Computational analysis and modelling of regulatory networks controlling embryonic development

Swann Floc'Hlay

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École normale supérieure

Computational analysis and modelling of regulatory networks controlling embryonic development

Soutenue par
Swann FLOC'HLAY
Le 25 Mai 2020

École doctorale n°515
Complexité du vivant

Spécialité
Génomique

Composition du jury :

Nathalie DOSTATNI Dynamique du noyau, Institut Curie	*Présidente*
Nicolas GOMPEL Genetics of phenotype evolution, Université de Munich	*Rapporteur*
Carl HERRMANN Health Data Science Unit, Université de Heidelberg	*Rapporteur*
Delphine POTIER Genomic instability and human hemopathies, CIML	*Examinatrice*
Denis THIEFFRY Ecole normale supérieure	*Directeur*
Morgane THOMAS-CHOLLIER Ecole normale supérieure	*Co-directrice*
Eileen FURLONG European Molecular Biology Laboratory, Heidelberg	*Co-encadrante*

1794
ENS

The whole matter of the world must have been present at the beginning, but the story it has to tell may be written step by step.

G. Lemaître, The Beginning of the World from the Point of View of Quantum Theory, Nature, 1931

The path that led me to this manuscript was a curious adventure, where I was lucky enough to meet wonderful people at every step.

I warmly thank the members of my thesis committee and jury for their time, advice and kindness, with special thanks to my two referees Carl and Nicolas. I am particularly grateful to my thesis directors Denis, Morgane and Eileen for welcoming me on board. The scientific and human flourishing of your laboratories truly reflects your convictions, your caring and your strength of character.

I also heartily thank all the past and present members of the CSB, Furlong and Lepage laboratories I had the pleasure of meeting, with special thanks to Céline, Aurélien, David and Bingqing for their mentoring and kindness. I also thank the members of GBCS, SysInfo and the Functional Genomics Section for their invaluable help and support.

As an eternal teaching enthusiast, I especially thank Denis, Morgane and all the teaching teams from IBENS and the Roscoff Bioinformatics Schools for giving me the unique opportunity to join their world.

Of course, this PhD would not be the same without the friends you make all along the trip. To all of you, I say a tremendous thank you, and I will be ready when you need me! I am particularly grateful to those with whom I shared the same boat, as a PhD student from IBENS, as a master student from ENS and AIV, and as a Pitchoune from Roscoff.

Lastly, I would not have grown the strength to accomplish this journey without my family, on whom I can count *contre vents et marées*. Your confidence and strength of character are the most invaluable things I own. I especially thank my parents and my brothers Hubert and Tristan in this regard. Last but not least, I heartily thank Ely for accompanying me in this adventure, and my godmother Martine for introducing me to the Parisian way of life!

Be cunning, and full of tricks, and your people shall never be destroyed.

R. Adams, Watership down, 1972

Abstract

In recent years, scientific advances have improved the treatment of genetic diseases through personalized medicine. This consists in detecting and measuring genetic variations specific to each patient, in order to better target the deregulated mechanism. These mechanisms, linking a genetic variation to a disease, remain to be elucidated. Although we now have access to the complete genomes of thousands of individuals, establishing this link requires the understanding of complex genetic regulatory mechanisms.

Indeed, the majority of known mutations are not located in coding regions of the genome. Their impact therefore affects indirectly gene expression, via epigenomic mechanisms. These *cis* mechanisms can significantly impact the level of gene expression. However, there are also *trans* feedback mechanisms that can limit the effect of these genetic variations. This feedback control stems from the structure of the gene regulatory network.

In my thesis, I have studied both *cis* and *trans* mechanisms of transcription regulation. As these mechanisms are fundamental processes shared by all organisms, it is possible to study them through systems that are less complex than human. I used the model organisms *Drosophila melanogaster* (fruit fly) and *Paracentrotus Lividus* (purple sea urchin). I focused on embryonic development, as it is a temporal window where transcriptional activity is particularly dynamic. Indeed, the development of the organizational plan of an embryo is a process which involves precise control of transcription in time and space.

Cis regulation derives from several closely related parameters: the level of DNA compaction, epigenomic markers and the affinity of the region for transcription factors. The accessibility of a region is regulated by the density of nucleosomes wrapped around the DNA molecule. Each of the subunits of a nucleosome, the histones, may have epigenomic markers (acetylation, methylation) that label the level of DNA compaction locally. Transcriptional factors are molecules present in the nucleus of the cell, which bind to DNA and recruit the elements specific and necessary for the activation or repression of the activity of the polymerase, and therefore of gene transcription.

In order to better understand the interactions between each actor of *cis* regulation, I analysed genetic, epigenomic and transcriptomic data in collaboration with the Furlong laboratory (EMBL, Heidelberg). These different data types reflect the level of DNA accessibility, epigenomic marker composition, and the level of gene expres-

sion. These data are also derived from heterozygous Drosophila embryos with a large number of genetic variations between alleles. It is therefore possible to test the impact of a mutation by directly comparing measurements between pairs of alleles. These analyses enabled the inference of direct interactions between regulatory layers, and suggest distinct actions of the two epigenomic markers H3K27ac and H3K4me3 on gene expression.

In a second step, *trans* regulation takes place on a different scale. Indeed, it comes from the interactions embedded within a gene regulatory network. Positive and negative feedback circuits allow to stabilize or amplify a signalling cascade by modulating the activation or repression of a gene, according to its expression level. Gene regulatory networks are highly interconnected, making them often complex to analyse and predict.

In order to better understand the dynamics of regulation in *trans*, I have modelled a gene network integrating the mechanisms of the dorsal-ventral axis specification in the urchin embryo, in collaboration with the Lepage laboratory (iBV, Nice). This model relies on a logical formalism, where the activity of a gene is described by a discrete variable and its regulation by a logical rule. The logical formalisation of a network allows to study in detail its dynamics and to make predictions based on the simulation results. The use of multicellular and stochastic modelling tools enabled the identification of the key interactions necessary for the development of the dorsal-ventral axis, in particular the mutual repression of the two TGF-β pathways Nodal and BMP.

In conclusion, my thesis focuses on the study of transcription regulation at several scales and from multiple angles. Allele-specific data analysis and logical modelling allowed me to study the mechanisms of transcriptional regulation from two complementary perspectives. This way, I contributed to assess the impacts of genetic variation and gene network structure on transcription. These regulatory links are of potential interest in biomedical applications related to genetic diseases.

Résumé

Ces dernières années, des avancées scientifiques visent à améliorer les traitements de maladies génétique grâce à la médecine personnalisée. Ceci consiste à détecter et mesurer les variations génétiques propres à chaque patient, afin de mieux cibler le mécanisme dérégulé. Ces mécanismes, liant une variation génétique à une maladie, restent à élucider. Bien que nous ayons aujourd'hui accès aux génomes complets de milliers d'individus, établir ce lien nécessite la compréhension des mécanismes de régulation génétique complexes.

En effet, la majorité des mutations connues ne se situent pas dans les régions codantes du génome. Leur impact porte donc indirectement sur l'expression des gènes, via des mécanismes épigénomiques. Ces mécanismes en *cis* peuvent impacter de manière conséquente le niveau d'expression génique. Néanmoins, il existe également des mécanismes en *trans* de rétrocontrôle permettant de limiter l'effet de ces variations génétiques. Ce rétrocontrôle émane de la structure du réseau de régulation génique.

Au cours de ma thèse, je me suis intéressée à la fois aux mécanismes en *cis* et en *trans* de la régulation transcriptionnelle. Comme ces mécanismes sont des processus fondamentaux partagés par l'ensemble des organismes, il est possible de les étudier à travers des systèmes moins complexes que l'humain. Dans mon cas, j'ai utilisé les organismes modèles *Drosophila melanogaster* (mouche du vinaigre) et *Paracentrotus Lividus* (oursin violet). Je me suis concentrée sur leur développement embryonnaire, car c'est un intervalle temporel où l'activité transcriptionnelle est particulièrement dynamique. En effet, l'élaboration du plan d'organisation d'un embryon est un processus qui implique un contrôle précis de la transcription dans le temps et dans l'espace.

La régulation en *cis* dérive de plusieurs paramètres étroitement liés : le niveau de compaction de l'ADN, les marques épigénomiques et l'affinité de la région pour les facteurs de transcriptions. L'accessibilité d'une région est régulée par la densité de nucléosomes enroulés autour de la molécule d'ADN. Chacune des sous-unités d'un nucléosome, les histones, peut présenter des marques épigénomiques (acétylation, méthylation) qui balisent le niveau de compaction de l'ADN localement. Les facteurs de transcription sont des molécules présentes dans le noyau de la cellule qui, en se fixant à l'ADN, vont pouvoir recruter les éléments nécessaires et spécifiques à l'activation ou la répression de l'activité de la polymérase, et donc du gène.

Afin de mieux comprendre les interactions entre les différents acteurs de la régulation

en *cis*, j'ai analysé des données génétiques, épigénomiques et transcriptomiques en collaboration avec le laboratoire Furlong (EMBL, Heidelberg). Ces différents types de données reflètent le niveau d'accessibilité de l'ADN, la composition en marque épigénomique, et le niveau d'expression des gènes. Ces données ont été obtenues à partir d'embryons de Drosophile hétérozygotes, présentant un nombre important de variations génétiques entre allèles. Il est donc possible de tester l'impact d'une mutation en comparant directement les mesures entre paires d'allèles. Ces analyses ont permis d'inférer les interactions directes entrant en jeu entre les niveaux de régulation, et suggèrent des actions distinctes des deux marques épigénomiques H3K27ac et H3K4me3 sur l'expression des gènes.

