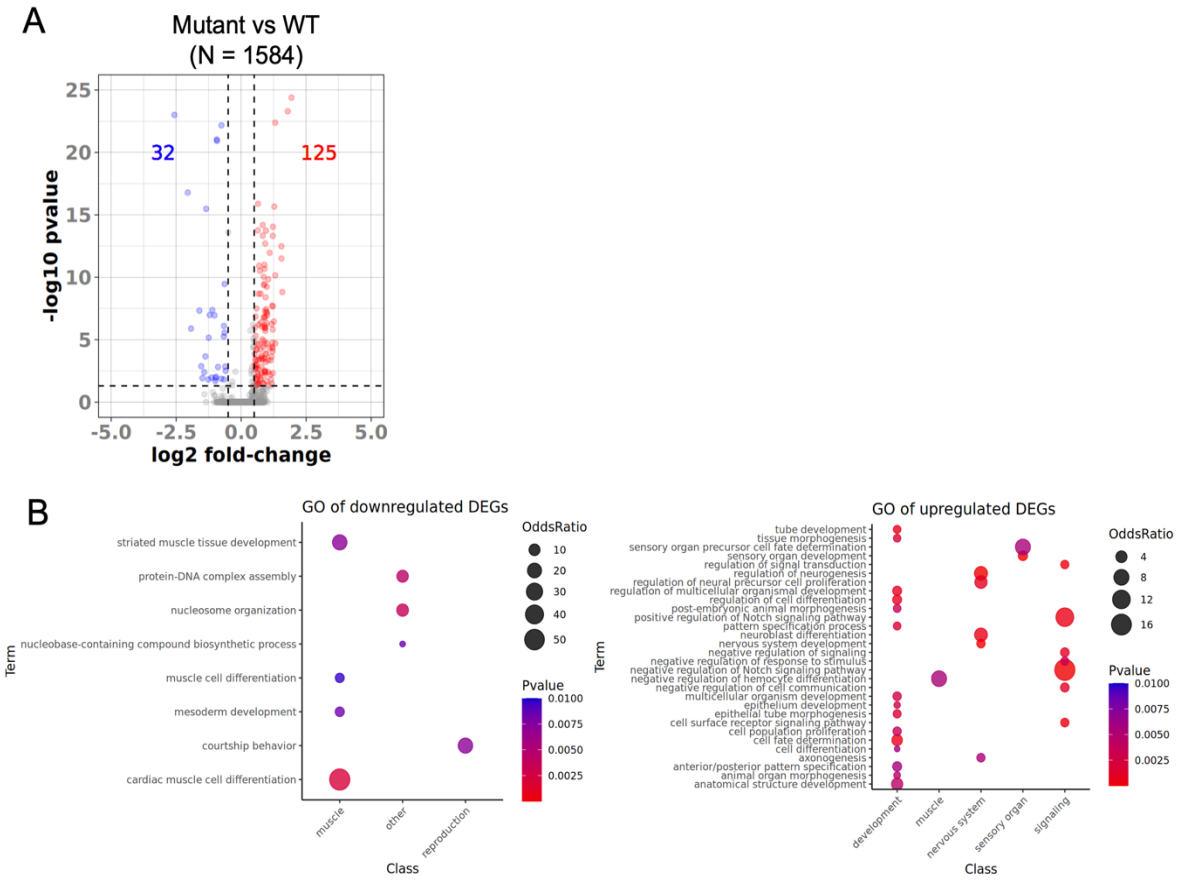


Figure 39: DA genes are mostly upregulated and enriched for general development and nervous system terms.

(A) Volcano plot of DA genes in the *twi-iLEXY* mutant vs the WT mesoderm. Differential genes have FDR-adjusted p value < 0.05, | log2FC | > 0.5 and signal 5x above background. DA genes are mostly upregulated.

(B) GO of downregulated (left) and upregulated (right) genes in *twi-iLEXY*. Downregulated genes are related to muscle terms as expected, whereas upregulated genes are enriched for broader developmental terms and signalling.

3.6.3 Relationship between changes in TF binding, accessibility and transcription

To address the role of single TFs (in this case either Mef2 or Twist) in the accessibility and activity of regulatory regions, I examined changes in accessibility at peaks depleted in bulk CUT&Tag (section 3.4, Figure 18), and their link to DA genes (sections 3.6.1.7 and 3.6.2.4 , Figure 31, Figure 39). First, I assessed how a single TF contributes to co-binding of other TFs by linking depletion to accessibility. Then, by comparing accessibility at enhancers and in the gene body, I derived some preliminary data on the enhancer state's impact on transcription.

Among all potential combinations of changes in TF binding, accessibility and transcription, I focused on four main cases linked to known TF behaviour at enhancers:

1. Decreased binding, accessibility and transcription (DDD) at direct TF targets. If the site were bound only by the depleted TF, the gene would be a direct target; if the site is bound by multiple TFs, it would be compatible with cooperative interactions, whereby TF depletion impairs binding of other TFs, or with additive interactions whereby the depleted TF contributes the most to accessibility and gene expression

2. Decreased binding, no change in accessibility and decreased transcription (DND). This may happen if other TFs could still bind to the site, but were unable to activate transcription to the same extent. The set of enhancer-bound TFs upon depletion may differ from the WT

3. Decreased binding, increased accessibility and no change in transcription (DIN). Here, the depleted TF could compete with other TFs, which could bind more effectively upon its removal; the depleted TF may either be dispensable for transcription, or compensated by the newly bound TFs

4. Decreased binding, increased accessibility and increased transcription (DII). As in 3., this may be due to competition between the depleted and other co-bound TFs, which would increase transcriptional output

Genes with multiple enhancers could be classified into multiple categories, according to the accessibility at each enhancer peak.

3.6.3.1 Mef2

For Mef2-iLEXY, 881 ATAC peaks were depleted in bulk CUT&Tag at stage 9, 11 or both for Mut-Lexy-1; among them, 317 peaks (35.98%) had decreased accessibility in sci-ATAC-seq, and 54 (6.13%) increased accessibility. For Mut-Lexy-2, 829 ATAC peaks were depleted in CUT&Tag. Of those, 367 (44.27%) had reduced accessibility, and one had increased accessibility.

Mut-Lexy-1 and Mut-Lexy-2 had a high overlap and a similar signature in DDD and DND genes. There were 25 DDD genes in Mut-Lexy-1 and 29 in Mut-Lexy-2, 19 of them in common, including *Act57B*, *Act87E*, *lmd*, *zfh1*. They are associated to development in general, and especially muscle development

and (in Mut-Lexy-2) myoblast fate specification, which were indeed hindered in the *Mef2* mutant. DDD genes were also enriched in terms related to tissue polarity migration and cell motility, consistent with defects in mesoderm morphogenesis and cytoskeleton reorganization; moreover, such genes were linked to cell proliferation and nucleic acid metabolism, suggesting that cells may enter a quiescent state. Similarly, DND genes (19 in Mut-Lexy-1 and 28 in Mut-Lexy-2; 12 in common) are involved in tissue and muscle development, and in proliferation.

Only Mut-Lexy-1 had multiple peaks with increased accessibility; among them, *brother of odd with entrails limited* (*bowl*) was transcriptionally upregulated, and 15 genes involved in head and neuronal development showed no change.

3.6.3.2 Twi

twi-iLEXY had 2708 ATAC peaks depleted in CUT&Tag at stage 9; among them, 108 were less accessible (3.99%) and 117 (4.32%) more accessible.

The 12 DDD genes included the known Twi targets *tin* and *Mef2*, and they had a strong mesoderm and muscle development signature, including myoblast fusion. Heterochromatin formation and negative regulation of transcription were also enriched in this set. Only two genes were exclusively classified as DND, *CG8281* and *Act87E.*

Among genes with reduced Twi binding and increased accessibility, there were 16 DIN and 20 DII genes, with clear differences in the involved biological processes. DIN genes are involved in development, proliferation and cell-cell communication, whereas DII genes have a strong neuronal signature: they are involved in CNS development, glial cell fate determination, regulation of neurogenesis (*SoxN*, *dpn*, *Krüppel* [*Kr]*, *nemo* [*nmo*], *Enhancer of split* [*E(spl)*]). Epidermal growth factor (Egf) and Wnt signalling, in particular repression of Notch, were also highly represented, further supporting a neuronal fate. *Rhomboid* (*rho*) and *Egf receptor (Egfr)* activate neuronal differentiation (Yasugi 2010), and *Egfr* is involved in glia formation and maintenance along with *gcm* (also DII); moreover, Notch downregulation promotes neuroblast specification at stage 10 (Skeath and Thor, 2003). Despite not being mesodermal, these genes are associated to regions bound by Twi in the dark and with reduced binding upon exposure to blue light; therefore, they are likely to be direct targets. Overall, this result supports the hypothesis that Twi acts as a repressor of neuronal identity at stage 9. The upregulation of *gcm*, which promotes plasmatocyte as well as glial fate, suggests that Twi could also repress plasmatocyte identity. Twi repression of the neuronal and plasmatocyte fate is in line with the higher proportion of *twi-iLEXY* mutants having these cell types (Figure 33C); thus, cells in the mutant cluster may initiate but not complete their reprogramming. This also shows how the development of one tissue can indirectly influence the development of another.

4 Discussion I

Molecular and cellular phenotypes reflect the TF's role and position in the GRN

In my project, I examined how single TFs within the mesoderm GRN contribute to muscle morphogenesis over time by direct transcriptional activation, or by promoting or hindering binding of other TFs at the same enhancer, with variable effects on target gene expression. I combined stage-specific perturbation of the upstream TFs Twi, Mef2 and Tin via iLEXY with (1) a bulk CUT&Tag readout of the depletion efficiency at different enhancers, and single-cell accessibility in (2) enhancers and (3) genes.

4.1 Mutant phenotypes in sc-ATAC-seq

Although incomplete, the iLEXY perturbations, either alone or in a sensitized background, could largely recapitulate gross mutant phenotypes. Having a single cell phenotypic readout (sci-ATAC-seq) allowed me to capture additional granularity in cell fate changes that were not apparent in previous bulk experiments, and to link them to enhancer and gene accessibility.

The effects of Mef2 depletion were limited to the mesoderm. As *Mef2* loss-of-function mutants have a block in myoblast fusion (Bour *et al.*, 1995; Lilly *et al.*, 1995), the somatic muscle FCs typically remain (the FCM were reported to undergo apoptosis) (Ranganayakulu *et al.*, 1995). This was also apparent in the Mef2-iLEXY experiments under blue light – some cells, presumably the FCs, still followed the WT developmental trajectory, but most cells after Mef2 depletion gave rise to two mutant clusters, both detected at stage 9 and 11 (Figure 23, Figure 25). The two mutant clusters showed higher accessibility at sites bound by TFs involved in gastrulation (Leptin, 1991; Ip, Maggert and Levine, 1994) and FC/FCM specification (Carmena, Bate and Jiménez, 1995), but showed some biological differences: one (Mut-Lexy-2) had almost exclusively direct effects from Mef2 loss, and was compatible with stalling in FCM specification, while the other (Mut-Lexy-1) had some additional indirect effects linked to the action of the pioneer factor Zld and fitting to a more stem-like identity (Figure 24). The two clusters included cells from both stages, but cells depleted at stage 9 were enriched in Mut-Lexy-2 (Figure 25), suggesting that it could be the first identity established in time and that a second step is required to trigger reactivation of a stem-like fate, or that more differentiated cells (stage 11) tend to show a more aberrant response. To further explore this hypothesis, it would be useful to derive the true cell age with higher resolution than the collection window; however, this is challenging since we cannot assume that the actual age matches the developmental progression within mutant cluster(s), and it unclear to what extent non-perturbed peaks would be sufficient to derive the developmental age calculated on WT data (Calderon *et al.*, 2022). The emergence of two distinct mutant clusters and the presence of residual Mef2-depleted cells on the WT trajectory highlights three responses of different muscle cell types to Mef2 depletion, a granularity that was masked in bulk experiments. Such a complex

response points to the contribution of additional factors in establishing or maintaining cell identity, which could act simultaneously or through epistasis.

Conversely, Twi depletion had a more widespread effect, leading to a global shift in the identity of cells belonging to the trunk mesoderm, which acquire an aberrant state with similarities to epidermal and neural tissue (Figure 33). The WT nervous system and plasmatocytes were also enriched in *twi-iLEXY* cells. Although Twi was depleted after mesoderm gastrulation, the absence of its late roles in establishing domains for the specification of the SM (high Twi) and the heart/VM/fat body (low Twi) (Dunin-Borkowski and Brown, 1995; Baylies and Bate, 1996) has a huge impact on mesoderm identity. The importance of Twi's late roles is supported by previous viability assays and RNA-seq (Kögler *et al.*, 2021), which showed that 1 h depletion at stage 9 is sufficient to reduce embryonic viability by 90%, as well as the mRNA levels of Twi targets *tin* and *Mef2*, despite enabling their initial transcriptional activation. These results were also confirmed by my sci-ATAC-seq data, showing accessibility on the gene body. This suggests that Twi is not only required to initiate *tin* and *Mef2* expression, but also to maintain their expression and perhaps also their ability to regulate downstream enhancers, in keeping with their extensive combinatorial binding (Sandmann *et al.*, 2007; Zinzen *et al.*, 2009). The negligible number of residual *twi-iLEXY* cells in the trunk mesoderm could be due to the absence of high-Twi domains, and Twi upstream position in the mesoderm GRN, combined with the decrease in Tin and Mef2, is compatible with a global effect on mesoderm development, as seen by the negligible number of remaining *twi*-positive cells in the trunk mesoderm. This is in stark contrast to the impact of Mef2 depletion, which is lower down in the GRN.

The apparent shift of trunk mesoderm identity, in the absence of Twi, to an epidermal and neural like fate is an exciting result that is very surprising. This suggests a shift in identity of cells across a germlayer, from mesoderm to ectoderm, which is normally never observed. However, as Twi is the earliest identify factor for the entire mesoderm, even being essential for the gastrulation of the presumptive mesoderm, it is conceivable in this case. It is reminiscent of a study in mouse ESCs showing that removal of Brm, an ATP dependent chromatin remodeler, leads to pre-cardiac mesoderm cells adopting a neuronal-like fate (Hota et al., 2022).

4.2 Differential peaks and genes are consistent with Twi and Mef2 roles

The different position of Twi and Mef2 in the GRN architecture was also apparent in the molecular phenotypes of the mutant clusters, as evaluated by the differential peaks and genes.

After Twi depletion, down DA peaks were strongly associated to high Twi occupancy (i.e. ChIP signal), with the top 10% of Twi peaks accounting for > 60% of down DA peaks (Figure 37). This indicates that Twi is the main driver of accessibility at these peaks, consistent with its role as a master mesoderm TF. On the contrary, the top 10% of Mef2 ChIP peaks account for ~20% of DA sci-ATAC-seq peaks (Figure 28), indicating that additional factors beyond Mef2 contribute to the change in accessibility: indeed, down DA peaks after Mef2 depletion tended to be co-bound by Tin or Twi (Figure 29). This is in agreement with the observation that Twi co-binds at 42% of all Mef2 enhancers (Sandmann *et al.*,

2007) and especially at early Mef2 enhancers (active at 4–6 hrs but not at later timepoints) (Sandmann *et al.*, 2006). At these enhancers, Mef2 and Twi could bind cooperatively, or the binding of Twi might act as a place holder to facilitate Mef2 binding, and then Mef2 could further stabilize Twi binding.

Besides distal peaks (putative enhancers), the affected direct target genes also reflect the role of Twi and Mef2: Twi depletion strongly impacts mesoderm identity, upregulating genes related to non-mesodermal identities (in particular neural fate), while Mef2 depletion is mostly associated to defects in muscle differentiation and myoblast fusion, and the upregulated targets have less defined roles.

4.3 Stage-specific perturbation dissects TF activity at precise time-points, enriching for direct effects

The molecular phenotypes of loss-of-function KO mutants for *Mef2*, *tin*, *bap* and *bin* at 10–12 h, were previously characterized by sci-ATAC-seq by a former Ph.D. student (Secchia *et al.*, 2022). This time point is during muscle terminal differentiation, and later than the time points that I studied here. KO mutants ensured complete TF depletion, thus maximizing phenotypic effects; moreover, the choice of a late timepoint allowed to investigate the presence or absence of specific cell populations, which might not be efficiently isolated before terminal differentiation. On the other hand, the assessment of cell identity at late stages may have masked intermediate cell identities which undergo apoptosis before 10–12 h; which was the case for *tin*, *bap* and *bin* which all showed reduced muscle cells numbers but no new mutant cluster (Secchia *et al.*, 2022). Furthermore, the stage at which those identities arise is unclear. Mutant clusters could emerge either as a direct result of the TF KO, or from cell reprogramming. In the latter case, this could be due to indirect effects at many genomic sites not directly bound by the KO TF.

Linking cellular and embryonic phenotypes to their molecular basis does not disentangle direct and indirect effects of the KO on chromatin accessibility. In particular, the indirect effects here may have three main sources: (1) failed activation by target TF of the depleted TF, (2) potential compensation from other TFs, which could be favoured by the prolonged absence of the KO TF, and (3) activation of independent cell differentiation programs (Secchia *et al.*, 2022). The accumulation of epistatic effects and compensation over time could have a strong impact on the resulting phenotype. Therefore, the KO mutant is limited in drawing causal links between enhancer state and phenotype.

A short (1 h) TF perturbation at selected time points, and assessing the cells state immediately after the end of the perturbation, ensured that all observed effects were due to depletion at the specific time. This way, I could achieve a precise mapping of TF roles in time, which is crucial to meet the fast temporal dynamics of GRN activation. I also minimized indirect effects, which generally require a longer time to arise, although some could still be observed (for instance in Mut-Lexy-1).

Short time-specific perturbations are particularly useful when examining TF co-binding, which is extremely widespread in the mesoderm GRN and dynamic in time (Zinzen *et al.*, 2009). Co-binding is especially prominent at stage 9 and 11, when the mesoderm is sub-divided into different muscle primordia. A combination of time differences and the use of time-specific (in my dataset) versus whole-

embryogenesis data (in Secchia *et al.*, 2022) may explain why down DA sites are mostly associated with co-binding of Mef2, Tin and Twi in my dataset at stage 9 and 11, while in Secchia *et al.* (2022) Mef2 alone was mainly associated to down DA sites. Indeed, the percentage of down DA sites was negatively correlated to the number of co-bound factors in Secchia *et al.* (2022), while after Mef2 optogenetic depletion it had a bell shape with a peak at 4 TFs.

In addition, the two sets have different DA peaks and genes, which reflects the temporal progression in mesoderm development. Among the down DA peaks, only four overlap between Secchia *et al.* (2022) and my Mef2-depleted clusters, which have 621 peaks in common; as for the genes, 5 are in common between Mut-Lexy-1 and Secchia *et al.* (2022). The overall low numbers may be due to biological differences in enhancer usage across time, which may impact both the mutant and the WT to which it is compared. Indeed, the *Mef2-iLEXY* mutant clusters are compared against the WT trunk mesoderm at stage 9–11, which includes the future somatic and heart mesoderm and is less differentiated, whereas in Secchia *et al.* (2022) the mutants are specifically compared to the somatic muscle at 10–12 h. To avoid confounders from potential differences in enhancer usage, I also checked the target genes. Among genes with accessibility above the background, 70 are downregulated only in Secchia *et al.* (2022), 109 in both the *Mef2-iLEXY* mutant clusters but not in Secchia *et al.* (2022), and six are in common. Upregulated genes on the contrary have only one overlap, being much more numerous in Secchia *et al.* (2022) (76) than in the Mef2-iLEXY dataset (8 in Mut-Lexy-1 and 2 in Mut-Lexy-2). The low overlap between the two datasets suggests that time specificity in both perturbation and collection is a key determinant that captures dynamic enhancer usage and gene expression during these developmental transitions. Moreover, the higher ratio of down DA compared to up DA peaks and genes in the Mef2-iLEXY dataset suggests that the iLEXY perturbation captures more direct effects, avoiding the accumulation of aberrations of a KO mutant over all subsequent stages of embryogenesis.

4.4 Incomplete depletion: from a limitation to a resource

4.4.1 Incomplete depletion in bulk may result from cell-cell heterogeneity, which can be disentangled in single-cell data

Spike-in CUT&Tag shows that the depletion with iLEXY is incomplete and variable for different TFs: Twi has the highest depletion, followed by Mef2 and Tin (Figure 18). As a bulk assay, the baseline signal and magnitude of the perturbation are an average across all TF-expressing cells in the whole embryo. Since the mesoderm cannot be pre-sorted due to the perturbation of its identity marker genes, the total number of cells expressing each mesoderm TFs and its expression level change according to the TF and the timepoint. At stage 9, *twi* has the highest expression (Jenkins, Larkin and Thurmond, 2022) and the largest expression domain, being pan-mesodermal (Dunin-Borkowski and Brown, 1995), accounting for ~30% of cells; *Mef2* is expressed in a similar number of cells at both stage 9 and 11, and at higher levels at stage 11 (Jenkins, Larkin and Thurmond, 2022); *tin* instead has a narrower expression domain at stage 11, being restricted to the cardioblasts and a very small population of dorsal somatic muscle cells (Azpiazu and Frasch, 1993), which is reflected in a lower number of peaks in CUT&Tag.

TFs expressed at lower levels, or in a smaller domain, have a lower baseline signal, diminishing the chance to call differential peaks by DESeq2 (Anders and Huber, 2010).

Importantly, depletion at a peak may either result from a global and uniform reduction of TF binding at the peak, or from complete peak loss in some cells, and no change in others. By microscopy, I observed that FCs have higher Mef2, Tin and Bin nuclear retention, supporting heterogenous TF binding. Additional cell types and expression domains within the mesoderm might exhibit some heterogeneity not as easily detected by imaging, but more apparent in the chromatin state: for instance, the mesodermal cells of the *Mef2-iLEXY* mutant are divided between the WT mesoderm (which could harbour FCs) and two distinct mutant clusters. Thus, bulk measurements do not reflect the heterogenous response of single cells. However, to date no single-cell CUT&Tag technique is able to capture TFs with few binding sites in specific cell populations – it has mostly been applied to chromatin marks (Kaya-Okur *et al.*, 2019) (reviewed in Fu *et al.*, 2024) and to abundant proteins like the cohesin component Rad21 (Bartosovic, Kabbe and Castelo-Branco, 2021) and to RNA pol II (Janssens *et al.*, 2022), whereas the profiling of a glia-specific TF (Olig2) in sorted brain cells was less successful, with very low average number of reads per cell (< 50) (Bartosovic, Kabbe and Castelo-Branco, 2021). To infer a relationship between potential binding in single cells and global binding, I examined TF binding by bulk CUT&Tag under sci-ATAC-seq peaks which are DA between the mutant clusters and the closest WT (section 3.6.3). The peaks directly affected upon TF depletion are those with decreased binding in CUT&Tag and differential accessibility in sci-ATAC-seq. Using a set of perturbed CUT&Tag peaks instead of a WT-based set of bound sites, I was able to restrict the search space to only include regions specifically linked to the cells with an aberrant development in the mutant potentially driving this phenotypic change. This facilitates the assessment of enhancer-to-phenotype links and overcomes the issue of heterogeneity in response to TF depletion.

4.4.2 Incomplete depletion can be leveraged to infer enhancer affinity

Different enhancers may be bound by the same TF but with different affinity across the genome, leading to different responses and sensitivity to TF levels (Ramos and Barolo, 2013; Long, Prescott and Wysocka, 2016; Reiter, Wienerroither and Stark, 2017; Syed, Duan and Lim, 2023). In KO mutants, since the TF is not expressed, there is no residual TF-DNA binding and the affinity must be deduced from the TF ChIP signal under enhancer peaks in WT embryos. Despite leading to less exacerbated phenotypes, a milder perturbation allows a more direct measurement by comparing binding in WT and treated embryos under ChIP (or Cut&Tag) peaks.

The change in TF binding upon depletion is indicative of enhancer affinity. Enhancers with low affinity motifs for a given TF will have a strong reduction in binding, unless that factor is bound cooperatively, and those with high affinity may either not be affected or only slightly perturbed. In the *Drosophila* mesoderm, the time point-specific binding of a TF to enhancers (i.e. at only early or only late stages) could not be explained by differences in the affinity of the TF's binding site, but rather by the co-occupancy of stage specific TFS (Wilczyski and Furlong, 2010); by TF perturbation via the iLEXY system I could assess this directly. After Mef2 depletion, sites co-bound by Mef2, Tin and Twi are the

most likely to decrease in accessibility at stage 9 and 11 (Figure 29), suggesting that Mef2 stabilizes the binding of the other TFs or engages in a cooperative interaction. To assess the contribution of each TF to the total binding, it is then sufficient to integrate binding and accessibility data from all examined TFs at the common enhancer. So far, I examined single TFs and I binned binding and accessibility into classes (down DA, no change, up DA), but an integrated quantitative approach could more precisely estimate the weights of single contributions. This is something that I plan to do during the completion of this study for publication.

4.5 The combination of binding and accessibility data suggests how TF perturbation impacts enhancer accessibility and transcription

To link chromatin state to phenotypes and to identify the direct targets of a TF, I first integrated TF binding with accessibility at the enhancer, followed by inspection of the transcriptional state of the potential target gene. Since iLEXY lines recapitulate most of the KO mutant phenotype (Figure 19), the sites affected by depletion should be those responsible for the phenotypic aberrations in the absence of the TF, and for correct development in the WT embryo; therefore, I selected sites that are significantly depleted in mutant clusters and inspected the cases described in section 3.6.3.

This approach is easily interpretable, but involves some technical and conceptual simplifications. First, accessibility on the promoter and gene body was used as a proxy for nascent transcription, since it is more informative to have both the accessibility and transcription from the same cell, and to date no protocols combining sc-ATAC-seq with single-cell nascent RNA-seq are available. Alternatively, the two modalities could be acquired independently and integrated during data analysis, but this has not been attempted yet. A simpler bulk nascent RNA-seq may either not detect or underestimate the changes, since they affect only part of the mesoderm. For promoter and intronic peaks, the signal from the ATAC peak and from the gene is mixed, which could lead to some false positives for proximal peaks; thus, distal enhancers are more reliable, as they are fully separate from the target gene. Intronic enhancers are extensively present in the mesoderm GRN alongside promoter-proximal enhancers, as in *tin*'s enhancers *tin* A and *tin* B (Yin, Xu and Frasch, 1997). Therefore, many biologically meaningful connections may be discarded by excluding proximal peaks. The set already had a limited number of connections, since the assignment of peaks to enhancers is based on BGDP *in situ* (Tomancak *et al.*, 2007; Hammonds *et al.*, 2013), which is inherently limited in the number of tested regions. Thus, both proximal and distal peaks were leveraged to avoid depleting the signal.

This approach has the advantage of limiting the search space for direct targets, by pre-filtering for peaks with a depletion in CUT&Tag. However, it has some limitations in sensitivity: since it relies on known enhancer-gene pairs, it does not identify *de novo* connections and is strongly dependent on previous knowledge of the system, unlike GRN inference methods like SCENIC+ (Bravo González-Blas *et al.*, 2023) or GRaNIE (Kamal *et al.*, 2023). The identified connections could increase by using a list of tissue-specific enhancer-gene pairs inferred from WT data using those algorithms. Moreover, my approach does not capture subpopulations below the resolution of sci-ATAC-seq clustering, unless cells

are manually specified according to other markers, or small populations which would not significantly reduce the bulk CUT&Tag signal.

Despite these limitations, the network recapitulates known Mef2 and Twi target genes in regions with reduced binding, accessibility and gene expression, consistent with cell identity and observed accessibility for other TFs. For Twi, the genes with reduced binding but upregulated accessibility on the putative enhancer peak and upregulated expression (inferred from accessibility) were in accordance with the increase in neurons and plasmatocytes in the clustering, suggesting that Twi might actively repress these fates. Remarkably, N-Twist was shown to inhibit activation of a mammalian homolog of the proneural *achaete-scute* complex in mouse cells (Verzi *et al.*, 2002), and Twist-1 ablation in the cardiac neural crest promotes differentiation of ectopic neuronal cells (Vincentz *et al.*, 2013). Although Twi has different roles in myogenesis in vertebrates and in *Drosophila* (Castanon and Baylies, 2002), this function might be a conserved ancient feature that was masked in *Drosophila* to Twi's earlier function in gastrulation. Additional phenotypic validation after Twi stage-specific depletions after gastrulation or at stage 9 by immunostaining with neuronal markers could to help address its role in neurogenesis.

Twi may also benefit from the co-binding of other TFs beyond Mef2 and Tin to regulate cell fate: for instance, Twi and Sna co-bind and directly regulate a pool of mesodermal genes, where Sna can act as a repressor or an activator (Rembold *et al.*, 2014), and some of those genes – *rho*, *Mef2*, *tin*, *stumps* – are recapitulated in my analysis (section 3.6.3). Sna was shown to potentiate the activity of Twi at co-bound enhancers (Rembold *et al.*, 2014); thus, Twi depletion will reduce Sna binding (and potentially accessibility) and transcriptional output, consistent with my observations. Sna is an activator also for *stumps*, and Twi depletion has the same effect. On the contrary, *rho* is repressed by Sna which seems to compete with Twi (Ip *et al.*, 1992), but it is not clear how this can be reconciled with the change in accessibility. Since Twi and Sna have very similar motifs, inspecting the motif syntax will be important to derive their interactions at *rho* (Rembold *et al.*, 2014). In general, the motif syntax at co-bound sites could help to stratify TF behavior and potentially derive rules for TF-TF interactions.

4.6 Future directions

The project is still ongoing, and the generation of additional datasets for Tin depletion will be important for a global view of TF interactions in the mesoderm GRN and their effect. For a more complete picture of the mesoderm GRN at stage 9, and of its 3 more upstream TFs at stage 11, I intend to perform sci-ATAC-seq on the *tin-iLEXY/tin*^{EC40} mutant with Tin depletion at stage 9 and 11. The sci-ATAC-seq data from the Twi, Mef2 and Tin iLEXY depletions will be integrated together with the WT stage 9–11 data to facilitate direct comparisons and examine their overlap. Despite the diversity in the markers of the Mef2 and Twi mutant clusters, a direct comparison on the same peak set could extract subpopulation with common markers.

Depleting Tin would also allow characterizing the stability of TF-TF interactions at the same enhancer across stages. For Twi, only one depletion stage was available from previous collections, so I could not

resolve Mef2 and Twi co-binding in time, as only stage 9 data can be compared. With the addition of the Tin-iLEXY dataset, it would be possible to examine the progression of Mef2 and Tin co-binding, and if the binding requirements for enhancer accessibility and activity change in time. Apart from TF-TF interactions, I will address feed-forward loops involving Twi and Tin, or Twi and Mef2, and refine the enhancer-to-gene assignment in a future collaboration. Phenotypic validation by RNA FISH will be important to confirm the insights from single-cell data.

Finally, a limitation of the current readout of enhancer state is that although active and primed enhancers might be indistinguishable based on accessibility, they have different effects on the transcription of the target gene. Enhancers with decreased accessibility could either (1) be bound by other TFs (albeit at lower levels), have active chromatin marks and promote transcription, though at lower levels, or (2) be primed or closed, harbor repressive chromatin marks and not trigger transcription. To distinguish between active and primed/closed enhancers, I am planning to perform single-cell CUT&Tag for the active mark H3K27ac, present at active enhancers, and for the repressive mark H3K27me3, both at close and primed enhancers. While the chromatin mark can distinguish active and primed/closed enhancers, closed and primed enhancers could be separated by the level of residual accessibility. Thus, by evaluating these histone modifications at single-cell level, I could disentangle heterogeneity in mark deposition and, based on the active enhancers, integrate single-cell CUT&Tag data with sci-ATAC-seq data. For the integration of all data together to form a 'functional' GRN of mesoderm development, I plan to collaborate with a computer scientist to model the TF inputs into the GRN.

5 Results II

Finding rules for TAD boundary function in *Drosophila*

The features of TAD boundaries are well-known in vertebrates, where the motif for the insulator protein CTCF in a convergent orientation is sufficient to create functional boundaries (de Wit *et al.*, 2015), and motif inversion disrupts chromatin loops (Guo *et al.*, 2015; Sanborn *et al.*, 2015). However, in *Drosophila* although the CTCF protein is enriched at TAD boundaries (Hou *et al.*, 2012; Sexton *et al.*, 2012), convergent CTCF motifs have not been uncovered to date (Rowley and Corces, 2018). On the other hand, a plethora of additional direct and indirect DNA-binding insulator proteins colocalize with TAD boundaries in the fly, binding mostly in combination (Hou *et al.*, 2012; Sexton *et al.*, 2012; Van Bortle *et al.*, 2012). Some combinations of insulator proteins have been linked to specific boundary features, like active transcription (Ramírez *et al.*, 2018). Despite multiple efforts to understand the importance of single proteins in establishing and maintaining a TAD boundary, so far none of the known insulator proteins has been proven necessary for a large proportion of TAD boundaries (Kaushal *et al.*, 2021; Cavalheiro *et al.*, 2023; Kahn *et al.*, 2023), and the contribution of single and multiple combinatorial proteins remains elusive. Moreover, the mammalian CTCF-based mechanism of TAD formation does not apply to all mammalian TAD boundaries (Bonev *et al.*, 2017; Nora *et al.*, 2017), and it is evolutionarily recent (Kabirova *et al.*, 2023), suggesting that other actors may be involved in establishing topology. In my project, I aimed to investigate the rules underlying the interactions between TAD boundaries in *Drosophila* to uncover not only *Drosophila*-specific mechanisms but more general principles conserved in evolution. In particular, I focused on the following questions:

- Can TAD boundaries function in an ectopic context?
- What combinations of insulator proteins are required?
- Is a promoter required?
- Is the orientation of boundary elements important for their function, i.e. for their ability to pair and form a TAD? If so, what are the directional elements that explain orientation specificity?
- Can we deduce rules for the interaction between boundaries?

To understand the contribution of boundary elements to boundary function and the role of context, I selected endogenous boundaries with different combinations of insulator proteins that occur with high frequency in the *Drosophila* genome, we then inserted each boundary into the middle of ectopic acceptor TADs and tested if the inserted element could form a new boundary. Since the insertion occurs in a random orientation, I could also assess the impact of boundary orientation on boundary function. After measuring the ability of the insertion to form a boundary by DNA FISH and capture Hi-C, I searched for features correlating with working and non-working boundaries. In this way, I could assess

sufficiency of insulator protein combinations and other elements to establish a boundary in an endogenous context.

This project is the result of a collaboration with multiple colleagues, led by a formed Ph.D. student, Dr. Gabriel Rodrigues Cavalheiro, and myself. A research scientist in the Furlong group, Dr. Rebecca Rodriguez Viales, performed the capture Hi-C and the ChIP-seq experiments, Dr. Adam Rabinowitz designed the capture Hi-C probes, and Dr. Charles Girardot preprocessed the capture Hi-C data and performed the ChIP-seq data analysis. Dr. Aaron Comeault performed the alignments to find evolutionary conserved sequences. Dr. Raquel Marco-Ferreres, Dr. Rebecca Rodriguez Viales and Dr. Katharina Bender helped with fly work. Finally, an intern, Songjie Feng, was involved in the generation of some fly lines and Dr. Alessandra Reversi performed all embryo injections as part of the EMBL injection service. Throughout the results section, I use 'I' to describe work that I have done and 'we' to describe work that was done together with one or more of the collaborators listed above.

5.1 Genome-wide TAD boundaries show different insulator and promoter combinations

To select the endogenous boundaries to insert into the acceptor TADs, I first examined the genome-wide distribution of co-bound insulator proteins at TAD boundaries, using ChIP-seq data. The definition of TADs and TAD boundaries is strongly dependent on the resolution: at higher resolution (1-2kb), more TAD boundaries are called; in addition, the identified TADs are smaller and could include a large percentage of sub-TADs; on the other hand, lower resolutions (for example 5-10kb) can detect bigger TADs but less boundaries. In Hi-C Explorer and other algorithms for TAD calling, TAD boundaries are defined at a specific resolution, and each domain between two consecutive boundaries is considered a TAD. However, for a sub-TAD, one or both of the boundaries could be involved in additional interactions to form a bigger TAD; this interaction is not detected by high-resolution TAD calls. Moreover, disordered domains are called TADs, despite the absence of a clear interaction beyond the linear genome. Whereas decreasing the resolution could solve these issues, it would also obliterate sub-TAD interactions and decrease the boundary number, leading to information loss. Although the resolution of the data (depth of sequencing) has a huge impact, part of the issue is also the biological nature of TADs in that they are hierarchical nested structures.

To overcome the dependency of TAD boundary calls from the resolution, I manually annotated TAD boundaries and TADs using TAD boundary calls at 1, 2, 5 and 10 kb, using previously generated Hi-C data in the lab (Pollex, Marco-Ferreres, *et al.*, 2024). TAD boundaries and TADs had to be present at least at one resolution. Moreover, TADs were further filtered by manual inspection of the Hi-C map of WT *D. melanogaster* embryos at 6–8 h to capture only visible interactions. With this method, I obtained a high-confidence set of 1852 TADs and their corresponding interacting TAD boundary pairs, consisting of 1567 single boundaries. For consistency, I resized all TAD boundaries to 2 kb, the shortest sequence including all ChIP and promoter features in almost all boundaries.

With respect to the insulator composition of the boundaries, I focused on the binding of four main known insulators proteins – BEAF-32, CTCF, CP190 and Su(Hw) – as assessed by ChIP-seq in 6–8 h embryos previously generated in the lab (BioProject accession number PRJEB38618). After TAD establishment (2-3 h), insulator binding and TAD boundaries are largely stable across tissues and time (Hug *et al.*, 2017; Ing-Simmons *et al.*, 2021; Pollex, Marco-Ferreres, *et al.*, 2024; Pollex, Rabinowitz, *et al.*, 2024); therefore, I chose 6–8 h data to match the stage used for the DNA FISH-based readout. The most frequent insulator combination at TAD boundaries is BEAF-32, CTCF and CP190 (221 boundaries), closely followed by BEAF-32 alone (187 boundaries) and CTCF and CP190 (162 boundaries) (Figure 40A).

Alongside insulators, TAD boundaries are generally enriched in promoters, in *Drosophila* as well as in other species, including mammals (Dixon *et al.*, 2012; Szabo, Bantignies and Cavalli, 2019). My analysis shows that almost 80% of TAD boundaries during embryogenesis have at least one promoter, and among them, 37.5% have bidirectional transcription (Figure 40B).

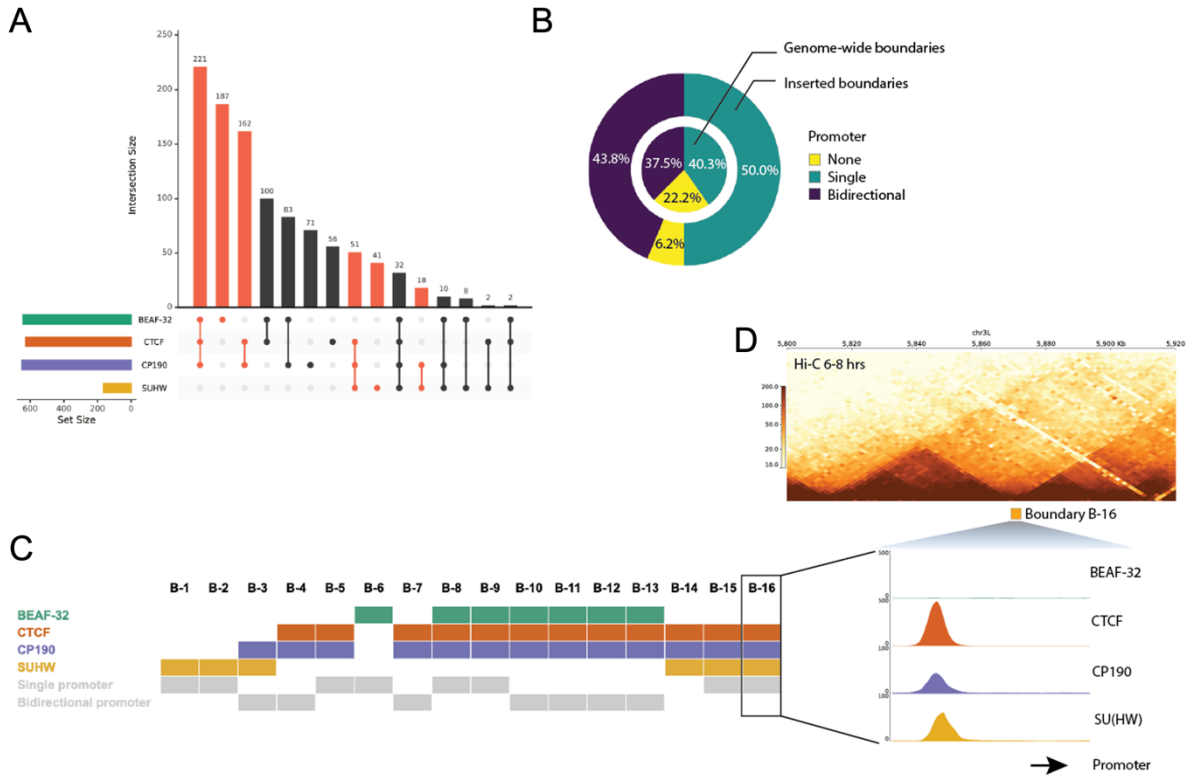


Figure 40: The selection of inserted boundaries recapitulates the diversity of TAD boundaries genome-wide.

(A) Frequency of insulator protein combinations genome-wide. Combinations tested in inserted boundaries are highlighted in red.

(B) Percentage of TAD boundaries with no promoter, a single or a bidirectional promoter in the genome (inner circle) and in the inserted boundaries (outer circle).

(C) Composition of inserted boundaries. For each boundary and feature (insulator protein or promoter), a coloured rectangle indicates that feature is present in the boundary, and a white rectangle that it is absent.

(D) Example of an inserted boundary (boundary B-16). Hi-C map in 6-8 h WT *D. melanogaster* embryos at 5 kb resolution, 6-8 h ChIP tracks for the insulator proteins BEAF-32, CTCF, CP190 and Su(Hw) and arrow indicating the direction of the promoter.

5.2 Rationale for the choice of inserted boundaries

Among the inserted boundaries, Dr. Gabriel Rodrigues Cavalheiro and I included the most frequent combinations observed genome-wide, but also combinations occurring with lower frequency (Figure 40A, in red). This approach broadens the number of combinations to test, and could show if the most frequent combinations also have a higher chance of working in an ectopic context. To understand if known insulator proteins are sufficient to explain boundary function or if additional elements not included in our model are required, we selected multiple endogenous boundaries for the same combination of insulators, except for the BEAF-32 -only and the Su(Hw) & CP190 boundary. All selected boundaries were called at 5 kb resolution and validated by manual inspection. The boundary size was chosen to include all ChIP peaks, and it ranges between 2–3 kb. A total of 16 boundaries were selected, among which the most represented insulator protein combinations include BEAF-32 + CTCF + CP190, CTCF + CP190 + Su(Hw) and CTCF + CP190 (Figure 40 C-D, Table 7).

In addition, the selected boundaries had either none, one or a bidirectional promoter, to explore the role of the transcriptional machinery on TAD boundary function (Figure 40B). In the selection, boundaries with at least one promoter were slightly overrepresented with respect to the level in the genome to assess the requirement of a promoter for boundary function.

5.3 Rationale for the choice of acceptor TADs

A combination of technical and biological factors guided the selection of acceptor TADs.

From a technical point of view, Dr. Gabriel Cavalheiro and I considered the TAD size and the availability of a recombination 'landing' acceptor site within the TAD, made available by fly resource, MiMIC (Venken *et al.*, 2011).

TAD size is crucial for a robust readout of the perturbation both by imaging and genomics. In *Drosophila*, TADs are generally smaller than in vertebrates, spanning 20–60 kb on average (Ramírez *et al.*, 2018; Ulianov *et al.*, 2021); to reliably detect if the inserted boundary disrupts the original TAD by DNA FISH, the probes should be ≥ 50 kb apart (Giorgetti and Heard, 2016). Otherwise, the proximity on the linear genome would result in colocalization, regardless of the 3D structure. Larger TADs also allowed using longer probes, improving the signal-to-noise ratio and facilitating spot detection and analysis. Moreover, genomic readouts like capture Hi-C would be more affordable, since lower resolutions are sufficient to detect the effect of the insertion.

On the other hand, the availability of the MiMIC fly lines greatly simplified the design, as we could leverage commercially available fly lines with the desired cassette. Based on both criteria, TAD size and an available acceptor site in roughly the middle of the TAD, we selected two TADs that are over 200kb.

In terms of the biological insights, we aimed at maximizing heterogeneity at various levels: TAD boundary composition, structural complexity and gene regulation throughout development.

5.3.1 TAD boundary composition

To effectively disrupt the acceptor TAD, the inserted boundary must hijack the interaction between endogenous TAD boundaries by creating a new pair with higher stability than the original one. Therefore, to discover the rules of boundary interaction, it is important to maximize diversity in both inserted and acceptor TAD boundaries. Since the same acceptor TAD is used for a large number of insertions and contributes less to the overall diversity, we selected TADs with different insulator combinations at their boundaries to test a larger number of interaction partners. At a genome-wide level, a specific combination of insulators does not interact univocally with another combination; thus, we aimed at reproducing this potential multi-way interaction in the experimental design to discover the features driving the interaction.

5.3.2 Structural complexity

In addition to the boundary composition, a second element that may influence TAD boundary function is TAD structure, especially, the presence of sub-TADs and chromatin loops; to address this point, we selected a TAD without additional structural elements and one with a complex internal arrangement including sub-TADs and loops. The presence of such structures may either stabilize the overall architecture or offer more interaction partners for the inserted boundaries.

5.3.3 Gene regulation

Lastly, developmental genes often require tight regulation of their expression, which can be highly correlated with the TAD structure. TAD disruption may have a significant impact of the transcription of those genes, leading to developmental arrest and lethality; this could facilitate the study of the link between chromatin topology and transcription, by comparing the magnitude of the structural and transcriptional alterations.

Considering these three sources of heterogeneity, we identified potential landing acceptor sites by visual inspection of *Drosophila* Hi-C data in 6–8 h embryos and ChIP-seq data for BEAF-32, CTCF, CP190 and Su(Hw) at the same stage. Among all regions fulfilling our criteria, we selected two TADs: the *scyl-chrb* TAD (TAD A) and the *esg-wor-sna* TAD (TAD B) (Figure 41).

5.4 Properties of acceptor TADs

5.4.1 TAD A

The first acceptor site is a ~310 Kb TAD containing the apoptotic genes *scylla* (*scyl*) and *charybde* (*chrb*) (Figure 41A). *scyl* and *chrb* are paralogous genes with the same temporal and spatial profile (Ghavi-Helm *et al.*, 2014): they are expressed from gastrulation (stage 5, 2 h) to terminal embryonic differentiation (stage 16–17, 16–24 h) with a peak of expression at 2–8 h, and have the same dynamic spatial expression domain, with strong signals in the dorsal part of gastrulating embryos, and segmental expression in the nervous system, the thoracic segments and the cardiogenic mesoderm at stages 9–13

(Scuderi *et al.*, 2006; Ghavi-Helm *et al.*, 2014). *scyl* and *chrb* co-expression indicates the presence of a shared regulatory program, as often reported for paralogous genes: both are targets of homeobox factors and induced by Dpp- and Zen-mediated signalling (Scuderi *et al.*, 2006). The double mutant is embryonic lethal and shows defects in head involution, whereas the single mutant does not cause any observable phenotype (Scuderi *et al.*, 2006). Therefore, in our experimental system, embryonic lethality can only occur if the expression of both genes is significantly perturbed by the insertion. TAD A is a gene-poor TAD, with only 12 other protein-coding genes besides *scyl* and *chrb* and multiple unannotated transcripts, with very low or no expression in embryogenesis (Jenkins, Larkin and Thurmond, 2022). Thus, the main phenotypic contributions will likely come from *scyl* and *chrb* only.

TAD A has a complex internal structure, with a three-way looping interaction, that is bound by CTCF: the left anchor (LA), upstream of *scyl*, interacts with the middle anchor (MA), situated in a non-coding region, and the right anchor (RA) at *chrb*, resulting in two strong loop interactions (Figure 41A). The right anchor also interacts with the middle anchor, forming a sub-TAD with a loop. CTCF binding strength is high in the left loop anchor, and significantly lower in the middle and right anchor. Additionally, the left loop anchor includes a bidirectional promoter. Outside the left and right anchor, there are two additional boundaries (Figure 41A): the left boundary (LB) is just downstream of BEAF-32 and CTCF binding sites and contains the forward promoter of *CG42671*, and the right boundary (RB) is bound by CTCF, CP190 and Su(Hw), overlapping the 3' UTR of *Mob2*. The outer boundaries do not engage in a looping interaction. The MiMIC landing site is located between the left and middle loop anchors (Figure 41A).

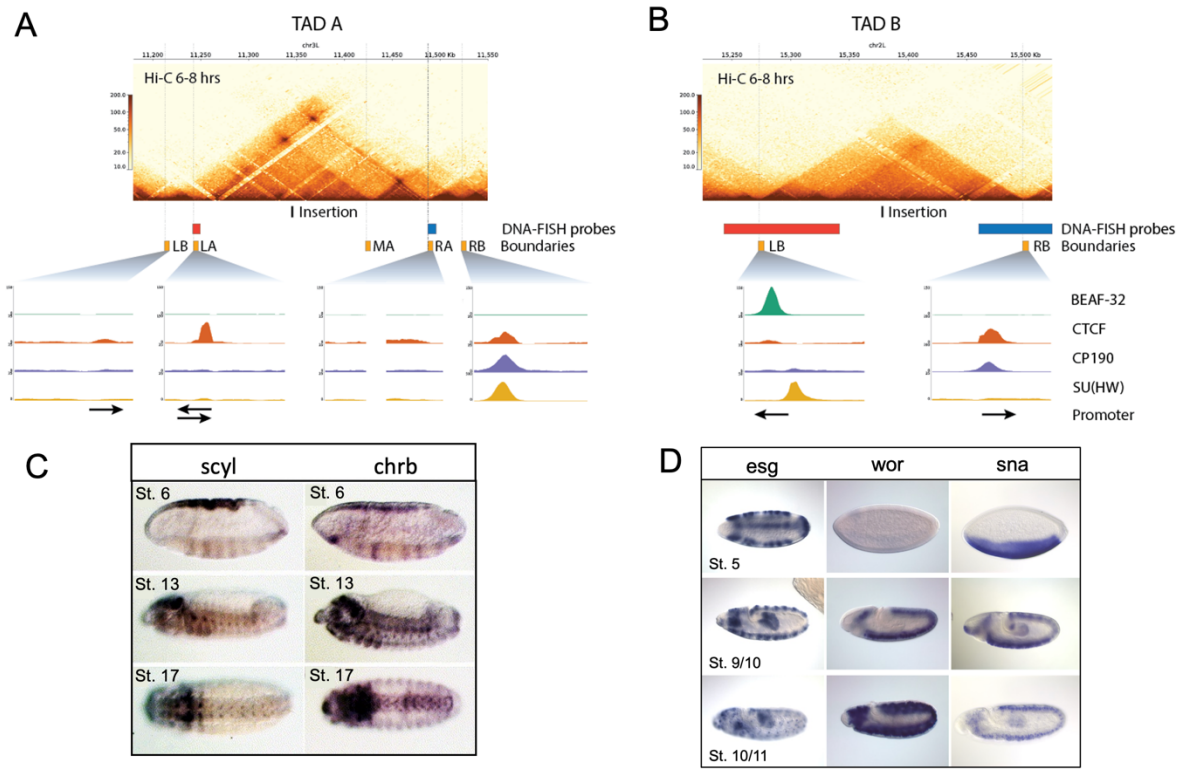


Figure 41: Acceptor TADs are diverse in their architecture, boundary composition and gene regulation.

(A) Hi-C of the *scyl-chrb* TAD (TAD A) in 6-8 h WT embryos, showing the position of the MiMIC insertion site, of the DNA FISH probes and of the endogenous boundaries. Below, ChIP tracks at 6-8 h and promoters at TAD boundaries.
(B) Same as A, for the *esg-sna-wor* TAD (TAD B).
(C) Expression of genes in TAD A, adapted from Scuderi *et al.* (2006).
(D) Expression of genes in TAD B, from BDGP (Wilczynski *et al.*, 2012).

5.4.2 TAD B

TAD B spans ~230 Kb and contains three developmental genes, *escargot* (*esg*), *worniu* (*wor*) and *snail* (*sna*), which are embryonic lethal and required for gastrulation (*sna*) (Simpson 1983) and for mesoderm and neuron development (Ashraf 2004, Hekmat-Scafe 2005) (Figure 41B). *sna*, *esg* and *wor* are highly homologous but have distinct temporal and spatial patterns in the embryo: *sna* is first expressed ventrally in the mesoderm at stage 4–8, and segmentally in the nerve cord from stage 9 to 16 (Alberga *et al.*, 1991; Tomancak *et al.*, 2007; Hammonds *et al.*, 2013); *esg* is expressed in the epidermis and neuroectoderm from stage 4 to 16 with a segmental and more extended pattern (Whiteley *et al.*, 1992; Tomancak *et al.*, 2007; Hammonds *et al.*, 2013); and *wor* is expressed in the brain and nerve cord from stage 9 to 16 (Ashraf *et al.*, 2004; Tomancak *et al.*, 2007; Hammonds *et al.*, 2013). All three genes are required together for correct neuroblast development; therefore, defects in the expression of just one of the genes could impact embryonic viability. TAD B includes eight additional protein-coding genes not expressed in embryogenesis (Jenkins, Larkin and Thurmond, 2022).

The structure of TAD B is very simple, with only a left (LB) and a right (RB) boundary, and no looping interactions (Figure 41B). This is the most common TAD structure in *Drosophila*. The left boundary is bound by BEAF-32, CTCF and Su(Hw) and includes the promoter of *Cul3* in reverse orientation; the right boundary has CTCF and CP190 binding and overlaps the *lace* promoter (Figure 41B). The MiMIC site is close to the middle of the TAD, between *esg* and *wor* (Figure 41B).

5.5 Generation of fly lines with boundary insertions

To insert the selected boundary into the acceptor TAD, we used phiC31 integrase-mediated recombination between the attP sites in the landing site within the acceptor TAD, and the attB sites in the donor DNA cassette (Venken *et al.*, 2011). The donor cassette includes an endogenous TAD boundary (2–3 kb) and an excisable *dsRed* marker to identify flies with the insertion, being flanked by convergent attB sites. The donor cassette was cloned into a plasmid, which was injected into *Drosophila* embryos having the landing site and expression of the phiC31 integrase, leading to recombination-mediated cassette exchange (RMCE) and integration of the donor cassette in the acceptor site (section 2.2.2). The insertion can occur in a random orientation, enabling us to insert a boundary in both orientations with a single injection. Even after multiple injection rounds, some insertions could not be obtained in both orientations, as one orientation was much more favoured than the other. All fly lines were genotyped to confirm the insertions' location, identity and orientation as described in sections 2.3.3 and 2.3.4.

Out of 32 possible insertions for each acceptor TAD (16 boundaries in two orientations), I obtained 27 insertions in TAD A and 26 in TAD B. Of the 27 insertions in TAD A, three are from boundaries which could only be inserted in one orientation, whereas the remaining 12 boundaries could be inserted in both. In TAD B, four boundaries were inserted in one orientation and 11 in both (Table 7).

5.6 Viability of boundary insertions

To assess the impact of the insertions, all fly lines were made homozygous, if possible.

All 27 fly lines with an insertion in TAD A were homozygous viable and fertile. This is consistent with the fact that the single KO of *scyl* or *chrb* is homozygous viable, and only the double KO is lethal; therefore, the expression of both genes must be affected in order to elicit a viability phenotype, which is less likely than the effect on a single gene. It is still possible that either *scyl* or *chrb* have minor alterations in their expression, which do not affect viability but may be detected via more in-depth phenotypic screenings or RNA-seq.

In TAD B, the three genes are embryonic lethal and leverage different regulatory elements. This increases the chance of observing effects on viability after a topological perturbation. Indeed, about 50% of the lines with an insertion (14 out of 26) were embryonic lethal (Table 7). Four inserted boundaries impacted viability in both orientations, and the other four in an orientation-dependent manner; among those, three had a topological change only in the lethal orientation, and one in both. Overall, embryonic lethality tends to be associated to topological changes (13/14 lines), but not vice-versa: embryos can be viable even after TAD disruption (5/12 lines) or compaction (3/12 lines). Thus, lethality may be due to the change in expression of ≥1 genes; notably, embryos had a normal morphology at stage 11, suggesting that gastrulation occurs normally and the embryonic defects emerge later, most probably during neurogenesis, due to altered regulation of *esg* and/or *wor*.

Table 7: Composition of inserted boundaries, effect on topology and on viability.

Boundary	***Composition***					***Effect in DNA FISH***				***Viability***	
	BEAF-32	***CTCF***	***CP190***	***Su(Hw)***	***Promoter***	***TAD A Or1***	***TAD A Or2***	***TAD B Or1***	***TAD B Or2***	***TAD B Or1***	***TAD B Or2***
B-1					Single	No	No	Yes	No	Viable	Viable
B-2					Single	No	No	Yes	Yes	Viable	Viable
B-3					Divergent	No	Yes	Yes	Yes	Lethal	Lethal
B-4					Divergent	No	Yes	Yes	No	Lethal	Viable
B-5					Single	Yes	Yes		Yes		Lethal
B-6					Single		No		Yes		Viable
B-7					Divergent	No	No	Yes	Yes	Viable	Viable
B-8					Single	No		Yes		Viable	
B-9					Single	Yes	No	No	Yes	Viable	Lethal
B-10					Divergent	No	No				
B-11					Divergent		No		Yes		Lethal
B-12					Divergent			Yes	No	Lethal	Lethal
B-13					Divergent	Yes	Yes	Yes	Yes	Viable	Lethal

B-14					Yes	Yes	Yes	Yes	Lethal	Lethal
B-15				Single	Yes	Yes	Yes	No	Lethal	Viable
B-16				Single	Yes	Yes	Yes	Yes	Lethal	Lethal

5.7 DNA FISH shows that most inserted boundaries have context- or orientation- specific effects

To detect topological changes across multiple lines, I first performed a DNA FISH-based screening on TAD A and TAD B loci in embryos in all insertion lines (Figure 42A). I used probes for the outer loop anchor of TAD A (~235 Kb, from the left to the right loop anchor) and for the two boundaries of TAD B, to maximize the inter-probe linear distance and get a clear readout (Figure 41). Thanks to the close contact between the left and right loop anchors, I could use small 8 kb probes in TAD A. Since TAD B does not have focal loops as in TAD A, the interaction may occur over larger genomic distances; therefore, I used 80 kb probes centred at the boundaries to achieve a stronger signal and evaluate TAD compactness as a whole.

The median distance between DNA FISH probes in WT embryos reflects differences in contact frequency from the Hi-C data: in TAD A, the looping interaction brings the two probes extremely close, at 248 nm distance, at the resolution limit of confocal microscopy; in TAD B, the median distance is higher, around 330 nm, consistent with the absence of strong looping interactions (Figure 41). This is independent from the linear distance of the probes, which is lower for TAD B compared to TAD A (205 kb versus 235 kb between the centre of the probes).

Of note, for all insertion lines and the WT control of TAD B, DNA FISH was performed in combination with RNA FISH to detect GFP expression. In fact, the homozygous lethal lines were maintained over a GFP balancer chromosome, and homozygous embryos were selected for measurements based on GFP absence.

To ensure uniformity, we performed all measurements in the central ventral ectoderm of 6–8 h embryos, matching the stage of our Hi-C and ChIP data. By choosing a position on the surface of the embryo, we minimized optical aberrations and the loss of fluorescence signal due to refraction and reduced probe penetrance, and we could gather a signal from a precisely defined cell monolayer.

Among the 16 selected boundaries, 10 were successfully inserted in both TADs in both orientations (Table 7). Most insertions had an observed effect by DNA FISH: only one boundary never had an effect, three were orientation-dependent, and the rest equally distributed between always working, context-dependent and orientation- and context- dependent boundaries, each accounting for 25% of the boundaries (4/16) (Figure 42B). Therefore, the inserted boundaries preserve the boundary function in a different TAD, out of their endogenous context: this suggests that they harbour specific features for this role, which is not only determined by the genomic context or by surrounding chromatin marks. However, the context still influences boundary function, creating a permissive state, as not all boundaries work in the acceptor TADs (Figure 42B).

Among the 15 boundaries inserted in TAD A in one or two orientations, four had no effect in either orientation, three had no effect in the single orientation tested, five an effect in both orientations and three an orientation-dependent effect. Overall, about 50% of the inserted boundaries could trigger topological alterations in TAD A in at least one orientation (Figure 42C).

Of the 15 boundaries inserted in TAD B, 11 were inserted in both orientations; among them, 5 had a topological effect in both orientation, and 6 were orientation-dependent. The 4 boundaries inserted in one orientation always had an effect (Figure 42D). All inserted boundaries altered topology in at least one orientation: the median distance between the DNA FISH probes increased in 19 of the 26 insertions, pointing to loss of contacts within the TAD, and decreased in three insertions, hinting at an increase in TAD compaction (Figure 42D). Notably, the same boundary (B-5) can have a different effect depending on the orientation of the insertion, leading to either TAD compaction or decompaction.

The higher sensitivity of TAD B to insertions, and the potential for compaction as well as disruption, may be due to a) the type of insulators found at the endogenous boundaries, which may engage in a larger variety of interactions, or b) to a more flexible context, with no internal loops. Indeed, sub-TADs and loops in TAD A could stabilize existing interactions, reducing the impact of the new insertion or confining it to a sub-TAD, while preserving the overall structure of the broader TAD.

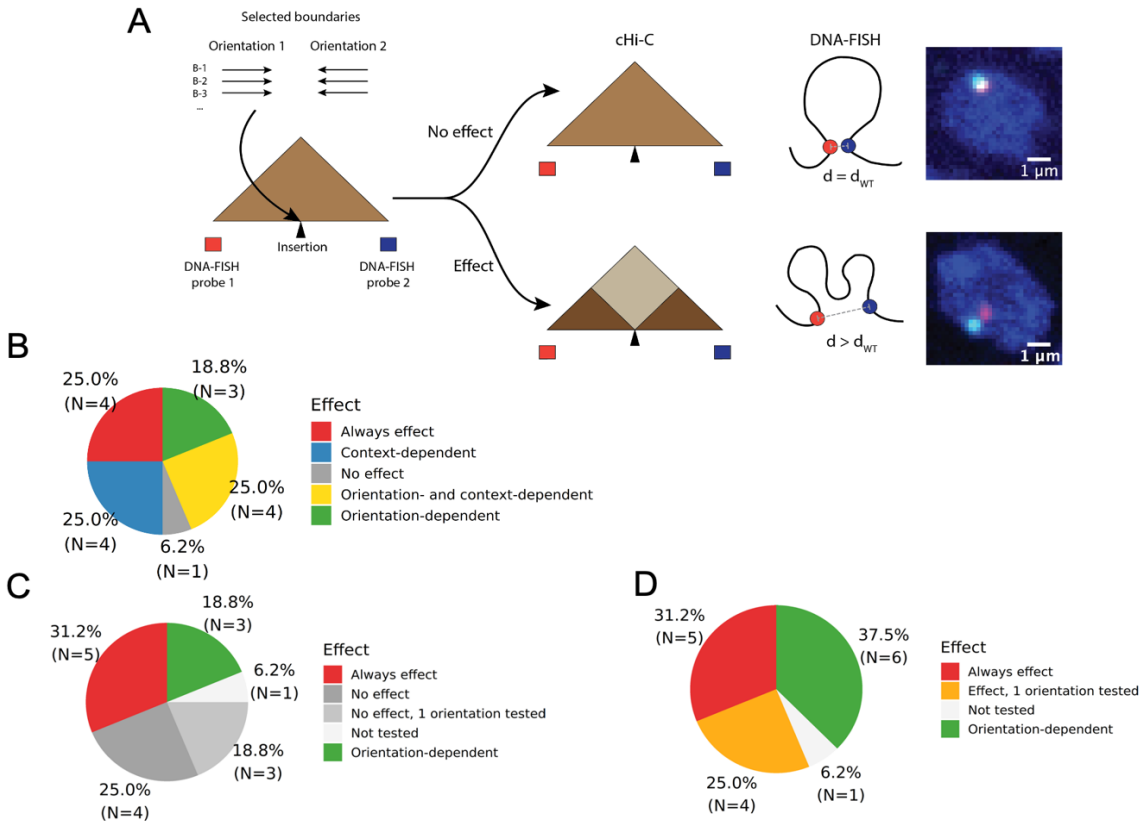


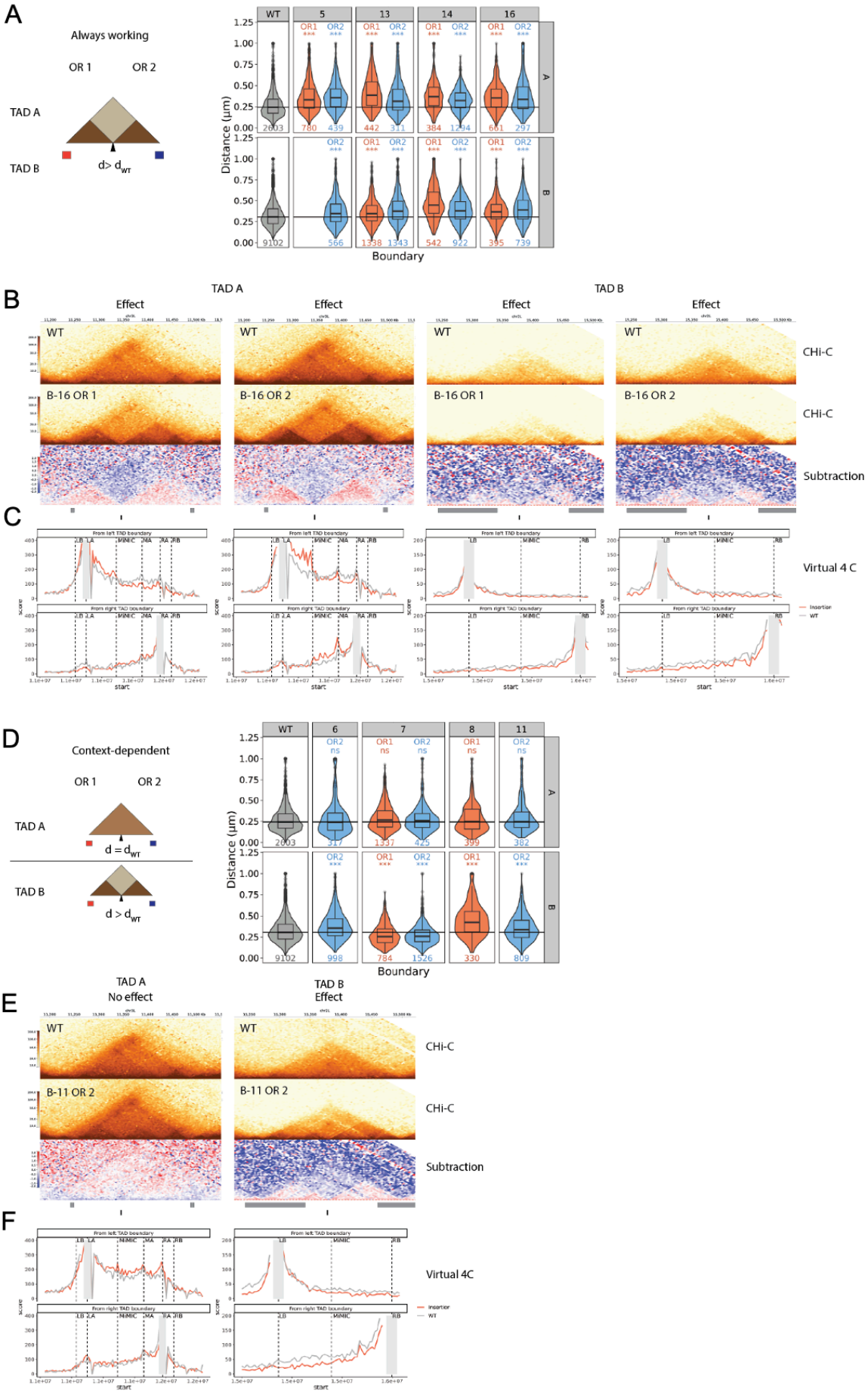
Figure 42: Most inserted boundaries have an effect, which can be orientation- and context- specific.

(A) Scheme of the effect of the inserted boundaries in capture Hi-C and DNA FISH, and DNA FISH images for a non-working (top) and a working (bottom) insertion.

(B) Pie chart with the proportion of boundaries tested in both orientations and in both TADs which have an effect always or depending on the orientation- and/or the context.

(C) Pie chart with the proportion of boundaries by effect in TAD A.

(D) Pie chart with the proportion of boundaries by effect in TAD B.


A
Always working
OR 1
OR 2
TAD A
TAD B
d> d WT
WT
5
13
14
16
Distance (µm)
Boundary
B
TAD A
TAD B
Effect
WT
B-16 OR 1
B-16 OR 2
CHi-C
Subtraction
C
Virtual 4 C
D
Context-dependent
d = d WT
d > d WT
6
7
8
11
E
TAD A
No effect
TAD B
Effect
B-11 OR 2
F
Virtual 4C

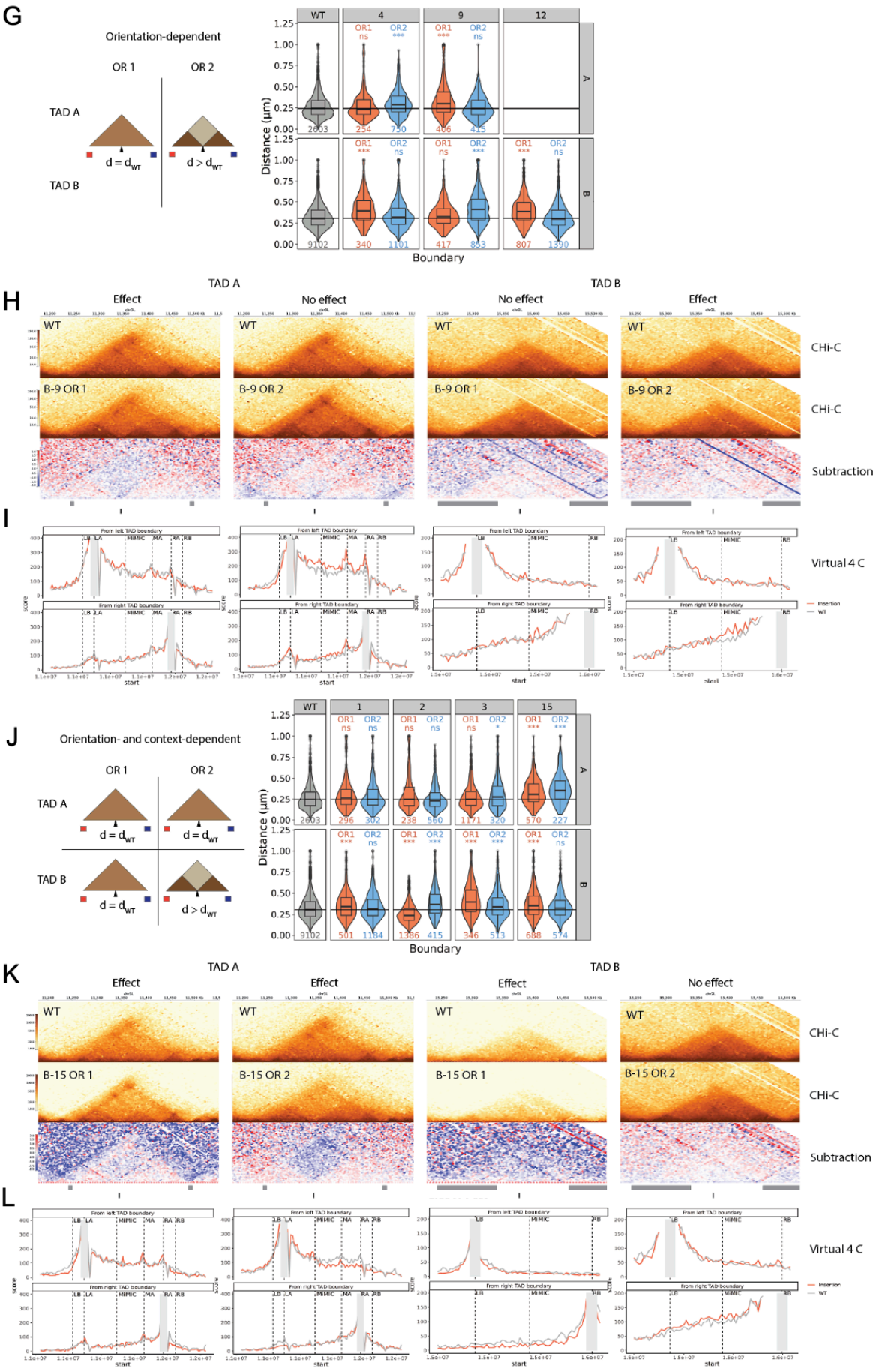

G
Orientation-dependent
OR 1
OR 2
TAD A
TAD B
d = d_WT
d > d_WT
WT
4
9
12
Distance (µm)
Boundary
H
TAD A
TAD B
Effect
No effect
WT
B-9 OR 1
B-9 OR 2
CHi-C
Subtraction
I
Virtual 4 C
J
Orientation- and context-dependent
K
B-15 OR 1
B-15 OR 2
L

Figure 43: DNA FISH, capture Hi-C and virtual 4C for boundaries with different effects.

(A) Scheme of the effect of the insertion, and DNA FISH measurements in TAD A (first row) and TAD B (second row) for always working boundaries. Insertion in the forward orientation (Or1) are in red, in the reverse orientation (Or2) in blue. The horizontal line is the median distance in the WT. 3-4 embryos were measured for each insertion; the total number of DNA FISH spots is indicated below each distribution. P values from Kolmogorov Smirnov test: * = BH-corrected p value < 0.01, **= BH-corrected p value < 0.001, *** = BH-corrected p value < 0.0001 .
(B) Capture Hi-C for a selected always working boundary at 5 kb resolution in 4-20 h embryos. Matched WT at the top, selected boundary in the middle and subtraction map on the bottom.
(C) Virtual 4C for the boundary in B, using as viewpoints the left (top) and right (bottom) acceptor TAD boundaries detected by the DNA FISH probes. Insertion line in red, WT in grey.
(D) Same as A, for context-dependent boundaries.
(E) Same as B, for context-dependent boundaries. Only one orientation (Or2) is shown.
(F) Same as C, for context-dependent boundaries. Only one orientation (Or2) is shown.
(G) Same as A, for orientation-dependent boundaries.
(H) Same as B, for orientation-dependent boundaries.
(I) Same as C, for orientation-dependent boundaries.
(J) Same as A, for orientation- and context-dependent boundaries.
(K) Same as B, for orientation- and context-dependent boundaries.
(L) Same as C, for orientation- and context-dependent boundaries.

5.8 Capture Hi-C reveals insight into the topological changes induced by the new boundaries

As shown above, DNA FISH, using probes at the TAD boundaries, gives information on the general compaction and proximity of the two TAD boundaries. When a new boundary is placed in between, it reveals if that proximity is perturbed or not, but provides no information on how that change is made – i.e. are there contacts between the new boundary and the endogenous boundary on the left or the right, or both? After the first screening by DNA FISH, we used Capture Hi-C to determine what type of topological changes is occurring in all insertion lines that had a positive effect: we assessed 24 insertions in TAD A and 13 in TAD B. The capture Hi-C data confirmed the findings from DNA FISH in close to 90% of the lines (33/37) and enabled a finer dissection of the topological disruptions upon boundary insertion.

In principle, insertions with an effect should lead to splitting of the acceptor TAD into two smaller TADs in TAD A and B; additionally, in TAD A they should obliterate the loop connecting the left and right loop anchor. However, I observed a broader spectrum of topological effects, involving both formation of sub-TADs and perturbations of any loop, as summarized in Table 8.

In TAD A, insertions with no effect by DNA FISH show either no or very mild effects in capture Hi-C, with intra-TAD interactions and loops comparable to the WT (Figure 43E). Insertions with an orientation-specific effect show fuzzier loops in the working orientation (B-3 OR2, B-4 OR2) (not shown), or a reduction in the outer loop (B-9 OR1) (Figure 43H). On the other hand, the effect of orientation-dependent boundaries was generally milder than that of 'always working' boundaries, which are mostly associated to splitting into sub-TADs and stable or reduced loops (Figure 43B). The reduction in intra-TAD interactions was also variable, and higher in boundaries working in both TADs and in both orientations (Figure 43B, Figure 43C): for instance, boundaries B-14 and B-16 completely disrupted the original intra-TAD interactions (Figure 43B).

Table 8: Effect of inserted boundaries in capture Hi-C.