Dans un second temps, la régulation en *trans* prend place à une échelle différente. En effet, elle dérive des interactions regroupées au sein d'un réseau de régulation génique. Les circuits de rétrocontrôle positifs et négatifs permettent de stabiliser ou d'amplifier une cascade de signalisation en modulant l'activation ou la répression d'un gène selon son niveau d'expression. Les réseaux de régulations géniques sont fortement interconnectés, ce qui les rends souvent complexes à analyser et prédire.

Afin de mieux comprendre la dynamique de régulation en *trans*, j'ai modélisé un réseau de gènes intégrant les mécanismes de spécification de l'axe dorso-ventral chez l'embryon d'oursin, en collaboration avec le laboratoire Lepage (iBV, Nice). Ce modèle utilise un formalisme logique, où l'activité d'un gène est décrite par une variable discrète et sa régulation par règle logique. La formalisation logique d'un réseau permet d'étudier en détail sa dynamique et de formuler des prédictions basées sur les résultats de simulations. L'utilisation d'outils de modélisation multicellulaire et stochastique ont permis de caractériser les interactions clés nécessaires à l'élaboration de l'axe dorso-ventral, notamment la répression mutuelle des deux voies de signalisation TGF-β Nodal et BMP.

En conclusion, ma thèse porte sur l'étude de la régulation de la transcription à plusieurs échelles et sous plusieurs angles. L'analyse de données allèle-spécifique ainsi que la modélisation logique m'ont permis de d'étudier les mécanismes de la régulation transcriptionnelle sous deux perspectives complémentaires. Ainsi, j'ai contribué à évaluer les impacts des variations génétiques et de la structure du réseau génique sur la transcription. Ces liens de régulation ont un intérêt potentiel dans les applications biomédicales liées aux maladies génétiques.

Summary

1 Introduction 11

1.1 Historical perspectives . . . 11
1.1.1 Sea urchin embryology in the 19th century . . . 11
1.1.2 Fruit fly genetics in the 20th century . . . 12
1.1.3 Drafting a gene regulatory circuit . . . 13
1.1.4 Emergence of computational systems biology . . . 14

1.2 Current view of gene regulatory logic . . . 16
1.2.1 The DNA regulatory modules of transcription . . . 16
1.2.2 The enhancer-promoter regulatory dialog . . . 16
1.2.3 The regulation of enhancer activity in space and time . . . 17
1.2.4 Enhancer activity screening and annotation . . . 18

1.3 Methods and challenges to assay genome complexity . . . 20
1.3.1 Pan-genomic characterisation of gene expression and epigenomic status . . . 20
1.3.2 Computational methods to harness genome-wide data . . . 21
1.3.3 Using perturbation to assess functionality . . . 24

1.4 A network perspective on gene regulation . . . 25
1.4.1 Systems Biology concepts . . . 25
1.4.2 Probabilistic network inference . . . 26
1.4.3 Mechanistic network modelling . . . 26

1.5 Aims of my PhD . . . 28

2 Deciphering *cis*-regulation using genetic variation 31

2.1 Study summary . . . 33

2.2 Methodological background . . . 34
2.2.1 The Drosophila Genetic Reference Panel . . . 34
2.2.2 Allelic ratio measures in F1 hybrids . . . 35
2.2.3 Mapping strategies for F1 hybrids . . . 36
2.2.4 Controlling for mapping bias . . . 36
2.2.5 Controlling for genotyping bias . . . 38

2.3 Contribution to the published work . . . 39

2.4 The mechanisms and evolutionary relevance of regulatory variants in embryonic development . . . 40
2.4.1 Abstract . . . 40
2.4.2 Introduction . . . 40
2.4.3 Results . . . 43
2.4.4 Discussion . . . 59
2.4.5 Methods . . . 61
2.4.6 Supplementary figures . . . 66
2.4.7 Supplementary methods . . . 75
2.4.8 References . . . 85

2.5 Complementary results . . . 92
2.5.1 Construction of the mappability mask . . . 92
2.5.2 Impact of the synthetic mask . . . 93
2.5.3 Impact of using F1 genomic data to discard genotyping errors . . . 94
2.5.4 Impact of using egg data to discard maternal transcripts . . . 96
2.5.5 Delineation of genomic regions with overlapping signals . . . 97
2.5.6 Probing direct interactions with partial correlation . . . 99
2.5.7 Exploring allelic imbalance at the SNP level . . . 99
2.5.8 Script availability . . . 101

3 Depicting *trans*-regulation using logical modelling 105

3.1 Study summary . . . 107

3.2 Methodological background . . . 107
3.2.1 The regulatory role of TGF-β signalling . . . 107
3.2.2 Draw me a TGF-β map . . . 108
3.2.3 Feedback circuits . . . 109
3.2.4 Logical formalism . . . 110
3.2.5 Dynamical simulation . . . 111

3.3 Contribution to the published work . . . 113

3.4 Deciphering and modelling the TGF-β signalling interplays specifying the dorsal-ventral axis of the sea urchin embryo . . . 114
3.4.1 Abstract . . . 114
3.4.2 Introduction . . . 114
3.4.3 Results . . . 117
3.4.4 Discussion . . . 128
3.4.5 Materials and methods . . . 131
3.4.6 Acknowledgements . . . 135
3.4.7 References . . . 136

3.5 Complementary results . . . 139
3.5.1 Script availability . . . 139

4 Conclusions and perspectives 141

4.1 Biological aspects . . . 141
4.1.1 Coupling between epigenetic and transcriptional regulatory mechanisms . . . 141
4.1.2 The mechanims of TGF-β cross-inhibition . . . 143

4.2 Methodological aspects . . . 144
4.2.1 Allele-specific measurements can contrast *cis*- versus *trans*- effects . . . 144
4.2.2 DNA binding motifs from ChIP-seq targeting histone marks . . . 145
4.2.3 The iterative process of network modelling . . . 148

4.3 Prospects . . . 149
4.3.1 Aiming at a system-wide model . . . 149
4.3.2 Towards the inference of regulatory networks . . . 150
4.3.3 Qualitative inference of network dynamics . . . 151

Chapter 1

Introduction

> *Genetic power is the most awesome force ever seen on this planet*
>
> **Ian Malcolm, Jurassic Park, 1993**

In 1993, the Jurassic Park movie was staging scientists filling genetic gaps in dinosaur DNA with frog DNA. Although we now know that cloning a *Tyrannosaurus Rex* is very unlikely [1], the idea of exploring genetics with *"thinking machine supercomputers and gene sequencers"* [2] has now become reality. Over the last decades, it even seems that, the more we learn about DNA, the more complex it appears to be. Today, we are still far from mastering the unfolding of the string of oligonucleotides into a living organism.

This introductory chapter is structured as follows :

⋈ I will first outline some important works in embryology and genetics, specifically focusing on the two model organisms studied in this thesis (fruit fly and sea urchin). Then, I will describe the emergence of the regulatory network theory, which is a key concept in my work.

⋈ Secondly, I will give a brief overview of the current understanding of enhancer logic, and the tools used to assay its activity. Although gene regulation involves a plethora of molecular actors and processes to convey the genetic signal at the DNA, RNA and protein levels [3], I have chosen to focus this introduction on the transcriptional level, and specifically outline the regulatory layers that I will further explore in chapter 2, namely the chromatin accessibility, the histone modifications and the transcription factor binding.

⋈ Lastly, I will outline the different network modelling strategies, with a specific focus on the two methods I applied : the probabilistic network inference (chapter 2) and the mechanistic network modelling (chapter 3).

1.1 Historical perspectives

1.1.1 Sea urchin embryology in the 19th century

In the late 19th century, the mechanisms triggering the development of a new organism from a resting egg cell were still obscure, and many biologists aimed at determining the nature of the elements controlling embryogenesis.

Experiments conducted in the sea urchins *Paracentrotus Lividus* (Lamarck, 1816) and *Strongylocentrotus purpuratus* (Stimpson, 1857) were of critical importance in the development of experimental embryology, in par-

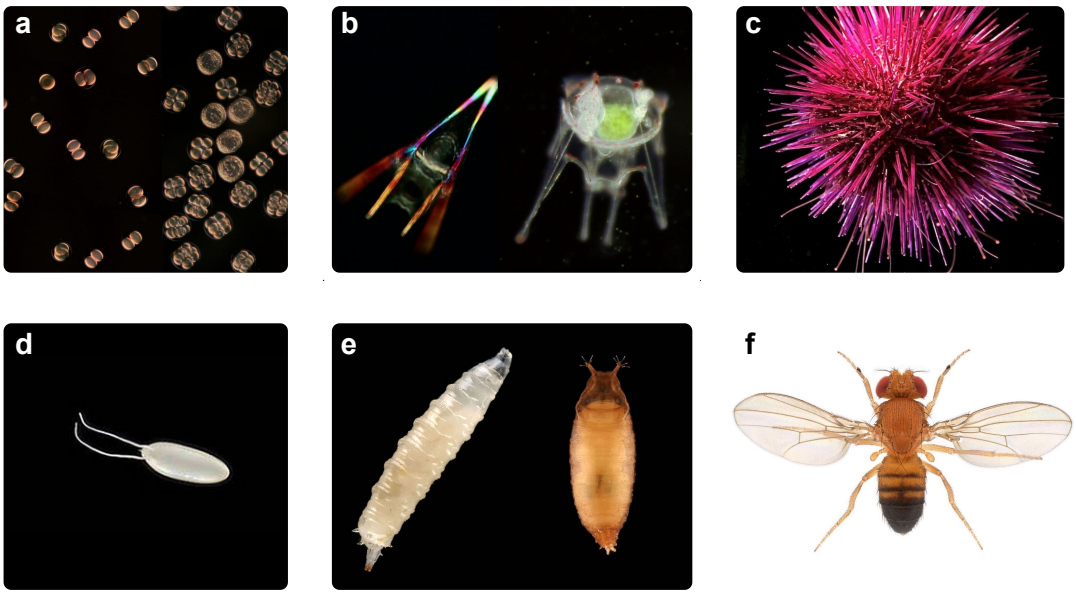

Figure 1.1: Model organisms considered in this study. Images of sea urchin (top panels, photo credits to N. and C. Sardet, planktonchronicles.org) and fruit fly individuals (bottom panels, photo credits to N. Gompel, gompel.org) at three developmental stages : embryos (a,d), larvae (b,e) and adult (c,f). a: embryos at different stages of blastomere segmentation and blastulas ; b: pluteus larvae ; c: adult sea urchin ; d: early stage embryo ; e: late stage larva and pupa ; f: adult fruit fly.

ticular with respect to the delineation of the relative implications of the cell nucleus and cytoplasm. The sea urchin (Fig. 1.1) quickly became a model system, as it offered various advantages for experimental embryology. Notably, the simple morphology, the short developmental time, the size and the transparency of the embryos were valuable characteristics for developmental studies [3].

In 1891, Hans Driesch performed blastomere dissociation in sea urchin embryos, in order to test the Weismann-Roux hypothesis of an intracellular determinant of development [3]. The experiment resulted in a full embryo for each isolated blastomere and lead Driesch to argue in favour of the existence of some regulation of development.

In parallel, works on artificial parthenogenesis and nuclear transplant by Jacques Loeb in Woods Hole and Yves Delage in Roscoff helped to further discriminate between the necessity of nucleus and cytoplasm for embryonic development [4].

At the end of the 19th century, Theodor Boveri observed morphological hybrids obtained from the fertilisation of sea urchin eggs with the sperm of a different species, and concluded in the individuality of each chromosomes [4]. Although his observations were first discredited by sceptics, notably Thomas Hunt Morgan [5], originally more prone to the epigenesis theory, the re-discovery of Mendel's law brought Boveri to gain recognition and to establish the theory of chromosomal inheritance [3].

1.1.2 Fruit fly genetics in the 20th century

At the beginning of the 20th century, Thomas Hunt Morgan further explored Boveri's theory of elementary particles

inheritance with his work on a new model organism, previously studied by the entomologist Charles Woodworth in Berkeley: the fruit fly *Drosophila melanogaster* (Meigen, 1830) [3]. He chose this new model organism for its propensity to generate spontaneous mutations, easily detectable by morphological examination. With this feature, the Drosophila (Fig. 1.1) was an ideal model to study the inheritance of new traits across generations [3].

Additionally, this organism is easy to raise. Its fast generation time, the diversity of morphological features, the small genome size, the ease for cross-breeding and maintenance of isogenic lines were all valuable features that contributed to its large use in genetic laboratories [6]. With his work on heredity, Thomas Hunt Morgan and collaborators pioneered the genetic mapping of inherited traits, and characterised the crossing-over mechanism [3, 6].

In the light of these pioneering works in genetic and embryology, model organisms such as the fruit fly and the sea urchin spread quickly in biology laboratories. Together with other model systems (eg. bacteria, yeast [7]), they paved the way for the discovery of the molecular structure of this hereditary particle, the deoxyribonucleic acid (DNA).

1.1.3 Drafting a gene regulatory circuit

The potential role of DNA as a inheritance driver, with both replicating and pairing mechanisms was notably suggested by Nikolai Koltsov in 1927 [8]. Its molecular structure as a double helix was then observed and modelled by Rosalind Franklin, Maurice Wilkins, James Watson and Francis Crick between 1952 and 1953 [9].

A decade later, André Lwoff, Jacques Monod and François Jacob established several key concepts: the distinction between regulatory and structural genes, the notion of repressor, an the idea of genetic program [10–12]. Based on their studies on the lactose operon and on the lysis/lysogeny decision of the bacteriophage lambda in *Escherichia Coli* (Escherich, 1885), they postulated the existence of a regulatory program, connecting DNA and protein synthesis *via* a factor X. This factor will later be characterised as a messenger ribonucleic acid molecule (mRNA) [12].

They demonstrated that enzymatic activity is regulated by the action of proteins, binding to DNA, thereby controlling the activity of the adjacent gene(s). With this discovery, Jacob and Monod established the basis of transcriptional regulation and made a major contribution to the domain of molecular biology.

Already in 1961, Jacob and Monod drafted explicit schemes of regulatory circuits [10, 12] (Fig. 1.2a), representing the DNA as a simple line, bearing contiguous operator and structural gene regions. The regulatory gene was represented on another DNA region, hence acting in *trans*. Chemical operations were represented as directed arcs, showing the synthesis of a repressor from the regulatory gene. This repressor was able to interact with the operon region and trigger the production of proteins via a messenger RNA, stemming from the structural genes. Together, all these components formed a network, analogous to an electronic circuit. Based on the consideration of different regulatory circuits, they concluded that the capacity of a cell to regulate protein biosynthesis could be the key mechanism enabling cells with the same genome to differentiate into various cell types [13].

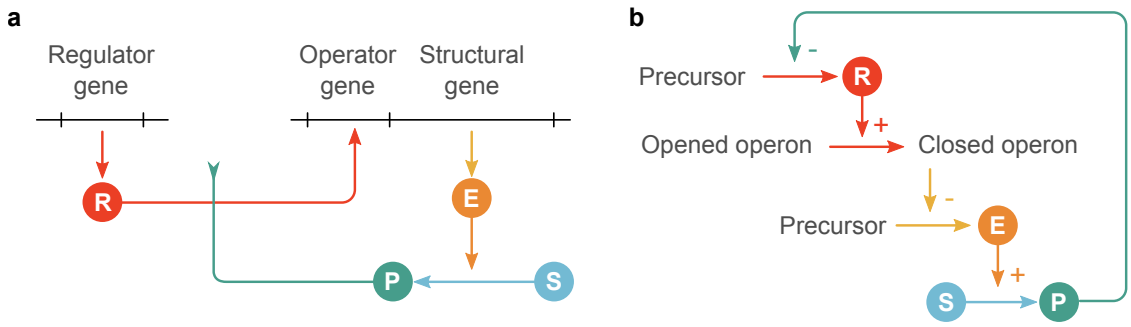

Figure 1.2: Contrasting views of the operon model. The lactose operon model, as pictured by Jacob and Monod (a) and Sugita (b). The repressor R can bind to the operator region and close the operon. The product P can hinder the action of repressor R and promote the activation of the structural gene, resulting in the activation of the enzyme E. As the enzyme E catalyses the substrate S into the product P, this reaction is self-maintained until S is depleted. In Sugita's perpective, Boolean rules define the action of each component as a signed arrow (+ for activation, - for inhibition).

In 1967, the isolation and characterisation by Mark Ptashne of the repressor of the bacteriophage lamba led to the delineation of the molecular mechanisms enabling the repressor to switch off target genes. He discovered that repressors were able to block gene transcription by binding the operon DNA region with high specificity and affinity [14]. This constituted the first instances of Transcription Factors (TF). In parallel, studies of bacteriophage lamba gene Q by René Thomas, William Dove and colleague concluded in the existence of positive regulators, capable of inducing gene transcription [7].

The key roles of gene regulatory circuits was not fully accepted by embryologists [15]. Indeed, several embryologists believed that a global mechanism impacting the majority of the genes was necessary for the early steps of embryo development. Based on the discovery of histone-mediated DNA compaction and their associated histone tail modifications (acetylation and methylation), the theory of a higher order gene regulation by DNA modification was elaborated by Robin Holliday in 1975 [3, 15]. This theory was consistent with the influential work of Conrad Waddington and his concept of epigenetic regulation [16], where additional mechanisms around the genome could draw the specification landscape of the cell.

1.1.4 Emergence of computational systems biology

In the years following the discovery of Jacob and Monod, the operon model served as a basis for key advances in the emerging field of molecular biology. In 1969, Eric Davidson further explored the concept of gene regulatory network and leveraged it to high-order systems. In his work, he suggested that regulatory circuits in metazoans were more complex and involved larger batteries of genes than in bacteria [17, 18]. Consequently, it required the existence of another class of genes, the integrator genes, which could simultaneously regulate the activity of a large number of receptor genes. In order for these new class of genes to be capable of regulating multiple receptor genes, Davidson suggested the presence of redundant *cis*-regulatory sequences upstream of the receptor genes. These sequences

could specifically recognise the signal perceived from the integrator genes by protein-DNA interaction [17]. He further studied this theory by dissecting the *cis*-regulatory sequences of the CyIIIa cytoskeletal actin gene in the sea urchin *Strongylocentrotus purpuratus*. He characterised 20 sites of specific protein-DNA interaction, together with their individual functions for space-time regulation of gene expression [17].

The evidence for integrator genes acting as master regulators also appeared with the work of Edward B. Lewis on Drosophila. In 1978, he suggested that the segmentation pattern of the embryo was governed by the concentration gradients stemming from a limited number of genes [19]. In 1980, Eric Wieschaus and Christiane Nüsslein-Volhard performed a systematic genetic screen to identify the genes involved in the patterning of the fly body [19]. By phenotyping almost 30,000 inbred fly lines following UV mutagenesis, they identified 600 mutants with defects in the embryo patterning, including 15 loci soon demonstrated to encode master transcription regulators. This study settled a considerable landmark in the precise determination of regulatory genes, along with the introduction of flowery gene names still in use today, such as *armadillo*, *stardust* and *basket*.

The operon model and the analogy with electronic circuits fostered studies at the interface between biology, mathematics and electronics. Already in the early 1960s, Motoyosi Sugita explored the parallel between genetic networks and electronic circuits [20]. He proposed to formalise the dynamics of biological circuits as done for electronic chips, introducing the concept of cellular automaton (Fig. 1.2b). He used the Boolean algebra to formulate the operon model as a set of logical equations and binary gene states, either closed or opened for protein-DNA interaction (cf. section 3.2.4).

At the end of the 1960s, Stuart Kauffman applied this Boolean formalism to study the behavior of larger, randomly generated gene networks [21]. Based on sets of logical equations, he simulated the temporal evolution of such Boolean networks using a synchronous updating strategy. He further characterised the asympototic dynamic trends, and concluded that simulations could give raise to two different kinds of attractors: stable states and dynamical cycles (cf. section 3.2.5).