TAD	*Effect*	*Lines*	*# Lines*
TAD A	No change	B-2 OR1 B-2 OR2 B-4 OR1 B-6 OR2 B-7 OR2 B-9 OR2 B-10 OR1 B-10 OR2 B-11 OR2	9
	Small local change	B-3 OR1	1
	Sub-TADs, loops strengthened	B-3 OR2	1
	Sub-TADs without loss of interaction in the original TAD, loops not perturbed	B-13 OR1 B-13 OR2	2
	Sub-TADs, loops reduced or lost	B-5 OR2 B-9 OR1 B-14 OR1 B-15 OR2 B-16 OR1	5
	Complete TAD disruption, loops lost	B-5 OR1 B-14 OR2	2
	Discordant with DNA FISH: change in DNA FISH but no clear global change in capture Hi-C	B-4 OR2 B-15 OR1	2
	Discordant with DNA FISH: change in capture Hi-C but not in DNA FISH, due to TAD disruption and loop strengthening	B-7 OR1	1
TAD B	No change	B-9 OR1 B-15 OR2	1
	Decreased distance in DNA FISH: slight increase in interactions, mild loop	B-2 OR1 B-7 OR1 B-7 OR2	3
	Interaction loss without splitting	B-15 OR1	1
	Local splitting at the insertion site	B-3 OR1 B-9 OR2	2
	2 sub-TADs	B-3 OR2	1
	4 sub-TADs + loop within the TAD	B-6 OR2	1
	Clear and sharp TAD splitting into 2 sub-TADs	B-11 OR2 B-16 OR2	2
	Clear and sharp TAD splitting into 2 sub-TADs, original TAD destroyed	B-16 OR1	1

Notably, boundaries with an effect in both orientations could display different mechanisms of topological disruption: for instance, the insertion of boundary B-15 in TAD A created sub-TADs and fuzzier (but stronger) loops in orientation 1, but led to a general decrease in intra-TAD interactions and loss of loops in orientation 2 (Figure 43K). In all instances, there was TAD splitting at the insertion site.

A notable case where topological differences are detected by capture Hi-C but not by DNA FISH is boundary B-7 OR1. Here, the insertion created two new sub-TADs between the left loop anchor and the middle anchor, but all loops within the TAD, including the loops between the left and right anchor, were strengthened (Figure 44). Since this looping interaction is still present, no change was detected by DNA FISH, highlighting the important of using Capture Hi-C.

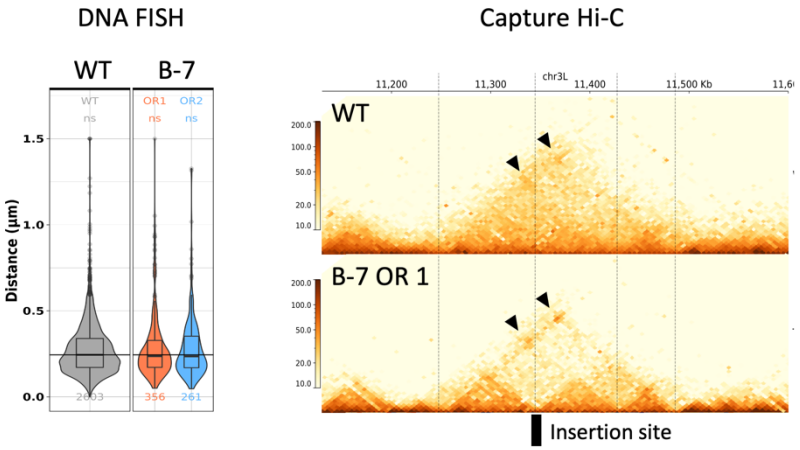


Figure 44: TAD disruption can be combined with loop strengthening, leading to no detectable change in DNA FISH. DNA FISH (left) and capture Hi-C of WT and B-7 OR 1 in TAD A. In the Hi-C matrix, the black arrows point to the loops, and the dotted vertical lines to the sites involved in the interactions.

In TAD B, insertions causing an increase in the median distance of DNA FISH probes led to a less defined TAD, with fuzzier borders and lower interaction frequency, especially towards the endogenous boundaries. However, the TAD does not split into two, in contrast to TAD A (Figure 43 B, E, K; Table 8). On the other hand, insertions which reduced the distance between probes (B-2 OR 1, B-7 OR1, B-7 OR2) strengthened the interactions within the TAD, especially between its boundaries (Table 8).

Overall, the capture Hi-C results mostly confirmed DNA FISH results, while the higher resolution could detect changes not visible by DNA FISH, and showed that the effect of insertions can be mediated by different mechanisms altering TAD compaction as a whole, or splitting it into sub-TADs or two separated TADs, or even altering its loops (Figure 43; Table 8). Moreover, they validate that not only TAD decompaction, but also compaction is possible, as shown for TAD B (B-7 in Figure 43D, B-2 OR1 in Figure 43J).

5.9 Well characterised insulators are not predictive of the inserted boundaries effect

As a working hypothesis, we assume that the inserted boundaries have intrinsic features correlated to their function in an ectopic context, and that the combination of features on the inserted and on the endogenous boundaries determines if they interact, thus causing topological effects. The presence of insulator proteins would not explain context- or orientation- dependent boundaries, but it could be sufficient for always working boundaries. Since always working boundaries generally have a stronger

effect in capture Hi-C, it would point to the most important features required to establish a strong boundary.

To test this hypothesis, I examined the occupancy of the classic well studied *Drosophila* insulator proteins BEAF-32, CTCF, CP190 and Su(Hw) based on 6–8 h ChIP data from the lab (BioProject accession: PRJEB38618), and the presence of promoters, in boundaries with different effects. All 'always working' boundaries were marked by CTCF and CP190; however, these factors are also present at all orientation-dependent boundaries, in a large fraction (75%) of context-dependent boundaries and in the only non-working boundary (Figure 45). Therefore, the presence of CTCF and CP190 is not sufficient or predictive to define a context- and orientation- independent boundary.

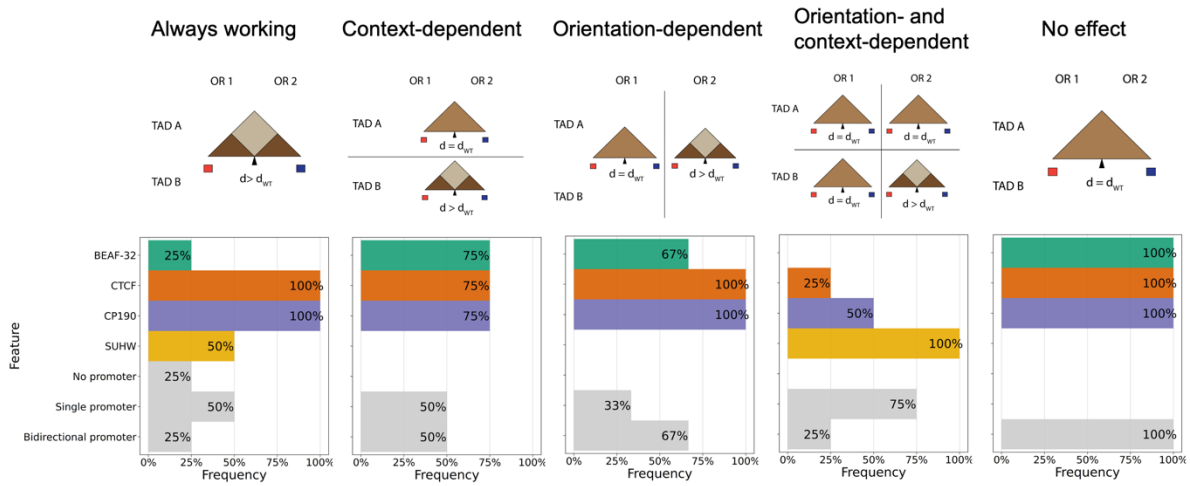


Figure 45: The binding of insulator proteins does not explain the effect of inserted boundaries.
Frequency of insulator proteins and promoter types in inserted boundaries with different effects.

I then investigated the effect of boundaries with the same combination of insulators in a single TAD. The insulator combination was not sufficient to define its effect in a given context: for instance, in TAD A the combination of CTCF and CP190 can work in an orientation-specific way (boundary B-4), in both orientations (B-5), or not at all (B-7) (Table 9), suggesting that there must be other factors binding to these boundaries. In TAD B, boundary B-4 and B-5 act as in TAD A, whereas boundary B-7 works in both orientations (Table 9). Therefore, the simple combination of known insulator binding is not predictive of an effect in a genomic location, and boundaries with the same composition can act differently in different genomic contexts.

Table 9: Boundaries with the same composition have different effects in different contexts.

Boundary	***Composition***					***Effect***			
	BEAF-32	***CTCF***	***CP190***	***Su(Hw)***	***Promoter***	***TAD A Or1***	***TAD A Or2***	***TAD B Or1***	***TAD B Or2***
B-4					Divergent	**No**	**Yes**	**Yes**	**No**
B-5					Single	**Yes**	**Yes**		**Yes**
B-7					Divergent	**No**	**No**	**Yes**	**Yes**

Some combinations and factors were more linked to a context-specific effect, but not sufficiently to define predictive rules. For instance, the combination of BEAF-32, CTCF and CP190 (boundaries B-8, B-9, B-10, B-11, B-12, B-13) had a higher chance of working in at least one orientation in TAD B, and generally did not work in TAD A (4/6 insertions, and 1 orientation-dependent). Similarly, Su(Hw) alone or in combination with CP190 (boundaries B-1, B-2, B-3) tended to work in TAD B but not in TAD A (Table 10).

Table 10: Insulator protein combinations which tend to work in a specific context (i.e. in TAD B, but not in TAD A).

Boundary	*Composition*					*Effect*			
	BEAF-32	*CTCF*	*CP190*	*Su(Hw)*	*Promoter*	*TAD A Or1*	*TAD A Or2*	*TAD B Or1*	*TAD B Or2*
B-8					Single	No		Yes	
B-9					Single	Yes	No	No	Yes
B-10					Divergent	No	No		
B-11					Divergent		No		Yes
B-12					Divergent			Yes	No
B-13					Divergent	Yes	Yes	Yes	Yes
B-1					Single	No	No	Yes	No
B-2					Single	No	No	Yes	Yes
B-3					Divergent	No	Yes	Yes	Yes

Despite the context sensitivity of most boundaries, a combination of insulators almost always worked in both TADs, regardless of the promoter and the orientation of the insertion: CTCF, CP190 and Su(Hw), present in boundaries B-14, B-15 and B-16, effectively split both acceptor TADs almost always (Table 11). This type of boundary appeared highly active, regardless of endogenous boundaries; this may be due to the high number of components, which could engage in functional interactions with a large variety of factors in the endogenous boundaries, as well as the accumulation of many insulators at the same site, which could work as a scaffold to load or block other components.

Table 11: The CTCF-CP190-Su(Hw) combination is associated to working boundaries.

Boundary	*Composition*					*Effect in DNA FISH*			
	BEAF-32	*CTCF*	*CP190*	*Su(Hw)*	*Promoter*	*TAD A Or1*	*TAD A Or2*	*TAD B Or1*	*TAD B Or2*
B-14						Yes	Yes	Yes	Yes
B-15					Single	Yes	Yes	Yes	No
B-16					Single	Yes	Yes	Yes	Yes

Taken together, this study demonstrates that TAD boundaries can work in ectopic contexts by interacting with other elements in the genomic context, and some boundary compositions are more effective at establishing those interactions. However, these cases do not explain the underlying rules.

5.10 The promoter does not establish the orientation specificity of inserted boundaries

Apart from insulator proteins, promoters are extensively present at TAD boundaries: 80% of *Drosophila* boundaries have a promoter (Ramírez *et al.*, 2018), and the relationship between TAD boundaries and the transcriptional machinery is conserved throughout evolution (Szabo, Bantignies and Cavalli, 2019). Therefore, I examined if the presence of promoters was correlated to boundary activity, and if their orientation could explain orientation-dependent boundaries.

Our data does not support a major role of promoters in the establishment of a boundary: a non-promoter boundary (B-14) worked in all TADs and orientations. Furthermore, a promoter in a single orientation was not necessary or sufficient to determine orientation specificity: boundaries working in both orientations could have a single promoter (boundary B-15 in TAD A, boundary B-2 in TAD B, boundary B-16 in both), and some orientation-specific boundaries had divergent promoters (boundary B-4 in TAD A and B, boundary B-12 in TAD B) (Table 12).

Table 12: The promoter does not determine orientation specificity.

Boundary	*Composition*					*Effect in DNA FISH*			
	BEAF-32	*CTCF*	*CP190*	*Su(Hw)*	*Promoter*	*TAD A Or1*	*TAD A Or2*	*TAD B Or1*	*TAD B Or2*
B-2					Single	No	No	**Yes**	**Yes**
B-3					Divergent	**No**	**Yes**	Yes	Yes
B-4					Divergent	**No**	**Yes**	**Yes**	**No**
B-12					Divergent			**Yes**	**No**
B-14						**Yes**	**Yes**	**Yes**	**Yes**
B-15					Single	**Yes**	**Yes**	Yes	No
B-16					Single	**Yes**	**Yes**	**Yes**	**Yes**

In summary, the presence of a promoter did not have a major contribution on boundary function or on orientation specificity. Of note, we included all promoters in this analysis regardless of their activity in the endogenous location. We note that we could not assess if their activity was preserved after insertion in the acceptor TAD as it is impossible to distinguish short RNA reads coming from the ectopic or the endogenous location.

The failure of classic insulators or a promoter to explain the effect of the insertion suggests that more contextual elements are required for boundary activity: for the 'always working' boundaries the basic conditions could be satisfied in both TADs, whereas for context- and orientation-dependent boundaries they may be satisfied only in specific cases. Moreover, since orientation dependency is not due to the promoter, it may lie in other directional elements like non-palindromic motifs or syntax between multiple motifs.

5.11 Combinations of well characterised insulators binding to the inserted boundaries and endogenous boundaries are not predictive of a working boundary pair

Since the capture Hi-C data showed that inserted boundaries have a higher interaction with some acceptor TAD boundaries (Figure 43), I focused on the composition of acceptor TAD boundaries to evaluate the genomic context.

In TAD A, which has multiple boundaries on the left and on the right of the insertion, the inserted boundary did not interact with all available boundaries, but had one or more preferential partners. To identify the partners of the inserted boundary, I performed virtual 4C from the normalized capture Hi-C of the matched WT control and insertion line, using as reference sites the boundaries at the centre of the DNA FISH probes (left and right loop anchor) (Figure 43 C, F, I, L) and the MiMIC insertion site. By comparing the profile of the WT and the insertion line and integrating it with the interaction frequency map and the insulation score, I identified a total of 79 working boundary pairs, out of 187 potential pairs, formed by the inserted boundary and one of the acceptor TAD A or TAD B boundaries (Figure 46A). The working boundary pairs include only insertions with an effect, as those with no effect do not interact with acceptor TAD boundaries.

Next, I annotated the inserted and acceptor TAD boundaries for the presence of insulator protein binding (BEAF-32, CTCF, CP190 and Su(Hw)) at 6–8 h, and the presence of a promoter. I considered all potential boundary pairs and the actual (working) pairs, and classified them based on their insulator and promoter composition, grouping them by the composition of the inserted boundary (Figure 46B). If a specific combination of insulator proteins and promoter in the acceptor TAD boundary and in the inserted boundary always worked, it could explain why the insertion had an effect. Since this approach is agnostic to features outside the inserted and the acceptor TAD boundaries, the same combination of insulator proteins must always work in TAD A and B.

This method only resulted in four potential combinations, all tested only in TAD B: on the first boundary, (1) BEAF-32, CTCF, Su(Hw) and a promoter or (2) CTCF, CP190 and a promoter, and on the second boundary (1) CP190, Su(Hw) and a promoter, or (2) CTCF, CP190 and Su(Hw) (Figure 46B). These combinations are present in only two boundaries and one TAD, suggesting that this feature could be selected by chance; thus, it is not a strong predictor. In general, the presence of insulator proteins and a promoter in single boundaries or boundary pairs is not sufficient to explain TAD boundary function – i.e. the ability to form a new interaction.

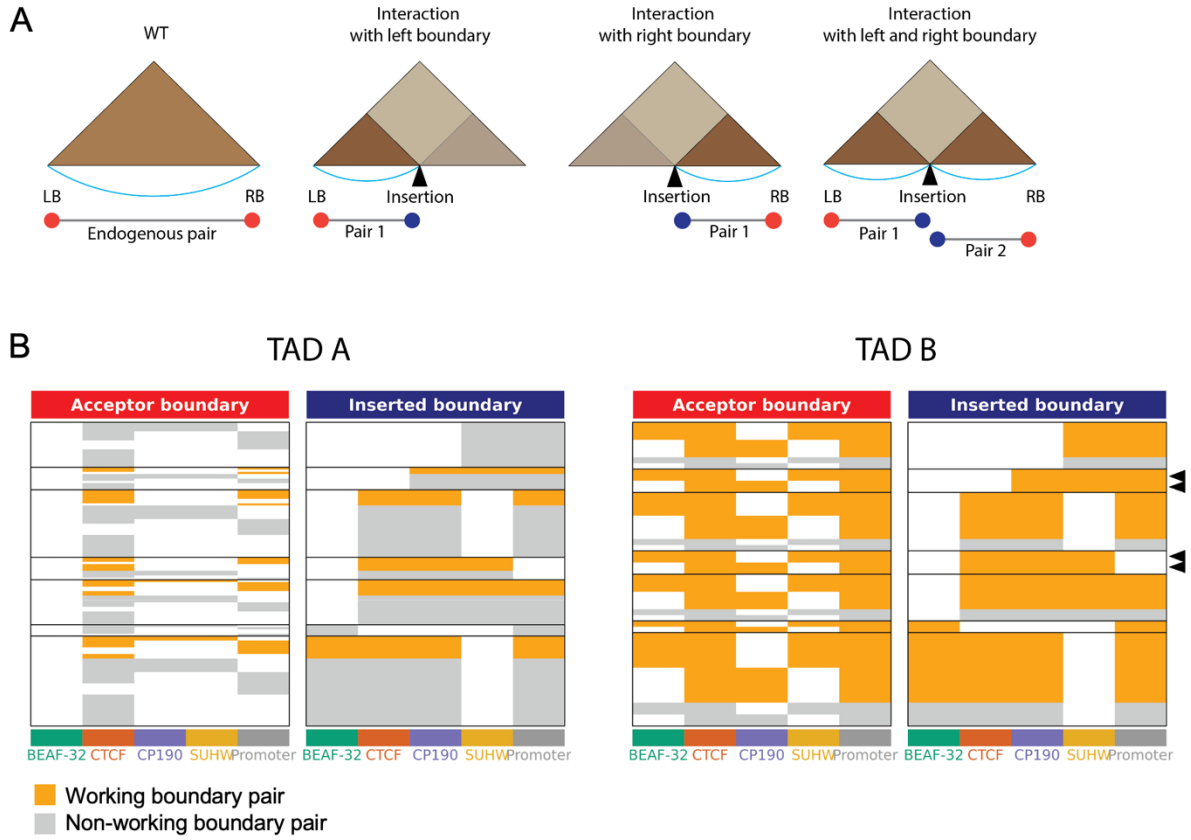


Figure 46: The pairing of boundaries with specific insulator proteins combinations does not explain boundary effect.

(A) Scheme of the detection of boundary pairs formed by the insertion, based on the capture Hi-C. If the interaction frequency between the insertion and one of the acceptor TAD boundaries is higher, the inserted boundary and this acceptor TAD boundary form a pair. Acceptor TAD boundaries are in red, inserted boundaries in blue.

(B) Composition of pairs of acceptor and inserted TAD boundaries. Each row is one pair, and each column one feature (insulator/promoter) of the acceptor (in red) or the inserted (in blue) TAD boundary; a coloured space indicates that the boundary has the feature. True boundary pairs, i.e. pairs formed by working boundaries, have feature coloured in orange; potential pairs which are not established and do not have an effect are in grey. The rows are grouped by the features of the inserted boundary. For a composition to be associated with working boundaries, all the pairs with that composition must be working (i.e. in orange). In TAD A (left), there is no composition associated to a working boundary pair. In TAD B, some compositions (indicated by black arrows) are associated exclusively to working pairs, and present in more than 1 case.

5.12 Other known insulator features do not explain boundary function

Since the presence of insulators and promoters was not sufficient to deduce rules for the working pairing of TAD boundaries, I expanded the annotation for the left and right boundary of each pair by including additional features for the known insulator proteins BEAF-32, CTCF, CP190 and Su(Hw) and also included more recently described insulators Ibf1 (Cuartero *et al.*, 2014), Ibf2 (Cuartero *et al.*, 2014), Pita (Maksimenko *et al.*, 2015), ZIPIC (Maksimenko *et al.*, 2015), Zw5 (Gaszner, Vazquez and Schedl, 1999), GAF (Ohtsuki and Levine, 1998; Melnikova *et al.*, 2004) and M1BP (Bag *et al.*, 2021). Such features included:

- the number of ChIP-seq peaks within the boundary, the coverage under the peak summit and the signal in the whole boundary region
- the presence, number and orientation of motifs
- the presence, number, orientation and shape (broad or focused) of all promoters, and of active promoters (Mikhaylichenko *et al.*, 2018).

Moreover, I recorded the chromatin state by various metrics:

- the number of DHSs (Reddington *et al.*, 2020), as a proxy for binding of other TFs
- the presence of chromatin marks in the endogenous location (Bonn, Zinzen, Girardot, *et al.*, 2012)
- the type of TAD at the two sides of the boundary (active chromatin, null, PcG, heterochromatin) (Sexton *et al.*, 2012).

This way, I included the features present in promoter-dependent and promoter-independent boundaries (Ramirez 2018), and also additional elements that may further stratify these two classes. In addition, I reported the distance between boundaries, which contributes to the baseline contact frequency (Smith *et al.*, 2016).

Given the limited number of boundary pairs assessed experimentally, to increase the power I expanded the analysis to all boundary pairs annotated at a genome-wide level. This extensive set of features across genome-wide boundaries did not lead to the discovery of rules for boundary function.

5.13 Using conservation to identify new motifs in functional boundaries

Since known boundary features were not sufficient to explain boundary function, I hypothesized that additional factors must be involved. Due to the orientation- and context- dependency of a large set of boundaries, the new features should be directional and should interact with features on the other element of the boundary pair: since we excluded the contribution of promoters, they could be motifs for other TFs not annotated as insulator proteins, or for unknown binding partners.

Due to the large number of TFs, a motif search under all available genome-wide ChIP-seq data would lead to overestimating TF binding, and it would introduce a large number of features largely absent from the small experimental set. Moreover, it would not detect motifs with unknown binding partners. An alternative approach based on the embryonic DHSs is agnostic to the identity of the bound TF, and points to regions accessible at our stages of interest; however, being a very diverse set from contribution from many TFs, it would be too heterogeneous for motif search.

To overcome this, I searched for motifs within DHS regions (from Reddington *et al.*, 2020), expanded by ± 100 bp, that overlap the inserted and acceptor TAD boundaries. The DHS peak expansion was performed to ensure that the DHS regions include the peaks of known insulators. Since insulators and the sequences of TAD boundaries are more evolutionarily conserved than other regulatory sequences

(Nègre *et al.*, 2010; Acemel and Lupiáñez, 2023), we reasoned that the most conserved bases within those regions should correspond to functional motifs, since they would be under positive selection.

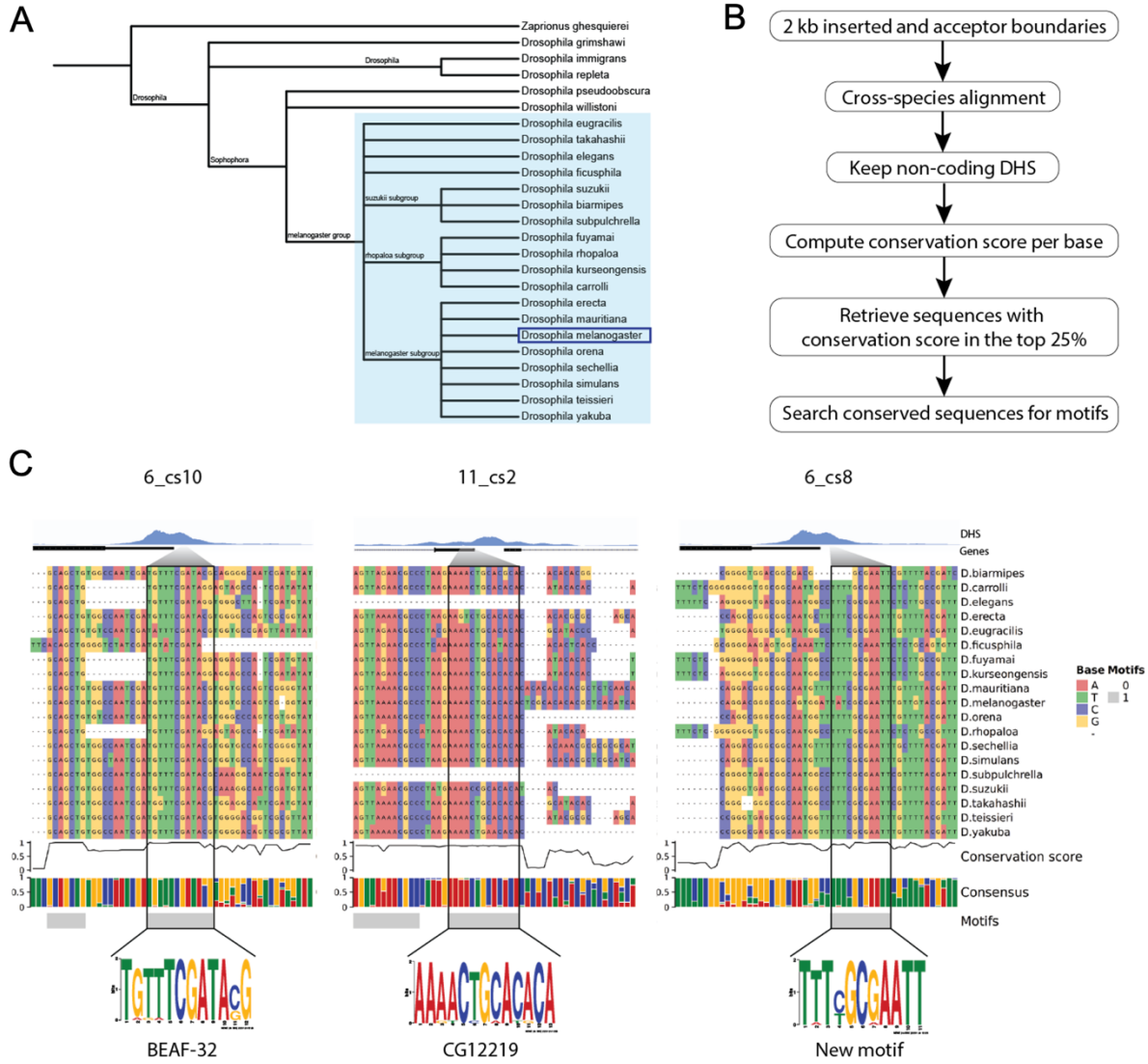


Figure 47: Conservation-based* de novo *motif discovery.

(A) Phylogenetic tree of the *Drosophila* genus. The *D. melanogaster* clade is indicated in light blue.

(B) Workflow for the conservation-based *de novo* motif discovery.

(C) Examples of known and new motifs found based on the conservation score. Motifs are searched in non-coding DHS regions (as shown by the DHS and gene track above), by aligning the 19 species in D. *melanogaster* clade and computing the conservation score and the consensus. The regions with the highest conservation across ≥ 6 bases are potential motifs (grey track). The PWM is calculated based on the consensus.

Based on this hypothesis, I evaluated evolutionary conservation in the *D. melanogaster* clade (19 species) (Figure 47A) in collaboration with Dr. Aaron Comeault (Bangor University), who performed the alignments. As expected, the most conserved regions were the coding sequences; however, non-coding regions show different degrees of conservation, which allowed me to select the most conserved bases (Figure 47C). The conservation score for each position of the alignment was defined as the mean of the probability of having the same base as *D. melanogaster* and that of having any base in that

position (Methods section 2.16.1). To account for differences in the overall conservation of the boundaries, I selected bases with a conservation score in the top 75% of the conservation score in DHSs in non-coding sequences separately for each boundary (Figure 47B). Each group of ≥ 6 consecutive conserved bases was considered a conserved sequence, and a potential motif (Figure 47C). Since many known motifs have up to three consecutive bases with a low score in the Position-Weight Matrix (PWM), I allowed up to three consecutive bases with lower conservation score within a conserved sequence. This selection yielded 279 conserved sequences, mostly between 6 and 30 nucleotides in length.

5.14 Known and new motifs are retrieved among conserved sequences in inserted and acceptor TAD boundaries

To identify the potential factors that could be binding to the 279 conserved sequences, I used FIMO (Grant, Bailey and Noble, 2011) to do a motif search using motifs from cisBP (Weirauch *et al.*, 2014), in addition to motifs from our insulator ChIP-seq data (BioProject accession PRJEB38618) and a collection of promoter motifs (Ohler, 2006). Among the 279 conserved sequences, 214 (76%) contained known motifs for (i) insulators (11%), (ii) insulators and promoters (9.6%), (iii) other TFs alone (37%) or (iv) in combination with promoters (15%), or (v) promoters alone (11%) (Figure 48A). Despite the large collection of motifs used, 65 conserved sequences could not be assigned to known motifs.

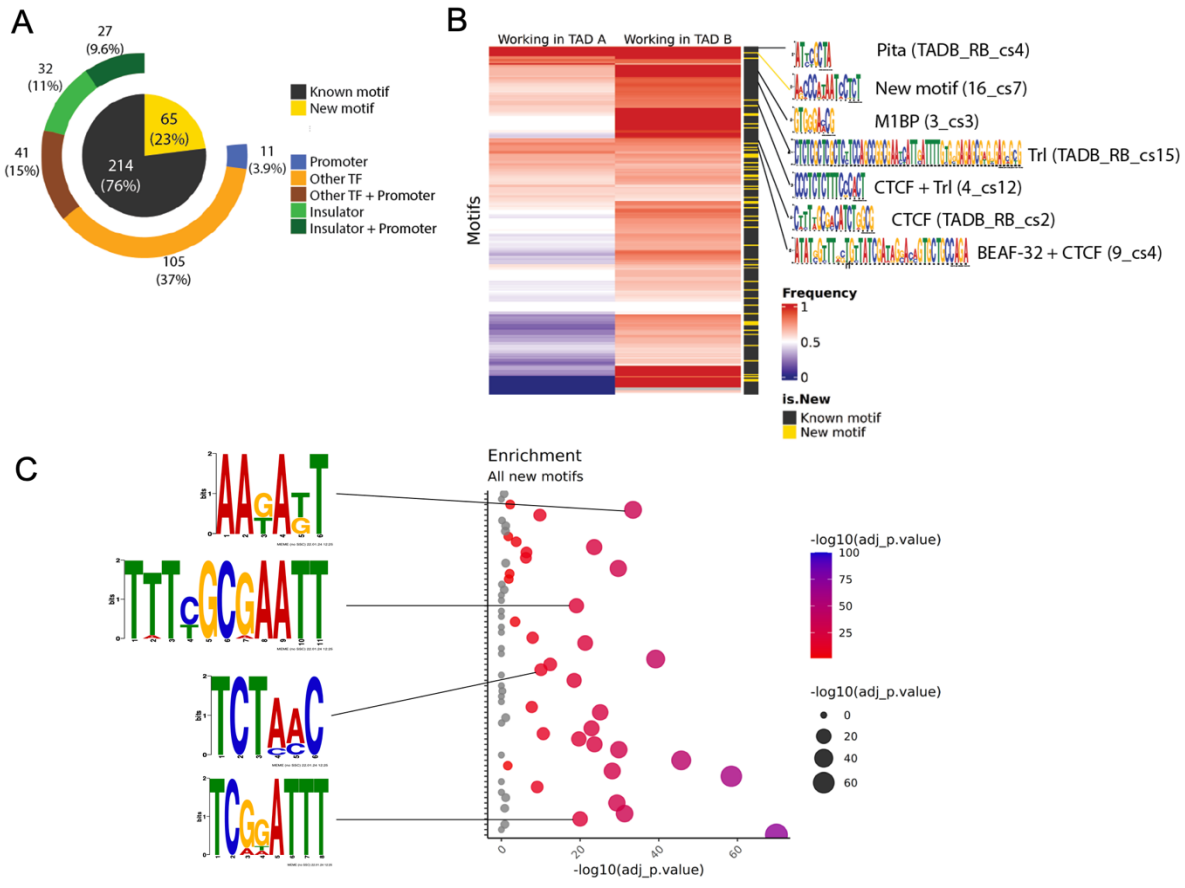


Figure 48: **De novo** ***motifs include known and new motifs enriched at genome-wide TAD boundaries and in working experimental boundaries.***

(A) Composition of conservation-based motifs. For sequences containing one or more known motifs, the type of motif(s) (insulator, promoter, generic TF) is reported.
(B) Frequency of working boundaries among those with a specific motif (on the rows) in TAD A (left) and TAD B (right). Red indicates that all the inserted boundaries with the selected motif have an effect in the TAD, blue that none has an effect, and white that 50% of the boundaries have an effect. On the right, known motifs are indicated in black and new motifs in yellow; some examples of motifs are reported on the far right (PWM and motif name).
(C) Genome-wide enrichment of the 65 new motifs, with some example PWM.

To understand the potential contributions of single motifs, I performed a motif search on all DHSs within the tested boundaries using motifs matching to ≥ 1 conserved sequence, and the 65 non-assigned sequences as potential new motifs. Motifs potentially impacting boundary function should be associated with working boundaries in both acceptor TADs. Very few motifs, such as the known insulator protein Pita and other new motifs, are present in elements that always work in both TADs; the vast majority – including BEAF-32, CTCF, Trl, M1BP – appeared to work in TAD B more often than in TAD A, and about 1/3 to work only in TAD B (Figure 48B). This is consistent with TAD B being more sensitive to perturbations. The distribution of known and new motifs was balanced across working and non-working boundaries, with a slight enrichment for known motifs among the always working boundaries.

If a motif is truly required to make an active TAD boundary, it should display genome-wide enrichment at endogenous TAD boundaries. In fact, I confirmed that all known insulator proteins, with the exception of Trl, are enriched at TAD boundaries. Among all the non-insulator motifs (495 in 279 conserved sequences), 177/412 TF motifs from cisBP, 14/18 promoter motifs and 34/65 new motifs were significantly enriched in DHSs overlapping endogenous TAD boundaries genome-wide ($p < 0.05$). The enrichment of new motifs in TAD boundaries genome-wide is displayed in Figure 48C. Among the conserved 279 sequences, 83.8% contained at least one motif enriched at TAD boundaries.

Upon refinement of the initial motif selection, I obtained a total of 174 known and 15 new motifs enriched at TAD-boundaries genome-wide, and linked to working boundaries in our experiments.

5.15 Pairs of conserved sequences in a defined orientation are linked to working TAD boundaries in our experiment and genome-wide

Since only the reciprocal orientation of two motifs on both elements of a boundary pair could explain orientation-dependent boundary function, I examined how often each motif pair from the conserved motif set was present in a convergent, divergent or tandem orientation in working vs non-working boundary pairs. I then selected the motifs and orientations always working in TAD B, where the baseline frequency of a working boundary pair was 80%, and in > 80% of the cases for TAD A, where the baseline working probability was 32%. This way, I identified 57 motif pairs associated to working boundaries in a specific orientation, 33 of which were always working both in TAD A and TAD B (Figure 49, Table 13). I also imposed that those motif pairs must be present genome-wide in the selected orientation for > 25% of the cases (i.e., more than random). Among them, 34 motif pairs were significantly enriched in the selected orientation genome-wide (Fisher test, adjusted p value < 0.05), and 19 of them had it as the main orientation, which suggests that these interactions play a role in genome topology.