During the 1970s, René Thomas refined the Boolean approach by using the asynchronous updating and multilevel variables, enabling more realistic simulations of cell specifications processes [22].

In parallel to the emergence of Boolean modelling of the gene regulatory circuits, several approaches using differential equations also appeared, notably the works of Brian Goodwin [21]. In their models, the production of proteins was quantitatively defined by kinetic rates and molecular concentrations. Their approach enabled more quantitative simulations of the system dynamics in a continuous time frame.

1.2 Current view of gene regulatory logic

1.2.1 The DNA regulatory modules of transcription

The DNA sequence is a stretch of four nucleotide types, precisely ordered to form genes and regulatory regions [3, 6]. Regulatory regions encode sequences targeted by transcription factors (TF). The binding of one or multiple transcription factors to a specific regulatory region controls the transcription of the surrounding gene(s) by promoting or repressing the recruitment of the polymerase [3, 23]. The regulatory region and the targeted genes are chiefly in close or direct proximity, on the same DNA strand. Consequently, we consider these TF-mediated interactions to be *cis*-driven, and define the regulatory regions involved as *cis*-regulatory modules (CRM). In general, two main types of CRMs are distinguished [24, 25] (Fig. 1.3). On the one hand, promoters are characterised by their capacity to recruit polymerase and trigger gene transcription; they are located next to the gene transcription start site (TSS). On the other hand, enhancers tend to be located further away from TSS ; they have the capacity to recruit TFs. Although, evidences accumulate on CRMs showing both promoter and enhancer characteristics [24]. It is therefore not clear whether such classification is biologically relevant. Instead, recent studies suggest that CRMs rather ranges according to a continuum between pure promoters and pure enhancers [26].

1.2.2 The enhancer-promoter regulatory dialog

In order to trigger gene transcription, enhancers physically interact with promoters with the help of other proteins, forming the so-called mediator complex (Fig. 1.3). The detailed mechanisms underlying this looping mechanism bringing promoter and enhancer together (P-E interaction) remain to be deciphered [23, 29].

With TFs and polymerase binding, we can see that gene regulation does not solely involve the *cis*-regulation of CRMs, but also requires the action of *trans*-acting molecules (Fig. 1.3). Consequently, the regulation of gene expression does not only stem from the 2D sequence of DNA ; it is rather driven by a complex 3D structure of molecules bound together [29].

The shape of the DNA molecule is controlled by multiple factors. Firstly, the nucleotide composition of DNA itself will affect the helix groove [30]. Secondly, in the nucleus, the DNA molecule is densely packed by nucleosomes, forming the chromatin [31]. Each nucleosome is formed by an octamer of four pairs of histones. They function like molecular spools for DNA strand, providing a tight compaction. This compaction capacity is critical for the cell cycle, as it permit the formation of chromosome during cell division. This conformation is also necessary to control the CRMs activity. Indeed, a local chromatin compaction on an enhancer can prevent TF from binding if the target site is not accessible [23, 31]. This additional layer of regulation formed by the molecules around the DNA represent the epigenomic landscape [16].

The regulation of chromatin compaction

Table 1.1: Main histone modifications and the associated regulatory states in mammalians, compiled on the basis of Zhou, Goren *et al.* [27] and Rivera and Ren [28]

CRM type	Associated histone modifications	Regulatory state
Promoter	H3K9me3 (stable) or H3K27me3 (transient)	inactive
	H3K4me3 only or H3K4me3 and H3K27me3	poised
	H3K4me3 and H3K27ac	active
Enhancer	H3K9me3 (stable) or H3K27me3 (transient)	inactive
	H3K4me1 only or H3K4me1 and H3K27me3	poised
	H3K4me1 and H3K27ac	active

involves histone tail post-translational modifications. As the N-terminal tails of histone protrude from the nucleosome, some of their amino-acids can be modified by biochemical reactions, such as methylation and acetylation [27, 32].

Some of these modifications have been shown to co-vary with transcriptional regulation and chromatin compaction changes [27, 33, 34] (Table 1.1). For example, the acetylation of the 27th lysine residue from the histone H3 (H3K27ac) is associated with the presence of active promoter or enhancer activity. It is therefore suggested that histone modifications define a code for transcription regulation.

A more global regulation of gene expression is imposed by a higher scale 3D organisation. In particular, topologically associated domains (TAD) are formed by clusters of long-range contacts between regions from the same DNA molecule [29, 35] ; their boundary are defined by insulators, characterised by the binding of the protein CTCF. These regions of higher physical interactions are known to favour transcription regulation by increasing the chance of promoter-enhancer contact within a TAD. In contrast, the presence of an insulator tend to limit interactions between the flanking regions [36].

In summary, the spatial-temporal tuning of transcription involves a complex interplay of different actors regulating the accessibility of regulatory regions.

1.2.3 The regulation of enhancer activity in space and time

During development, the activation of specific gene must be perfectly controlled in space and time to properly pattern the embryo [23]. Thus, it is crucial to tightly regulate the activation and repression of enhancers and promoters. In this respect, both long-term and short-term actions are taking place.

Firstly, the DNA compaction can be adapted by changing nucleosome positioning on DNA [34]. Specific enzymes can read, erase or write the histone modification code to flag a region as a target for chromatin remodelling [33, 34]. For example, the methyltransferases and demethylases can respec-

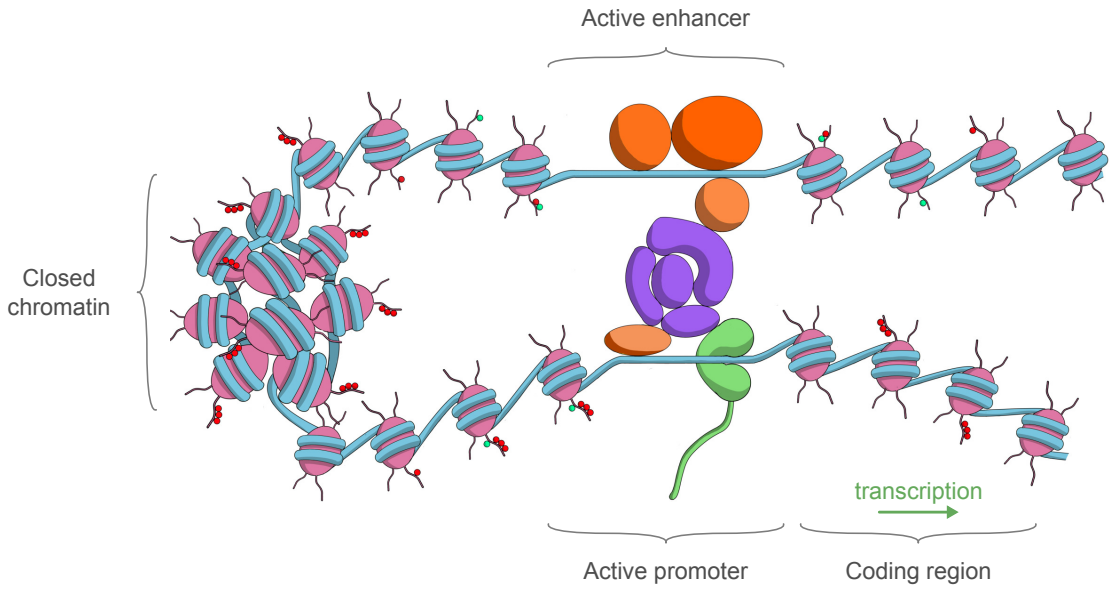

Figure 1.3: Enhancer-promoter looping. Schematic of the molecular interactions taking place during gene transcription. The DNA (blue) is coiled around nucleosomes (pink), forming open and closed chromatin. Protruding histone tails bear acetylation (green) and methylation (red) marks on their lysine residues. Transcription factors (orange) bind specific sites of the enhancer region and recruit the mediator complex (purple). This complex recruits the polymerase (green) and forms a bridge between the enhancer and the promoter. The polymerase initiates transcription and starts synthesising mRNA. Image credits to T. Floc'hlay.

tively add and remove methylation on the histone tails. Following such histone post-translational modifications, nucleosomes can relocate and modify the accessibility of surrounding CRMs. Secondly, pioneer transcription factors have the ability to bind closed chromatin regions and promote local nucleosome release [37].

These mechanisms of chromatin remodelling take time, as they require nucleosome re-positioning. Transcription factor binding further refine the spatial-temporal resolution of transcriptional regulation. Indeed, the diffusion of TFs and their recruitment at enhancer regions are comparatively fast. They can rapidly adapt the gene expression level, while keeping a high signal sensitivity and specificity [23]. The specificity of transcription factor signal can be notably achieved by cooperative binding, where multiple TFs need to join force to trigger the gene activation [23].

Within an embryo, the regulation of the activity of transcription factors is therefore the key mechanism for rapid regulatory changes. Additionally, slower changes in chromatin accessibility can help to maintain transcriptional control at broader scales [38].

1.2.4 Enhancer activity screening and annotation

Following decades of work on model organisms, extensive annotation resources on genes and regulatory regions are now available, together with the characterisation of the corresponding spatial-temporal activity patterns.