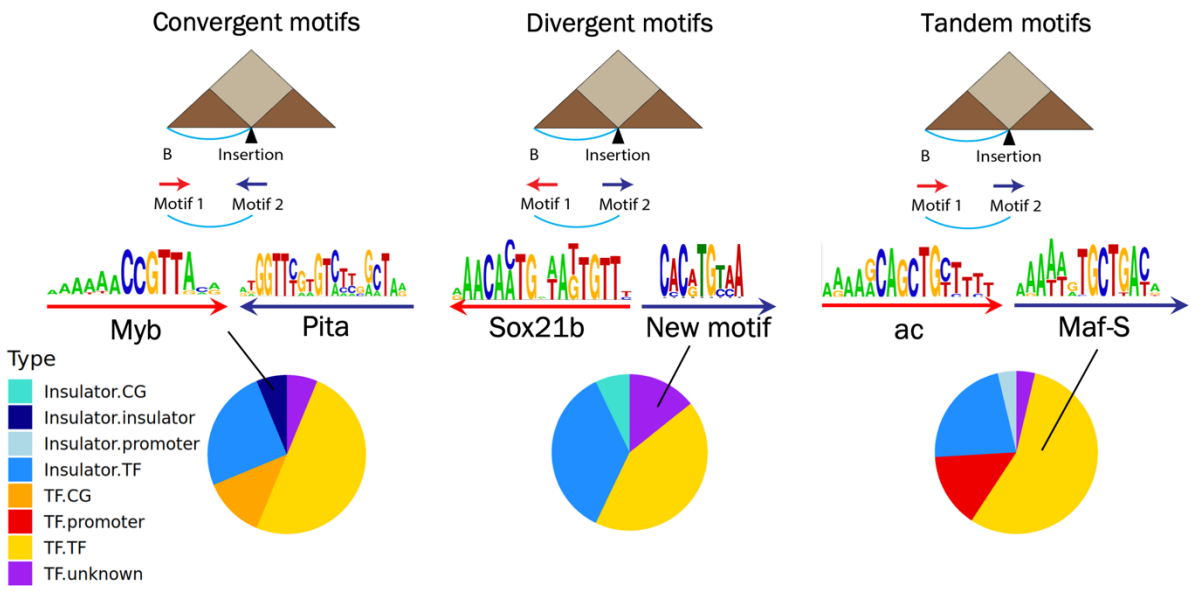
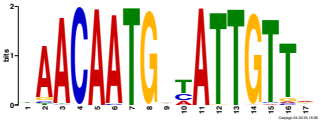
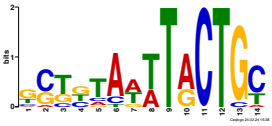
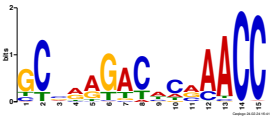
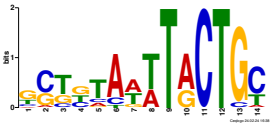
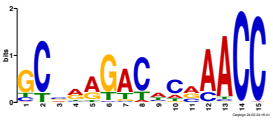
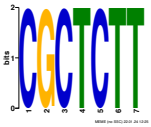
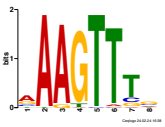
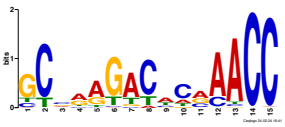
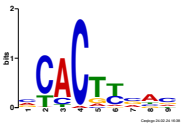
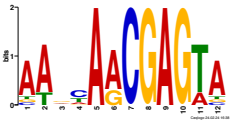
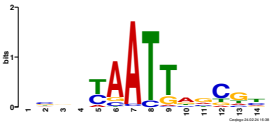
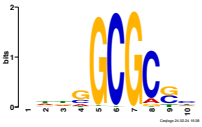
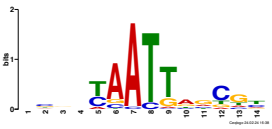
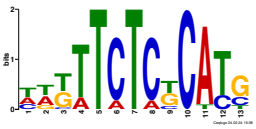
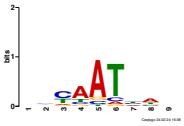
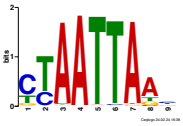
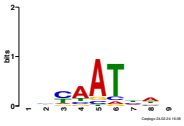
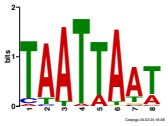
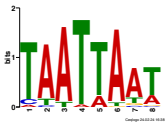
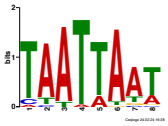
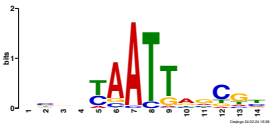
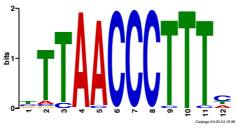
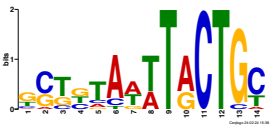
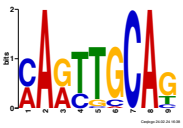
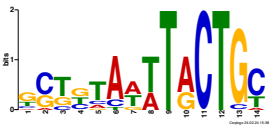
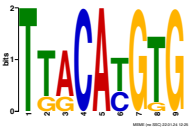
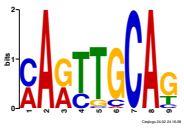
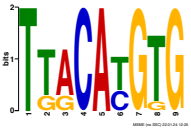
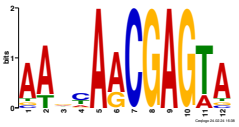
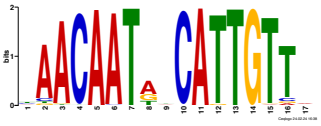
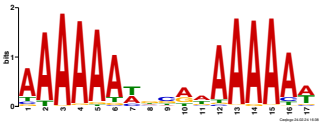
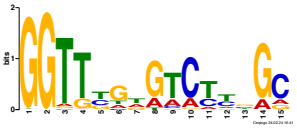
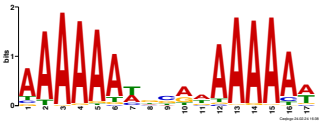
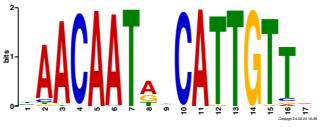
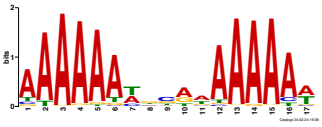
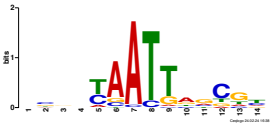
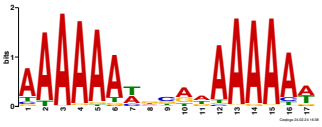
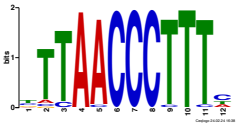
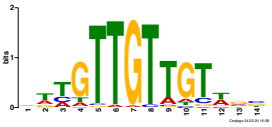
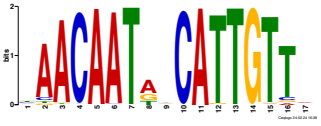
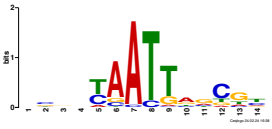
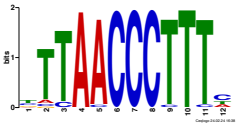


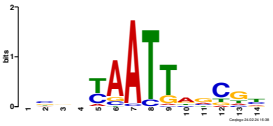
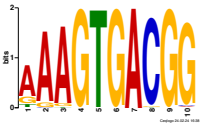
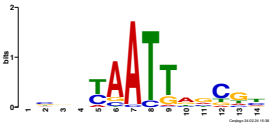
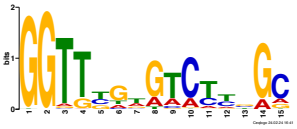
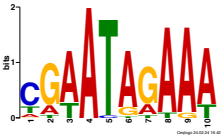
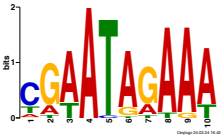
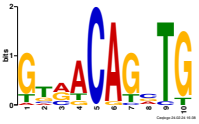
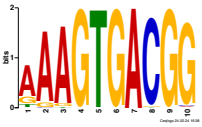
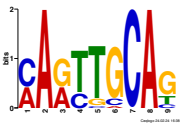
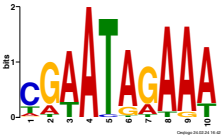
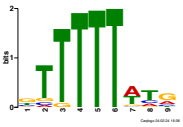
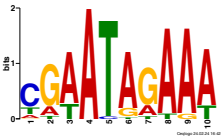
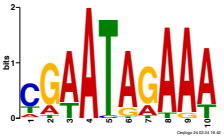
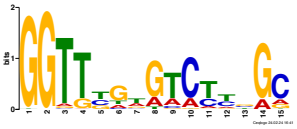
Figure 49: Examples and composition of oriented motif pairs.

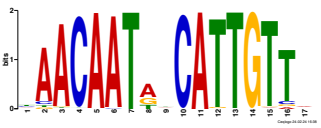
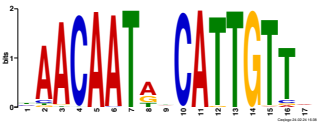
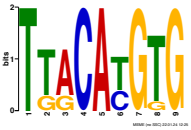
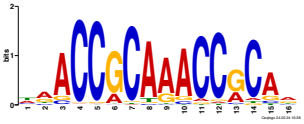
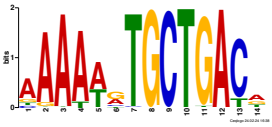
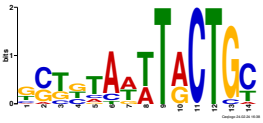
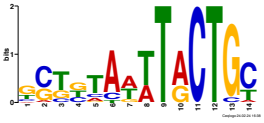
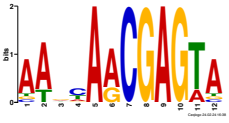
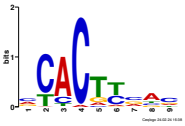
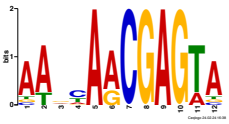
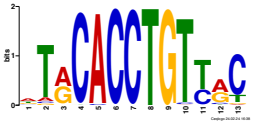
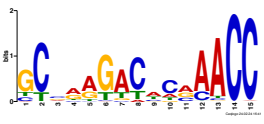
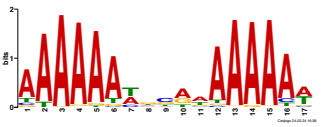
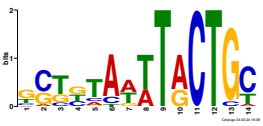
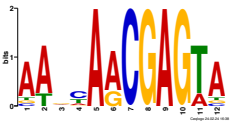
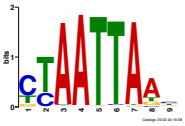
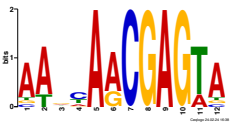
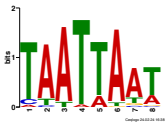
Table 13: Motif pairs associated to working boundary pairs.

PWM 1	*PWM 2*	*TF 1*	*TF 2*	*Orientation*
		ac	bbx	Convergent
		ac	CG4404	Convergent
		Q8IP13	Pita	Convergent
		CG12605	CG4404	Convergent
		Myb	Pita	Convergent
		Vsx1	TADA_LA_cs13	Convergent

		CG43689	Pita	Convergent
		bbx	scro	Convergent
		bbx	CG12155	Convergent
		exex	E2f2	Convergent
		exex	CG10904	Convergent
		CG15696	Vsx1	Convergent
		CG15696	Dbx	Convergent
		Vsx1	Dbx	Convergent
		Dbx	Dbx	Convergent
		bin	exex	Divergent
		bin	Kr	Divergent
		CG4404	Myb	Divergent
		CG4404	TADA_LA_cs14	Divergent

		Sox21b	Myb	Divergent
		Sox21b	TADA_LA_cs14	Divergent
		CG12155	Pita	Divergent
		Su(Hw)	bbx	Divergent
		CG2052	Pita	Divergent
		CG2052	bbx	Divergent
		CG2052	exex	Divergent
		CG2052	Kr	Divergent
		Aef1	bbx	Divergent
		bin	maf-S	Divergent
		ac	bbx	Tandem plus
		ac	exex	Tandem plus
		ac	Kr	Tandem plus

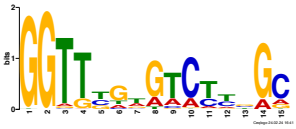
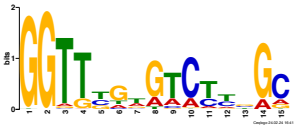
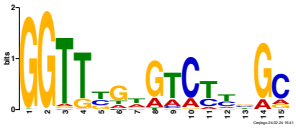
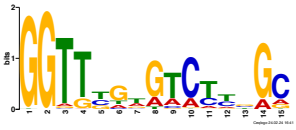
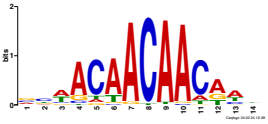
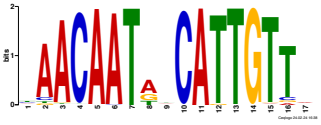
		exex	CG7818	Tandem plus
		exex	Pita	Tandem plus
		Q8IP13	GAATAGAAA	Tandem plus
		CG12605	GAATAGAAA	Tandem plus
		Fer3	CG7818	Tandem plus
		Myb	GAATAGAAA	Tandem plus
		rn	GAATAGAAA	Tandem plus
		Vsx1	GAATAGAAA	Tandem plus
		Dbx	Pita	Tandem plus
		ac	Maf-S	Tandem plus
		bbx	CG4404	Tandem plus
		bbx	Myb	Tandem plus

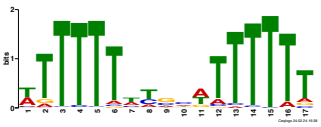
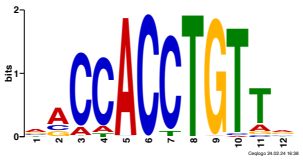
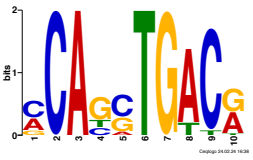
		bbx	vnd	Tandem plus
		bbx	TADA_LA_cs14	Tandem plus
		CG4404	RunxA	Tandem plus
		CG4404	Maf-S	Tandem plus
		bin	CG4404	Tandem minus
		CG4404	CG12155	Tandem minus
		Sox21b	scro	Tandem minus
		Sox21b	CG12155	Tandem minus
		wor	Pita	Tandem minus
		CG2052	CG4404	Tandem minus
		CG12155	Vsx1	Tandem minus
		CG12155	Dbx	Tandem minus

5.16 The presence of non-oriented sequences is not sufficient to infer rules

Since one of the motif pairs (ac-bbx) was present with high frequency in two different orientations (Table 13), I examined if some motif combinations could work independently of their relative orientation. Thus, I calculated how frequently boundary pairs with one motif on the left side and another on the right could work in TAD A and B. Using the same cut-offs as in 5.15, only 10 motif pairs were strongly associated to working boundaries (Table 14). The insulator protein Pita was strongly represented among those pairs, including homotypic Pita-Pita interaction as well as many heterotypic interactions (Table 14). However, the genomic frequency of those pairs was not significantly higher than the product of the frequencies of the two TFs alone, which implies that those interactions may not have a role at a genome-wide level.

Table 14: Non-directional motif pairs enriched in working boundaries.

PWM 1	*PWM 2*	*TF 1*	*TF 2*
		Pita	Pita
		Pita	Promoter motif GAATAGAAA
		Myb	Pita
		wor	Pita
		CG4404	Pita
		ac	bbx
		Aef1	bbx
		bbx	vnd

		CG2052	bbx
		Kah1	Fer2

5.17 Motif directionality explains most experimental boundaries

To assess the specificity and sensitivity of the set of 57 oriented motif pairs in explaining our working experimental boundary pairs, I examined how many working and non-working boundary pairs had one of the motif pairs. The motif set could explain 80% (63/79) of the working pairs (Figure 50A), and only 8% (9/108) of the non-working pairs (Figure 50B). This indicates that the motif set can explain most boundaries, with < 10% of false positives.

Based on the rules from our experimental data, I also investigated how many boundary pairs I could predict genome-wide. Given the high number of motif combinations, and the fact that I could test experimentally only a small part, the motif selection might have predicted only a very small number of genome-wide boundary pairs. Notably, the motif set explained 47% of TAD boundary pairs genome-wide (878/1852) (Figure 50C).

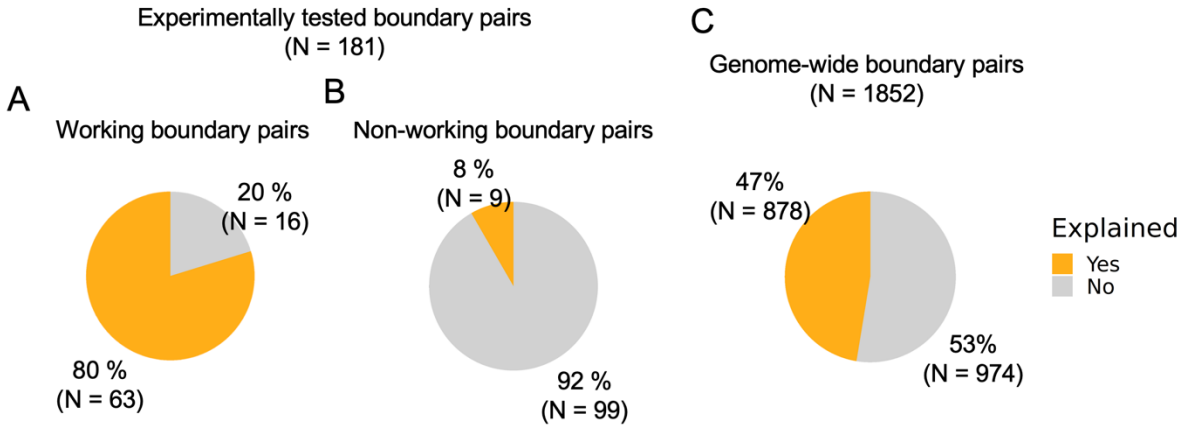


Figure 50: Motif pairs explain the effect of experimentally inserted boundaries and they are found in ~50% of genome-wide boundaries.

(A) Sensitivity of the set of motif pairs. Percentage of working experimental boundary pairs correctly predicted as working (with ≥ 1 motif pair, in orange), and with no motif pairs (in grey).

(B) Specificity of the set of motif pairs. Percentage of non-working experimental boundary pairs incorrectly predicted as working (with ≥ 1 motif pair, in orange), and with no motif pairs (in grey).

(C) Percentage of genome-wide boundaries with ≥ 1 motif pair (in orange).

5.18 Three-way interactions contribute to TAD boundaries to a small extent

After investigating the contribution of motif pairs with one motif per boundary, I examined if combinations of > 2 motifs could be important for pairing of TAD boundaries. To derive genome-wide rules, I assessed how often two motif pairs from section 5.15 were both present at the same boundary pair. Motif pairs with a co-occurrence close to 100% could contain very similar motifs. For instance, the motif pairs Bbx-Myb and Bbx-Scro, and Sox21-Myb and Sox21-Scro, have 98% co-occurrence due to the high similarity of Myb and Scro motifs, as both contain a CAATT sequence, and most likely represent the same factor. Motif pairs with 50–90% co-occurrence were more likely to form a triplet: for example, convergent Exex (a homeodomain TF) and CG10904 motifs were present with convergent Exex and CG7818 in 71.4% of the cases, and CG10904 and CG7818 have different motifs, pointing to a triplet with Exex in the forward orientation on the left boundary, and CG10904 and CG7818 in reverse orientation on the right boundary (Figure 51A). Other cases include:

- Ac in forward orientation and CG2052 in reverse orientation on the left boundary, and Kr in forward orientation on the right boundary (60% co-occurrence) (Figure 51B)
- CG12605 in forward orientation on the left boundary, and Pita in reverse orientation and the promoter motif GAATAGAAA in forward orientation on the right boundary (62.5% co-occurrence) (Figure 51C).

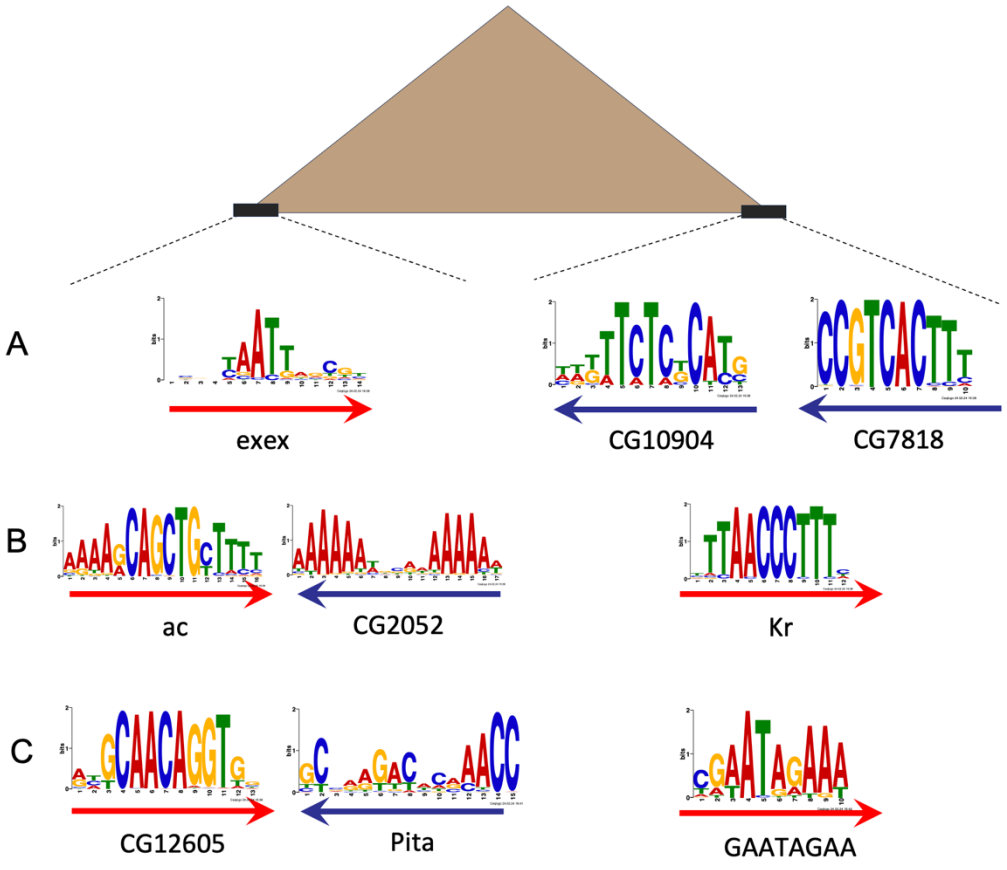


Figure 51: Combinations of 3 motifs in the left and right TAD boundary.

In cases of co-occurrence close to 50%, the motif triplet could help stabilize boundary interactions, but may not be strictly required. Overall, there were few motif triplets involving a small set of single motifs, which indicated that pairs are the main contributors.

5.19 Motifs involved in boundary function include known insulators and transcriptional regulators

Motif pairs mostly included insulators and other TFs. Among the well characterised insulators, Pita is the most prominent, followed by Su(Hw) and the GATA family (Trl), whereas BEAF-32 and CTCF were not present in a sufficient number of working boundaries to be selected among the top motifs. However, the motif for the Myb protein, which colocalizes with BEAF-32, CTCF and CP190 and shown to be necessary for BEAF-32 loading (Santana *et al.*, 2020), was highly represented in the motif pairs in working boundaries, along with its cofactor E2f2 (Lewis *et al.*, 2004) (Figure 49, Table 13). As well as being more predictive than the motifs of co-localizing insulator proteins, the Myb motif also had a clear directionality preference both in the experimental set and genome-wide, not observed for the directional motif of CTCF (Rowley *et al.*, 2017). Such directionality preference for the Myb motif can intrinsically not be present for the palindromic BEAF-32 motif or in the indirect DNA binding protein CP190. Due its role as a loading and stabilizing factor (Santana *et al.*, 2020), Myb may enhance the magnitude and residency time of its recruited factors, fostering boundary function in our experimental setup and genome-wide.

As well as contributing to insulator function, Myb stabilizes the repressive H3K27me3 PcG domain in *Drosophila* (Santana *et al.*, 2020), which tends to have strong transcription-independent boundaries separating them from active TADs (Ramírez *et al.*, 2018). The association with PcG domains was also suggested by the presence of the Aef-1 motif, enriched in Polycomb response elements (Orsi *et al.*, 2014), and of CG7878 (also known as Mettl14), which colocalizes with H3K27me3 and mediates its demethylation by recruiting the lysine demethylase KDM6B (Dou *et al.*, 2023).

Other elements detected in motif pairs include motifs for ubiquitous (Ac, CG4404, MafS) and cell- and tissue-specific TFs (Dbx, Vsx, Sox21b, Exex, Bin, Rn, RunxA, Vnd, Scro, Wor, Kr) (Table 13). Despite the fact that no specific boundary-related roles are known for these TFs, their motifs are enriched at TAD boundaries genome-wide and in the working experimental set. This indicates that the motifs could be shared by other uncharacterized proteins with boundary functionality.

Finally, the motif pair selection includes two new conserved motifs with unknown binding partners (TADA_LA_cs13, TADA_LA_cs14) (Table 13, Figure 52).

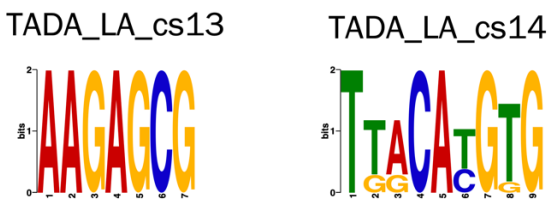


Figure 52: PWM of the two new conserved motifs.

6 Discussion II

Finding rules of TAD boundaries with a perturbation-based approach in an endogenous setting

In this project, our aim was to derive rules for TAD boundary pairing in *D. melanogaster*, based on the outcome of the insertion of endogenous boundaries into ectopic acceptor TADs. A selection of boundaries representative of the natural boundary diversity *in vivo* was inserted in two different orientations into two TADs with different structure, boundary and gene expression, and the effect of the insertion was measured by DNA FISH and capture Hi-C. The effect was then linked to features of single boundaries and boundary pairs, to find the drivers of boundary interactions, which I confirmed by inspection of endogenous boundaries genome-wide.

Most of the ectopic boundaries had a topological effect in the insertion site, and in some cases this was in an orientation- and context-specific manner, showing that boundary function depends both on intrinsic boundary features and on the context, and that it involves directional elements. The directional elements in the inserted boundaries should recognize other directional elements in the acceptor TAD boundaries. The binding of known insulator proteins on the single boundary or on a pair, being orientation-invariant, does not explain orientation-dependent boundaries. Among oriented elements, promoters, which are present at ~ 80% of *Drosophila* boundaries (Ramírez *et al.*, 2018), could also not distinguish between working and not working boundaries, being neither necessary nor sufficient to establish a boundary. Using evolutionarily conserved sequences to discover further elements involved in boundary pairing, I was able to identify motif pairs associated to working boundary pairs in the experimental set. The motif pairs are mostly composed of insulators and TFs, and they also include two new motifs not associated to known elements. This set of motif pairs accurately predicts experimentally tested boundaries, and explains close to 50% of the boundary pairs genome-wide.

6.1 Pros and cons of using an endogenous setting

6.1.1 Why use endogenous boundaries?

Arguably, using a set of synthetic boundaries with a defined arrangement of insulator motifs would simplify the search for minimal elements responsible for boundary function. However, to date no single insulator protein has been shown to be widely responsible for boundary function in *Drosophila* (Ramírez *et al.*, 2018; Kaushal *et al.*, 2021; Cavalheiro *et al.*, 2023; Kahn *et al.*, 2023), and the binding of insulator proteins is not restricted to TAD boundaries (Schwartz *et al.*, 2012; Cavalheiro *et al.*, 2023; Messina *et al.*, 2023). Therefore, a fully synthetic boundary would have a low chance of being functional, and it would still require a considerable investment in the design, with the combination of

multiple insulator motifs, spaced by an appropriate number of bases from either inert DNA, or DNA with a similar base composition to TAD boundaries but no known motifs.

By using endogenous boundaries, I had a positive control for their boundary activity since they work in their endogenous location, and I could bypass more complex and risky design choices. To mitigate confounders due to the properties of a single boundary, I stratified the boundaries by the binding of well characterized insulator proteins (BEAF-32, CTCF, CP190, Su(Hw)), and selected multiple boundaries for each combination, especially for the most frequent ones (Figure 40). If all the boundaries with the same combination behaved similarly, I could conclude that the effect is due to the insulator combination alone and not to additional as yet unknown factors; otherwise, it would have pointed to the presence of additional motifs. However, our data shows that boundaries with presumably similar composition in terms of the known insulator binding proteins have different effects out of their endogenous locus, suggesting that they harbour additional motifs responsible for their function (Figure 45).

The potential to discover additional factors is a huge advantage of using endogenous boundaries, especially since multiple insulator proteins have been reported in *Drosophila* so far (Özdemir *et al.*, 2019), and probably the list is not complete. Moreover, the annotation of endogenous boundaries is more complete than what could be achieved in synthetic boundaries: in the endogenous locus, the distance from partner boundaries, the insulation score, the chromatin state and the activity of any promoters are known, and they can be used as additional feature to characterize the boundary. Of note, the insulation score and distance will not be preserved in the ectopic location, and the other variables may change; however, they could be correlated to other conserved properties. Properties embedded in the sequence, like motifs and conservation, can also be extracted. For this project, it was crucial to leverage such additional information beyond the current knowledge base.

6.1.2 Why use an endogenous landing site?

The choice of endogenous acceptor TAD partially reflects the reasoning for the single inserted boundaries, since the annotation of the TAD boundaries and of the whole genomic region within the TAD are known. Moreover, by exploiting two TADs with very different structure, looping and gene expression (Figure 41), we could examine if the context had an influence on the function of the inserted boundaries. The choice of two TADs limits the generalizability; to compensate for this, the properties of the working boundaries were later verified genome-wide.

6.2 Readouts by DNA FISH and capture Hi-C focus on different aspects of the interaction and they are highly concordant

The double DNA FISH and capture Hi-C -based readouts of the effect of the insertions provided a relatively fast and cost-effective screening (DNA FISH) with a more detailed examination of topological effects (capture Hi-C).

By DNA FISH, we measured the distance between the two loop anchors in TAD A, and between the left and right boundary in TAD B. Since the TAD A probes are 8 kb in size, they detect distance changes

between the two anchors upon insertion, but they do not capture internal reorganization of TAD contacts. Due to the complex structure of TAD A, which internal sub-TADs and loops, assessing the global structure with longer probes could lead to less interpretable results. On the contrary, the simple structure of TAD B, with not additional structural elements, allowed for the detection of global TAD compaction by using long 80 kb probes, which provide very bright signals.

Capture Hi-C showed the changes in contacts within the full TAD, enabling us to observe additional interactions not detected by DNA FISH, especially in TAD A, and to evaluate the type and magnitude of the effect of the insertion. In particular, in TAD A the insertions had a spectrum of effects, ranging from a mild perturbation of intra-TAD loops, to the formation of sub-TADs, to complete TAD disruption (Figure 43, Table 8). For the current analysis, the insertions were classified into working and not working, but the magnitude of the perturbation in working boundaries was not considered, since it may be locus-specific and not informative for other TADs with different boundaries and structure. However, it could still be used to compare boundary strength within the TAD A locus.

Most of the insertions in TAD A affected contacts at the loop anchors, which could also be detected by DNA FISH; only 3 insertions had discordant results, due to interactions with other elements internal to the loop anchors (Table 8). In 2 of those insertions, the effect was detected in DNA FISH, but very mild in capture Hi-C. In the case of boundary B-7 OR1, with no change in DNA FISH, capture Hi-C showed a combination of TAD splitting into sub-TADs and strengthening of the loop, which could have compensated each other in the DNA FISH. In TAD B, on the contrary, there was complete agreement as expected. Discordance between Hi-C and DNA FISH measurement has been extensively reported (Williamson *et al.*, 2014; Giorgetti and Heard, 2016; Fudenberg and Imakaev, 2017), and attributed to fixation differences, dynamic interactions and single-cell versus averaged readout; in our case, the high retention time of insulator proteins, the stability of TAD boundaries, the long linear distance between the probes and their genomic location may have minimized the discrepancies.

6.3 The orientation specificity of some boundaries points to a role of directional motifs in boundary pairing

Orientation dependency was observed previously at *Drosophila* insulators (Kyrchanova *et al.*, 2007; Kyrchanova, Chetverina, *et al.*, 2008; Kyrchanova, Toshchakov, *et al.*, 2008; Krivega *et al.*, 2010), but never linked to the function of TAD boundaries in *Drosophila*. The much larger insulator repertoire than in mammals (Özdemir *et al.*, 2019), the inability to detected enriched convergent CTCF sites (Rowley *et al.*, 2017) and the overlap of TADs with chromatin domains and transcriptional states (Hou *et al.*, 2012; Ulianov *et al.*, 2016; Rowley *et al.*, 2017) have masked the intrinsic boundary contribution, in favor of models where TADs mostly result from the context (Ulianov *et al.*, 2016; Rozenwald *et al.*, 2020).

From our functional data, it is clear that boundaries have intrinsic features which enable them to work in an ectopic context with different chromatin states and gene expression. The context dependency was suggestive of the presence of interacting DNA elements via the recruitment of specific proteins, and the

orientation dependency imposed the additional requirement of directional DNA motifs in interacting boundaries (Figure 42). Overall, the results pointed to the presence of motifs in a specific orientation (convergent, divergent or tandem) in the two interacting boundaries of a pair (Table 13). As for the boundaries working in both TADs and in both orientations, they are the result of the presence of multiple motifs which pair with all the acceptor TAD boundaries. Indeed, the mere insulator protein composition is not sufficient to explain whether an ectopically inserted boundary will have an effect, and if it does which type of effect (Figure 46).

6.4 Boundary pairs are not just consecutive boundaries

When shifting to a paradigm based on boundary pairs instead of single boundaries, it is pivotal to correctly define the pairs. By capture Hi-C combined to virtual 4C, I was able to detect the interactions formed by the inserted boundary and use true boundary pairs for the analysis, without assuming that consecutive boundaries would automatically interact. Indeed, I observed a specificity in boundary interactions in TAD A, whereby different inserted boundaries could pair with different boundaries in the acceptor TAD.

For the same reason, I manually annotated genome-wide boundary pairs, as to include multi-way interactions (as in sub-TADs and loops within loops) and pairs detected at different resolutions. It is common practice in current Hi-C studies to focus on a specific resolution, possibly small, and to define TADs any region between two boundaries (Ramírez *et al.* 2018 for an example); this is feasible to examine the composition of single boundaries, but incompatible with finding pair-based rules, or with an accurate TAD characterization. In fact, it would mostly detect sub-TADs and smaller enhancer-promoter contacts, while underestimating the contributions from larger TADs. My manually annotated set of boundaries and boundary pairs in a 1-10 kb resolution range integrates multiple resolution levels, and most importantly includes curated boundary pairs which would not be fully covered by other methods. In addition, it includes all the genome-wide boundaries, in contrast to other high-confidence sets that focus on a smaller subset (Stadler, Haines and Eisen, 2017); this ensures that the proportion of boundaries with a specific set of features matches exactly the genomic distribution, and creates a complementary set of non-boundary sequences which can be used as a negative control.

6.5 Experimental perturbations facilitate detection of features of boundary pairs by restricting the feature search space

Combinations of insulators at TAD boundaries have been extensively described (Hou *et al.*, 2012; Sexton *et al.*, 2012; Van Bortle *et al.*, 2014; Cavalheiro *et al.*, 2023). However, due to the high number of insulator protein (Özdemir *et al.*, 2019), the large space of all their possible and actual combinations (Ramírez *et al.*, 2018; Cavalheiro *et al.*, 2023; Kahn *et al.*, 2023), and the relatively small number of boundaries in a small genome such as *D. melanogaster*'s (from 701 high-confidence boundaries in Torosin *et al.* (2020) to 2852 in Ramírez *et al.* (2018)), it is challenging to move away from a descriptive approach to finding causal links based on genome-wide data. Attempts based on machine learning are limited by the small genome size, by the diversity across species (Dekker and Heard, 2015; Szabo,

Bantignies and Cavalli, 2019) which prevents the usage of available mouse or human data to train a model for *Drosophila*, and by the imbalance between the large number of features and the small repertoire of boundaries (Huber, 1981; Johnstone and Titterington, 2009). Whereas single boundaries can still be predicted against a random background of sequences internal to the TADs (Henderson *et al.*, 2019), there is no report of boundaries against a matched background (i.e. with some insulator protein binding, but not in a TAD boundary), and boundary pairs have never been addressed. With respect to single boundaries, finding the driving features of pairs is even more challenging, since the features are at least double, and the directional elements could each be present in a convergent, divergent or tandem orientation. With the explosion in the feature number and a slight reduction in the number of pairs with respect to single boundaries, the imbalance between the samples and features is even more extreme, leading to a situation where the choice of top features could be mainly by chance (Huber, 1981; Johnstone and Titterington, 2009).

In such a scenario, moving to a perturbative approach allowed us to test boundary combinations directly and to restrict the feature space, by a preliminary selection of features associated with working boundary pairs only. Of note, the selection was influenced by the choice of the acceptor TADs and of the inserted boundaries, which does not capture the full diversity of genome-wide TAD boundaries. Testing a larger number of inserted boundaries and acceptor TADs would be helpful to inspect more combinations and identify the most important motif pairs for boundary pairing, but would also require a considerable investment of time and resources. As an alternative solution to identify the most important motif pairs, we can refine the motif selection that emerged from our 'working boundary pairs' by only keeping the motif pairs strongly enriched in the preferred orientation genome-wide.

6.6 Conservation reveals new motifs associated to boundary pairs

Since the motifs of known insulators are not sufficient to explain TAD boundary pairing, I searched for additional motifs within the DHSs at working TAD boundary pairs, using conservation within chromatin accessible sites to identify potential boundary-associated motifs.

TAD boundaries and insulators are on average more conserved than other regulatory regions in *Drosophila* (Nègre *et al.*, 2010; Acemel and Lupiáñez, 2023), pointing to the presence of purifying selection, generally linked to the functional importance of the sequence (Jordan *et al.*, 2002; Aerts, 2012). Since coding regions are much more conserved than regulatory regions (Bejerano *et al.*, 2004; Cai, Hu and Li, 2009; Goodman *et al.*, 2023), and single boundaries may have different global conservation, I searched for the most conserved sequences within regulatory regions in each boundary, using a boundary-specific cutoff depending on the overall boundary conservation, following a similar approach as described in Cai, Hu and Li (2009). I then inferred a conservation-based PWM for each sequence via phylogenetic footprinting (Tagle *et al.*, 1988; Fickett and Wasserman, 2000). This enables *de novo* motif discovery exclusively on conserved sequences, which are more likely to have a functional role; however, it creates a large set of motifs (279) (Figure 48), which must be reduced to independent elements and prioritized.

To prioritize the motifs using functional information, I required each motif to be associated to working boundaries and enriched at TAD boundaries genome-wide (Figure 48). The enrichment-based selection may exclude motifs that contribute to boundaries but are not the main driver of boundary identity, like Trl. However, this pre-filtering step is crucial as the number of boundary pair features scales as the square of the number of motifs: by considering oriented motif pairs where the A-B motif combination is not equivalent to B-A, for N motifs there would be $4N^2$ motif pairs. After that, motif pairs were filtered by selecting for those most associated to working boundary pairs, and present in endogenous genome-wide pairs. The filtering based on working pairs is naturally sensitive to the choice of boundaries that we experimentally tested; however, by choosing a large set of inserted boundaries, it should still capture a good representation of the genome-wide interactions. Even just using 2 acceptor TADs, with the 57 motif pairs identified in our working boundary pairs, I was able to explain almost 50% of the genome-wide boundaries (Figure 50).

6.7 Limitations of the assignment of conserved sequences to TF motifs

The conservation-based approach enables the identification of sequence features *de novo*, and should enrich for potential functional sequences. In principle, it should incorporate motif syntax, and find new motifs that stabilize known TFs in close proximity to the main motif.

However, it is challenging to assign a TF, or its PWM, to *de novo* sequence-based words, leading to caveats in the interpretation of the motif assignment. First, conserved sequences assigned to known TF motifs may be bound by other TFs with a very similar motif; this was mitigated by searching for similar motifs using Tomtom (Gupta *et al.*, 2007) and FIMO (Grant, Bailey and Noble, 2011) and by comparing the co-occurrence of motif pairs genome-wide, but the sensitivity of those approaches may still be limited. In fact, only one motif, Scro, was found to coincide with the more recurrent Myb motif.

Moreover, some motifs may be alternative motifs for another TF which is stabilized by a co-binding hub, or motifs for non-annotated or poorly known TFs. To evaluate the presence of alternative motifs, I am planning to assess the percentage of motifs under the ChIP signal for each TF genome-wide, for those TF for which data is available in modERN (Kudron *et al.*, 2018), and to derive a set of alternative motifs for known TFs. For unknown TFs, the sequence itself can be used as a temporary reference.

6.8 Future directions

The emergence of motif pairs to explain TAD boundaries in *Drosophila* is a paradigm shift with respect to the current view. Most of the motif pairs are heterotypic, and have a specific reciprocal orientation of the motifs within the boundaries (Table 13). This is compatible with a model where the direction of both motifs can be read by proteins which could have a protein-protein interaction between them, either homotypic or heterotypic. For example, a loop extruding complex, such as cohesin, could be blocked by a direct DNA binding TF or a protein recruited by the TF's protein-binding domain, and then extrude DNA up to the second interactors. This model is purely theoretical, and additional experiments beyond the scope of the project would be needed to test it.

For this project, it is pivotal to refine and thoroughly validate the motif set. First, I am planning to check the motif-to-TF link by inspecting motif enrichment under ChIP peaks for each motif-associated TF, and to include alternative motifs from MEME-ChIP (Machanick and Bailey, 2011) on the modERN database (Kudron *et al.*, 2018). This could potentially reduce the initial list of motifs and assign some of the new motifs to known TFs. I am also collaborating with Dr. Miguel Hernandez in Dr. Michiel Vermuelen lab to find TFs binding to the new motifs by Affinity purification-Mass spectrometry on *Drosophila* nuclear extracts, using the new motifs as bait. I will also focus on the motif pairs mostly enriched in the specific orientation genome-wide.

Once the identity of the TF bound to the motifs is known, I am planning to validate some of the motif pairs experimentally or computationally, by examining how many motif pairs in the correct orientation make a boundary or not, and considering both boundary and non-boundary regions.