These patterns have been characterised in multiple ways. Notably, reporter assays have been largely used to study the CRM involved in embryogenesis [25]. In this type of experiment, the DNA sequence of a candidate *cis*-regulatory re-

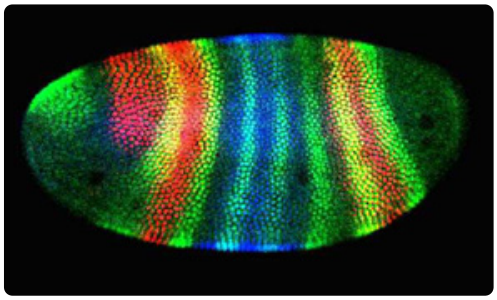

Figure 1.4: Drosophila in-situ. Microscope image of an in-situ hybridisation, obtained by fluorescently labelling the master TF hairy (red), krüppel (green) and giant (blue) in a Drosophila embryo. This photo was published in the 1996 easter edition of the New York Times Magazine, entitled "Identified Flying Objects". Photo credits to S. Paddock.

gion is combined with a minimal promoter and a reporter gene (e.g. luciferase). In 2014, the Stark laboratory has used systematic reporter assays to characterise 7,793 *cis*-regulatory regions and their respective transcriptional activity throughout the space and developmental time of Drosophila embryogenesis [39]. The Furlong laboratory also contributed to these annotation efforts with 525 manually curated regions documented in a CRM Activity database (CAD) [40].

Another method to study the space-time dynamic of gene expression is the use of in-situ hybridisation [41, 42]. This type of experiment consists in designing a DNA or RNA probe, complementing the sequence of a target gene. By adding a label to the probe (enzymes, antibody, fluorescent label) and injecting this construct into a fixed embryo, it becomes possible to visualise the corresponding gene expression pattern. The Davidson laboratory combined this approach with gene perturbations to infer the gene regulatory network governing the endomesoderm specification of the sea urchin *Strongylocentrotus Purpuratus* [43]. Similarly, the Lepage laboratory produced a considerable amount of in-situ experiments to characterise the regulatory network of the ectoderm specification in another sea urchin, *Paracentrotus Lividus* [41].

To have an idea of the considerable work done by the Drosophila and Sea urchin communities, one can look at the available wealth of curated data.

In the REDfly database [44], 24,415 Drosophila CRMs are curated from 1,058 publications. Importantly, with more than 60% of its protein-coding genes having one or more homologs in human [6], the Drosophila knowledge is a valuable resource to study gene regulation for both fundamental and biomedical research. For example, the FlyBase Human disease model index links 1451 Drosophila disease models to specific human diseases.

In the Echinobase database [45], the Gene Regulatory Network of the endomesoderm initially started by Eric Davidson now gather the results of multiple research laboratories and offer an impressive granularity in space and time of the network structure, documenting over a hundred genes and their interactions.

The wealth of data gathered from decades of genetic screens on the fruit fly and the sea urchin opens the possibility to build novel hypotheses regarding the regulatory control of gene expression. These low throughput approaches are now supplemented by high-throughput approaches to detect novel regulatory interactions and characterise gene expression, which are introduced in the next section.

1.3 Methods and challenges to assay genome complexity

1.3.1 Pan-genomic characterisation of gene expression and epigenomic status

Studying gene regulation implies to tackle two scale problems. Firstly, DNA is a molecule compacted inside the nucleus which size typically ranges between 2 and 10 microns [6]. Observing physical interactions occurring at such small scale requires advanced imaging technics. Secondly, the DNA sequence can reach several Giga base pairs (Gbp) in length [6], calling for for high-throughput reading/sequencing technologies.

Sequencing DNA became reality with the Sanger sequencing in 1977 [47] and reached a high-throughput capacity with the technical advances of Roche and Illumina sequencing [47]. The Illumina sequencing technology sequentially incorporate fluorescently labelled nucleotides on short DNA strands (also called reads or tags), generated by the polymerase from a template strand.

To specifically study the regulatory regions, several types of experiments have been designed using sequencing methods (Table 1.2).

All these techniques bring precious information to better delineate each of the regulatory layers described in the previous section. Indeed, gDNA-seq detects genetic variants (nucleotide polymorphism, insertion, deletion) ; ATAC-seq provides a direct measure of chromatin accessibility and thus highlights potential CRMs [48] ; ChIP-seq targeting histone modifications depicts the activation state of these CRMs [49] ; ChIP-seq targeting TFs can identify the targets of these factors for specific cell types, tissues and conditions [49] ; Hi-C-seq delineates the 3D organisation of the genome [50] ; RNA-seq reveals the expression level of each individual gene [51]. As a result, high-throughput sequencing technologies help to unfold the 3D structure back on the DNA.

Nevertheless, these methods still have limitations. Indeed, they do not always enable the precise definition of active regions [52]. For example, it has been shown that chromatin accessibility is not a perfect proxy for enhancer activity [38, 53, 54]. Moreover, ChIP-seq methods inherently display high noise and do not detect histone or TF location at a base pair resolution [55, 56]. Additionally, the enzyme cleavage bias [57] and a prolonged formaldehyde fixation [58] generate signal artifacts. Lastly, these experiments are based on pools of cells, consequently flattening the cell-specific signals into an average measure. In order to overcome these limitations, new sequencing methods are rapidly emerging, such as long read sequencing [59], native ChIP [55] and single-cell technique [60, 61].

In parallel to high-throughput sequencing method, there is also a fast development of tools based on high resolution microscopy [62]. These techniques have the advantage of being informative regarding both the space and time dimensions. Yet, all these different mentioned techniques are just the tip of the iceberg, as we currently experience a sharp increase in the number of newly developed methods (cf. Sequencing Method explorer from Illumina).

Table 1.2: Main high-throughput sequencing applications in functional genomics, compiled on the basis of Elkon *et al.* [46] and Gasperini *et al.* [25].

Target	Assay	Principle
Genome	gDNA-seq	Whole genomic DNA sequencing
Open chromatin	FAIRE-seq	Phenol-chloroform extraction of unbound DNA following formaldehyde fixation
	DNase-seq	Excision of unprotected DNA by DNase digestion
	ATAC-seq	Excision of unprotected DNA by Tn5 transposase
Nucleosome	MNase-seq	MNase digestion of unprotected DNA
Transcriptome	RNA-seq	Capture of mRNA poly-A 3' ends using poly-T beads.
	CAGE-seq	Capture of RNA transcripts caps on their 5' ends
	GRO-seq	RNA labelling and capture using BrUTP-labelled nucleotide, blocking of transcription initiation with sarkosyl
Histone marks	ChIP-seq	Formaldehyde fixation, labelling of the histone modification with specific antibody, followed by immunoprecipitation.
Protein-binding	ChIP-seq	Formaldehyde fixation, labelling of the transcription factor with specific antibody, followed by immunoprecipitation.
	CUT&RUN	Labelling of the transcription factor with specific antibody bound to MNase, followed by DNA cleavage by MNase digestion.
	CUT&TAG	Labelling of the transcription factor with specific antibody bound to Tn5 transposase, followed by DNA excision by Tn5 digestion.
3D proximity	Hi-C	Formaldehyde fixation followed by DNA fragmentation and random ligation based on spatial proximity.

1.3.2 Computational methods to harness genome-wide data

Although new sequencing techniques allow for a refined characterisation of the transcription regulatory landscape, they still need to be carefully processed with adapted bioinformatic tools to eliminate potential biases and extract relevant functional information. The main read processing steps and their respective biases are listed in Table 1.3.

Sequenced reads may be subjected to sequencing errors, stemming from technical noise (eg. weak fluorescence, overlapping probes). To grade the se-

Table 1.3: Processing of sequence reads and potential pitfalls, compiled on the basis of Landt *et al.* [49], Dilies *et al.* [63], Bailey *et al.* [64] and Robinson and Oshlack [65]

Processing step	Potential bias	Correction
Quality check: assessing the reads sequencing quantity and quality.	PCR duplicates and sequencing errors	Remove identical reads and tream reads with low sequencing quality score
Mapping: aligning the read on a reference genome sequence.	Duplicated regions, genotyping differences	Remove multi mapping reads, allow for a limited number of mismatches.
Peak calling: For intergenic signal (ChIP-seq, ATAC-seq), detect the regions enriched in aligned reads.	Signal artifacts from technical (fixation) and biological origin (copy number variants)	Comparison with a control sample (input), apply ENCODE masks.
Genic quantification (RNA-seq): count the number of reads aligned to each gene.	Difference in initial quantities between libraries, outlier highly expressed genes hogging the sequencing power.	Library scaling to equal sequencing depth, TMM normalisation.
Intergenic quantification (ChIP-seq, ATAC-seq): count the number of reads aligned to each peak.	Difference in initial quantities between libraries, outlier highly expressed peaks hogging the sequencing power.	Library scaling to equal sequencing depth, TMM normalisation by peak or by genomic bins.

quencing quality, sequencer machines assign to each nucleotide base call a Phred score, based on the probability of incorrect base identification. Reads with an average low Phred score are chiefly discarded using quality-check tools such as FastQC (bioinformatics.babraham.ac.uk/projects/fastqc).

Read mapping constitutes one of the first processing steps. It consists in localising, within the genome sequence, the genomic coordinates corresponding to the region of origin of each read. Although intuitive at first glance, it requires a efficient implementation and precise parameter tuning to be correctly optimised. Indeed, as a sequencing experiment produces several million of reads, there is a clear need for efficient mapping algorithms. Mapping software such as Bowtie 2 [66] and STAR [67] have relatively short running times thanks to their genome indexing method.

Mapping algorithms align the reads on a pre-existing reference genome. This method avoids the need of a *de-novo*

genome assembly for each new sequencing assay. However, the reference genome does not exactly correspond to the probed genome. Consequently, mismatches may exist between the read and its genomic source on the reference. Together with sequencing errors, these mismatches must be taken into account to avoid discarding properly-mapped read. In that respect, mapping algorithms can accept imperfectly aligned read if they do not exceed a certain penalty score, based on the number of mismatches and gaps in the alignment.

Reads generally do not exceed 300bp length with Illumina machines. A genomic sequence of the same size is usually only found once within the genome, excepted for regions with low complexity and/or series of short repeats. Read originating from such regions may equally align to multiple genomic coordinates. As the real region of origin cannot be distinguished from the others, multi-mapping reads are generally discarded.