7 Conclusion and future perspectives

In my two projects, I used experimental perturbations to investigate different aspects of gene regulation; on the one hand dissecting the functional inputs of TFs within the GRN driving mesoderm development, and on the other hand dissecting the functional requirements to make TAD boundaries, and therefore chromatin domains. While gene regulation occurs within the context of chromatin domains, integrating the two projects was beyond the scope of my Ph.D., but would be an exciting direction going forward. To assess the contribution of chromatin domains for GRN function, for example, or to improve enhancer-promoter connections, it would be important to integrate the two into a single predictive model which could estimate their relative weight and any interactions between the two. Since the active GRN is different across cells and tissues, these predictions should recapitulate tissue- and possibly cell-specificity.

To date, the prediction of gene transcription, enhancer-to-gene relationships and TADs have been tackled separately and mainly in mammals (human and mouse).

7.1 Prediction of gene transcription

Many sequence-based models have been developed for the prediction of gene transcription. The most advanced sequence-based model at the moment, Enformer, captures contributions from enhancers at up to 20 kb distance and from CTCF sites at mammalian TAD boundaries (Avsec *et al.*, 2021), but likely underestimates their contribution. Despite using a wide receptive field (± 98 kb from the TSS), it achieves 2/3 of the explained variance from only 1001 bp around the TSS, and the contributions between 30 and 100 kb have negligible impact, despite containing ~25% of validated human enhancers (Karollus, Mauermeier and Gagneur, 2023). The decay is even steeper in an endogenous context (Fulco *et al.*, 2019; Gasperini *et al.*, 2019), with enhancers > 10 kb accounting for < 5% of the variance (Karollus, Mauermeier and Gagneur, 2023). As expected, such models fall short in predicting enhancer strength and enhancer-driven variations across conditions and cells, even for validated enhancers (60% of validated enhancers have < 10% predicted impact on expression according to the model) (Fulco *et al.*, 2019; Gasperini *et al.*, 2019; Karollus, Mauermeier and Gagneur, 2023). This could be due to the imbalance between the relevant regulatory elements and all the potential candidate genes with increased distance (Karollus, Mauermeier and Gagneur, 2023). This strong distance dependence hinders the incorporation of TADs in the model: due to their scale, the model would not be able to detect their effect. Due to its compact genome (~25% smaller than mice), *Drosophila* could have an advantage for predictions as there is a more complete coverage of the regulatory elements from a smaller genomic window; on the other hand, the amount of training data is also smaller. Moreover, the density of potential regulatory elements may lead to a steeper distance decay.

The Activity By Contact (ABC) model (Fulco *et al.*, 2019) could facilitate the detection of contributions of distal enhancers and overcome the distance decay. It is a feature-based model which incorporates Hi-C information of the E-P contacts, along with accessibility and active chromatin marks at enhancers.

Tissue-specific genes are predicted relatively well (Spearman correlation to experimental data = –0.63). However, CTCF slightly worsened the prediction (AUPR 0.64 vs 0.67 for all genes and 0.76 for tissue-specific genes) (Fulco *et al.*, 2019), suggesting that this model may not be suitable for integrating topology and GRN data. Actually, when the chromatin topology data (Hi-C) was removed, it had minimal impact on the model's predictions. The most predictive feature of the ABC model was clusters of DHSs, which likely reflects the number of enhancers that a gene has.

7.2 Prediction of enhancer-to-gene connections

The reconstruction of enhancer-to-gene connections and regulation has been tackled by tools based on single cell expression and/or accessibility, like SCENIC+ , GRaNIE, and CellOracle (Bravo González-Blas *et al.*, 2023; Kamal *et al.*, 2023; Kamimoto *et al.*, 2023). Such methods could be applied on WT data to construct the base GRN and then used to predict the mutant phenotypes *in silico* and compare them to the experimental perturbation. These are typically 'two step' models, first assigning a TF to an enhancer (using motif activity scores within the sc-ATAC-seq peak) and then linking the enhancer to a gene that is also expressed in the same single cell, typically within some pre-defined genomic window. These methods make a number of assumptions and do not explicitly incorporate time information; other tools like SCRIBE, TENET and SINGE (Qiu *et al.*, 2020; Kim *et al.*, 2021; Deshpande *et al.*, 2022) use pseudotime, which does not take into account sample variations (Hou *et al.*, 2023) and may fail to capture trajectories between WT and mutants. Among those models, the most appropriate to address GRNs in my data would be SCENIC+ and GRaNIE, which were tested in *Drosophila* and do not make assumptions on pseudotime, and CellOracle for TF perturbations *in silico*.

7.3 Prediction of TADs

The current methods for the predictions of 3D interactions (Dali and Blanchette, 2017; Kai *et al.*, 2018; Chyr *et al.*, 2021; Abbas *et al.*, 2024) mostly focus on mammals and rely on CTCF and chromatin marks. Among them, PredTAD can model changes in the boundaries (Chyr *et al.*, 2021), and 3DPredictor, developed for E-P interactions, can also be used with boundary information to reconstruct 3D organization and rearrangements (Belokopytova *et al.*, 2020).

These models are promising to integrate topology information in gene regulation; however, due to the strong focus on mammals and CTCF, if and how this could be transferable to *Drosophila* needs to be assessed.

In *Drosophila*, prediction on WT TADs have been made from a combination of 18 chromatin marks and insulators which can be reduced to 5 (H3K27ac, H3K27me3, CTCF, Su(Hw), Chriz/Chromator) with little impact on the performance (Rozenwald *et al.*, 2020). However, this machine learning algorithm can predict if single bins are inside a TAD, in an inter-TAD region or in a boundary, but it does not address interactions between bins or between boundaries.

7.4 How experimental data can inform predictions and vice versa

At the moment, experimental data is crucial to understand which features should be prioritized in computational methods. This is even more relevant for *Drosophila*, since the current approaches are mainly based on mammals. In general, the complexity of the system – for instance, a large number of features as in the case of TAD boundaries and TFs binding to enhancers, and connections of multiple enhancers to genes – is a huge challenge for all models. Even sequence-based predictions might benefit from the incorporation of additional features to improve in recapitulating experimental data, especially when they involve tissue- and time- specificity and mutants. A preliminary feature selection will favour accurate and interpretable predictions.

Moreover, most of the models rely on the ground truth from the WT situation, and to grasp their power and limitations it is useful to assess their performance on mutants, for which less experimental data are available. My datasets can be used as non-WT test set.

Once the predictions have been optimized, they could help in quantifying contributions from different features – TF binding and accessibility at enhancers, chromatin state, expression, topology – recorded via multiple modalities, and to assess their overall interaction. This would provide a fundamental understanding of their importance and roles at the genome-wide scale, and identify global patterns without the need for the dissection of single loci, which could be biased by the context. Notably, the estimated contributions will be based on genome-wide data, and at single genes there may be deviations from the expected behaviour; this could be leveraged to extract specific properties to be addressed experimentally.

Finally, my data can contribute a set of biologically meaningful features and ground truth on non-WT embryos, and they can be further explored with predictive models or used to select features to be integrated in such models, with the final goal of building an integrated framework for linking enhancers and topology to gene expression. This genome-wide view will aid our understanding of basic rules in all the steps from TF-enhancer binding and TF-TF interactions, to enhancer usage, E-P communication and finally gene expression.

8 Appendix

Supplementary table 1: List of oligonucleotides used for iLEXY cloning

Name	*Sequence (5'-3')*	*Line*	*Purpose*
mCherry-iLEXY fwd	GCAGCACCTGCGATCGATTCCGCCGG AAGCGC		Amplification of mCherry-iLEXY from plasmid pAct-MCS-XhoI-mCherry-LEXY-BstBI
mCherry-iLEXY rev	TCTTTCTAGGGTTAACTAGTCCAGGTT CAGGTCG		Amplification of mCherry-iLEXY from plasmid pAct-MCS-XhoI-mCherry-LEXY-BstBI
lin Scarless-mChe_fw	TTAACCCTAGAAAGATAATCATATTG		Linearization of pHD-Scarless dsRed-mChe-LEXY to insert STOP-TTAA
lin Scarless-mChe_rev	CTAGTCCAGGTTCAGGTCG		Linearization of pHD-Scarless dsRed-mChe-LEXY to insert STOP-TTAA
tin_sgRNA_fwd	CTTCGCTGCGATTGCTGCATGGTT	*tin-iLEXY*	gRNA
tin_sgRNA_rev	AAACAACCATGCAGCAATCGCAGC	*tin-iLEXY*	gRNA
tin_lha_fwd	tatagtgtcttcggggccgaGGTGAGTACCAGAAT CCCCACC	*tin-iLEXY*	Amplification of left HA
tin_lha_rev	cggcggaatcCATGTGCTGCATCTGTTGCTG CTGTGCATGATG	*tin-iLEXY*	Amplification of left HA
tin_stop-ttaa_fwd	gaacctggacTAGTGGACATCTCGCAGGACG	*tin-iLEXY*	Amplification of STOP-TTAA linker
tin_stop-ttaa_rev	gattatctttctagggTTAAGCTTTATTTATTGTC CAATATATACATAAAACAAAGAACTAG AAC	*tin-iLEXY*	Amplification of STOP-TTAA linker
tin_rha_fwd	gactatctttctagggTTAAGACCAATCAGTTCA GACTATATATTTCATTTTCAACG	*tin-iLEXY*	Amplification of right HA
tin_rha_rev	atatggtcttcttttcccggATCTTAAAGCTTAAAAT CTGAAATTTGTATCTCCTAAGTTCA	*tin-iLEXY*	Amplification of right HA
tin_PAMmut_fwd	CGTGATGTGGCCaCCAACCATGCAGCAA TCG	*tin-iLEXY*	PAM mutation
tin_PAMmut_rev	CGATTGCTGCATGGTTGGtGGCCACATC ACG	*tin-iLEXY*	PAM mutation
bin_sgRNA_fwd	CTTCGGGGGCGGTGGATACTACGA	*bin-iLEXY*	gRNA
bin_sgRNA_rev	AAACTCGTAGTATCCACCGCCCCC	*bin-iLEXY*	gRNA
bin_lha_fwd	tatagtgtcttcggggccgaCGGCTCATCTCTTCGA GGAC	*bin-iLEXY*	Amplification of left HA
bin_lha_rev	gcttccggcggaatcGTTGGCGTACTTGAGTCtCt c	*bin-iLEXY*	Amplification of left HA and PAM mutation
bin_stop-ttaa_fwd	atattacctcaaaggacacaataatcacgtTTAACCCTAG AAAGATAATCATATTG	*bin-iLEXY*	Amplification of STOP-TTAA linker
bin_stop-ttaa_rev	tgatgcacgccatcccaagttgaggatggctaGTCCAGGT TCAGGTCG	*bin-iLEXY*	Amplification of STOP-TTAA linker
bin_rha_fwd	gactatctttctagggTTAATTAGGCTCTAAGTA GTGGCACAGTAAAAG	*bin-iLEXY*	Amplification of right HA
bin_rha_rev	atatggtcttcttttcccggATTCGCATCGATTGAG ATAATCCATTTACC	*bin-iLEXY*	Amplification of right HA
mef2_sgRNA_fwd	CTTCGGCCGAGAATATCGGGCGGA	*Mef2-iLEXY*	gRNA
mef2_sgRNA_rev	AAACTCCGCCCGATATTCTCGGCC	*Mef2-iLEXY*	gRNA
mef2_lha_fw	tgtcttcggggccgaAGCAGCACCAGTAGCAAC GC	*Mef2-iLEXY*	Amplification of left HA
mef2_lha_rev	gcttccggcggaatcTGTGCCCaATCCGCCCGA	*Mef2-iLEXY*	Amplification of left HA and PAM mutation
mef2_stop-ttaa_fwd	ctgaacctggacTAGCCGGCACGAGCTGC	*Mef2-iLEXY*	Amplification of STOP-TTAA linker
mef2_stop-ttaa_rev	tctttctagggTTAATCCGAAAACAAAACAAA ACTGC	*Mef2-iLEXY*	Amplification of STOP-TTAA linker
mef2_rha_fwd	gactatctttctagggTTAAGCGAACAGTGACAG TGTCATATCTATAGT	*Mef2-iLEXY*	Amplification of right HA
mef2_rha_rev	atatggtcttcttttcccggCTCTATTCAAATATGCA CAGCTGATGCAG	*Mef2-iLEXY*	Amplification of right HA

Supplementary table 2: List of oligonucleotides used for iLEXY genotyping

Name	*Sequence (5'-3')*	*Line*	*Purpose*
tin_g1_fwd	GATCCCTTCTGGGCTGGTC	*tin-iLEXY*	Genotyping PCR 1
mef2_g1_fwd	TTTCCCAGTCTCTCGCACCT	*Mef2-iLEXY*	Genotyping PCR 1
bin_g1_fwd	TGAGCGTTGAAGGGACAGAG	*bin-iLEXY*	Genotyping PCR 1
LEXY_g1_rev	AAAGTTCTTCTCAATACGTTCAAGTG	All	Genotyping PCR 1
LEXY_g2_fwd	TTGGCTACTACACTTGAACGTAT	All	Genotyping PCR 2, genotyping DsRed removal
LEXY_g2_rev	CCGTCGCTGTGCATTTAGGA	All	Genotyping PCR 2
LEXY_g3_fwd	TTCTAGTTGTGGTTTGTCCAAACTCATC	All	Genotyping PCR 3
tin_g3_rev	ACTCCACGGACTCCTCCTTT	*tin-iLEXY*	Genotyping PCR 3, genotyping DsRed removal
mef2_g3_rev	ATGAAAACCATTGGGGGTCAAC	*Mef2-iLEXY*	Genotyping PCR 3, genotyping DsRed removal
bin_g3_rev	AGCTGCTGATTTGCATAAAACGC	*bin-iLEXY*	Genotyping PCR 3, genotyping DsRed removal

Supplementary table 3: List of fly lines used for the mesoderm GRN project

Line Name	*Full Genotype*	*Source*	*Description*
Bin-mCherry-iLEXY(V416L) DsRed+	w;;Bin[mCherry-iLEXY-V416L, 3XP3-DsRed], Pbac{y[+mDint2]=vas-Cas9, 3XP3-GFP}VK00027	This study	*bin-iLEXY* line
Bin-mCherry-iLEXY(V416L) DsRed-	w;;Bin[mCherry-iLEXY-V416L], Pbac{y[+mDint2]=vas-Cas9, 3XP3-GFP}VK00027	This study	*bin-iLEXY* line without DsRed
Mef2-mCherry-iLEXY(V416L)	w;Mef2[mCherry-iLEXY-V416L, 3XP3-DsRed];;	This study	*Mef2-iLEXY* line
Mef2-mCherry-iLEXY(V416L) DsRed-	w;Mef2[mCherry-iLEXY-V416L];;	This study	*Mef2-iLEXY* line without DsRed
Tin-mCherry-iLEXY(V416L) DsRed-	w;;Tin[mCherry-iLEXY-V416L], Pbac{y[+mDint2]=vas-Cas9, 3XP3-GFP}VK00027	This study	*tin-iLEXY* line without DsRed
Twi-mCherry-iLEXYs	yw/+; twi[mCherry-LEXYV416L](DsRed); +	This study	*twi-iLEXY* line
*bin*KO / LacZ	;;bin 3.48.2.1/TM3, sb, AbdA-lacZ	Furlong lab	*bin* KO line for immunostaining
*bin*KO / GFP	;;bin 3.48.2.1/TM3, twi-Gal4, UAS-GFP	Furlong lab	*bin* KO line for immunostaining
*Mef2*KO / GFP	;dMef22.21/CyO, twi-Gal4, UAS-GFP;	Furlong lab	*mef2* KO lines for immunostainings
*tin*KO / GFP	;;tin 1.35.1.10/TM3, twi-Gal4, UAS-GFP	This study	*tin* KO line for genomics
*tin*KO / LacZ	;;tinEC40/TM3, sb, ubx-lacZ	Mary Baylies lab	*tin* KO line for immunostainings
piggyBac, DB	w1118; If/CyO; Herm{3xP3-eCFP, αtub-piggyBacK10}M10/TM6C, Sb1	Furlong lab	Double-balancer line for piggyBac-mediated DsRed removal

Supplementary table 4: List of oligonucleotides used for TAD boundary cloning

Name	*Sequence (5'--> 3')*
Boundary_B-1_Fw	ccctcgaggaattccTGAATGCTCGAATCAATTGCGG
Boundary_B-1_Rv	actctagaactagtcTACGTGCCCAACATGGACAG
Boundary_B-2_Fw	ccctcgaggaattccCTTTCAGTGCAGATGGCGTTG
Boundary_B-2_Rv	actctagaactagtcGGGACAGACAGTTGAAGTTCCTG
Boundary_B-3_Fw	ccctcgaggaattccTTGGAGCCGAAGTATCGCTC
Boundary_B-3_Rv	actctagaactagtcTATGCCAAATGCTGCCATACG
Boundary_B-4_Fw	ccctcgaggaattccTCGAACCGAACTTGGAGTTCC
Boundary_B-4_Rv	actctagaactagtcACTTGTGGATGTCGTCTTCGTG
Boundary_B-5_Fw	ccctcgaggaattccCTAGACACTGGAAGAGAGCCAC
Boundary_B-5_Rv	actctagaactagtcCTCTCTATCTGGTTGAGGTCATACG
Boundary_B-6_Fw	ccctcgaggaattccAAGAGGCACAGGTCCATAAGC
Boundary_B-6_Rv	actctagaactagtcGCGGCTATTACTTCGCTGTC
Boundary_B-7_Fw	ccctcgaggaattccGGTGACAATTGGAGCTACAGC
Boundary_B-7_Rv	actctagaactagtcGATGATAGTGCCGTACACATCG
Boundary_B-8_Fw	ccctcgaggaattccCCAAGAAGTACGATGTCTGCG
Boundary_B-8_Rv	actctagaactagtcTTTATGGTTAGAGGCAAGCGC
Boundary_B-9_Fw	ccctcgaggaattccTGGTCAATATGAGGCGTATGGAC
Boundary_B-9_Rv	actctagaactagtcGAACCAATAGCAGAAGTGGCAC
Boundary_B-10_Fw	ccctcgaggaattccGCGGATTTGCTTGGGTTACAC
Boundary_B-10_Rv	actctagaactagtcAGAAGGTACTTGACCGACTGG
Boundary_B-11_Fw	ccctcgaggaattccGCACCTATGATGTATCGCATGC
Boundary_B-11_Rv	actctagaactagtcGGCACGAGATACAATGTGCAC
Boundary_B-12_Fw	ccctcgaggaattccAGTGACGCCCAAGCATAACTG
Boundary_B-12_Rv	actctagaactagtcTGTTCATGACCTGCTGCTGAG
Boundary_B-13_Fw	ccctcgaggaattccTCACTATTCACGTCCTCGATGG
Boundary_B-13_Rv	actctagaactagtcAGCTCGCTCAGGTATGACTTG
Boundary_B-14_Fw	ccctcgaggaattccACCGATTGGATGGAGCAACTG
Boundary_B-14_Rv	actctagaactagtcAAGTGCCTGAGCATGATCTCC
Boundary_B-15_Fw	ccctcgaggaattccATCAGATACCCGTTACTCATCTGG
Boundary_B-15_Rv	actctagaactagtcGGATCGGTCGCTTGAATCCTAC
Boundary_B-16_Fw	ccctcgaggaattccTCATCCGCAAGGATAACAGGC
Boundary_B-16_Rv	actctagaactagtcACTGGTCGAAAGTTCATCTCCAC

Supplementary table 5: List of oligonucleotides used for TAD boundary genotyping

Name	*Sequence (5'- 3')*
MiL.F	GCGTAAGCTACCTTAATCTCAAGAAGAG
MiL.R	CGCGGCGTAATGTGATTTACTATCATAC
pw35_seq10	GGAATTCCTCGAGGGATCCAAG
TAD_A_Fw	CTTTCGTGCACGCCCAATTA
TAD_A_Rv	TTTGCACGGAACAAACGACAA
TAD_B_Fw	TCGACAGAGCTTGAGTTGGTTT
TAD_B_Rv	GTCTGCCGCTATTTGTTGCATA

Supplementary table 6: Coordinates of inserted boundaries

Chr	*Start*	*End*	*Boundary name*
chrX	17283782	17286719	B-1
chr2L	808691	811086	B-2
chr3R	17810546	17812852	B-3
chr2R	21479673	21482576	B-4
chr3L	10897481	10900569	B-5
chr3R	4820372	4822479	B-6
chrX	8572475	8575127	B-7
chr3L	17556384	17558888	B-8
chr2R	14301700	14304020	B-9
chr3R	24944115	24946221	B-10
chr3L	12001928	12004244	B-11
chr3R	20007647	20009928	B-12
chr2R	7915420	7918221	B-13
chr3L	6993614	6996923	B-14
chr3R	5732341	5734625	B-15
chr3L	5869139	5872197	B-16

Supplementary table 7: List of fly lines used for the TAD boundary insertion project

Line name	*Full genotype*	*Source*	*Origin id*	*Description*
WT	yw;;	BDSC	#6598	
MiMIC acceptor site (TAD A)	y[1] w[*]; Mi{y[+mDint2]=MIC}MI02296	BDSC	#33194	Carries an acceptor site at TAD A
MiMIC acceptor site (TAD B)	y[1] w[*]; Mi{y[+mDint2]=MIC}MI14731	BDSC	#60957	Carries an acceptor site at TAD B
Integrase B, Flp	P{ry[+t7.2]=hsFLP}12, y[1] w[*] M{vas-int.B}ZH-2A; Pri[1]/TM3, Sb[1]	BDSC	#33216	Carries integrase B and Flp transgene; those are combined with MiMIC acceptor sites to generate final acceptor lines
TAD A - B-1 OR1	yw;;TAD A {Boundary B-1 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-1 OR2	yw;;TAD A {Boundary B-1 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-2 OR1	yw;;TAD A {Boundary B-2 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-2 OR2	yw;;TAD A {Boundary B-2 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-3 OR1	yw;;TAD A {Boundary B-3 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-3 OR2	yw;;TAD A {Boundary B-3 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-4 OR1	yw;;TAD A {Boundary B-4 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-4 OR2	yw;;TAD A {Boundary B-4 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-5 OR1	yw;;TAD A {Boundary B-5 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-5 OR2	yw;;TAD A {Boundary B-5 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-6 OR2	yw;;TAD A {Boundary B-6 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-7 OR1	yw;;TAD A {Boundary B-7 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-7 OR2	yw;;TAD A {Boundary B-7 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-8 OR1	yw;;TAD A {Boundary B-8 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-9 OR1	yw;;TAD A {Boundary B-9 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-9 OR2	yw;;TAD A {Boundary B-9 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-10 OR1	yw;;TAD A {Boundary B-10 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-10 OR2	yw;;TAD A {Boundary B-10 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-11 OR2	yw;;TAD A {Boundary B-11 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-13 OR1	yw;;TAD A {Boundary B-13 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-13 OR2	yw;;TAD A {Boundary B-13 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-14 OR1	yw;;TAD A {Boundary B-14 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-14 OR2	yw;;TAD A {Boundary B-14 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-15 OR1	yw;;TAD A {Boundary B-15 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-15 OR2	yw;;TAD A {Boundary B-15 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-16 OR1	yw;;TAD A {Boundary B-16 OR1-DsRed}	This thesis		Boundary insertion into TAD A
TAD A - B-16 OR2	yw;;TAD A {Boundary B-16 OR2-DsRed}	This thesis		Boundary insertion into TAD A
TAD B - B-1 OR1	yw;TAD B {Boundary B-1 OR1-DsRed}	This thesis		Boundary insertion into TAD B
TAD B - B-1 OR2	yw;TAD B {Boundary B-1 OR2-DsRed}	This thesis		Boundary insertion into TAD B
TAD B - B-2 OR1	yw;TAD B {Boundary B-2 OR1-DsRed}	This thesis		Boundary insertion into TAD B
TAD B - B-2 OR2	yw;TAD B {Boundary B-2 OR2-DsRed}	This thesis		Boundary insertion into TAD B
TAD B - B-3 OR1	yw;TAD B {Boundary B-3 OR1-DsRed}/CyO, GFP	This thesis		Boundary insertion into TAD B

TAD B - B-3 OR2	yw;TAD B {Boundary B-3 OR2-DsRed}/CyO, GFP	This thesis	Boundary insertion into TAD B
TAD B - B-4 OR1	yw;TAD B {Boundary B-4 OR1-DsRed}/CyO, GFP	This thesis	Boundary insertion into TAD B
TAD B - B-4 OR2	yw;TAD B {Boundary B-4 OR2-DsRed}	This thesis	Boundary insertion into TAD B
TAD B - B-5 OR2	yw;TAD B {Boundary B-5 OR2-DsRed}/CyO, GFP	This thesis	Boundary insertion into TAD B
TAD B - B-6 OR2	yw;TAD B {Boundary B-6 OR2-DsRed}	This thesis	Boundary insertion into TAD B
TAD B - B-7 OR1	yw;TAD B {Boundary B-7 OR1-DsRed}	This thesis	Boundary insertion into TAD B
TAD B - B-7 OR2	yw;TAD B {Boundary B-7 OR2-DsRed}	This thesis	Boundary insertion into TAD B
TAD B - B-8 OR1	yw;TAD B {Boundary B-8 OR1-DsRed}	This thesis	Boundary insertion into TAD B
TAD B - B-9 OR1	yw;TAD B {Boundary B-9 OR1-DsRed}	This thesis	Boundary insertion into TAD B
TAD B - B-9 OR2	yw;TAD B {Boundary B-9 OR2-DsRed}/CyO, GFP	This thesis	Boundary insertion into TAD B
TAD B - B-11 OR2	yw;TAD B {Boundary B-11 OR2-DsRed}/CyO, GFP	This thesis	Boundary insertion into TAD B
TAD B - B-12 OR1	yw;TAD B {Boundary B-12 OR1-DsRed}/CyO, GFP	This thesis	Boundary insertion into TAD B
TAD B - B-12 OR2	yw;TAD B {Boundary B-12 OR2-DsRed}/CyO, GFP	This thesis	Boundary insertion into TAD B
TAD B - B-13 OR1	yw;TAD B {Boundary B-13 OR1-DsRed}	This thesis	Boundary insertion into TAD B
TAD B - B-13 OR2	yw;TAD B {Boundary B-13 OR2-DsRed}/CyO, GFP	This thesis	Boundary insertion into TAD B
TAD B - B-14 OR1	yw;TAD B {Boundary B-14 OR1-DsRed}/CyO, GFP	This thesis	Boundary insertion into TAD B

Supplementary table 8: List of oligonucleotides used for DNA FISH probe cloning

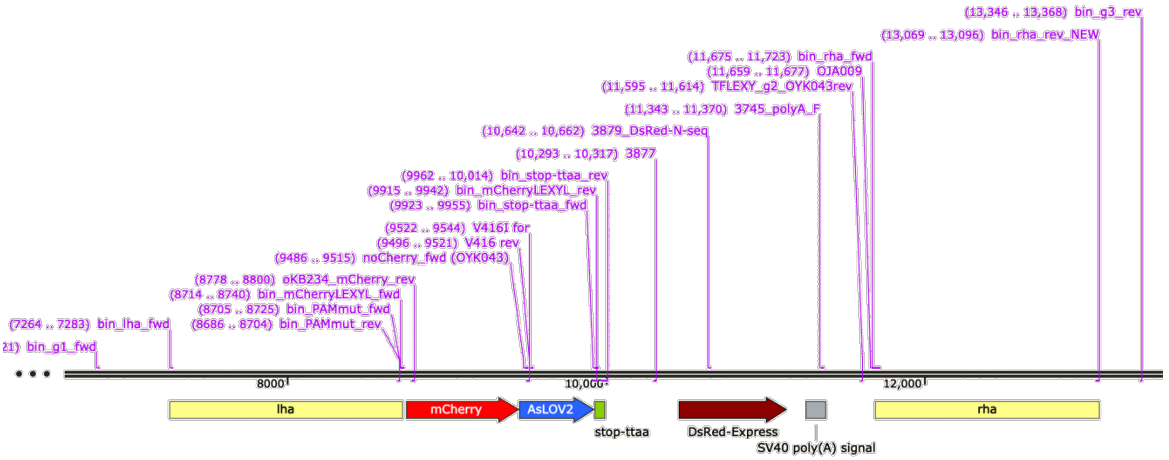
Name	*Sequence (5'- 3')*
scylchrb_DNAFISH_Probe1_Fw	GTCTACGCGATTCGGGACAT
scylchrb_DNAFISH_Probe1_Rv	CTTTGCGATACATGGCCAGC
scylchrb_DNAFISH_Probe2_Fw	GAGAGTTCACGGCCAAGGAA
scylchrb_DNAFISH_Probe2_Rv	CTTCCATCCGGGGGTTTTGA

Supplementary table 9: List of vectors used for DNA FISH probe cloning

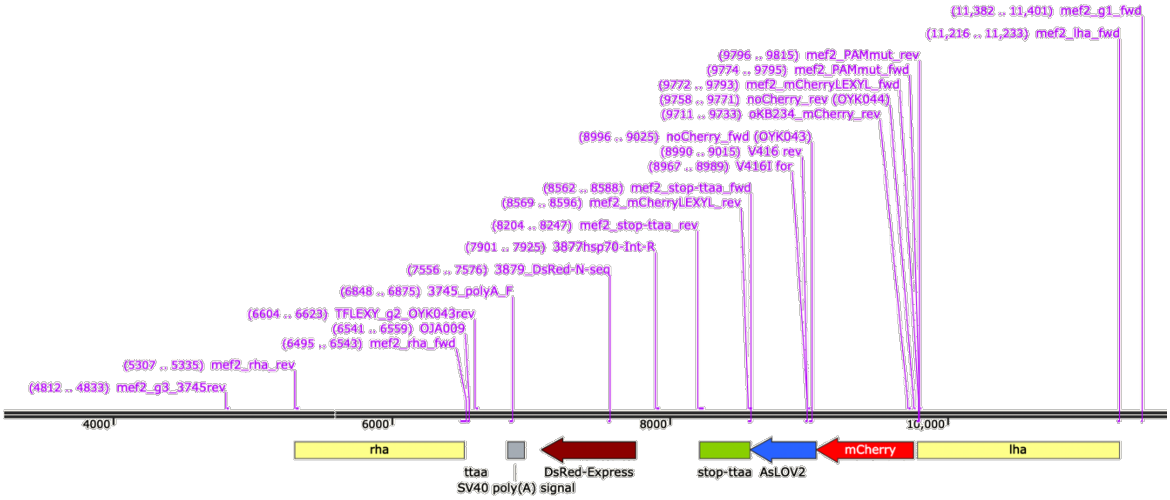
Code	*Name*	*Source*	*Purpose*
2487	pGEM-T-easy-444-445		DNA FISH probe TAD A left loop anchor
2488	pGEM-T-easy-446-447		DNA FISH probe TAD A right loop anchor
CH321-92M21	attB-Pacman-CmR-BW-TADB-LB	DGRC #2840	DNA FISH probe TAD B left boundary
CH321-76O20	attB-Pacman-CmR-BW-TADB-RB	DGRC #2841	DNA FISH probe TAD B right boundary

Supplementary figure S1: iLEXY constructs

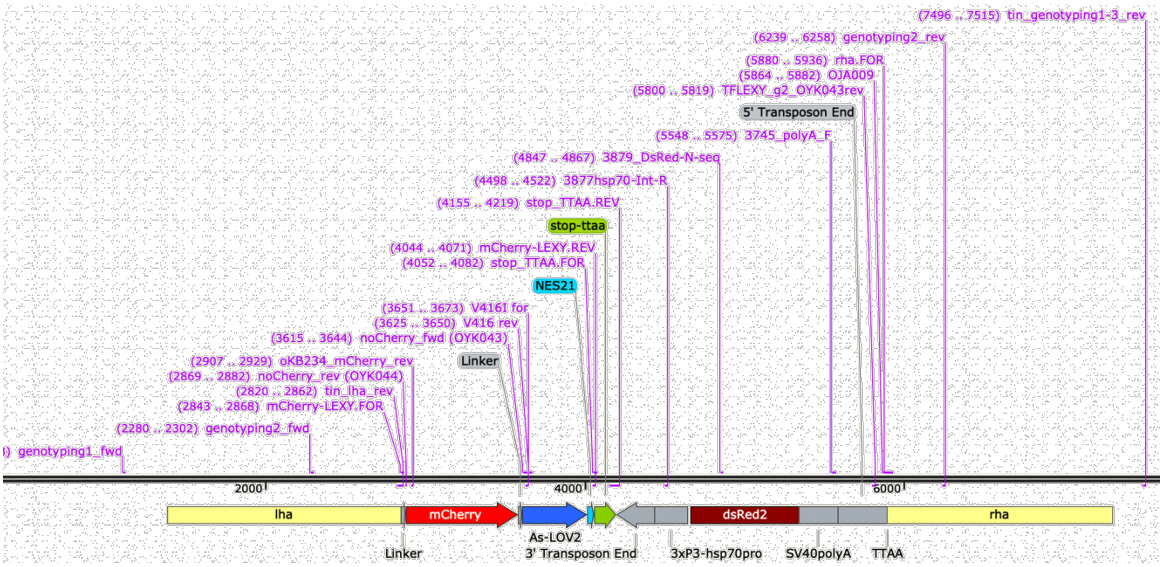
bin-mCherry-iLEXY dsRed



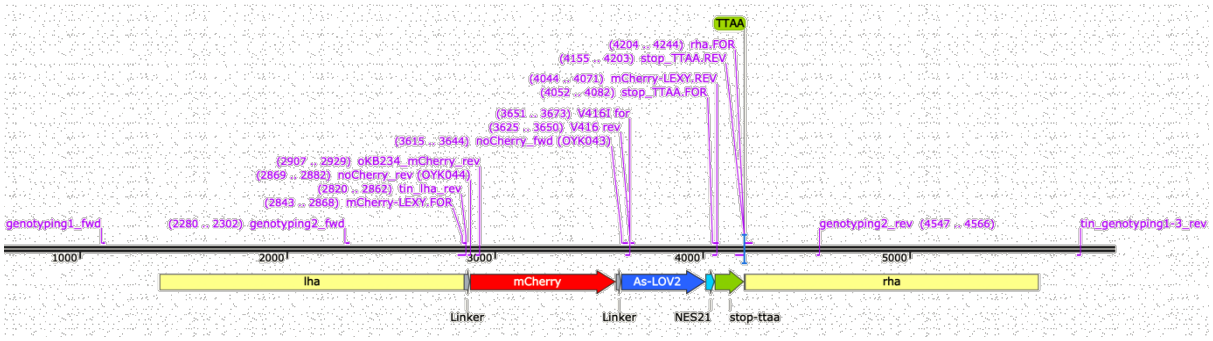
Mef2-mCherry-iLEXY dsRed



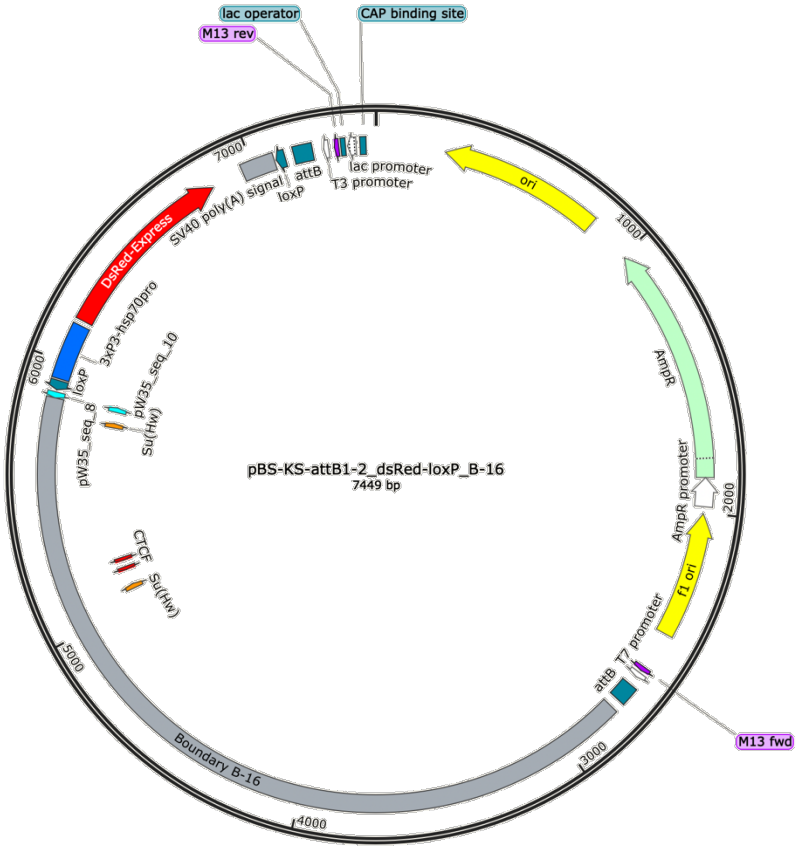
tin-mCherry-iLEXY dsRed



tin-mCherry-iLEXY (after DsRed removal)



Supplementary figure S2: Example of a vector for TAD boundary insertion



9 References

Abbas, A. *et al.* (2024) 'ChIPr: accurate prediction of cohesin-mediated 3D genome organization from 2D chromatin features', *Genome Biology*, 25(1), pp. 1–27. doi: 10.1186/S13059-023-03158-7/FIGURES/12.

Abraham, J. *et al.* (1984) 'Regulation of mating-type information in yeast. Negative control requiring sequences both 5' and 3' to the regulated region', *Journal of molecular biology*, 176(3), pp. 307–331. doi: 10.1016/0022-2836(84)90492-3.

Acemel, R. D. and Lupiáñez, D. G. (2023) 'Evolution of 3D chromatin organization at different scales', *Current Opinion in Genetics and Development*. doi: 10.1016/j.gde.2022.102019.

Aerts, S. (2012) 'Computational Strategies for the Genome-Wide Identification of cis-Regulatory Elements and Transcriptional Targets', in *Current Topics in Developmental Biology*, pp. 121–145. doi: 10.1016/B978-0-12-386499-4.00005-7.

Afgan, E. *et al.* (2016) 'The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update', *Nucleic Acids Research*, 44(W1). doi: 10.1093/nar/gkw343.

Ahn, J. H. *et al.* (2021) 'Phase separation drives aberrant chromatin looping and cancer development', *Nature*, 595(7868), pp. 591–595. doi: 10.1038/S41586-021-03662-5.

Aizarani, N. *et al.* (2019) 'A human liver cell atlas reveals heterogeneity and epithelial progenitors', *Nature*, 572(7768), pp. 199–204. doi: 10.1038/S41586-019-1373-2.

Akdemir, K. C. *et al.* (2020) 'Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer', *Nature Genetics 2020 52:3*, 52(3), pp. 294–305. doi: 10.1038/s41588-019-0564-y.

Alberga, A. *et al.* (1991) 'The snail gene required for mesoderm formation in Drosophila is expressed dynamically in derivatives of all three germ layers', *Development (Cambridge, England)*, 111(4), pp. 983–992. doi: 10.1242/DEV.111.4.983.

Alon, U. (2007) 'Network motifs: Theory and experimental approaches', *Nature Reviews Genetics*. doi: 10.1038/nrg2102.

Anania, C. *et al.* (2022) 'In vivo dissection of a clustered-CTCF domain boundary reveals developmental principles of regulatory insulation', *Nature genetics*, 54(7), pp. 1026–1036. doi: 10.1038/S41588-022-01117-9.