Following read mapping, we usually aim at comparing the signal within and between samples. HTseq [68] and STAR enable the quantification of read counts per genomic feature. However, as two samples may not have the same sequencing depth, the raw number of read mapping on the target region may not reflect the same level of signal. Consequently, it is crucial to apply a library scaling [63] to adjust sequencing depth prior to the comparison.

For high-throughput data targeting non-coding regions, one additional step is the definition of peaks (Fig. 1.5). Indeed, the features used for read count in RNA-seq stem from gene annotation databases. For ATAC-seq and ChIP-seq data, there is not predefined set of regions to assess. It is therefore necessary to define, for each sample, the enriched non-coding regions. This peak calling step is implemented in multiple algorithms, such as MACS2 [69]. Peak-calling algorithms usually account for false positive detection by comparing the ChIP-seq measures with signal coming from untargeted DNA fragments extracted from the same sample (*input*).

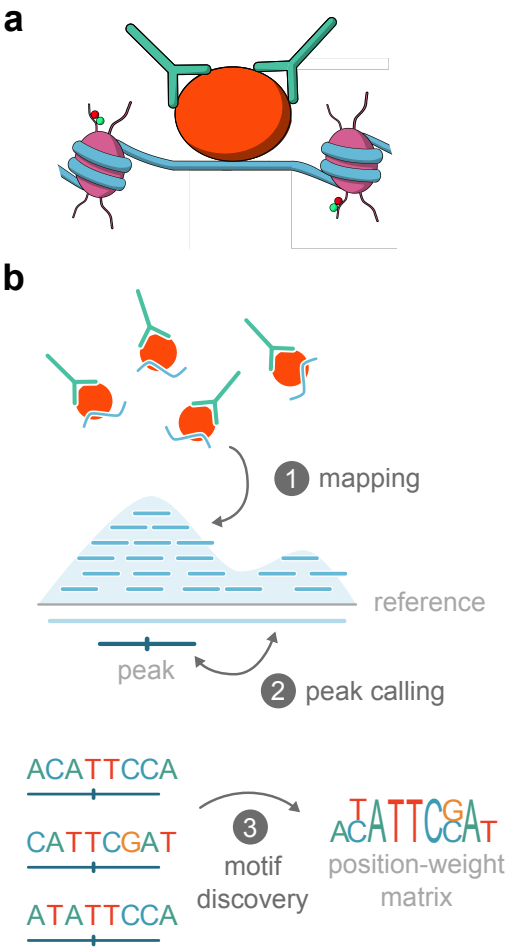

Figure 1.5: ChIP-seq TF processing. a: ChIP-seq TF assay consists in targeting a TF of interest (red) with antibodies (green), followed by immunoprecipitation of the TF and the bound DNA fragment (blue). b: The extracted DNA fragments (blue) are sequenced and mapped on a reference genome (1). Signal enrichment comparison for a given region versus the surrounding signal (larger windows, total genome background) and the input signal enables to call ChIP-seq peaks (2). Within the peaks, enriched short sequences are detected and combined into a position-weight matrix, mirroring the TF preferences profile (3).

Following feature count and library scaling, the data are still susceptible to bias stemming from very highly expressed features (genes or peaks), which monopolise a large fraction of the sequencing effort in a subset of the samples. In such case, even though all samples are scaled to the same sequencing depth, read will not be equally distributed across the genome [63]. Several strategies specifically normalise the signal of the outlier features. The most widely used method is the trimmed mean of M values [65], implemented in the software DESeq2 [70] and EdgeR [71] for genic signal and in csaw for intergenic signal [72].

After careful scaling and normalisation, the comparison of the signal between two samples can be tested within a statistical framework. In order to cope with the large dispersion of count data, the beta-binomial distribution with estimated over-dispertion parameter is chiefly used to test for differential feature expression. This statistical test is implemented in DESeq2 [70] and EdgeR [71]. On key requirement of the study design to greatly improve the power of the test is the inclusion of biological replicates for each condition.

ChIP-seq data targeting transcription factors also open the possibility to search for motifs of transcription binding site [73]. Peaks detected from ChIP-seq targeting a given transcription factor should be enriched in sequences matching its binding motif and the one of its potential co-factor (Fig. 1.5). One can infer the corresponding binding motifs with computational suites, such as the RSAT suite [73–75]. Motifs are usually represented as Position Weight Matrices (PWM), giving the likelihood of observing one of the four possible nucleotides at each base pair position [76]. Although such analyses are extremely powerful for studying the actors of transcription regulation, they still require the consideration of large amounts of data to delineate tissue or co-binding specificities [77]. Additionally, such method are still lacking detection power for assays yielding less specific and broader signal, such as ChIP-seq targeting histone [78].

To conclude, there is a vast diversity of tools to process functional genomics sequencing data. They each come with their specific specificity, advantages and challenge. However, due to the diversity of possible analysis design, there is still no clear consensus in the "best" pipeline to use. This situation can lead to differences in analysis results and hinder reproducibility when associated to poor documentation. To tackle this problem, consortium such as ENCODE [79] and ROADMAP [80] are documenting and making publicly available processing guidelines, although they might not always be completely flexible (eg. ROADMAP is chiefly targeted on human data).

1.3.3 Using perturbation to assess functionality

Each of the experiments aforementioned in section 1.3.1 are chiefly used within control-treatment or time course designs, in order to contrast signal between conditions. Indeed, cells are in perpetual action, balancing between their internal states and external environments [6] ; these permanent kinetic adaptations can blur the signal from underlying regulatory processes. Thus, performing a molecular essay for a perturbed condition (treatment) and compare the results with those obtained for a wild-type condition (control) enable to detect regulatory changes, while controlling for inherent biological noise.

Perturbation conditions can take various forms and affect the cell at different space-time scales. At the DNA level, mutations can be generated by UV screen [19], CRISPR-Cas9 technology [25], or obtained from natural populations (cf. DGRP in section 2.2.1). At the gene level, gain-of or loss-of function perturbations can be achieved by respectively injecting mRNA or morpholino [43]. Perturbation at a larger scale can also be performed by changing the environmental conditions (e.g. by transplanting a micromere in a different embryonic region).

However, perturbations are generally affecting multiple levels of gene regulation. Indeed, the impact of a regulatory gene perturbation can propagate across a regulatory network of tightly interconnected genes, which drastically complexify the search for causal mechanisms [81, 82]. In that respect, network modelling offers a powerful framework to help disentangling the *cis-* and *trans-* regulatory interactions taking place.

1.4 A network perspective on gene regulation

1.4.1 Systems Biology concepts

Systems biology emerged in response to the ever growing wealth of biological available data. A complex living system can be pictured as a jigsaw, where each piece might be well characterised individually, but still, it is only when the pieces are associated correctly that new patterns emerge, making the whole greater than the sum of its part. In order to draw this larger picture, Systems Biology aims at modelling the regulatory signal as a Gene Regulatory Network (GRN) [82].

A GRN represents a gene regulatory pathway as a graph, where each protein or other molecular identity is represented as a vertex (node) and each pairwise interaction as an arc. Consequently, this formal representation of gene regulatory logic offers a powerful framework to study a regulatory system. Additionally, the specification of a mathematical function to each vertex, mirroring its regulatory logic, enable the construction of a dynamical model.

GRN is of particular interest to study transcription regulation. Indeed, transcription is a tightly controlled and buffered process, involving multiple intertwined regulatory circuits (cf. section 3.2.3). Additionally, transcription factors govern gene expression, with varying level of specificity, cooperativity and effect size (e.g. small effect size eQTL, shadow enhancers, dosage response). For these reasons, studying transcription regulation through network modelling can help to deepen our understanding of the dynamical properties of GRNs. For example, GRN modelling both in the fruit fly [83] and the sea urchin [43] have contributed to gain a better mechanistic view of the molecular processes.

In addition, the construction of dynamic GRN model enables *in-silico* simulations (cf. section 3.2.3). Consequently, it becomes possible to infer the key regulatory circuits within the network. However, the delineation of GRNs can be challenging, especially when it requires the integration of a large number of datasets of heterogeneous origin, size and specificity [84]. Multiple quantitative and qualitative approaches exist to overcome this problem. We will describe them in the next section.

1.4.2 Probabilistic network inference

Probabilistic network inference is similar to a reverse-engineering approach: based on the observed data, we can construct a network starting with limited prior knowledge based on quantitative assessment of interactions between variables [85, 86] (Fig. 1.6b).

The most intuitive way to infer component interactions consists in assessing their level of co-variability, similarly to eQTL detection in GWAS (cf. section 2.2.1). If the correlation is significantly high, we can define an interaction (negative or positive as a function of the regression slope sign). However, this method is hindered by its inability to contrast direct from indirect interactions. For example, two genes regulated by the same upstream component (confounding factor) may co-vary without being directly connected.

This limitation is addressed by the Generalised Linear Models (GLM) and Gaussian Graphical Models (GGM) [86, 87]. In these approaches, the interaction likelihood between a given pair of components is conditioned by interaction likelihoods with all the other components of the network. An example of conditional approach is the partial correlation analysis [86], as implemented in the software GeneNet [88]. In this approach, the correlation between two elements is computed with residuals values, obtained from the linear regression of the confounding factors against the variables of interest. As a result, only the variance unexplained by correlations with confounding factors is taken into account.

However, these methods show limitations for the inference of high-dimension networks [87]. Indeed, the sum of each individually characterised gene-by-gene interaction may not fully reflect higher order network patterns. A possible approach to obtain a broader view is to visualise the network into a new coordinate space of latent variables, similar to the dimension reduction strategy of Principal component analysis [89]. In the case of Matrix Factorisation [90], a multi-dimensional matrix is approximated into a product of two sub-matrices, one common to all dimensions and the other dimension specific. The dimension shared by the two matrices represents the latent feature space, which depicts global regulation pattern. The matrix approximation is solved by a Bayesian optimisation framework, similar to a Monte-Carlo Markov Chain [91]. With such decomposition capacities, these algorithms are particularly well suited for large single-cell datasets.