Anant, S., Roy, S. and Vijayraghavan, K. (1998) 'Twist and Notch negatively regulate adult muscle differentiation in Drosophila', *Development (Cambridge, England)*, 125(8), pp. 1361–1369. doi: 10.1242/DEV.125.8.1361.

Anders, S. and Huber, W. (2010) 'Differential expression analysis for sequence count data', *Genome biology*, 11(R106), pp. 1–12. doi: 10.1186/gb-2010-11-10-r106.

Anders, S., Pyl, P. T. and Huber, W. (2015) 'HTSeq-A Python framework to work with high-throughput sequencing data', *Bioinformatics*, 31(2), pp. 166–169. doi: 10.1093/bioinformatics/btu638.

Andersson, R. *et al.* (2014) 'An atlas of active enhancers across human cell types and tissues', *Nature 2014 507:7493*, 507(7493), pp. 455–461. doi: 10.1038/nature12787.

Andrews, S. *et al.* (2015) *FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics*, *Babraham Institute*.

Artero, R. *et al.* (1998) 'The muscleblind gene participates in the organization of Z-bands and epidermal attachments of Drosophila muscles and is regulated by Dmef2', *Developmental Biology*, 195(2). doi: 10.1006/dbio.1997.8833.

Arzate-Mejía, R. G. *et al.* (2020) 'In situ dissection of domain boundaries affect genome topology and gene transcription in Drosophila', *Nature Communications*, 11(1). doi: 10.1038/s41467-020-14651-z.

Ashraf, S. I. *et al.* (2004) 'Worniu, a Snail family zinc-finger protein, is required for brain development in Drosophila', *Developmental Dynamics*, 231(2), pp. 379–386. doi: 10.1002/DVDY.20130.

Avsec, Ž. *et al.* (2021) 'Effective gene expression prediction from sequence by integrating long-range interactions', *Nature methods*, 18(10), pp. 1196–1203. doi: 10.1038/S41592-021-01252-X.

Avva, S. V. S. P. and Hart, C. M. (2016) 'Characterization of the Drosophila BEAF-32A and BEAF-32B Insulator Proteins', *PLoS ONE*, 11(9). doi: 10.1371/JOURNAL.PONE.0162906.

Azpiazu, N. and Frasch, M. (1993) 'tinman and bagpipe: two homeo box genes that determine cell fates in the dorsal mesoderm of Drosophila.', *Genes & Development*, 7(7 PART B), pp. 1325–1340. doi: 10.1101/GAD.7.7B.1325.

Bag, I. *et al.* (2021) 'M1BP cooperates with CP190 to activate transcription at TAD borders and promote chromatin insulator activity', *Nature communications*, 12(1). doi: 10.1038/S41467-021-24407-Y.

Bajic, V. B. *et al.* (2006) 'Mice and Men: Their Promoter Properties', *PLOS Genetics*, 2(4), p. e54. doi: 10.1371/JOURNAL.PGEN.0020054.

Balasubramanian, D. *et al.* (2024) 'Enhancer–promoter interactions can form independently of genomic distance and be functional across TAD boundaries', *Nucleic Acids Research*, 52(4), pp. 1702–1719. doi: 10.1093/NAR/GKAD1183.

Balleza, E., Kim, J. M. and Cluzel, P. (2018) 'Systematic characterization of maturation time of fluorescent proteins in living cells', *Nature methods*, 15(1), p. 47. doi: 10.1038/NMETH.4509.

Banerji, J., Rusconi, S. and Schaffner, W. (1981) 'Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences', *Cell*, 27(2 Pt 1), pp. 299–308. doi: 10.1016/0092-8674(81)90413-X.

Bartosovic, M., Kabbe, M. and Castelo-Branco, G. (2021) 'Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues', *Nature biotechnology*, 39(7), pp. 825–835. doi: 10.1038/S41587-021-00869-9.

Barutcu, A. R. *et al.* (2018) 'A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus', *Nature Communications*, 9(1). doi: 10.1038/S41467-018-03614-0.

Bate, M. (1993) 'The Mesoderm and Its Derivatives', in *The Development of Drosophila melanogaster*.

Bate, M. and Arias, A. M. (1993) *The Development of Drosophila melanogaster*, *Cold Spring Harbor Laboratory Press*.

Bate, M., Rushton, E. and Currie, D. A. (1991) 'Cells with persistent twist expression are the embryonic precursors of adult muscles in Drosophila', *Development (Cambridge, England)*, 113(1), pp. 79–89. doi: 10.1242/DEV.113.1.79.

Baylies, M. K. and Bate, M. (1996) 'Twist: a myogenic switch in Drosophila', *Science*, 272(5267), pp. 1481–1485. Available at: https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=00368075&v=2.1&it=r&id=GALE%7CA18405823&sid=googleScholar&linkaccess=fulltext.

Baylies, M. K., Bate, M. and Gomez, M. R. (1998) 'Myogenesis: A View from Drosophila', *Cell*, 93(6), pp. 921–927. doi: 10.1016/S0092-8674(00)81198-8.

Bejerano, G. *et al.* (2004) 'Ultraconserved elements in the human genome', *Science*, 304(5675), pp. 1321–1325. doi: 10.1126/science.1098119.

Belokopytova, P. S. *et al.* (2020) 'Quantitative prediction of enhancer-promoter interactions', *Genome research*, 30(1), pp. 72–84. doi: 10.1101/GR.249367.119.

van Bemmel, J. G. *et al.* (2010) 'The insulator protein SU(HW) fine-tunes nuclear lamina interactions of the Drosophila genome', *PloS one*, 5(11), p. 15013. doi: 10.1371/JOURNAL.PONE.0015013.

Benabdallah, N. S. *et al.* (2019) 'Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation', *Molecular cell*, 76(3), pp. 473-484.e7. doi: 10.1016/J.MOLCEL.2019.07.038.

Benedetti, F. *et al.* (2017) 'Transcription-induced supercoiling explains formation of self-interacting chromatin domains in S. pombe', *Nucleic Acids Research*, 45(17), pp. 9850–9859. doi: 10.1093/NAR/GKX716.

Bernardoni, R., Vivancos, V. and Giangrande, A. (1997) 'glide/gcm is expressed and required in the scavenger cell lineage', *Developmental Biology*, 191(1), pp. 118–130. doi: 10.1006/dbio.1997.8702.

Bernstein, B. E. *et al.* (2006) 'A bivalent chromatin structure marks key developmental genes in embryonic stem cells', *Cell*, 125(2), pp. 315–326. doi: 10.1016/J.CELL.2006.02.041.

Bintu, B. *et al.* (2018) 'Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells',

Science (New York, N.Y.), 362(6413). doi: 10.1126/SCIENCE.AAU1783.

Bodmer, R. (1993) 'The gene tinman is required for specification of the heart and visceral muscles in Drosophila', *Development (Cambridge, England)*, 118(3), pp. 719–729. doi: 10.1242/DEV.118.3.719.

Boija, A. *et al.* (2018) 'Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains', *Cell*, 175(7), pp. 1842-1855.e16. doi: 10.1016/J.CELL.2018.10.042.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Boltsis, I. *et al.* (2021) 'Chromatin Conformation in Development and Disease', *Frontiers in Cell and Developmental Biology*. Frontiers Media S.A., p. 723859. doi: 10.3389/fcell.2021.723859.

Bonev, B. *et al.* (2017) 'Multiscale 3D Genome Rewiring during Mouse Neural Development', *Cell*, 171(3), pp. 557-572.e24. doi: 10.1016/j.cell.2017.09.043.

Bonn, S., Zinzen, R. P., Perez-Gonzalez, A., *et al.* (2012) 'Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP', *Nature protocols*, 7(5), pp. 978–994. doi: 10.1038/NPROT.2012.049.

Bonn, S., Zinzen, R. P., Girardot, C., *et al.* (2012) 'Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development', *Nature genetics*, 44(2), pp. 148–156. doi: 10.1038/NG.1064.

Bonn, S. and Furlong, E. E. M. M. (2008) 'cis-Regulatory networks during development: a view of Drosophila', *Current opinion in genetics & development*, 18(6), pp. 513–520. doi: 10.1016/J.GDE.2008.09.005.

Boomer, R., Jan, L. Y. and Jan, Y. N. (1990) 'A new homeobox-containing gene, msh-2, is transiently expressed early during mesoderm formation of Drosophila', *Development (Cambridge, England)*, 110(3), pp. 661–669. doi: 10.1242/DEV.110.3.661.

Van Bortle, K. *et al.* (2012) 'Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains', *Genome Research*, 22(11), pp. 2176–2187. doi: 10.1101/GR.136788.111.

Van Bortle, K. *et al.* (2014) 'Insulator function and topological domain border strength scale with architectural protein occupancy', *Genome biology*, 15(6), p. R82. doi: 10.1186/GB-2014-15-5-R82/FIGURES/6.

Bour, B. A. *et al.* (1995) 'Drosophila MEF2, a transcription factor that is essential for myogenesis', *Genes & development*, 9(6), pp. 730–741. doi: 10.1101/GAD.9.6.730.

Bour, B. A. *et al.* (2000) 'Drosophila SNS, a member of the immunoglobulin superfamily that is essential for myoblast fusion', *Genes and Development*, 14(12). doi: 10.1101/gad.14.12.1498.

Bourbon, H. M. G. *et al.* (2022) 'A shared ancient enhancer element differentially regulates the bric-a-brac tandem gene duplicates in the developing Drosophila leg', *PLoS genetics*, 18(3). doi: 10.1371/JOURNAL.PGEN.1010083.

Bowman, S. K. *et al.* (2014) 'H3K27 modifications define segmental regulatory domains in the Drosophila bithorax complex', *eLife*, 3(July2014). doi: 10.7554/eLife.02833.

Bozek, M. *et al.* (2019) 'ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the Drosophila blastoderm', *Genome Research*, 29(5), pp. 771–783. doi: 10.1101/GR.242362.118/-/DC1.

Bravo González-Blas, C. *et al.* (2019) 'cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data', *Nature Methods*, 16(5), pp. 397–400. doi: 10.1038/s41592-019-0367-1.

Bravo González-Blas, C. *et al.* (2023) 'SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks', *Nature Methods 2023 20:9*, 20(9), pp. 1355–1367. doi: 10.1038/s41592-023-01938-4.

Brignall, R. *et al.* (2019) 'Considering Abundance, Affinity, and Binding Site Availability in the NF-κB Target Selection Puzzle', *Frontiers in Immunology*, 10(MAR), p. 609. doi: 10.3389/FIMMU.2019.00609.

Britten, R. J. and Davidson, E. H. (1969) 'Gene regulation for higher cells: a theory', *Science (New York, N.Y.)*, 165(3891), pp. 349–357. doi: 10.1126/SCIENCE.165.3891.349.

Buecker, C. and Wysocka, J. (2012) 'Enhancers as information integration hubs in development: lessons from genomics', *Trends in Genetics*, 28(6), pp. 276–284. doi: 10.1016/J.TIG.2012.02.008.

Buenrostro, J. D. *et al.* (2013) 'Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position', *Nature Methods*, 10(12), pp. 1213–1218. doi: 10.1038/nmeth.2688.

Burke, T. W. and Kadonaga, J. T. (1997) 'The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila', *Genes & development*, 11(22), pp. 3020–3031. doi: 10.1101/GAD.11.22.3020.

Bushey, A. M., Ramos, E. and Corces, V. G. (2009) 'Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions', *Genes and Development*, 23(11), pp. 1338–1350. doi: 10.1101/GAD.1798209.

Busser, B. W. *et al.* (2012) 'Integrative analysis of the zinc finger transcription factor Lame duck in the Drosophila myogenic gene regulatory network', *Proceedings of the National Academy of Sciences of the United States of America*, 109(50), pp. 20768–20773. doi: 10.1073/PNAS.1210415109.

Butler, J. E. F. and Kadonaga, J. T. (2001) 'Enhancer–promoter specificity mediated by DPE or TATA core promoter motifs', *Genes & Development*, 15(19), pp. 2515–2519. doi: 10.1101/GAD.924301.

Cai, X., Hu, H. and Li, X. (2009) 'A new measurement of sequence conservation', *BMC Genomics*, 10(1), pp. 1–13. doi: 10.1186/1471-2164-10-623/FIGURES/7.

Calderon, D. *et al.* (2022) 'The continuum of Drosophila embryonic development at single-cell resolution', *Science (New York, N.Y.)*, 377(6606), p. eabn5800. doi: 10.1126/SCIENCE.ABN5800.

Campos-Ortega, J. A. and Hartenstein, V. (2013) *The embryonic development of Drosophila melanogaster*. Springer Science & Business Media.

Cannavò, E. *et al.* (2016) 'Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks', *Current biology : CB*, 26(1), pp. 38–51. doi: 10.1016/J.CUB.2015.11.034.

Canver, M. C. *et al.* (2015) 'BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis', *Nature 2015 527:7577*, 527(7577), pp. 192–197. doi: 10.1038/nature15521.

Carmena, A., Bate, M. and Jiménez, F. (1995) 'Lethal of scute, a proneural gene, participates in the specification of muscle progenitors during Drosophila embryogenesis', *Genes & development*, 9(19), pp. 2373–2383. doi: 10.1101/GAD.9.19.2373.

Carnesecchi, J. *et al.* (2020) 'Multi-level and lineage-specific interactomes of the Hox transcription factor Ubx contribute to its functional specificity', *Nature Communications 2020 11:1*, 11(1), pp. 1–17. doi: 10.1038/s41467-020-15223-x.

Carninci, P. *et al.* (2006) 'Genome-wide analysis of mammalian promoter architecture and evolution', *Nature Genetics 2006 38:6*, 38(6), pp. 626–635. doi: 10.1038/ng1789.

Carvajal, J. J., Keith, A. and Rigby, P. W. J. (2008) 'Global transcriptional regulation of the locus encoding the skeletal muscle determination genes Mrf4 and Myf5', *Genes & Development*, 22(2), pp. 265–276. doi: 10.1101/GAD.442408.

Castanon, I. *et al.* (2001) 'Dimerization partners determine the activity of the Twist bHLH protein during Drosophila mesoderm development', *Development*, 128(16). doi: 10.1242/dev.128.16.3145.

Castanon, I. and Baylies, M. K. (2002) 'A Twist in fate: evolutionary comparison of Twist structure and function', *Gene*, 287(1–2), pp. 11–22. doi: 10.1016/S0378-1119(01)00893-9.

Castro-Mondragon, J. A. *et al.* (2022) 'JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles', *Nucleic Acids Research*, 50(D1), pp. D165–D173. doi: 10.1093/NAR/GKAB1113.

Cavalheiro, G. R. *et al.* (2023) 'CTCF, BEAF-32, and CP190 are not required for the establishment of TADs in early Drosophila embryos but have locus-specific roles', *Science Advances*, 9(5). doi: 10.1126/SCIADV.ADE1085/ASSET/A9801EF5-C368-4E60-8574-B217D0F554C9/ASSETS/IMAGES/LARGE/SCIADV.ADE1085-F3.JPG.

Cavallo, R. A. *et al.* (1998) 'Drosophila Tcf and Groucho interact to repress wingless signalling activity', *Nature*, 395(6702). doi: 10.1038/26982.

Chalkley, G. E. and Verrijzer, C. P. (1999) 'DNA binding site selection by RNA polymerase II TAFs: a TAF II 250-TAF II 150 complex recognizes the Initiator', *The EMBO Journal*, 18(17), pp. 4835–4845. Available at: https://www.embopress.org.

Chang, L. H. *et al.* (2023) 'Multi-feature clustering of CTCF binding creates robustness for loop extrusion blocking and Topologically Associating Domain boundaries', *Nature Communications 2023 14:1*, 14(1), pp. 1–19. doi: 10.1038/s41467-023-41265-y.

Chang, T. H. T. *et al.* (2004) 'An enhancer directs differential expression of the linked Mrf4 and Myf5 myogenic regulatory genes in the mouse', *Developmental Biology*, 269(2), pp. 595–608. doi: 10.1016/J.YDBIO.2004.02.013.

Chathoth, K. T. *et al.* (2022) 'The role of insulators and transcription in 3D chromatin organization of flies', *Genome Research*, 32(4), pp. 682–698. doi: 10.1101/GR.275809.121/-/DC1.

Chen, Y. and Belmont, A. S. (2019) 'Genome organization around nuclear speckles', *Current opinion in genetics & development*, 55, pp. 91–99. doi: 10.1016/J.GDE.2019.06.008.

Choudhary, M. N. K. *et al.* (2023) 'Widespread contribution of transposable elements to the rewiring of mammalian 3D genomes', *Nature Communications*, 14(1). doi: 10.1038/S41467-023-36364-9.

Chung, J. H., Whiteley, M. and Felsenfeld, G. (1993) 'A 5′ element of the chicken β-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila', *Cell*, 74(3). doi: 10.1016/0092-8674(93)80052-G.

Chyr, J. *et al.* (2021) 'PredTAD: A machine learning framework that models 3D chromatin organization alterations leading to oncogene dysregulation in breast cancer cell lines', *Computational and Structural Biotechnology Journal*, 19, pp. 2870–2880. doi: 10.1016/J.CSBJ.2021.05.013.

Ciglar, L. *et al.* (2014) 'Coordinated repression and activation of two transcriptional programs stabilizes cell fate during myogenesis', *Development (Cambridge)*, 141(13), pp. 2633–2643. doi: 10.1242/DEV.101956/-/DC1.

Clark, I. B. N. N. *et al.* (2011) 'Fibroblast growth factor signalling controls successive cell behaviours during mesoderm layer formation in Drosophila', *Development*, 138(13), pp. 2705–2715. doi: 10.1242/DEV.060277/-/DC1.

Corces, M. R. *et al.* (2017) 'An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues', *Nature Methods*, 14(10), pp. 959–962. doi: 10.1038/nmeth.4396.

Core, L. J. *et al.* (2014) 'Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers', *Nature Genetics 2014 46:12*, 46(12), pp. 1311–1320. doi: 10.1038/ng.3142.

Costea, J. *et al.* (2023) 'A de novo transcription-dependent TAD boundary underpins critical multiway interactions during antibody class switch recombination', *Molecular cell*, 83(5), pp. 681-697.e7. doi: 10.1016/J.MOLCEL.2023.01.014.

Courey, A. J. and Jia, S. (2001) 'Transcriptional repression: the long and the short of it', *Genes & development*, 15(21), pp. 2786–2796. doi: 10.1101/GAD.939601.

Crane, E. *et al.* (2015) 'Condensin-driven remodelling of X chromosome topology during dosage compensation', *Nature*, 523(7559), pp. 240–244. doi: 10.1038/NATURE14450.

Cremer, T. and Cremer, C. (2001) 'Chromosome territories, nuclear architecture and gene regulation in mammalian cells', *Nature reviews. Genetics*, 2(4), pp. 292–301. doi: 10.1038/35066075.

Creyghton, M. P. *et al.* (2010) 'Histone H3K27ac separates active from poised enhancers and predicts developmental state', *Proceedings of the National Academy of Sciences of the United States of America*, 107(50), pp. 21931–21936. doi: 10.1073/PNAS.1016071107/-/DCSUPPLEMENTAL/PNAS.201016071SI.PDF.

Cripps, R. M. *et al.* (1998) 'The myogenic regulatory gene Mef2 is a direct target for transcriptional activation by Twist during Drosophila myogenesis', *Genes & development*, 12(3), pp. 422–434. doi: 10.1101/GAD.12.3.422.

Cripps, R. M., Lovato, T. A. L. and Olson, E. N. (2004) 'Positive autoregulation of the Myocyte enhancer factor-2 myogenic control gene during somatic muscle development in Drosophila', *Developmental Biology*, 267(2), pp. 536–547.

Cripps, R. M. and Olson, E. N. (1998) 'Twist is required for muscle template splitting during adult Drosophila myogenesis', *Developmental biology*, 203(1), pp. 106–115. doi: 10.1006/DBIO.1998.9040.

Cripps, R. M. and Olson, E. N. (2002) 'Control of Cardiac Development by an Evolutionarily Conserved Transcriptional Network', *Developmental Biology*, 246(1), pp. 14–28. doi: 10.1006/DBIO.2002.0666.

Cripps, R. M., Zhao, B. and Olson, E. N. (1999) 'Transcription of the myogenic regulatory gene Mef2 in cardiac, somatic, and

visceral muscle cell lineages is regulated by a Tinman-dependent core enhancer', *Developmental Biology*, 215(2), pp. 420–430. doi: 10.1006/dbio.1999.9446.

Cuartero, S. *et al.* (2014) 'Ibf1 and Ibf2 are novel CP190-interacting proteins required for insulator function', *The EMBO journal*, 33(6), pp. 637–647. doi: 10.1002/EMBJ.201386001.

Cubas, P. *et al.* (1991) 'Proneural clusters of achaete-scute expression and the generation of sensory organs in the Drosophila imaginal wing disc', *Genes & development*, 5(6), pp. 996–1008. doi: 10.1101/GAD.5.6.996.

Cunha, P. M. F. *et al.* (2010) 'Combinatorial Binding Leads to Diverse Regulatory Responses: Lmd Is a Tissue-Specific Modulator of Mef2 Activity', *PLoS Genetics*, 6(7), pp. 1–11. doi: 10.1371/JOURNAL.PGEN.1001014.

Cusanovich, D. A. *et al.* (2015) 'Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing', *Science*, 348(6237), pp. 910–914. doi: 10.1126/SCIENCE.AAB1601/SUPPL_FILE/CUSANOVICH-SM.PDF.

Cusanovich, D. A. *et al.* (2018) 'The cis-regulatory dynamics of embryonic development at single-cell resolution', *Nature 2018 555:7697*, 555(7697), pp. 538–542. doi: 10.1038/nature25981.

Dali, R. and Blanchette, M. (2017) 'A critical assessment of topologically associating domain prediction tools', *Nucleic Acids Research*, 45(6), p. 2994. doi: 10.1093/NAR/GKX145.

Damm, C. *et al.* (1998) 'Independent regulatory elements in the upstream region of the Drosophila 133 tubulin gene (βTub60D) guide expression in the dorsal vessel and the somatic muscles', *Developmental Biology*, 199(1), pp. 138–149. doi: 10.1006/dbio.1998.8916.

Danecek, P. *et al.* (2021) 'Twelve years of SAMtools and BCFtools', *GigaScience*, 10(2). doi: 10.1093/gigascience/giab008.

Danino, Y. M. *et al.* (2015) 'The core promoter: At the heart of gene expression'. doi: 10.1016/j.bbagrm.2015.04.003.

Davidson, E. (2006) *The Regulatory Genome*, *The Regulatory Genome*. Elsevier Inc. doi: 10.1016/B978-0-12-088563-3.X5018-4.

Davidson, E. H. *et al.* (2002) 'A genomic regulatory network for development', *Science (New York, N.Y.)*, 295(5560), pp. 1669–1678. doi: 10.1126/SCIENCE.1069883.

Davidson, E. H. and Erwin, D. H. (2006) 'Gene Regulatory Networks and the Evolution of Animal Body Plans', *Science (New York, N.Y.)*, 311(5762), pp. 796–797. doi: 10.1126/SCIENCE.1113832.

Davidson, I. F. *et al.* (2023) 'CTCF is a DNA-tension-dependent barrier to cohesin-mediated loop extrusion', *Nature 2023 616:7958*, 616(7958), pp. 822–827. doi: 10.1038/s41586-023-05961-5.

Dehingia, B. *et al.* (2022) ' CTCF shapes chromatin structure and gene expression in health and disease ', *EMBO reports*, 23(9). doi: 10.15252/EMBR.202255146/ASSET/FAD20AAE-EE99-492E-A832-503F5BBF51B3/ASSETS/GRAPHIC/EMBR202255146-FIG-0003-M.PNG.

Dekker, J. and Heard, E. (2015) 'Structural and Functional Diversity of Topologically Associating Domains', *FEBS letters*, 589(20 0 0), p. 2877. doi: 10.1016/J.FEBSLET.2015.08.044.

Demmerle, J., Hao, S. and Cai, D. (2023) 'Transcriptional condensates and phase separation: condensing information across scales and mechanisms', *Nucleus*, 14(1). doi: 10.1080/19491034.2023.2213551.

Deshpande, A. *et al.* (2022) 'Network inference with Granger causality ensembles on single-cell transcriptomics', *Cell reports*, 38(6). doi: 10.1016/J.CELREP.2022.110333.

Despang, A. *et al.* (2019) 'Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture', *Nature Genetics 2019 51:8*, 51(8), pp. 1263–1271. doi: 10.1038/s41588-019-0466-z.

Le Dily, F. L. *et al.* (2014) 'Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation', *Genes & Development*, 28(19), p. 2151. doi: 10.1101/GAD.241422.114.

Dixon, J. R. *et al.* (2012) 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature*, 485(7398), pp. 376–380. doi: 10.1038/nature11082.

Dixon, J. R. *et al.* (2015) 'Chromatin architecture reorganization during stem cell differentiation', *Nature 2015 518:7539*,

518(7539), pp. 331–336. doi: 10.1038/nature14222.

Dobi, K. C., Schulman, V. K. and Baylies, M. K. (2015) 'Specification of the somatic musculature in Drosophila', *Wiley interdisciplinary reviews. Developmental biology*, 4(4), p. 357. doi: 10.1002/WDEV.182.

Dominguez, M. and Campuzano, S. (1993) 'asense, a member of the Drosophila achaete-scute complex, is a proneural and neural differentiation gene', *The EMBO journal*, 12(5), pp. 2049–2060. doi: 10.1002/J.1460-2075.1993.TB05854.X.

Dong, P. *et al.* (2017) '3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments', *Molecular Plant*, 10(12), pp. 1497–1509. doi: 10.1016/j.molp.2017.11.005.

Dong, Q. *et al.* (2018) 'Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice', *The Plant Journal*, 94(6), pp. 1141–1156. doi: 10.1111/tpj.13925.

Dong, Y. *et al.* (2020) 'Promoter-Proximal Chromatin Domain Insulator Protein BEAF Mediates Local and Long-Range Communication with a Transcription Factor and Directly Activates a Housekeeping Promoter in Drosophila', *Genetics*, 215(1), pp. 89–101. doi: 10.1534/GENETICS.120.303144.

Dou, X. *et al.* (2023) 'METTL14 is a chromatin regulator independent of its RNA N6-methyladenosine methyltransferase activity', *Protein & cell*, 14(9). doi: 10.1093/procel/pwad009.

Drechsler, M. *et al.* (2013) 'The Conserved ADAMTS-like Protein Lonely heart Mediates Matrix Formation and Cardiac Tissue Integrity', *PLoS Genetics*, 9(7), p. 1003616. doi: 10.1371/journal.pgen.1003616.

Duan, Z. *et al.* (2010) 'A three-dimensional model of the yeast genome', *Nature 2010 465:7296*, 465(7296), pp. 363–367. doi: 10.1038/nature08973.

Dufourt, J. *et al.* (2018) 'Temporal control of gene expression by the pioneer factor Zelda through transient interactions in hubs', *Nature communications*, 9(1). doi: 10.1038/S41467-018-07613-Z.

Dufourt, J. *et al.* (2021) 'Imaging translation dynamics in live embryos reveals spatial heterogeneities', *Science (New York, N.Y.)*, 372(6544), pp. 840–844. doi: 10.1126/SCIENCE.ABC3483.

Dunin-Borkowski, O. M. and Brown, N. H. (1995) 'Mammalian CD2 is an effective heterologous marker of the cell surface in Drosophila', *Developmental biology*, 168(2), pp. 689–693. doi: 10.1006/DBIO.1995.1115.

Durand, N. C. *et al.* (2016) 'Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom', *Cell systems*, 3(1), pp. 99–101. doi: 10.1016/J.CELS.2015.07.012.

Ewels, P. *et al.* (2016) 'MultiQC: Summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*, 32(19). doi: 10.1093/bioinformatics/btw354.

Falcon, S. and Gentleman, R. (2007) 'Using GOstats to test gene lists for GO term association', *Bioinformatics*, 23(2), pp. 257–258. doi: 10.1093/bioinformatics/btl567.

Feng, S. *et al.* (2014) 'Genome-wide Hi-C analyses in wild type and mutants reveal high-resolution chromatin interactions in Arabidopsis', *Molecular cell*, 55(5), p. 694. doi: 10.1016/J.MOLCEL.2014.07.008.

Fernandez-Valverde, S. L., Aguilera, F. and Ramos-Dıaz, R. A. (2018) 'Inference of Developmental Gene Regulatory Networks Beyond Classical Model Systems: New Approaches in the Post-genomic Era', *Integrative and Comparative Biology*, 58(4), pp. 640–653. doi: 10.1093/ICB/ICY061.

Fickett, J. W. and Wasserman, W. W. (2000) 'Discovery and modeling of transcriptional regulatory regions', *Current Opinion in Biotechnology*. doi: 10.1016/S0958-1669(99)00049-X.

FitzGerald, P. C. *et al.* (2006) 'Comparative genomics of Drosophila and human core promoters', *Genome Biology*, 7(7), pp. 1–22. doi: 10.1186/GB-2006-7-7-R53/FIGURES/12.

Flavahan, W. A. *et al.* (2016) 'Insulator dysfunction and oncogene activation in IDH mutant gliomas', *Nature*, 529(7584), pp. 110–114. doi: 10.1038/NATURE16490.

Fortini, M. E., Lai, Z. and Rubin, G. M. (1991) 'The Drosophila zfh-1 and zfh-2 genes encode novel proteins containing both zinc-finger and homeodomain motifs', *Mechanisms of development*, 34(2–3), pp. 113–122. doi: 10.1016/0925-4773(91)90048-B.

Frankel, N. *et al.* (2010) 'Phenotypic robustness conferred by apparently redundant transcriptional enhancers', *Nature*, 466(7305), pp. 490–493. doi: 10.1038/NATURE09158.

Frasch, M. (2016) 'Genome-Wide Approaches to Drosophila Heart Development', *Journal of Cardiovascular Development and Disease*, 3(2), p. 20. doi: 10.3390/jcdd3020020.

Frasch, M. and Leptin, M. (2000) 'Minireview Mergers and Acquisitions: Unequal Partnerships in Drosophila Myoblast Fusion', *Cell*, 102, pp. 127–129.

Friedman, R. Z. *et al.* (2021) 'Information content differentiates enhancers from silencers in mouse photoreceptors', *eLife*, 10. doi: 10.7554/ELIFE.67403.

Fu, Z. *et al.* (2024) 'Cut&tag: a powerful epigenetic tool for chromatin profiling', *Epigenetics*, 19(1). doi: 10.1080/15592294.2023.2293411.

Fudenberg, G. *et al.* (2016) 'Formation of Chromosomal Domains by Loop Extrusion', *Cell reports*, 15(9), pp. 2038–2049. doi: 10.1016/J.CELREP.2016.04.085.

Fudenberg, G. and Imakaev, M. (2017) 'FISH-ing for captured contacts: towards reconciling FISH and 3C', *Nature methods*, 14(7), p. 673. doi: 10.1038/NMETH.4329.

Fujioka, M. *et al.* (1999) 'Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients', *Development (Cambridge, England)*, 126(11), pp. 2527–2538. doi: 10.1242/DEV.126.11.2527.

Fujioka, M., Sun, G. and Jaynes, J. B. (2013) 'The Drosophila eve Insulator Homie Promotes eve Expression and Protects the Adjacent Gene from Repression by Polycomb Spreading', *PLoS Genetics*, 9(10). doi: 10.1371/journal.pgen.1003883.

Fukaya, T., Lim, B. and Levine, M. (2016) 'Enhancer Control of Transcriptional Bursting', *Cell*, 166(2), pp. 358–368. doi: 10.1016/J.CELL.2016.05.025.

Fulco, C. P. *et al.* (2019) *Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations*, *Nature Genetics*. Nature Research. Available at: https://doi.org/10.1038/s41588-019-0538-0 (Accessed: 6 May 2024).

Furlong, E. E. M. *et al.* (2001) 'Patterns of gene expression during Drosophila mesoderm development', *Science*, 293(5535). doi: 10.1126/science.1062660.

Furlong, E. E. M. and Levine, M. (2018) 'Developmental enhancers and chromosome topology', *Science (New York, N.Y.)*, 361(6409), pp. 1341–1345. doi: 10.1126/SCIENCE.AAU0320.

Gajewski, K. *et al.* (1997) 'D-mef2 is a target for Tinman activation during Drosophila heart development', *The EMBO Journal*, 16(3), pp. 515–522. doi: 10.1093/EMBOJ/16.3.515.

Gajewski, K. *et al.* (2000) 'Genetically Distinct Cardial Cells Within the Drosophila Heart'. doi: 10.1002/1526-968X(200009)28:1<36::AID-GENE50>3.0.CO;2-4.

Gajewski, K. *et al.* (2001) 'Pannier is a transcriptional target and partner of Tinman during Drosophila cardiogenesis', *Developmental Biology*, 233(2). doi: 10.1006/dbio.2001.0220.

Gallagher, M. D. *et al.* (2017) 'A Dementia-Associated Risk Variant near TMEM106B Alters Chromatin Architecture and Gene Expression', *American journal of human genetics*, 101(5), pp. 643–663. doi: 10.1016/J.AJHG.2017.09.004.

Galouzis, C. C. and Furlong, E. E. M. (2022) 'Regulating specificity in enhancer–promoter communication', *Current Opinion in Cell Biology*. Elsevier Ltd, p. 102065. doi: 10.1016/j.ceb.2022.01.010.

Ganji, M. *et al.* (2018) 'Real-time imaging of DNA loop extrusion by condensin', *Science (New York, N.Y.)*, 360(6384), pp. 102–105. doi: 10.1126/SCIENCE.AAR7831.

Gasperini, M. *et al.* (2019) 'A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens', *Cell*, 176(6), p. 1516. doi: 10.1016/J.CELL.2019.02.027.

Gaszner, M. and Felsenfeld, G. (2006) 'Insulators: exploiting transcriptional and epigenetic mechanisms', *Nature reviews. Genetics*, 7(9), pp. 703–713. doi: 10.1038/NRG1925.

Gaszner, M., Vazquez, J. and Schedl, P. (1999) 'The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction', *Genes & development*, 13(16), pp. 2098–2107. doi: 10.1101/GAD.13.16.2098.

Gaudet, J. and Mango, S. E. (2002) 'Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4', *Science (New York, N.Y.)*, 295(5556), pp. 821–825. doi: 10.1126/SCIENCE.1065175.

Gdula, D. A., Gerasimova, T. I. and Corces, V. G. (1996) 'Genetic and molecular analysis of the gypsy chromatin insulator of Drosophila', *Proceedings of the National Academy of Sciences of the United States of America*, 93(18). doi: 10.1073/pnas.93.18.9378.

Gerasimova, T. I. and Corces, V. G. (1996) 'Boundary and insulator elements in chromosomes', *Current opinion in genetics & development*, 6(2), pp. 185–192. doi: 10.1016/S0959-437X(96)80049-9.

Geyer, P. K. and Corces, V. G. (1992) 'DNA position-specific repression of transcription by a Drosophila zinc finger protein', *Genes & development*, 6(10), pp. 1865–1873. doi: 10.1101/GAD.6.10.1865.

Ghavi-Helm, Y. *et al.* (2014) 'Enhancer loops appear stable during development and are associated with paused polymerase', *Nature*, 512(1). doi: 10.1038/nature13417.

Ghavi-Helm, Y. *et al.* (2019) 'Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression', *Nature genetics*, 51(8), pp. 1272–1282. doi: 10.1038/S41588-019-0462-3.

Giebel, B. (1999) 'The Notch signaling pathway is required to specify muscle progenitor cells in Drosophila', *Mechanisms of Development*, 86(1–2), pp. 137–145. doi: 10.1016/S0925-4773(99)00128-8.

Gilbert, M. K., Tan, Y. Y. and Hart, C. M. (2006) 'The drosophila boundary element-associated factors BEAF-32A and BEAF-32B affect chromatin structure', *Genetics*, 173(3). doi: 10.1534/genetics.106.056002.

Giorgetti, L. and Heard, E. (2016) 'Closing the loop: 3C versus DNA FISH', *Genome Biology*, 17(1), pp. 1–9. doi: 10.1186/S13059-016-1081-2/FIGURES/1.

González-Blas, C. B. *et al.* (2020) 'Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics', *Molecular systems biology*, 16(5). doi: 10.15252/MSB.20209438.

Goodman, D. B. *et al.* (2023) 'Evolutionary constraint and innovation across hundreds of placental mammals', *Science*, 380(6643). doi: 10.1126/SCIENCE.ABN3943.

Gopalan, S. and Fazzio, T. G. (2022) 'Multi-CUT&Tag to simultaneously profile multiple chromatin factors', *STAR Protocols*, 3(1), p. 101100. doi: 10.1016/J.XPRO.2021.101100.

Grant, C. E., Bailey, T. L. and Noble, W. S. (2011) 'FIMO: Scanning for occurrences of a given motif', *Bioinformatics*, 27(7). doi: 10.1093/bioinformatics/btr064.

Gratz, S. J. *et al.* (2013) 'Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease', *Genetics*, 194(4), pp. 1029–1035. doi: 10.1534/GENETICS.113.152710/-/DC1/GENETICS.113.152710-4.PDF.

Gu, Z. (2022) 'Complex heatmap visualization', *iMeta*, 1(3). doi: 10.1002/imt2.43.

Gu, Z., Eils, R. and Schlesner, M. (2016) 'Complex heatmaps reveal patterns and correlations in multidimensional genomic data', *Bioinformatics*, 32(18). doi: 10.1093/bioinformatics/btw313.

Guelen, L. *et al.* (2008) 'Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions', *Nature*, 453(7197), pp. 948–951. doi: 10.1038/nature06947.

Guglielmi, G. *et al.* (2015) 'An Optogenetic Method to Modulate Cell Contractility during Tissue Morphogenesis', *Developmental cell*, 35(5), pp. 646–660. doi: 10.1016/J.DEVCEL.2015.10.020.

Gullaud, M., Delanoue, R. and Silber, J. (2003) 'A Drosophila model to study the functions of TWIST orthologs in apoptosis and proliferation', *Cell death and differentiation*, 10(6), pp. 641–651. doi: 10.1038/SJ.CDD.4401222.

Guo, Y. *et al.* (2015) 'CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function', *Cell*, 162(4), p. 900. doi: 10.1016/J.CELL.2015.07.038.

Gupta, S. *et al.* (2007) 'Quantifying similarity between motifs', *Genome Biology*, 8(2). doi: 10.1186/gb-2007-8-2-r24.

Gyurkovics, H. *et al.* (1990) 'A new homeotic mutation in the Drosophila bithorax complex removes a boundary separating two domains of regulation.', *The EMBO Journal*, 9(8), pp. 2579–2585. doi: 10.1002/j.1460-2075.1990.tb07439.x.

Haarhuis, J. H. I. *et al.* (2017) 'The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension', *Cell*, 169(4), pp. 693-707.e14. doi: 10.1016/J.CELL.2017.04.013.

Haberle, V. *et al.* (2019) 'Transcriptional cofactors display specificity for distinct types of core promoters', *Nature 2019 570:7759*, 570(7759), pp. 122–126. doi: 10.1038/s41586-019-1210-7.