A limitation of probabilistic network inference is the lack of predictive power and space-resolution. Indeed, only a few tools enable the exploitation of non-steady state datasets, such as time series and control-treatment experimental designs [92]. In order to get a better mechanistic and predictive network, the model-based solution offers great advantages and complements the *ab initio* approach.

1.4.3 Mechanistic network modelling

Mechanistic network modelling corresponds to some kind of re-engineering: based on pre-existing knowledge, a network model is built and simulated to verify its compatibility with existing dynamical data [21, 82] (Fig. 1.6c).

A common modelling approach uses Ordinary Differential Equations (ODE) [21, 82]. In this framework, the dif-

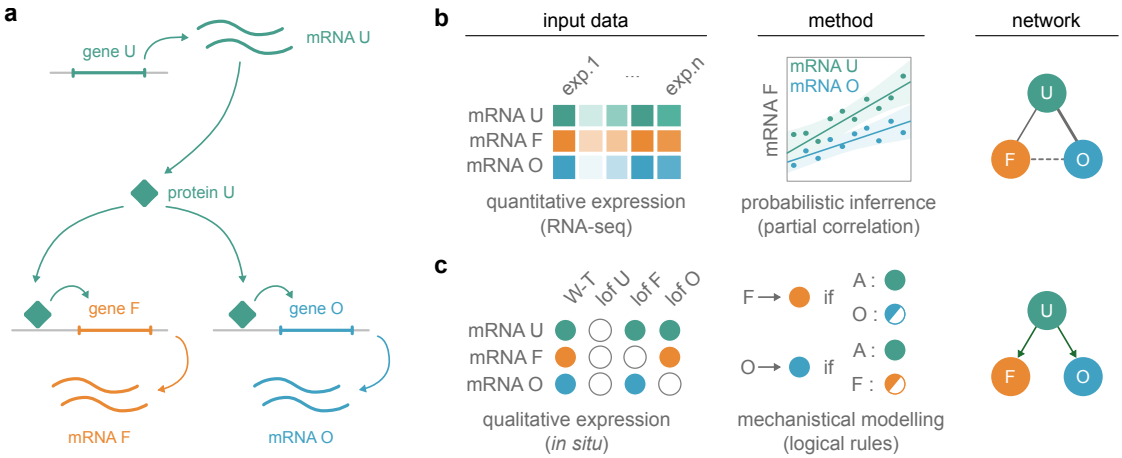

Figure 1.6: Two network modelling strategies. a: A toy example pathway to reconstruct, where a gene U produces a protein U which triggers the expression of both genes F and O. b: Using quantitative expression data, the interactions between the genes can be inferred by probabilistic analyses (eg. partial correlation), resulting in an undirected network. c: Using qualitative expression data of wild-type and perturbation conditions, the regulatory rules between the genes can be modelled into a mechanistic, directed, network (eg. logical model).

ferent molecular reactions taking place are modelled by differential equations, which are integrated to generate time-plots showing the evolution of protein concentrations or activation over time. A second approach consists in formulating Discrete Stochastic Equations (DSE) and simulate the model using a Gillespie algorithm [93]. Contrary to the first approach, DSE systems are non-deterministic and therefore better reflect the molecular noise. However, a limitation of these models is the need for strong assumptions on the structures of the equations and for precise data of the different reaction rates, which in practice are often lacking.

In contrast, Boolean modelling (cf. section 3.2.3) associates a binary variable with each component to reflect its activity level, as well as a logical rule (combining literals with the Boolean operators NOT, AND and OR) specifying when this component can be present or active [22]. This qualitative approximation greatly ease the derivation of the consistent rules and enable model-checking analyses to characterise the emerging global model dynamics (eg. to assess the existence of attractors and their reachability from given initial conditions) [94]. The simulation of Boolean models can be refined by considering probabilistic (up or down) transition rates [95]. Such stochastic extension enables the computation of relative state/path probabilities.

A drawback of the Boolean modelling approach is that it is sometimes too crude to represent subtle regulatory effects. For example, during development, it is known that morphogen gradients play a key role in the first step of embryogenesis. In such situations, different ranges of morphogen concentrations presumably trigger different sets of targets. Extensions of the Boolean approaches considering multilevel variables have been proposed to better model these situations [96].

In summary, the mechanistic modelling of GRNs enables the exploration of their dynamical properties in space and time [82]. However, the "re-engeering" strategy relies on assumptions regard-

ing pre-existing knowledge (regulatory rules, production rates, ...) and may be subjected to over-fitting.

Consequently, *ab initio* network inference methods and mechanistic modelling methods are complementary, with specific drawbacks and assets. The selection of a specific method must be based on the type of data available and on the regulatory insights sought [87].

1.5 Aims of my PhD

Transcription regulation is increasingly characterised, both dynamically and spatially, thanks to the advent of numerous novel techniques and methods. Still, the mechanical understanding of enhancer regulation and enhancer/promoter cooperativity remain poorly understood. In this context, I aimed to address the following general questions:

* How does genetic variation impact the epigenomic and transcriptomic levels?

* How does variation at the epigenomic level associates with variation at the transcriptomic level?

* How do gene regulatory circuits give rise to robust phenotypic patterning in the context of development?

* What are the determinants driving the choice of specification trajectory in the context of development?

Taking advantage of the existence of two complementary model systems, I aimed to advance our understanding of the organisation and functioning of developmental regulatory networks in two main directions:

⋈ First, using a statistical approach, I focused on the analysis of an extensive dataset of high-throughput allele-specific data targeting different layers of transcriptional regulation, generated by the Furlong laboratory. This work notably involves the design of bioinformatic methods to (i) control for mapping bias, (ii) control for confounding factor effects and (iii) integrate multiple omic layers together into a probabilistic interaction network.

⋈ Secondly, using a Boolean approach, I focused on the construction of a mechanistic model of the regulatory network controlling a specific embryo patterning process. Based on an extensive review of the literature and *in-situ* data generated by the Lepage laboratory, this modelling work includes the GRN delination, the definition of logical rules, as well as multiple dynamical simulations and analyses, at both unicellular and tissue levels.

In the next chapters, I will demonstrate how each of these approaches can contribute to gain a more comprehensive view of transcription regulation and hits to novel regulatory interactions, both in term of *cis*- and *trans*-acting mechanisms.

Chapter 2

Deciphering *cis*-regulation using genetic variation

2.1 Study summary . . . 33

2.2 Methodological background . . . 34
2.2.1 The Drosophila Genetic Reference Panel . . . 34
2.2.2 Allelic ratio measures in F1 hybrids . . . 35
2.2.3 Mapping strategies for F1 hybrids . . . 36
2.2.4 Controlling for mapping bias . . . 36
2.2.5 Controlling for genotyping bias . . . 38

2.3 Contribution to the published work . . . 39

2.4 The mechanisms and evolutionary relevance of regulatory variants in embryonic development . . . 40
2.4.1 Abstract . . . 40
2.4.2 Introduction . . . 40
2.4.3 Results . . . 43
2.4.4 Discussion . . . 59
2.4.5 Methods . . . 61
2.4.6 Supplementary figures . . . 66
2.4.7 Supplementary methods . . . 75
2.4.8 References . . . 85

2.5 Complementary results . . . 92
2.5.1 Construction of the mappability mask . . . 92
2.5.2 Impact of the synthetic mask . . . 93
2.5.3 Impact of using F1 genomic data to discard genotyping errors . . . 94
2.5.4 Impact of using egg data to discard maternal transcripts . . . 96
2.5.5 Delineation of genomic regions with overlapping signals . . . 97

2.5.6 Probing direct interactions with partial correlation 99
2.5.7 Exploring allelic imbalance at the SNP level 99
2.5.8 Script availability 101

When two flies make a child, it is not the child of a sycamore or a diplodocus.

François Jacob, 1979

The mechanisms and evolutionary relevance of regulatory variants in embryonic development

Swann Floc'hlay[1*], Emily Wong[2,3,4*], Bingqing Zhao[5*], Rebecca R Viales[5], Morgane Thomas-Chollier[1,6], Denis Thieffry[1], David A Garfield[5⋈] and Eileen EM Furlong[5⋈]

* equal contributions ; ⋈ corresponding authors

2.1 Study summary

In the presented manuscript, we aimed at better understanding the impact of natural genetic variation on transcriptional regulation. In this respect, we assayed the chromatin accessibility, histone modification and gene expression levels of *Drosophila melanogaster* embryos from eight heterozygous F1 lines (Table 2.1). After controlling for potential mapping and genotyping biases, we were able to measure the level of allelic imbalance across the genome and perform partial correlation analyses. As a result, we have inferred an interaction network depicting the direct *cis*-interactions between regulatory layers and further noted a difference in the interaction structure obtained from total count and allelic ratio partial correlations, notably between RNA and H3K4me3 signals.

Table 2.1: Samples analysed for this study.

Sample type	Assay and annotations	Number of samples
Reference	Dm3 r5.57 reference genome sequence and gene annotation from FlyBase	1
Virginizer	Genome sequence and variant annotation from Ghavi-Helm, Jankowski, Meiers *et al.*[97], egg RNA sequencing from our project.	8
DGRP	Genome sequence and Freeze2.0 variant call from the DGRP database.	8
F1 (Furlong lab)	Whole genome sequencing (gDNA-seq)	8
	Open chromatin profiling (ATAC-Seq)	60
	H3K27ac Histone modification profiling (ChIP-seq)	60
	H3K4me3 Histone modification profiling (ChIP-seq)	60
	Strand-specific RNA sequencing (RNA-seq)	60

2.2 Methodological background

2.2.1 The Drosophila Genetic Reference Panel

With the publication of the sequence of the *Drosophila melanogaster* genome in 2000 [98], the cartography of the genotype-phenotype relationships increased consequently in speed and span. We have now access to the sequence of the 14,000 protein-coding genes, spread along the 140 Mbp of the seven chromosome arms of the Drosophila genome.