Haberle, V. and Stark, A. (2018) 'Eukaryotic core promoters and the functional basis of transcription initiation', *Nature reviews. Molecular cell biology*, 19(10), pp. 621–637. doi: 10.1038/S41580-018-0028-8.

Hafner, A. *et al.* (2023) 'Loop stacking organizes genome folding from TADs to chromosomes', *Molecular cell*, 83(9), pp. 1377-1392.e6. doi: 10.1016/J.MOLCEL.2023.04.008.

Haghverdi, L. *et al.* (2018) 'Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors', *Nature biotechnology*, 36(5), pp. 421–427. doi: 10.1038/NBT.4091.

Hagstrom, K., Muller, M. and Schedl, P. (1996) 'Fab-7 functions as a chromatin domain boundary to ensure proper segment specification by the Drosophila bithorax complex.', *Genes & Development*, 10(24), pp. 3202–3215. doi: 10.1101/GAD.10.24.3202.

Halfon, M. S. *et al.* (2000) 'Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors', *Cell*, 103(1). doi: 10.1016/S0092-8674(00)00105-7.

Hammonds, A. S. *et al.* (2013) 'Spatial expression of transcription factors in Drosophila embryonic organ development', *Genome biology*, 14(12). doi: 10.1186/GB-2013-14-12-R140.

Hanssen, L. L. P. *et al.* (2017) 'Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo', *Nature cell biology*, 19(8), pp. 952–961. doi: 10.1038/NCB3573.

Hao, Y. *et al.* (2021) 'Integrated analysis of multimodal single-cell data', *Cell*, 184(13), pp. 3573-3587.e29. doi: 10.1016/J.CELL.2021.04.048.

Harafuji, N., Keys, D. N. and Levine, M. (2002) 'Genome-wide identification of tissue-specific enhancers in the Ciona tadpole', *Proceedings of the National Academy of Sciences*, 99(10), pp. 6802–6805. doi: 10.1073/pnas.052024999.

Harrison, M. M. *et al.* (2011) 'Zelda Binding in the Early Drosophila melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition', *PLOS Genetics*, 7(10), p. e1002266. doi: 10.1371/JOURNAL.PGEN.1002266.

Harrison, M. M., Botchan, M. R. and Cline, T. W. (2010) 'Grainyhead and Zelda compete for binding to the promoters of the earliest-expressed Drosophila genes', *Developmental biology*, 345(2), pp. 248–255. doi: 10.1016/J.YDBIO.2010.06.026.

Hartenstein, V. (1993) 'Atlas of Drosophila development', *The Development of Drosophila melanogaster*.

Hartenstein, V. and Campos-Ortega, J. A. (1985) 'Fate-mapping in wild-type Drosophila melanogaster - I. The spatio-temporal pattern of embryonic cell divisions', *Wilhelm Roux's Archives of Developmental Biology*, 194(4). doi: 10.1007/BF00848246.

He, H. H. *et al.* (2010) 'Nucleosome dynamics define transcriptional enhancers', *Nature genetics*, 42(4), pp. 343–347. doi: 10.1038/NG.545.

Heger, P. *et al.* (2012) 'The chromatin insulator CTCF and the emergence of metazoan diversity', *Proceedings of the National Academy of Sciences of the United States of America*, 109(43), pp. 17507–17512. doi: 10.1073/PNAS.1111941109/-/DCSUPPLEMENTAL/SAPP.PDF.

Heintzman, N. D. *et al.* (2009) 'Histone Modifications at Human Enhancers Reflect Global Cell Type-Specific Gene Expression', *Nature*, 459(7243), p. 108. doi: 10.1038/NATURE07829.

Heitz, E. (1928) 'Das Heterochromatin der Moose', *Jahrbücher für Wissenschaftliche Botanik*. Available at: https://books.google.ch/books/about/Das_Heterochromatin_der_Moose.html?id=3JwgSQAACAAJ&redir_esc=y.

Henderson, J. *et al.* (2019) 'Accurate prediction of boundaries of high resolution topologically associated domains (TADs) in fruit flies using deep learning', *Nucleic Acids Research*, 47(13), pp. e78–e78. doi: 10.1093/NAR/GKZ315.

Herz, H. M. *et al.* (2012) 'Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4', *Genes & Development*, 26(23), p. 2604. doi: 10.1101/GAD.201327.112.

Hobert, O. (2008) 'Gene regulation by transcription factors and microRNAs', *Science (New York, N.Y.)*, 319(5871), pp. 1785–1786. doi: 10.1126/SCIENCE.1151651.

Hogga, I. *et al.* (2001) 'Replacement of Fab-7 by the gypsy or scs insulator disrupts long-distance regulatory interactions in the Abd-B gene of the bithorax complex', *Molecular cell*, 8(5), pp. 1145–1151. doi: 10.1016/S1097-2765(01)00377-X.

Holland, A. J. *et al.* (2012) 'Inducible, reversible system for the rapid and complete degradation of proteins in mammalian cells', *Proceedings of the National Academy of Sciences of the United States of America*, 109(49). doi: 10.1073/PNAS.1216880109.

Hollfelder, D., Frasch, M. and Reim, I. (2014) 'Distinct functions of the laminin β LN domain and collagen IV during cardiac extracellular matrix formation and stabilization of alary muscle attachments revealed by EMS mutagenesis in Drosophila', *BMC developmental biology*, 14(1). doi: 10.1186/1471-213X-14-26.

Horikoshi, M. *et al.* (1988) 'Transcription factor ATF interacts with the TATA factor to facilitate establishment of a preinitiation complex', *Cell*, 54(7), pp. 1033–1042. doi: 10.1016/0092-8674(88)90118-3.

Horikoshi, M. *et al.* (1989) 'Cloning and structure of a yeast gene encoding a general transcription initiation factor TFIID that binds to the TATA box', *Nature 1989 341:6240*, 341(6240), pp. 299–303. doi: 10.1038/341299a0.

Hota, S. K. *et al.* (2022) 'Brahma safeguards canalization of cardiac mesoderm differentiation', *Nature*, 602(7895), pp. 129–134. doi: 10.1038/S41586-021-04336-Y.

Hou, C. *et al.* (2012) 'Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains', *Molecular cell*, 48(3), pp. 471–484. doi: 10.1016/J.MOLCEL.2012.08.031.

Hou, W. *et al.* (2023) 'A statistical framework for differential pseudotime analysis with multiple single-cell RNA-seq samples', *Nature Communications 2023 14:1*, 14(1), pp. 1–21. doi: 10.1038/s41467-023-42841-y.

Hsieh, T.-H. S. *et al.* (2015) 'Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C', *Cell*, 162(1), pp. 108–119. doi: 10.1016/j.cell.2015.05.048.

Hsieh, T. H. S. *et al.* (2020) 'Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding', *Molecular cell*, 78(3), pp. 539-553.e8. doi: 10.1016/J.MOLCEL.2020.03.002.

Huang, D. *et al.* (2019) 'Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression', *Genome research*, 29(4), pp. 657–667. doi: 10.1101/GR.247007.118.

Huber, P. J. (1981) 'Robust Statistics'. doi: 10.1002/0471725250.

Hug, C. B. *et al.* (2017) 'Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription', *Cell*, 169(2), pp. 216-228.e19. doi: 10.1016/j.cell.2017.03.024.

Hwang, B., Lee, J. H. and Bang, D. (2018) 'Single-cell RNA sequencing technologies and bioinformatics pipelines', *Experimental & Molecular Medicine 2018 50:8*, 50(8), pp. 1–14. doi: 10.1038/s12276-018-0071-8.

Iampietro, C. *et al.* (2008) 'Boundary swapping in the Drosophila Bithorax complex', *Development (Cambridge, England)*, 135(24), pp. 3983–3987. doi: 10.1242/DEV.025700.

Ing-Simmons, E. *et al.* (2021) 'Independence of chromatin conformation and gene regulation during Drosophila dorsoventral patterning', *Nature Genetics 2021 53:4*, 53(4), pp. 487–499. doi: 10.1038/s41588-021-00799-x.

Inukai, S., Kock, K. H. and Bulyk, M. L. (2017) 'Transcription factor–DNA binding: beyond binding site motifs', *Current opinion in genetics & development*, 43, p. 110. doi: 10.1016/J.GDE.2017.02.007.

Ip, Y. T. *et al.* (1992) 'The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the Drosophila embryo', *Genes & development*, 6(9), pp. 1728–1739. doi: 10.1101/GAD.6.9.1728.

Ip, Y. T., Maggert, K. and Levine, M. (1994) 'Uncoupling gastrulation and mesoderm differentiation in the Drosophila embryo', *The EMBO journal*, 13(24), pp. 5826–5834. doi: 10.1002/J.1460-2075.1994.TB06926.X.

Islam, S. *et al.* (2011) 'Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq', *Genome Research*, 21(7), p. 1160. doi: 10.1101/GR.110882.110.

Islam, Z. *et al.* (2023) 'Active enhancers strengthen insulation by RNA-mediated CTCF binding at chromatin domain boundaries', *Genome research*, 33(1), pp. 1–17. doi: 10.1101/GR.276643.122.

Ismat, A. *et al.* (2010) 'HLH54F is required for the specification and migration of longitudinal gut muscle founders from the caudal mesoderm of Drosophila', *Development*, 137(18), pp. 3107–3117. doi: 10.1242/dev.046573.

Iwafuchi-Doi, M. and Zaret, K. S. (2014) 'Pioneer transcription factors in cell reprogramming', *Genes & development*, 28(24), pp. 2679–2692. doi: 10.1101/GAD.253443.114.

Jack, J. *et al.* (1991) 'Expression of the cut locus in the Drosophila wing margin is required for cell type specification and is regulated by a distant enhancer', *Development*, 113(3). doi: 10.1242/dev.113.3.735.

Jackson, D. A. *et al.* (1993) 'Visualization of focal sites of transcription within human nuclei.; Visualization of focal sites of transcription within human nuclei.', *The EMBO Journal*, 1(3), pp. 1059–1065. doi: 10.1002/j.1460-2075.1993.tb05747.x.

Jakobsen, J. S. *et al.* (2007) 'Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network', *Genes and Development*, 21(19), pp. 2448–2460. doi: 10.1101/gad.437607.

Janssens, D. H. *et al.* (2022) 'CUT&Tag2for1: a modified method for simultaneous profiling of the accessible and silenced regulome in single cells', *Genome biology*, 23(1). doi: 10.1186/S13059-022-02642-W.

Jen, W. P. *et al.* (2020) 'Twist1 Plays an Anti-apoptotic Role in Mutant Huntingtin Expression Striatal Progenitor Cells', *Molecular neurobiology*, 57(3), pp. 1688–1703. doi: 10.1007/S12035-019-01836-X.

Jenkins, V. K., Larkin, A. and Thurmond, J. (2022) 'Using FlyBase: A Database of Drosophila Genes and Genetics', *Methods in Molecular Biology*, 2540, pp. 1–34. doi: 10.1007/978-1-0716-2541-5_1/TABLES/3.

Jiang, N. *et al.* (2009) 'Genome-wide mapping of boundary element-associated factor (BEAF) binding sites in Drosophila melanogaster links BEAF to transcription', *Molecular and cellular biology*, 29(13), pp. 3556–3568. doi: 10.1128/MCB.01748-08.

Jiang, Y. *et al.* (2020) 'Genome-wide analyses of chromatin interactions after the loss of Pol I, Pol II, and Pol III', *Genome biology*, 21(1). doi: 10.1186/S13059-020-02067-3.

Jin, H. *et al.* (2013) 'Genome-Wide Screens for In Vivo Tinman Binding Sites Identify Cardiac Enhancers with Diverse Functional Architectures', *PLoS Genetics*, 9(1). doi: 10.1371/journal.pgen.1003195.

Johnson, H. E. *et al.* (2017) 'The Spatiotemporal Limits of Developmental Erk Signaling', *Developmental cell*, 40(2), pp. 185–192. doi: 10.1016/J.DEVCEL.2016.12.002.

Johnstone, I. M. and Titterington, D. M. (2009) 'Statistical challenges of high-dimensional data', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906), pp. 4237–4253. doi: 10.1098/RSTA.2009.0159.

Jordan, I. K. *et al.* (2002) 'Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria', *Genome Research*, 12(6), pp. 962–968. doi: 10.1101/gr.87702.

Joung, J. *et al.* (2023) 'A transcription factor atlas of directed differentiation', *Cell*, 186(1), pp. 209-229.e26. doi: 10.1016/J.CELL.2022.11.026.

Junion, G. *et al.* (2012) 'A transcription factor collective defines cardiac cell fate and reflects lineage history', *Cell*, 148(3), pp. 473–486. doi: 10.1016/j.cell.2012.01.030.

Juven-Gershon, T., Hsu, J. Y. and Kadonaga, J. T. (2008) 'Caudal, a key developmental regulator, is a DPE-specific transcriptional factor', *Genes & Development*, 22(20), pp. 2823–2830. doi: 10.1101/GAD.1698108.

Juven-Gershon, T. and Kadonaga, J. T. (2010) 'Regulation of gene expression via the core promoter and the basal transcriptional machinery', *Developmental Biology*. Elsevier Inc., pp. 225–229. doi: 10.1016/j.ydbio.2009.08.009.

Kabirova, E. *et al.* (2023) 'Function and Evolution of the Loop Extrusion Machinery in Animals', *International Journal of Molecular Sciences*, 24(5). doi: 10.3390/IJMS24055017.

Kadonaga, J. T. (2012) 'Perspectives on the RNA polymerase II core promoter', *Wiley interdisciplinary reviews. Developmental biology*, 1(1), pp. 40–51. doi: 10.1002/WDEV.21.

Kahn, T. G. *et al.* (2006) 'Polycomb complexes and the propagation of the methylation mark at the drosophila Ubx gene', *Journal of Biological Chemistry*, 281(39). doi: 10.1074/jbc.M605430200.

Kahn, T. G. *et al.* (2023) 'Topological screen identifies hundreds of Cp190- and CTCF-dependent Drosophila chromatin insulator elements', *Science Advances*, 9(5). doi: 10.1126/SCIADV.ADE0090/SUPPL_FILE/SCIADV.ADE0090_SUPPLEMENTARY_CODE.ZIP.

Kai, Y. *et al.* (2018) 'Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features', *Nature Communications*, 9(1), pp. 1–14. doi: 10.1038/S41467-018-06664-6/FIGURES/7.

Kamal, A. *et al.* (2023) 'GRaNIE and GRaNPA : inference and evaluation of enhancer-mediated gene regulatory networks', *Molecular Systems Biology*, 19(6). doi: 10.15252/MSB.202311627.

Kamimoto, K. *et al.* (2023) 'Dissecting cell identity via network inference and in silico gene perturbation', *Nature 2023 614:7949*, 614(7949), pp. 742–751. doi: 10.1038/s41586-022-05688-9.

Karollus, A., Mauermeier, T. and Gagneur, J. (2023) 'Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers', *Genome Biology*, 24(1), pp. 1–29. doi: 10.1186/S13059-023-02899-9/FIGURES/6.

Kaur, P., Saunders, T. E. and Tolwinski, N. S. (2017) 'Coupling optogenetics and light-sheet microscopy, a method to study Wnt signaling during embryogenesis', *Scientific Reports 2017 7:1*, 7(1), pp. 1–11. doi: 10.1038/s41598-017-16879-0.

Kaushal, A. *et al.* (2021) 'CTCF loss has limited effects on global genome architecture in Drosophila despite critical regulatory functions', *Nature Communications 2021 12:1*, 12(1), pp. 1–16. doi: 10.1038/s41467-021-21366-2.

Kaya-Okur, H. S. *et al.* (2019) 'CUT&Tag for efficient epigenomic profiling of small samples and single cells', *Nature Communications 2019 10:1*, 10(1), pp. 1–10. doi: 10.1038/s41467-019-09982-5.

Kellum, R. and Schedl, P. (1991) 'A position-effect assay for boundaries of higher order chromosomal domains', *Cell*, 64(5), pp. 941–950. doi: 10.1016/0092-8674(91)90318-S.

Kellum, R. and Schedl, P. (1992) 'A group of scs elements function as domain boundaries in an enhancer-blocking assay.', *Molecular and Cellular Biology*, 12(5), pp. 2424–2431. doi: 10.1128/MCB.12.5.2424.

Kelly, K. K., Meadows, S. M. and Cripps, R. M. (2002) 'Drosophila MEF2 is a direct regulator of Actin57B transcription in cardiac, skeletal, and visceral muscle lineages', *Mechanisms of Development*, 110(1–2), pp. 39–50. doi: 10.1016/S0925-4773(01)00586-X.

Kennedy-Darling, J. and Smith, L. M. (2014) 'Measuring the formaldehyde protein-DNA cross-link reversal rate', *Analytical Chemistry*, 86(12), pp. 5678–5681. doi: 10.1021/AC501354Y/SUPPL_FILE/AC501354Y_SI_001.PDF.

Kerpedjiev, P. *et al.* (2018) 'HiGlass: Web-based visual exploration and analysis of genome interaction maps', *Genome Biology*, 19(1), pp. 1–12. doi: 10.1186/S13059-018-1486-1/FIGURES/6.

Khoueiry, P. *et al.* (2017) 'Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity', *eLife*, 6. doi: 10.7554/ELIFE.28440.

Kim, J. *et al.* (1996) 'The DNA-binding and enhancer-blocking domains of the Drosophila suppressor of Hairy-wing protein', *Molecular and cellular biology*, 16(7), pp. 3381–3392. doi: 10.1128/MCB.16.7.3381.

Kim, J. *et al.* (2021) 'TENET: Gene network reconstruction using transfer entropy reveals key regulatory factors from single cell transcriptomic data', *Nucleic Acids Research*, 49(1). doi: 10.1093/NAR/GKAA1014.

Kim, S. and Wysocka, J. (2023) 'Deciphering the multi-scale, quantitative cis-regulatory code', *Molecular Cell*, 83, pp. 373–392. doi: 10.1016/j.molcel.2022.12.032.

Kim, T. K. *et al.* (2010) 'Widespread transcription at neuronal activity-regulated enhancers', *Nature*, 465(7295), pp. 182–187. doi: 10.1038/NATURE09033.

Kim, Y. O. *et al.* (2004) 'A functional genomic screen for cardiogenic genes using RNA interference in developing Drosophila

embryos', *Proceedings of the National Academy of Sciences of the United States of America*, 101(1), pp. 159–164. doi: 10.1073/pnas.0307205101.

Knight, P. A. and Ruiz, D. (2013) 'A fast algorithm for matrix balancing', *IMA Journal of Numerical Analysis*, 33(3), pp. 1029–1047. doi: 10.1093/IMANUM/DRS019.

Kögler, A. C. *et al.* (2021) 'Extremely rapid and reversible optogenetic perturbation of nuclear proteins in living embryos', *Developmental cell*, 56(16), pp. 2348-2363.e8. doi: 10.1016/J.DEVCEL.2021.07.011.

Komuro, I. and Izumo, S. (1993) 'Csx: a murine homeobox-containing gene specifically expressed in the developing heart.', *Proceedings of the National Academy of Sciences of the United States of America*, 90(17), p. 8145. doi: 10.1073/PNAS.90.17.8145.

Korzelius, J. *et al.* (2014) ' Escargot maintains stemness and suppresses differentiation in Drosophila intestinal stem cells ', *The EMBO Journal*, 33(24). doi: 10.15252/embj.201489072.

Krivega, M. *et al.* (2010) 'Interaction between a pair of gypsy insulators or between heterologous gypsy and Wari insulators modulates Flp site-specific recombination in Drosophila melanogaster', *Chromosoma*, 119(4), pp. 425–434. doi: 10.1007/S00412-010-0268-7.

Kudron, M. M. *et al.* (2018) 'The modern resource: genome-wide binding profiles for hundreds of Drosophila and Caenorhabditis elegans transcription factors', *Genetics*, 208(3), pp. 937–949. doi: 10.1534/genetics.117.300657.

Kuhn, E. J., Hart, C. M. and Geyer, P. K. (2004) 'Studies of the role of the Drosophila scs and scs' insulators in defining boundaries of a chromosome puff', *Molecular and cellular biology*, 24(4), pp. 1470–1480. doi: 10.1128/MCB.24.4.1470-1480.2004.

Kurukuti, S. *et al.* (2006) 'CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2', *Proceedings of the National Academy of Sciences of the United States of America*, 103(28). doi: 10.1073/pnas.0600326103.

Kvon, E. Z. *et al.* (2014) 'Genome-scale functional characterization of Drosophila developmental enhancers in vivo', *Nature 2014 512:7512*, 512(7512), pp. 91–95. doi: 10.1038/nature13395.

Kvon, E. Z. (2015) 'Using transgenic reporter assays to functionally characterize enhancers in animals', *Genomics*, 106(3), pp. 185–192. doi: 10.1016/J.YGENO.2015.06.007.

Kyrchanova, O. *et al.* (2007) 'Study of the Functional Interaction between Mcp Insulators from the Drosophila bithorax Complex: Effects of Insulator Pairing on Enhancer-Promoter Communication', *Molecular and Cellular Biology*, 27(8), pp. 3035–3043. doi: 10.1128/MCB.02203-06.

Kyrchanova, O., Toshchakov, S., *et al.* (2008) 'Functional Interaction between the Fab-7 and Fab-8 Boundaries and the Upstream Promoter Region in the Drosophila Abd-B Gene', *Molecular and Cellular Biology*, 28(12), pp. 4188–4195. doi: 10.1128/MCB.00229-08.

Kyrchanova, O., Chetverina, D., *et al.* (2008) 'Orientation-dependent interaction between Drosophila insulators is a property of this class of regulatory elements', *Nucleic Acids Research*, 36(22), p. 7019. doi: 10.1093/NAR/GKN781.

Kyrchanova, O. *et al.* (2016) 'Functional Dissection of the Blocking and Bypass Activities of the Fab-8 Boundary in the Drosophila Bithorax Complex', *PLoS Genetics*, 12(7). doi: 10.1371/JOURNAL.PGEN.1006188.

Kyrchanova, O. *et al.* (2019) 'Complete reconstitution of bypass and blocking functions in a minimal artificial Fab-7 insulator from Drosophila bithorax complex', *Proceedings of the National Academy of Sciences of the United States of America*, 116(27), pp. 13462–13467. doi: 10.1073/PNAS.1907190116.

Kyrchanova, O. and Georgiev, P. (2014) 'Chromatin insulators and long-distance interactions in Drosophila', *FEBS Letters*, 588(1), pp. 8–14. doi: 10.1016/j.febslet.2013.10.039.

Laat, W. De and Duboule, D. (2013) 'Topology of mammalian developmental enhancers and their regulatory landscapes', *Nature*, 502(7472), pp. 499–506. doi: 10.1038/NATURE12753.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Larsson, A. J. M. *et al.* (2019) 'Genomic encoding of transcriptional burst kinetics', *Nature*, 565(7738), pp. 251–254. doi: 10.1038/S41586-018-0836-1.

Laugsch, M. *et al.* (2019) 'Modeling the Pathological Long-Range Regulatory Effects of Human Structural Variation with Patient-Specific hiPSCs', *Cell stem cell*, 24(5), pp. 736-752.e12. doi: 10.1016/J.STEM.2019.03.004.

Lawrence, M. *et al.* (2013) 'Software for Computing and Annotating Genomic Ranges', *PLOS Computational Biology*, 9(8), p. e1003118. doi: 10.1371/JOURNAL.PCBI.1003118.

Le, T. B. K. *et al.* (2013) 'High-resolution mapping of the spatial organization of a bacterial chromosome', *Science*, 342(6159), pp. 731–734. doi: 10.1126/SCIENCE.1242059/SUPPL_FILE/LE.SM.PDF.

Lebestky, T. *et al.* (2000) 'Specification of Drosophila hematopoietic lineage by conserved transcription factors', *Science (New York, N.Y.)*, 288(5463), pp. 146–149. doi: 10.1126/SCIENCE.288.5463.146.

Lee, C. *et al.* (2008) 'NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila', *Molecular and cellular biology*, 28(10), pp. 3290–3300. doi: 10.1128/MCB.02224-07.

Lee, H.-H., Zaffran, S. and Frasch, M. (2005) 'The development of the visceral musculature of Drosophila', in *Drosophila Muscle Development*. Sink H.

Lee, H. H. and Frasch, M. (2000) 'Wingless effects mesoderm patterning and ectoderm segmentation events via induction of its downstream target sloppy paired', *Development*, 127(24). doi: 10.1242/dev.127.24.5497.

Lee, H. H. and Frasch, M. (2005) 'Nuclear integration of positive Dpp signals, antagonistic Wg inputs and mesodermal competence factors during Drosophila visceral mesoderm induction', *Development (Cambridge, England)*, 132(6), pp. 1429–1442. doi: 10.1242/DEV.01687.

Lee, K.-H., Xu, Q. and Breitbart, R. E. (1996) 'A New tinman-Related Gene, nkx2.7, Anticipates the Expression of nkx2.5 and nkx2.3 in Zebrafish Heart and Pharyngeal Endoderm', *DEVELOPMENTAL BIOLOGY*, 180, pp. 722–731.

Lee, Y. M. *et al.* (1997) 'Twist-mediated activation of the NK-4 homeobox gene in the visceral mesoderm of Drosophila requires two distinct clusters of E-box regulatory elements', *The Journal of biological chemistry*, 272(28), pp. 17531–17541. doi: 10.1074/JBC.272.28.17531.

Leifer, I. *et al.* (2020) 'Circuits with broken fibration symmetries perform core logic computations in biological networks', *PLOS Computational Biology*, 16(6), p. e1007776. doi: 10.1371/JOURNAL.PCBI.1007776.

Lenhard, B., Sandelin, A. and Carninci, P. (2012) 'Metazoan promoters: Emerging characteristics and insights into transcriptional regulation', *Nature Reviews Genetics*, 13(4), pp. 233–245. doi: 10.1038/nrg3163.

Leptin, M. (1991) 'twist and snail as positive and negative regulators during Drosophila mesoderm development', *Genes & development*, 5(9), pp. 1568–1576. doi: 10.1101/GAD.5.9.1568.

Levine, M. and Davidson, E. H. (2005) 'Gene regulatory networks for development', *Proceedings of the National Academy of Sciences of the United States of America*, 102(14), pp. 4936–4942. doi: 10.1073/PNAS.0408031102/ASSET/F9FA3AD8-360A-48EB-9F89-DD59D2BDB2D7/ASSETS/GRAPHIC/ZPQ0120577430002.JPEG.

Lewis, P. W. *et al.* (2004) 'Identification of a Drosophila Myb-E2F2/RBF transcriptional repressor complex', *Genes and Development*, 18(23). doi: 10.1101/gad.1255204.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), p. 2078. doi: 10.1093/BIOINFORMATICS/BTP352.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics (Oxford, England)*, 25(14), pp. 1754–1760. doi: 10.1093/BIOINFORMATICS/BTP324.

Li, J. *et al.* (2020) 'Single-gene imaging links genome topology, promoter-enhancer communication and transcription control', *NATure STruCTurAL & MOLeCuLAr BiOLOGY* |, 27. doi: 10.1038/s41594-020-0493-6.

Li, Q. *et al.* (2011) 'Measuring reproducibility of high-throughput experiments', *Annals of Applied Statistics*, 5(3), pp. 1752–1779. doi: 10.1214/11-AOAS466.

Li, X. and Noll, M. (1994) 'Compatibility between enhancers and promoters determines the transcriptional specificity of

gooseberry and gooseberry neuro in the Drosophila embryo.', *The EMBO Journal*, 13(2), pp. 400–406. doi: 10.1002/J.1460-2075.1994.TB06274.X.

Liang, H. L. *et al.* (2008) 'The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila', *Nature*, 456(7220). doi: 10.1038/nature07388.

Liber, D. *et al.* (2010) 'Epigenetic Priming of a Pre-B Cell-Specific Enhancer through Binding of Sox2 and Foxd3 at the ESC Stage', *Cell Stem Cell*, 7(1), pp. 114–126. doi: 10.1016/J.STEM.2010.05.020.

Lieberman-Aiden, E. *et al.* (2009) 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome', *Science*, 326(5950), pp. 289–293.

Lilly, B. *et al.* (1995) 'Requirement of MADS domain transcription factor D-MEF2 for muscle formation in Drosophila', *Science*, 267(5198), pp. 688–693. doi: 10.1126/science.7839146.

Lin, M. H. *et al.* (1997) 'Ectopic expression of MEF2 in the epidermis induces epidermal expression of muscle genes and abnormal muscle development in Drosophila', *Developmental Biology*, 182(2). doi: 10.1006/dbio.1996.8484.

Lin, S. C. and Storti, R. V. (1997) 'Developement regulation of the Drosophila Tropomyosin I (TmI) gene is controlled by muscle activator enhancer region that contains multiple cis- elements and binding sites for multiple proteins', *Developmental Genetics*, 20(4). doi: 10.1002/(SICI)1520-6408(1997)20:4<297::AID-DVG1>3.0.CO;2-2.

Lints, T. J. *et al.* (1993) 'Nkx-2.5: a novel murine homeobox gene expressed in early heart progenitor cells and their myogenic descendants', *Development (Cambridge, England)*, 119(2), pp. 419–431. doi: 10.1242/DEV.119.2.419.

Liu, C. *et al.* (2017) 'Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis', *Nature plants,*. Available at: https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/s41477-017-0005-9&casa_token=bqcp0UcoI3wAAAAA:o6b_viVzuXsY4M_nwn8Wj3LlaxWTHtf6HRxpgmmq0hNDwU9YsfQ_Mm86toBt7_NC2JktMdlcc2-tlcE5Kjc.

Liu, Y. H. *et al.* (2009) 'A Systematic Analysis of Tinman Function Reveals Eya and JAK-STAT Signaling as Essential Regulators of Muscle Development', *Developmental Cell*, 16(2), pp. 280–291. doi: 10.1016/j.devcel.2009.01.006.

Lo, P. C. H. and Frasch, M. (2001) 'A role for the COUP-TF-related gene seven-up in the diversification of cardioblast identities in the dorsal vessel of Drosophila', *Mechanisms of Development*, 104(1–2), pp. 49–60. doi: 10.1016/S0925-4773(01)00361-6.

Long, H. (2016) 'Ever-changing landscapes: transcriptional enhancers in development and evolution', *Cell*, 167(5), pp. 248–268. doi: 10.1002/9780470696590.ch14.

Long, H. K., Prescott, S. L. and Wysocka, J. (2016) 'Leading Edge Review Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution', *Cell*, 167, pp. 1170–1187. doi: 10.1016/j.cell.2016.09.018.

Lovato, T. L. *et al.* (2015) 'The drosophila transcription factors Tinman and pannier activate and collaborate with Myocyte Enhancer Factor-2 to promote heart cell fate', *PLoS ONE*, 10(7), pp. 1–18. doi: 10.1371/journal.pone.0132965.

Lu, D. *et al.* (2020) 'Developmental regulation of cell type-specific transcription by novel promoter-proximal sequence elements', *Genes and Development*, 34(9–10), pp. 663–677. doi: 10.1101/gad.335331.119.

Ludwig, M. Z., Patel, N. H. and Kreitman, M. (1998) 'Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change', *Development (Cambridge, England)*, 125(5), pp. 949–958. doi: 10.1242/DEV.125.5.949.

Lupiáñez, D. G. *et al.* (2015) 'Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions', *Cell*, 161(5), pp. 1012–1025. doi: 10.1016/J.CELL.2015.04.004.

Lyko, F. *et al.* (1997) 'An imprinting element from the mouse H19 locus functions as a silencer in Drosophila', *Nature genetics*, 16(2), pp. 171–173. doi: 10.1038/NG0697-171.

M., Z., Zalokar, M. and Erk, I. (1976) 'Division and migration of nuclei during early embryogenesis of Drosophila melanogaster.', *J. Microsc. Biol. Cell.*, 25(2), pp. 97–106. Available at: https://cir.nii.ac.jp/crid/1573387449290484480.

Machanick, P. and Bailey, T. L. (2011) 'MEME-ChIP: Motif analysis of large DNA datasets', *Bioinformatics*, 27(12). doi:

10.1093/bioinformatics/btr189.

MacKay, T. F. C. *et al.* (2012) 'The Drosophila melanogaster Genetic Reference Panel', *Nature*, 482(7384). doi: 10.1038/nature10811.

MacNeil, L. T. and Walhout, A. J. M. (2011) 'Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression', *Genome Research*, 21(5), pp. 645–657. doi: 10.1101/GR.097378.109.

Maestro, R. *et al.* (1999) 'twist is a potential oncogene that inhibits apoptosis', *Genes & Development*, 13(17), p. 2207. doi: 10.1101/GAD.13.17.2207.

Maksimenko, O. *et al.* (2015) 'Two new insulator proteins, Pita and ZIPIC, target CP190 to chromatin', *Genome research*, 25(1), pp. 89–99. doi: 10.1101/GR.174169.114.

Marbach, D. *et al.* (2012) 'Wisdom of crowds for robust gene network inference', *Nature methods*, 9(8), pp. 796–804. doi: 10.1038/NMETH.2016.

Marbouty, M. *et al.* (2015) 'Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging', *Molecular Cell*, 59(4), pp. 588–602. doi: 10.1016/j.molcel.2015.07.020.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), p. 10. doi: 10.14806/ej.17.1.200.

Martín, M., Organista, M. F. and de Celis, J. F. (2016) 'Structure of developmental gene regulatory networks from the perspective of cell fate-determining genes', *Transcription*, 7(1), pp. 32–37. doi: 10.1080/21541264.2015.1130118.

Mateo, L. J. *et al.* (2019) 'Visualizing DNA folding and RNA in embryos at single-cell resolution', *Nature*, 568(7750), pp. 49–54. doi: 10.1038/S41586-019-1035-4.

Matthews, N. E. and White, R. (2019) 'Chromatin Architecture in the Fly: Living without CTCF/Cohesin Loop Extrusion?', *BioEssays*, 41(9), p. 1900048. doi: 10.1002/bies.201900048.

Mbodj, A. *et al.* (2016) 'Qualitative Dynamical Modelling Can Formally Explain Mesoderm Specification and Predict Novel Developmental Phenotypes', *PLoS Computational Biology*, 12(9), pp. 1–17. doi: 10.1371/journal.pcbi.1005073.

McKenna, A. *et al.* (2010) 'The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Research*, 20(9). doi: 10.1101/gr.107524.110.

McLeay, R. C. and Bailey, T. L. (2010) 'Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data', *BMC Bioinformatics*, 11. doi: 10.1186/1471-2105-11-165.

Melnikova, L. *et al.* (2004) 'Interaction between the GAGA factor and Mod(mdg4) proteins promotes insulator bypass in Drosophila', *Proceedings of the National Academy of Sciences of the United States of America*, 101(41), pp. 14806–14811. doi: 10.1073/PNAS.0403959101.

Melnikova, L. *et al.* (2017) 'Multiple interactions are involved in a highly specific association of the Mod(mdg4)-67.2 isoform with the Su(Hw) sites in Drosophila', *Open Biology*, 7(10). doi: 10.1098/RSOB.170150.

Melnikova, L. *et al.* (2018) 'Interactions between BTB domain of CP190 and two adjacent regions in Su(Hw) are required for the insulator complex formation', *Chromosoma*, 127(1), pp. 59–71. doi: 10.1007/S00412-017-0645-6.

Mercer, E. M. *et al.* (2011) 'Multilineage priming of enhancer repertoires precedes commitment to the B and myeloid cell lineages in hematopoietic progenitors', *Immunity*, 35(3), pp. 413–425. doi: 10.1016/J.IMMUNI.2011.06.013.

Messina, O. *et al.* (2023) '3D chromatin interactions involving Drosophila insulators are infrequent but preferential and arise before TADs and transcription', *Nature Communications 2023 14:1*, 14(1), pp. 1–14. doi: 10.1038/s41467-023-42485-y.

Mikhaylichenko, O. *et al.* (2018) 'The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription', *Genes and Development*, 32(1). doi: 10.1101/gad.308619.117.

Mizuguchi, T. *et al.* (2014) 'Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe', *Nature*, 516(7531), pp. 432–435. doi: 10.1038/nature13833.

Moon, H. *et al.* (2005) 'CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator',

EMBO reports, 6(2), pp. 165–170. doi: 10.1038/SJ.EMBOR.7400334.

Moreau, P. *et al.* (1981) 'The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants', *Nucleic acids research*, 9(22), pp. 6047–6068. doi: 10.1093/NAR/9.22.6047.

Mourad, R. (2022) 'TADreg: a versatile regression framework for TAD identification, differential analysis and rearranged 3D genome prediction', *BMC Bioinformatics*, 23(1), pp. 1–14. doi: 10.1186/S12859-022-04614-0/FIGURES/4.

Muller, H. J. (1930) 'Types of visible variations induced by X-rays in Drosophila', *Journal of Genetics*, 22(3), pp. 299–334. doi: 10.1007/BF02984195/METRICS.

Mulqueen, R. M. *et al.* (2021) 'High-content single-cell combinatorial indexing', *Nature biotechnology*, 39(12), pp. 1574–1580. doi: 10.1038/S41587-021-00962-Z.

Muravyova, E. *et al.* (2001) 'Loss of insulator activity by paired Su(Hw) chromatin insulators', *Science (New York, N.Y.)*, 291(5503), pp. 495–498. doi: 10.1126/SCIENCE.291.5503.495.

Nanni, L., Ceri, S. and Logie, C. (2020) 'Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries', *Genome Biology*, 21(1), pp. 1–25. doi: 10.1186/S13059-020-02108-X/FIGURES/6.

Narendra, V. *et al.* (2016) 'CTCF-mediated topological boundaries during development foster appropriate gene regulation', *Genes & development*, 30(24), pp. 2657–2662. doi: 10.1101/GAD.288324.116.

Nasmyth, K. and Haering, C. H. (2009) 'Cohesin: its roles and mechanisms', *Annual review of genetics*, 43, pp. 525–558. doi: 10.1146/ANNUREV-GENET-102108-134233.

Neely, G. G. *et al.* (2010) 'A Global In Vivo Drosophila RNAi Screen Identifies NOT3 as a Conserved Regulator of Heart Function', *Cell*, 141(1), pp. 142–153. doi: 10.1016/j.cell.2010.02.023.

Nègre, N. *et al.* (2010) 'A comprehensive map of insulator elements for the Drosophila genome', *PLoS Genetics*, 6(1). doi: 10.1371/journal.pgen.1000814.

Nevil, M. *et al.* (2017) 'Stable Binding of the Conserved Transcription Factor Grainy Head to its Target Genes Throughout Drosophila melanogaster Development'. doi: 10.1534/genetics.116.195685.

Ngoc, L. V. *et al.* (2017) 'The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters', *Genes and Development*, 31(1), pp. 6–11. doi: 10.1101/GAD.293837.116/-/DC1.

Nguyen, H. T. and Xu, X. (1998) 'Drosophila mef2 expression during mesoderm development is controlled by a complex array of cis-acting regulatory modules', *Developmental biology*, 204(2), pp. 550–566. doi: 10.1006/DBIO.1998.9081.

Ninova, M., Ronshaugen, M. and Griffiths-Jones, S. (2014) 'Conserved temporal patterns of microRNA expression in Drosophila support a developmental hourglass model', *Genome Biology and Evolution*, 6(9), pp. 2459–2467. doi: 10.1093/gbe/evu183.

Nishimura, K. *et al.* (2009) 'An auxin-based degron system for the rapid depletion of proteins in nonplant cells', *Nature methods*, 6(12), pp. 917–922. doi: 10.1038/NMETH.1401.

Nora, E. P. *et al.* (2012) 'Spatial partitioning of the regulatory landscape of the X-inactivation centre', *Nature*, 485(7398), pp. 381–385. doi: 10.1038/NATURE11049.

Nora, E. P. *et al.* (2017) 'Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization', *Cell*, 169(5), pp. 930-944.e22. doi: 10.1016/J.CELL.2017.05.004.

Nüsslein-Volhard, C., Wieschaus, E. and Kluding, H. (1984) 'Mutations affecting the pattern of the larval cuticle inDrosophila melanogaster : I. Zygotic loci on the second chromosome', *Wilhelm Roux's archives of developmental biology*, 193(5), pp. 267–282. doi: 10.1007/BF00848156.

Ohler, U. *et al.* (2002) 'Computational analysis of core promoters in the Drosophila genome', *Genome Biology*, 3(12), pp. 1–12. doi: 10.1186/GB-2002-3-12-RESEARCH0087/TABLES/6.

Ohler, U. (2006) 'Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction', *Nucleic acids research*, 34(20), pp. 5943–5950. doi: 10.1093/NAR/GKL608.

Ohtsuki, S. and Levine, M. (1998) 'GAGA mediates the enhancer blocking activity of the eve promoter in the Drosophila embryo', *Genes & Development*, 12(21), p. 3325. doi: 10.1101/GAD.12.21.3325.

Ohtsuki, S., Levine, M. and Cai, H. N. (1998) 'Different core promoters possess distinct regulatory activities in the Drosophila embryo', *Genes and Development*, 12(4), pp. 547–556. doi: 10.1101/gad.12.4.547.

Oliveri, P. and Davidson, E. H. (2007) 'Built to run, not fail', *Science*, 315(5818), pp. 1510–1511. doi: 10.1126/SCIENCE.1140979/ASSET/0CC85671-E8F8-4036-A885-4D0B6B5CB4D4/ASSETS/GRAPHIC/1510-1.GIF.

Orsi, G. A. *et al.* (2014) 'High-resolution mapping defines the cooperative architecture of Polycomb response elements', *Genome Research*, 24(5). doi: 10.1101/gr.163642.113.

Osterwalder, M. *et al.* (2018) 'Enhancer redundancy provides phenotypic robustness in mammalian development', *Nature 2018 554:7691*, 554(7691), pp. 239–243. doi: 10.1038/nature25461.

Özdemir *et al.* (2019) 'The Role of Insulation in Patterning Gene Expression', *Genes*, 10(10), p. 767. doi: 10.3390/GENES10100767.

Pang, B. *et al.* (2023) 'Identification of non-coding silencer elements and their regulation of gene expression', *Nature reviews. Molecular cell biology*, 24(6), pp. 383–395. doi: 10.1038/S41580-022-00549-9.

Panigrahi, A. and O'Malley, B. W. (2021) 'Mechanisms of enhancer action: the known and the unknown', *Genome biology*, 22(1). doi: 10.1186/S13059-021-02322-1.

Park, J. *et al.* (2018) 'Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease', *Science (New York, N.Y.)*, 360(6390), pp. 758–763. doi: 10.1126/SCIENCE.AAR2131.

Park, J. *et al.* (2019) 'Dissecting the sharp response of a canonical developmental enhancer reveals multiple sources of cooperativity', *eLife*, 8, pp. 1–25. doi: 10.7554/eLife.41266.001.

Peifer, M. and Bender, W. (1986) 'The anterobithorax and bithorax mutations of the bithorax complex', *The EMBO journal*, 5(9), pp. 2293–2303. doi: 10.1002/J.1460-2075.1986.TB04497.X.

Perry, M. W., Boettiger, A. N. and Levine, M. (2011) 'Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo', *Proceedings of the National Academy of Sciences of the United States of America*, 108(33), pp. 13570–13575. doi: 10.1073/PNAS.1109873108/SUPPL_FILE/PNAS.201109873SI.PDF.

Peter, I. S., Faure, E. and Davidson, E. H. (2012) 'Predictive computation of genomic logic processing functions in embryonic development', *Proceedings of the National Academy of Sciences of the United States of America*, 109(41), pp. 16434–16442. doi: 10.1073/PNAS.1207852109/-/DCSUPPLEMENTAL/PNAS.201207852SI.PDF.

Peters, J. M., Tedeschi, A. and Schmitz, J. (2008) 'The cohesin complex and its roles in chromosome biology', *Genes & development*, 22(22), pp. 3089–3114. doi: 10.1101/GAD.1724308.

Phillips-Cremins, J. E. *et al.* (2013) 'Architectural protein subclasses shape 3D organization of genomes during lineage commitment', *Cell*, 153(6), pp. 1281–1295. doi: 10.1016/J.CELL.2013.04.053.

Picard Toolkit (2019) *Broad Institute, GitHub repository*. Available at: https://broadinstitute.github.io/picard/.

Pollex, T., Marco-Ferreres, R., *et al.* (2024) 'Chromatin gene-gene loops support the cross-regulation of genes with related function', *Molecular Cell*, 84(5). doi: 10.1016/j.molcel.2023.12.023.

Pollex, T., Rabinowitz, A., *et al.* (2024) 'Enhancer–promoter interactions become more instructive in the transition from cell-fate specification to tissue differentiation', *Nature Genetics*, 56(4). doi: 10.1038/s41588-024-01678-x.

Pombo, A. and Dillon, N. (2015) 'Three-dimensional genome architecture: players and mechanisms', *Nature reviews. Molecular cell biology*, 16(4), pp. 245–257. doi: 10.1038/NRM3965.

Postika, N. *et al.* (2018) 'Boundaries mediate long-distance interactions between enhancers and promoters in the Drosophila Bithorax complex', *PLoS genetics*, 14(12). doi: 10.1371/JOURNAL.PGEN.1007702.

Pownall, M. E. *et al.* (2023) 'Chromatin expansion microscopy reveals nanoscale organization of transcription and chromatin', *Science (New York, N.Y.)*, 381(6653), pp. 92–100. doi: 10.1126/SCIENCE.ADE5308.

Purnell, B. A., Emanuel, P. A. and Gilmour, D. S. (1994) 'TFIID sequence recognition of the initiator and sequences farther downstream in Drosophila class II genes', *Genes & development*, 8(7), pp. 830–842. doi: 10.1101/GAD.8.7.830.

Qiu, X. *et al.* (2020) 'Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe', *Cell systems*, 10(3), pp. 265-274.e11. doi: 10.1016/J.CELS.2020.02.003.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: A flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Rach, E. A. *et al.* (2009) 'Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome', *Genome biology*, 10(7). doi: 10.1186/GB-2009-10-7-R73.

Rach, E. A. *et al.* (2011) 'Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level', *PLoS Genetics*, 7(1), pp. 1–6. doi: 10.1371/journal.pgen.1001274.

Rada-Iglesias, A. *et al.* (2011) 'A unique chromatin signature uncovers early developmental enhancers in humans', *Nature*, 470(7333), pp. 279–285. doi: 10.1038/nature09692.

Rajarajan, P. *et al.* (2018) 'Chromosomal Conformations and Epigenomic Regulation in Schizophrenia', *Progress in molecular biology and translational science*, 157, pp. 21–40. doi: 10.1016/BS.PMBTS.2017.11.022.

Ramani, V. *et al.* (2017) 'Massively multiplex single-cell Hi-C', *Nature Methods*, 14(3), pp. 263–266. doi: 10.1038/nmeth.4155.

Ramírez, F. *et al.* (2016) 'deepTools2: a next generation web server for deep-sequencing data analysis', *Nucleic Acids Research*, 44(W1), pp. W160–W165. doi: 10.1093/NAR/GKW257.

Ramírez, F. *et al.* (2018) 'High-resolution TADs reveal DNA sequences underlying genome organization in flies', *Nature Communications*, 9(1), pp. 1–15. doi: 10.1038/s41467-017-02525-w.

Ramos, A. I. and Barolo, S. (2013) 'Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1632). doi: 10.1098/rstb.2013.0018.

Rand, M. D. *et al.* (2010) 'Permeabilization of Drosophila embryos for introduction of small molecules', *Insect Biochemistry and Molecular Biology*, 40(11), p. 792. doi: 10.1016/J.IBMB.2010.07.007.

Ranganayakulu, G. *et al.* (1995) 'A series of mutations in the D-MEF2 transcription factor reveal multiple functions in larval and adult myogenesis in Drosophila', *Developmental biology*, 171(1), pp. 169–181. doi: 10.1006/DBIO.1995.1269.

Rao, S. S. P. *et al.* (2014) 'A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping', *Cell*, 159(7), pp. 1665–1680. doi: 10.1016/J.CELL.2014.11.021.

Rao, S. S. P. *et al.* (2017) 'Cohesin Loss Eliminates All Loop Domains', *Cell*, 171(2), pp. 305-320.e24. doi: 10.1016/J.CELL.2017.09.026.

Reddington, J. P. *et al.* (2020) 'Lineage-Resolved Enhancer and Promoter Usage during a Time Course of Embryogenesis', *Developmental Cell*, 55(5), pp. 648-664.e9. doi: 10.1016/j.devcel.2020.10.009.

Rehorn, K. P. *et al.* (1996) 'A molecular aspect of hematopoiesis and endoderm development common to vertebrates and Drosophila', *Development (Cambridge, England)*, 122(12), pp. 4023–4031. doi: 10.1242/DEV.122.12.4023.

Reiter, F., Wienerroither, S. and Stark, A. (2017) 'Combinatorial function of transcription factors and cofactors', *Current Opinion in Genetics and Development*. doi: 10.1016/j.gde.2016.12.007.

Rembold, M. *et al.* (2014) 'A conserved role for Snail as a potentiator of active transcription', *Genes & Development*, 28(2), p. 167. doi: 10.1101/GAD.230953.113.

Riechmann, V. *et al.* (1997) 'Control of cell fates and segmentation in the Drosophila mesoderm', *Development*, 124(15). doi: 10.1242/dev.124.15.2915.

Rivera, J. *et al.* (2019) 'REDfly: The transcriptional regulatory element database for Drosophila', *Nucleic Acids Research*, 47(D1), pp. D828–D834. doi: 10.1093/nar/gky957.

Rodríguez-Carballo, E. *et al.* (2019) 'Impact of genome architecture on the functional activation and repression of Hox

regulatory landscapes', *BMC Biology*, 17(1), pp. 1–18. doi: 10.1186/S12915-019-0677-X/FIGURES/6.

Roider, H. G. *et al.* (2009) 'CpG-depleted promoters harbor tissue-specific transcription factor binding signals—implications for motif overrepresentation analyses', *Nucleic Acids Research*, 37(19), p. 6305. doi: 10.1093/NAR/GKP682.

Rosenfeld, M. G., Lunyak, V. V and Glass, C. K. (2006) 'Sensors and signals: a coactivator/ corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response'. doi: 10.1101/gad.1424806.

Rosenfeld, N. and Alon, U. (2003) 'Response delays and the structure of transcription networks', *Journal of Molecular Biology*, 329(4), pp. 645–654. doi: 10.1016/S0022-2836(03)00506-0.

Rossi, M. J., Lai, W. K. M. and Pugh, B. F. (2018) 'Simplified ChIP-exo assays', *Nature Communications*, 9(1), pp. 1–13. doi: 10.1038/s41467-018-05265-7.

Rowley, M. J. *et al.* (2017) 'Evolutionarily Conserved Principles Predict 3D Chromatin Organization', *Molecular cell*, 67(5), pp. 837-852.e7. doi: 10.1016/J.MOLCEL.2017.07.022.

Rowley, M. J. and Corces, V. G. (2018) 'Organizational principles of 3D genome architecture', *Nature Reviews Genetics*, 19(12), pp. 789–800. doi: 10.1038/s41576-018-0060-8.

Roy, S. *et al.* (2010) 'Identification of functional elements and regulatory circuits by Drosophila modENCODE', *Science*, 330(6012), pp. 1787–1797. doi: 10.1126/science.1198374.

Rozenwald, M. B. *et al.* (2020) 'A machine learning framework for the prediction of chromatin folding in Drosophila using epigenetic features', *PeerJ Computer Science*, 6, pp. 2–21. doi: 10.7717/PEERJ-CS.307/SUPP-3.

Rudnicki, M. A. *et al.* (1993) 'MyoD or Myf-5 is required for the formation of skeletal muscle', *Cell*, 75(7), pp. 1351–1359. doi: 10.1016/0092-8674(93)90621-V.

Ruiz-Gómez, M. *et al.* (2000) 'Drosophila dumbfounded: A myoblast attractant essential for fusion', *Cell*, 102(2). doi: 10.1016/S0092-8674(00)00024-6.

Sabari, B. R. *et al.* (2018) 'Coactivator condensation at super-enhancers links phase separation and gene control', *Science*, 361(6400). doi: 10.1126/SCIENCE.AAR3958.

Sam, S., Leise, W. and Hoshizaki, D. K. (1996) 'The serpent gene is necessary for progression through the early stages of fat-body development', *Mechanisms of Development*, 60(2), pp. 197–205. doi: 10.1016/S0925-4773(96)00615-6.

Samee, M. A. H. *et al.* (2017) 'Quantitative Measurement and Thermodynamic Modeling of Fused Enhancers Support a Two-Tiered Mechanism for Interpreting Regulatory DNA', *Cell Reports*, 21(1). doi: 10.1016/j.celrep.2017.09.033.

San Martin, B. *et al.* (2001) 'A distinct set of founders and fusion-competent myoblasts make visceral muscles in the Drosophila embryo', *Development*, 128(17). doi: 10.1242/dev.128.17.3331.

Sanborn, A. L. *et al.* (2015) 'Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes', *Proceedings of the National Academy of Sciences of the United States of America*, 112(47), pp. E6456–E6465. doi: 10.1073/PNAS.1518552112.

Sandmann, T. *et al.* (2006) 'A Temporal Map of Transcription Factor Activity: Mef2 Directly Regulates Target Genes at All Stages of Muscle Development', *Developmental Cell*, 10(6), pp. 797–807. doi: 10.1016/J.DEVCEL.2006.04.009.

Sandmann, T. *et al.* (2007) 'A core transcriptional network for early mesoderm development in Drosophila melanogaster', *Genes and Development*, 21(4), pp. 436–449. doi: 10.1101/gad.1509007.

Santana, J. F. *et al.* (2020) 'The Dm-Myb Oncoprotein Contributes to Insulator Function and Stabilizes Repressive H3K27me3 PcG Domains', *Cell Reports*, 30(10). doi: 10.1016/j.celrep.2020.02.053.

Schep, A. N. *et al.* (2017) 'chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data', *Nature methods*, 14(10), pp. 975–978. doi: 10.1038/NMETH.4401.

Schindelin, J. *et al.* (2012) 'Fiji: An open-source platform for biological-image analysis', *Nature Methods*. doi: 10.1038/nmeth.2019.

Schultheiss, T. M., Xydas, S. and Lassar, A. B. (1995) 'Induction of avian cardiac myogenesis by anterior endoderm',

Development (Cambridge, England), 121(12), pp. 4203–4214. doi: 10.1242/DEV.121.12.4203.

Schulz, K. N. *et al.* (2015) 'Zelda is differentially required for chromatin accessibility, transcription factor binding, and gene expression in the early Drosophila embryo', *Genome Research*, 25(11), pp. 1715–1726. doi: 10.1101/GR.192682.115.

Schwartz, Y. B. *et al.* (2012) 'Nature and function of insulator protein binding sites in the Drosophila genome', *Genome Research*, 22(11), pp. 2188–2198. doi: 10.1101/gr.138156.112.

Schwarzer, W. *et al.* (2017) 'Two independent modes of chromatin organization revealed by cohesin removal', *Nature*, 551(7678), pp. 51–56. doi: 10.1038/NATURE24281.

Scuderi, A. *et al.* (2006) 'scylla and charybde, homologues of the human apoptotic gene RTP801, are required for head involution in Drosophila', *Developmental Biology*, 291(1). doi: 10.1016/j.ydbio.2005.12.014.

Secchia, S. *et al.* (2022) 'Simultaneous cellular and molecular phenotyping of embryonic mutants using single-cell regulatory trajectories', *Developmental Cell*, 57(4), pp. 496-511.e8. doi: 10.1016/j.devcel.2022.01.016.

Sexton, T. *et al.* (2012) 'Three-dimensional folding and functional organization principles of the Drosophila genome', *Cell*, 148(3), pp. 458–472. doi: 10.1016/j.cell.2012.01.010.

Seyres, D. *et al.* (2016) 'Identification and in silico modeling of enhancers reveals new features of the cardiac differentiation network', *Development (Cambridge, England)*, 143(23), pp. 4533–4542. doi: 10.1242/DEV.140822.

Shamir, M. *et al.* (2016) 'SnapShot: Timescales in Cell Biology', *Cell*, 164(6), pp. 1302-1302.e1. doi: 10.1016/J.CELL.2016.02.058.

Shashikant, T. and Ettensohn, C. A. (2019) 'Genome-wide analysis of chromatin accessibility using ATAC-seq', *Methods in cell biology*, 151, pp. 219–235. doi: 10.1016/BS.MCB.2018.11.002.

Shen, Y. *et al.* (2012) 'A map of the cis-regulatory sequences in the mouse genome', *Nature*, 488(7409), pp. 116–120. doi: 10.1038/NATURE11243.

Shir-Shapira, H. *et al.* (2019) 'Identification of evolutionarily conserved downstream core promoter elements required for the transcriptional regulation of Fushi tarazu target genes', *PLOS ONE*, 14(4), p. e0215695. doi: 10.1371/JOURNAL.PONE.0215695.

Shishido, E. *et al.* (1993) 'Two FGF-receptor homologues of Drosophila: One is expressed in mesodermal primordium in early embryos', *Development*, 117(2). doi: 10.1242/dev.117.2.751.

Shoval, O. and Alon, U. (2010) 'SnapShot: Network Motifs', *Cell*. doi: 10.1016/j.cell.2010.09.050.

Shrinivas, K. *et al.* (2019) 'Enhancer features that drive formation of transcriptional condensates', *Molecular cell*, 75(3), p. 549. doi: 10.1016/J.MOLCEL.2019.07.009.

Shu, H. *et al.* (2021) 'Modeling gene regulatory networks using neural network architectures', *Nature computational science*, 1(7), pp. 491–501. doi: 10.1038/S43588-021-00099-8.

Simpson, P. (1983) 'Maternal-Zygotic Gene Interactions during Formation of the Dorsoventral Pattern in Drosophila Embryos', *Genetics*, 105(3), pp. 615–632. doi: 10.1093/GENETICS/105.3.615.

Simpson, P. and Carteret, C. (1990) 'Proneural clusters: equivalence groups in the epithelium of Drosophila', *Development (Cambridge, England)*, 110(3), pp. 927–932. doi: 10.1242/DEV.110.3.927.

Skeath, J. B. and Thor, S. (2003) 'Genetic control of Drosophila nerve cord development', *Current Opinion in Neurobiology*, 13(1), pp. 8–15. doi: 10.1016/S0959-4388(03)00007-2.

Small, S., Blair, A. and Levine, M. (1992) 'Regulation of even-skipped stripe 2 in the Drosophila embryo.', *The EMBO Journal*, 11(11), pp. 4047–4057. doi: 10.1002/J.1460-2075.1992.TB05498.X.

Smith, E. M. *et al.* (2016) 'Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus', *American Journal of Human Genetics*, 98(1), p. 185. doi: 10.1016/J.AJHG.2015.12.002.

Sniegowski, J. A. *et al.* (2005) 'Base Catalysis of Chromophore Formation in Arg 96 and Glu 222 Variants of Green

Fluorescent Protein*'. doi: 10.1074/jbc.M412327200.

Song, S. *et al.* (2019) 'DTcf/Pangolin suppresses growth and tumor formation in Drosophila', *Proceedings of the National Academy of Sciences of the United States of America*, 116(28), pp. 14055–14064. doi: 10.1073/PNAS.1816981116/-/DCSUPPLEMENTAL.

Spahn, P. *et al.* (2014) 'Multiple regulatory safeguards confine the expression of the GATA factor Serpent to the hemocyte primordium within the Drosophila mesoderm', *Developmental Biology*, 386(1), pp. 272–279. doi: 10.1016/J.YDBIO.2013.12.012.

Spitz, F. and Furlong, E. E. M. M. (2012) 'Transcription factors: From enhancer binding to developmental control', *Nature Reviews Genetics*, 13(9), pp. 613–626. doi: 10.1038/nrg3207.

Splinter, E. *et al.* (2006) 'CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus', *Genes & Development*, 20(17), p. 2349. doi: 10.1101/GAD.399506.

Srivastava, D. and Mahony, S. (2020) 'Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns', *Biochimica et biophysica acta. Gene regulatory mechanisms*, 1863(6), p. 194443. doi: 10.1016/J.BBAGRM.2019.194443.

Stadhouders, R. *et al.* (2018) 'Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming', *Nature genetics*, 50(2), pp. 238–249. doi: 10.1038/S41588-017-0030-7.

Stadler, M. R., Haines, J. E. and Eisen, M. B. (2017) 'Convergence of topological domain boundaries, insulators, and polytene interbands revealed by high-resolution mapping of chromatin contacts in the early Drosophila melanogaster embryo', *eLife*, 6. doi: 10.7554/eLife.29550.

Stanojevic, D., Small, S. and Levine, M. (1991) 'Regulation of a Segmentation Stripe by Overlapping Activators and Repressors in the Drosophila Embryo | Enhanced Reader', *Science*, 254(5036), pp. 1385–1387. doi: 10.1126/SCIENCE.1683715.

Staudt, N. *et al.* (2006) 'Mutations of the Drosophila Zinc Finger-encoding Gene vielfältig Impair Mitotic Cell Divisions and Cause Improper Chromosome Segregation', *Molecular Biology of the Cell*, 17(5), p. 2356. doi: 10.1091/MBC.E05-11-1056.

van Steensel, B. and Belmont, A. S. (2017) 'Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression', *Cell*, 169(5), pp. 780–791. doi: 10.1016/J.CELL.2017.04.022.

Stronach, B. E. *et al.* (1999) 'Muscle LIM proteins are associated with muscle sarcomeres and require dMEF2 for their expression during Drosophila myogenesis', *Molecular biology of the cell*, 10(7), pp. 2329–2342. doi: 10.1091/MBC.10.7.2329.

Stuart, T. *et al.* (2021) 'Single-cell chromatin state analysis with Signac', *Nature Methods*, 18(11). doi: 10.1038/s41592-021-01282-5.

Suo, S. *et al.* (2018) 'Revealing the Critical Regulators of Cell Identity in the Mouse Cell Atlas', *Cell reports*, 25(6), p. 1436. doi: 10.1016/J.CELREP.2018.10.045.

Supek, F. *et al.* (2011) 'REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms', *PLOS ONE*, 6(7), p. e21800. doi: 10.1371/JOURNAL.PONE.0021800.

Syed, S., Duan, Y. and Lim, B. (2023) 'Modulation of protein-DNA binding reveals mechanisms of spatiotemporal gene control in early Drosophila embryos', *eLife*, 12, p. 85997. doi: 10.7554/ELIFE.85997.

Symmons, O. *et al.* (2014) 'Functional and topological characteristics of mammalian regulatory domains', *Genome research*, 24(3), pp. 390–400. doi: 10.1101/GR.163519.113.

Szabo, Q., Bantignies, F. and Cavalli, G. (2019) 'Principles of genome folding into topologically associating domains', *Science Advances*. American Association for the Advancement of Science. doi: 10.1126/sciadv.aaw1668.

Tagle, D. A. *et al.* (1988) 'Embryonic ε and γ globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints', *Journal of Molecular Biology*, 203(2). doi: 10.1016/0022-2836(88)90011-3.

Tajbakhsh, S. and Cossu, G. (1997) 'Establishing myogenic identity during somitogenesis', *Current Opinion in Genetics &*

Development, 7(5), pp. 634–641. doi: 10.1016/S0959-437X(97)80011-1.

Tanaka, M. *et al.* (1999) 'Complex modular cis-acting elements regulate expression of the cardiac specifying homeobox gene Csx/Nkx2.5', *Development (Cambridge, England)*, 126(7), pp. 1439–1450. doi: 10.1242/DEV.126.7.1439.

Tao, Y., Christiansen, A. E. and Schulz, R. A. (2007) 'Second chromosome genes required for heart development in Drosophila melanogaster', *Genesis*, 45(10), pp. 607–617. doi: 10.1002/dvg.20333.

Taylor, M. V. *et al.* (1995) 'Drosophila MEF2 is regulated by twist and is expressed in both the primordia and differentiated cells of the embryonic somatic, visceral and heart musculature', *Mechanisms of Development*, 50(1), pp. 29–41. doi: 10.1016/0925-4773(94)00323-F.

Taylor, M. V. (2000) 'A novel Drosophila, mef2-regulated muscle gene isolated in a subtractive hybridisation-based molecular screen using small amounts of zygotic mutant RNA', *Developmental Biology*, 220(1), pp. 37–52. doi: 10.1006/dbio.2000.9608.

Thisse, B., Messal, M. El and Perrin-Schmitt, F. (1987) 'The twist gene: isolation of a Drosophila zygotic gene necessary for the establishment of dorsoventral pattern', *Nucleic acids research*, 15(8), pp. 3439–3453. doi: 10.1093/NAR/15.8.3439.

Thurman, R. E. *et al.* (2012) 'The accessible chromatin landscape of the human genome', *Nature*, 489(7414), pp. 75–82. doi: 10.1038/NATURE11232.

Tomancak, P. *et al.* (2002) 'Systematic determination of patterns of gene expression during Drosophila embryogenesis', *Genome biology*, 3(12). doi: 10.1186/GB-2002-3-12-RESEARCH0088.

Tomancak, P. *et al.* (2007) 'Global analysis of patterns of gene expression during Drosophila embryogenesis', *Genome biology*, 8(7). doi: 10.1186/GB-2007-8-7-R145.

Torosin, N. S. *et al.* (2020) '3D genome evolution and reorganization in the Drosophila melanogaster species group', *PLOS Genetics*, 16(12), p. e1009229. doi: 10.1371/JOURNAL.PGEN.1009229.

Tsai, A., Alves, M. R. P. and Crocker, J. (2019) 'Multi-enhancer transcriptional hubs confer phenotypic robustness', *eLife*, 8. doi: 10.7554/ELIFE.45325.

TxDb.Dmelanogaster.UCSC.dm6.ensGene: Annotation package for TxDb object(s) (2019) *BC, Team BP, Maintainer*. doi: 10.18129/B9.bioc.TxDb.Dmelanogaster.UCSC.dm6.ensGene.

Udvardy, A., Maine, E. and Schedl, P. (1985) 'The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains', *Journal of molecular biology*, 185(2), pp. 341–358. doi: 10.1016/0022-2836(85)90408-5.

Ulianov, S. V. *et al.* (2021) 'Order and stochasticity in the folding of individual Drosophila genomes', *Nature Communications 2021 12:1*, 12(1), pp. 1–17. doi: 10.1038/s41467-020-20292-z.

Ulianov, S. V *et al.* (2016) 'Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains', *Genome research*, 26(1), pp. 70–84. doi: 10.1101/GR.196006.115.

Uyehara, C. M. *et al.* (2017) 'Hormone-dependent control of developmental timing through regulation of chromatin accessibility', *Genes and Development*, 31(9), pp. 862–875. doi: 10.1101/GAD.298182.117/-/DC1.

Venken, K. J. T. *et al.* (2011) 'MiMIC: A highly versatile transposon insertion resource for engineering Drosophila melanogaster genes', *Nature Methods*, 8(9), pp. 737–747. doi: 10.1038/nmeth.1662.

Verzi, M. P. *et al.* (2002) 'N-twist, an evolutionarily conserved bHLH protein expressed in the developing CNS, functions as a transcriptional inhibitor', *Developmental Biology*, 249(1), pp. 174–190. doi: 10.1006/dbio.2002.0753.

Vian, L. *et al.* (2018) 'The Energetics and Physiological Impact of Cohesin Extrusion', *Cell*, 173(5), pp. 1165-1178.e20. doi: 10.1016/J.CELL.2018.03.072.

Vietri Rudan, M. *et al.* (2015) 'Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture', *Cell Reports*, 10(8). doi: 10.1016/j.celrep.2015.02.004.

Vincentz, J. W. *et al.* (2013) 'Twist1 Controls a Cell-Specification Switch Governing Cell Fate Decisions within the Cardiac Neural Crest', *PLOS Genetics*, 9(3), p. e1003405. doi: 10.1371/JOURNAL.PGEN.1003405.

Wagh, K. *et al.* (2023) 'Transcription Factor Dynamics: One Molecule at a Time', *Annual Review of Cell and Developmental Biology*, 39, pp. 277–305. doi: 10.1146/ANNUREV-CELLBIO-022823-013847.

Wakabayashi-Ito, N. and Ip, Y. T. (2006) 'Mesoderm Formation in the Drosophila Embryo', *Muscle Development in Drosophila*, pp. 28–37. doi: 10.1007/0-387-32963-3_3.

Wallace, J. A. and Felsenfeld, G. (2007) 'We gather together: insulators and genome organization', *Current Opinion in Genetics & Development*, 17(5), pp. 400–407. doi: 10.1016/J.GDE.2007.08.005.

Wang, A. *et al.* (2015) 'Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates', *Cell stem cell*, 16(4), p. 386. doi: 10.1016/J.STEM.2015.02.013.

Wang, C. *et al.* (2015) 'Genome-wide analysis of local chromatin packing in Arabidopsis thaliana', *Genome research*, 25(2), pp. 246–256. doi: 10.1101/GR.170332.113.

Wang, P. *et al.* (2023) 'Deciphering driver regulators of cell fate decisions from single-cell transcriptomics data with CEFCON', *Nature Communications 2023 14:1*, 14(1), pp. 1–16. doi: 10.1038/s41467-023-44103-3.

Wang, Q. *et al.* (2018) 'Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells', *Nature Communications*, 9(1). doi: 10.1038/s41467-017-02526-9.

Wang, R. *et al.* (2023) 'Multiomic analysis of cohesin reveals that ZBTB transcription factors contribute to chromatin interactions', *Nucleic acids research*, 51(13), pp. 6784–6805. doi: 10.1093/NAR/GKAD401.

Waymack, R. *et al.* (2020) 'Shadow enhancers can suppress input transcription factor noise through distinct regulatory logic', *eLife*, 9, pp. 1–57. doi: 10.7554/ELIFE.59351.

Weigmann, K. *et al.* (2003) 'FlyMove--a new way to look at development of Drosophila', *Trends in genetics : TIG*, 19(6), pp. 310–311. doi: 10.1016/S0168-9525(03)00050-7.

Weirauch, M. T. *et al.* (2014) 'Determination and inference of eukaryotic transcription factor sequence specificity', *Cell*, 158(6), pp. 1431–1443. doi: 10.1016/J.CELL.2014.08.009.

Wendt, K. S. *et al.* (2008) 'Cohesin mediates transcriptional insulation by CCCTC-binding factor', *Nature*, 451(7180), pp. 796–801. doi: 10.1038/NATURE06634.

Wenick, A. S. and Hobert, O. (2004) 'Genomic cis-Regulatory Architecture and trans-Acting Regulators of a Single Interneuron-Specific Gene Battery in C. elegans', *Developmental cell*, 6(6), pp. 757–770. doi: 10.1016/J.DEVCEL.2004.05.004.

Whiteley, M. *et al.* (1992) 'The Drosophila gene escargot encodes a zinc finger motif found in snail-related genes', *Mechanisms of Development*, 36, pp. 117–127.

Wickham H, François R, Henry L, Müller K, V. D. (2023) 'dplyr: A Grammar of Data Manipulation. R package version 1.1.3', *Media*.

Wieschaus, E. and Nüsslein-Volhard, C. (2016) 'The Heidelberg Screen for Pattern Mutants of Drosophila: A Personal Account', *Annual Review of Cell and Developmental Biology*, 32(Volume 32, 2016), pp. 1–46. doi: 10.1146/ANNUREV-CELLBIO-113015-023138/CITE/REFWORKS.

Wilczynski, B. *et al.* (2012) 'Predicting Spatial and Temporal Gene Expression Using an Integrative Model of Transcription Factor Occupancy and Chromatin State', *PLoS Computational Biology*, 8(12), p. 1002798. doi: 10.1371/JOURNAL.PCBI.1002798.

Wilczyski, B. and Furlong, E. E. M. (2010) 'Dynamic CRM occupancy reflects a temporal map of developmental progression', *Molecular Systems Biology*, 6, p. 383. doi: 10.1038/MSB.2010.35.

Williams, J. *et al.* (1993) 'asense is a Drosophila neural precursor gene and is capable of initiating sense organ formation', *Development (Cambridge, England)*, 119(1), pp. 1–7. doi: 10.1242/DEV.119.SUPPLEMENT.1.

Williamson, I. *et al.* (2014) 'Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization', *Genes & Development*, 28(24), p. 2778. doi: 10.1101/GAD.251694.114.

Williamson, I. *et al.* (2019) 'Developmentally regulated Shh expression is robust to TAD perturbations', *Development*

(Cambridge, England), 146(19). doi: 10.1242/DEV.179523.

de Wit, E. *et al.* (2015) 'CTCF Binding Polarity Determines Chromatin Looping', *Molecular cell*, 60(4), pp. 676–684. doi: 10.1016/J.MOLCEL.2015.09.023.

Wong, M. C., Castanon, I. and Baylies, M. K. (2008) 'Daughterless dictates Twist activity in a context-dependent manner during somatic myogenesis', *Developmental Biology*, 317(2). doi: 10.1016/j.ydbio.2008.02.020.

Wu, W. *et al.* (2011) 'Vertebrate Paralogous MEF2 Genes: Origin, Conservation, and Evolution', *PLoS ONE*, 6(3). doi: 10.1371/JOURNAL.PONE.0017334.

Xi, W. and Beer, M. A. (2021) 'Loop competition and extrusion model predicts CTCF interaction specificity', *Nature Communications 2021 12:1*, 12(1), pp. 1–15. doi: 10.1038/s41467-021-21368-0.

Xu, X. *et al.* (1998) 'Smad proteins act in combination with synergistic and antagonistic regulators to target Dpp responses to the Drosophila mesoderm', *Genes & development*, 12(15), pp. 2354–2370. doi: 10.1101/GAD.12.15.2354.

Xu, Z. *et al.* (2014) 'Impacts of the ubiquitous factor Zelda on Bicoid-dependent DNA binding and transcription in Drosophila', *Genes & development*, 28(6), pp. 608–621. doi: 10.1101/GAD.234534.113.

Yáñez-Cuna, J. O. *et al.* (2012) 'Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding', *Genome Research*, 22(10), pp. 2018–2030. doi: 10.1101/GR.132811.111.

Yin, Z., Xu, X. L. and Frasch, M. (1997) 'Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development', *Development (Cambridge, England)*, 124(24), pp. 4971–4982. doi: 10.1242/DEV.124.24.4971.

Yu, G., Wang, L. G. and He, Q. Y. (2015) 'ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization', *Bioinformatics*, 31(14), pp. 2382–2383. doi: 10.1093/bioinformatics/btv145.

Yuen, K. C., Slaughter, B. D. and Gerton, J. L. (2017) 'Condensin II is anchored by TFIIIC and H3K4me3 in the mammalian genome and supports the expression of active dense gene clusters', *Science Advances*, 3(6). doi: 10.1126/sciadv.1700191.

Zabidi, M. A. *et al.* (2014) 'Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation', *Nature 2014 518:7540*, 518(7540), pp. 556–559. doi: 10.1038/nature13994.

Zaffran, S. *et al.* (2001) 'biniou (FoxF), a central component in a regulatory network controlling visceral mesoderm development and midgut morphogenesis in Drosophila', *Genes and Development*, 15(21), pp. 2900–2915. doi: 10.1101/GAD.917101.

Zaffran, S. *et al.* (2006) 'Cardioblast-intrinsic Tinman activity controls proper diversification and differentiation of myocardial cells in Drosophila', *Development*, 133(20), pp. 4073–4083. doi: 10.1242/dev.02586.

Zaret, K. S. and Carroll, J. S. (2011) 'Pioneer transcription factors: establishing competence for gene expression', *Genes & development*, 25(21), pp. 2227–2241. doi: 10.1101/GAD.176826.111.

Zehavi, Y. *et al.* (2014) 'Core Promoter Functions in the Regulation of Gene Expression of Drosophila Dorsal Target Genes', *The Journal of Biological Chemistry*, 289(17), p. 11993. doi: 10.1074/JBC.M114.550251.

Zhang, H. *et al.* (2023) 'CTCF and R-loops are boundaries of cohesin-mediated DNA looping', *Molecular Cell*, 83(16), pp. 2856-2871.e8. doi: 10.1016/j.molcel.2023.07.006.

Zhang, Y. *et al.* (2008) 'Model-based analysis of ChIP-Seq (MACS)', *Genome Biology*, 9(9), pp. 1–9. doi: 10.1186/GB-2008-9-9-R137/FIGURES/3.

Zhang, Y. *et al.* (2019) 'Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells', *Nature Genetics*, 51(9). doi: 10.1038/s41588-019-0479-7.

Zhong, X. P. and Krangel, M. S. (1997) 'An enhancer-blocking element between alpha and delta gene segments within the human T cell receptor alpha/delta locus', *Proceedings of the National Academy of Sciences of the United States of America*, 94(10), pp. 5219–5224. doi: 10.1073/PNAS.94.10.5219.

Zhou, J. *et al.* (1996) 'The Fab-7 element of the bithorax complex attenuates enhancer-promoter interactions in the Drosophila embryo.', *Genes & Development*, 10(24), pp. 3195–3201. doi: 10.1101/GAD.10.24.3195.

Zhou, J. *et al.* (1999) 'Characterization of the transvection mediating region of the Abdominal-B locus in Drosophila', *Development*, 126(14), pp. 3057–3065. doi: 10.1242/DEV.126.14.3057.

Zinzen, R. P. *et al.* (2009) 'Combinatorial binding predicts spatio-temporal cis-regulatory activity', *Nature*, 462(7269), pp. 65–70. doi: 10.1038/nature08531.

Zolotarev, N. *et al.* (2016) 'Architectural proteins Pita, Zw5,and ZIPIC contain homodimerization domain and support specific long-range interactions in Drosophila', *Nucleic acids research*, 44(15), pp. 7228–7241. doi: 10.1093/NAR/GKW371.

Zufferey, M. *et al.* (2018) 'Comparison of computational methods for the identification of topologically associating domains', *Genome biology*, 19(1). doi: 10.1186/S13059-018-1596-9.

Zuin, J. *et al.* (2022) 'Nonlinear control of transcription through enhancer–promoter interactions', *Nature 2022 604:7906*, 604(7906), pp. 571–577. doi: 10.1038/s41586-022-04570-y.