In 2012, MacKay *et al.* generated an impressive collection of nearly 200 fully sequenced genomes of fly lines collected from the wild : the Drosophila Genetic Reference Panel [99]. These lines have undergone a minimum of twenty generations of inbred crosses, making them highly homozygous. However, we still observe a residual amount of heterozygous sites, likely maintained in the population for their large deleterious effect in homozygous state.

The natural genetic variation present in these lines is a unique opportunity to perform genome-wide association study (GWAS, Fig. 2.1) [100]. By statistically testing the co-segregation of a polymorphism with a given phenotype, MacKay *et al.* could associate specific SNPs with starvation resistance. Such associations between genomic loci and phenotypes are called *quantitative trait loci* (QTL).

New GWAS analyses using the DGRP are still performed to detect new QTL and further dissect the *cis*-regulatory logic governing gene regulation [53, 101]. Yet, genome-wide studies have some limitations.

Even though GWAS studies are very powerful, they are susceptible to multiple testing issues [100]. Indeed, when testing a very large number of SNPs for co-segregation, one must adjust the false discovery rate. In consequence, GWAS studies must involve a large population in order to retain enough statistical power for detecting small-effect size QTL.

In addition, although GWAS studies and enhancer curation enable to depict relationships between genomic loci and phenotypes, this does not imply a mechanistical understanding. Indeed, a QTL associates a genetic variation with a given phenotype with a certain likelihood, but it does not imply causality nor direct interaction [100].

For example, multiple genetic variations can co-segregate when located close to each other, and it is not always possible to infer the one having a mechanical impact due to linkage disequilibrium [100, 102]. In addition, inter-individual variations can potentially complicate the detection of QTL in the case of *trans*-acting variants, breaking a regulatory link between the tested *cis*-acting variant and the phenotype of interest.

An important challenge in the field of genetics is to better characterise each regulatory step between the genotype and the phenotype. In that aim, one complementary approach to GWAS study is the analysis of allele-specific data [103, 104] (Fig. 2.1). This type of study involves the analysis of expression data from F1 individuals (cf. section 2.2.2) ; it has already been applied to plants model organisms [105], yeast [106, 107], mice [108–110], fruit fly [111–116] and human [117–121].

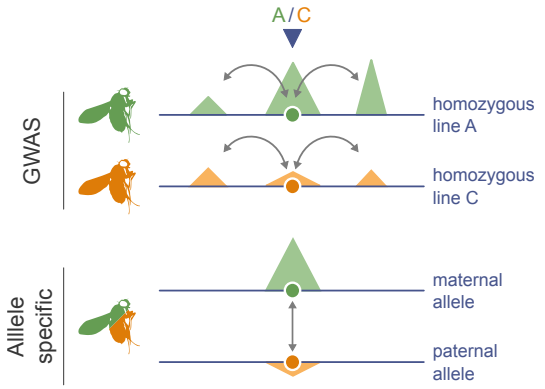

Figure 2.1: Linking genotypes and phenotypes. Top panel: Genome-Wide Association Studies (GWAS) compare multiple homozygous individuals to detect mutations (circles) co-segregating with surrounding regulatory signal (orange and green triangles) or phenotype. Bottom panel: Allele-specific studies compare the signal from each allele within a heterozygous individual to detect regions with imbalance in the regulatory signal.

2.2.2 Allelic ratio measures in F1 hybrids

GWAS analyses have unveiled a considerable number of SNPs significantly associated with transcriptional regulation. Yet, the transcription machinery intertwines multiple layers of regulations, and makes it challenging to reduce the noise arising from the complexity of cellular environment.

Indeed, a remaining challenge in genetics is to differentiate between the intramolecular *cis*-regulatory signal and the background *trans*-regulatory signal coming from the action of other molecules present within the nucleus [103, 104]. One solution to minimise this *trans*-regulation is the use of heterozygous hybrids.

Taking advantage of the natural genetic variation present in heterozygous diploid individuals, one can perform allele-specific measures (Fig. 2.2). In a sequencing assay, reads falling within one or several heterozygous sites will either bear the sequence of the maternal or the paternal allele. Knowing the parental genotypes, we can infer, for each non-coding region or gene, what is the fraction of reads coming from the paternal and the maternal alleles.

The computed allelic ratio ($\frac{maternal}{paternal+maternal}$) [117] provides an estimation of the relative activity yielded by each allele. In order to maximise the breath of the analysis, one needs to have a large number of known heterozygous sites spread along the genome. In this respect, it is necessary to use offspring obtained from crosses of homozygous lines with sufficient genetic divergence. Such hybrid individuals are called F1s [104].

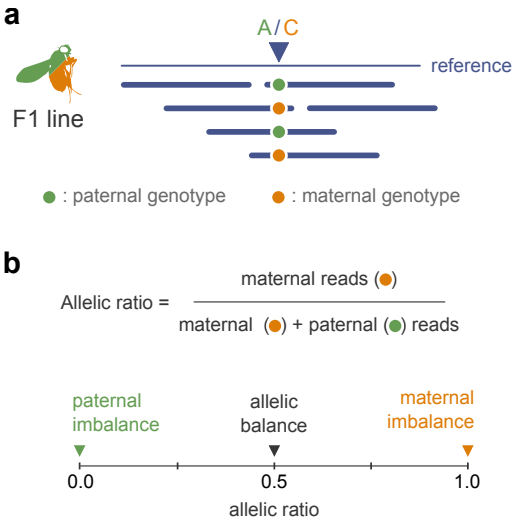

Figure 2.2: SNP-based read assignment. Schematic of the read processing used to generate allele-specific signal from F1 sample. a: Reads (blue lines) falling on a heterozygous SNP (blue triangle, A/C genotype) site will either match the paternal (green, A) or maternal (orange, C) genotype. Each read will be assigned to its parent of origin, based on its genotype at the SNP location. Reads not overlapping a SNP will be ignored. b: The allelic ratio represents the proportion of maternal reads among the total number of reads assigned to one of the two parents. Region with allelic balance are expected to have an allelic ratio of 0.5, whereas paternal or maternal imbalance are respectively reflected by an allelic ratio closer to 0 or 1.

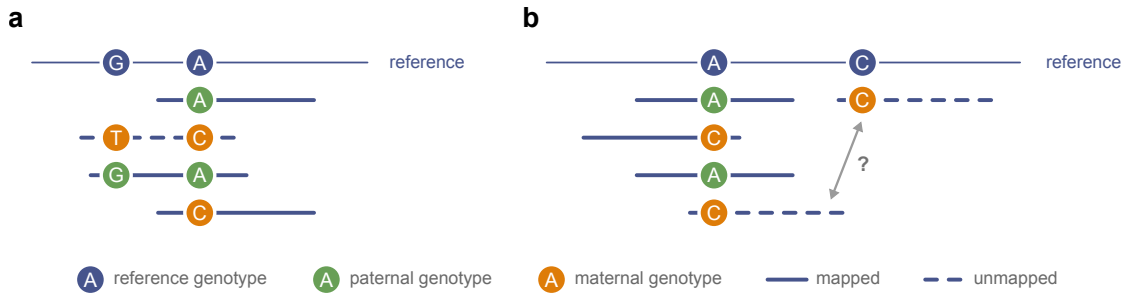

Figure 2.3: Reference mapping bias in read allelic assignment. In case of reference bias, one of the two parental genotypes (orange and green circles) is more distant from the reference sequence (blue circles). A higher proportion of reads from this parent may thus not be aligned due to too many mismatches (dashed read in panel a). Additionally, biases may also arise when genotype differences lead a single read to align at multiple positions in the genome (dashed reads in panel b), creating ambiguous mapping. These non-uniquely mapped reads are discarded from further analyses. Both of these reference mapping biases may affect the computation of allelic imbalance and must be carefully controlled.

A key feature of allelic-specific profiling is the output of ratio, i.e., whatever the type of signal, the allelic signal will range in the same finite intervals. This intrinsic normalisation allows to compare results across different signal types, such as RNA-seq, ATAC-seq and ChIP-seq data.

Allele-specific analysis is therefore a powerful tool, complementary to GWAS. It helps to depict the *cis*-regulatory mechanisms controlling transcription regulation. However, such analysis are based on the comparison of genetically diverse samples. It thus heavily relies on a careful processing of the reads, to account for potential genotyping errors and mapping bias, arising from the comparison of two divergent genomes.

To measure an allelic imbalance, one needs to compare the relative amount of sequenced reads obtained from each allele [117]. It is necessary to properly assign each read to the correct parent of origin, otherwise the computed allelic ratio may be misleading. In this respect, reads must be equally mappable, independently of the parent of origin, and SNPs used to assign the reads must be heterozygous and correctly annotated in the parental genomes.

2.2.3 Mapping strategies for F1 hybrids

Mapping reads from an heterozygous line directly to a reference genome is problematic. Indeed, depending on how much the alleles diverge from the reference genotype, a better mappability may be conferred to one of them [122]. This assymetry in allele mappability may lead the alignment algorithm to discard the reads with more divergent sequences. This occurs when too many mismatches relative to the reference are found in the sequences (Fig. 2.3a). Additionally, reads with multiple mismatches may also become mappable in multiple regions (Fig. 2.3b). The corresponding reads would then be discarded from further analyses.

2.2.4 Controlling for mapping bias

Several methods now exist to avoid mapping biases (Table 2.2). In the

Table 2.2: Mapping strategies to control for reference biases in F1 samples.

Strategy	Principle	Limitations	Ref.
N-masking	Masks heterozygous SNPs present in the F1 as 'N' in the reference genotype.	Unable to map regions with a high density of heterozygous sites.	[123]
Personalised parental genomes	Maps reads on both personalised parental reference genomes, each incorporating the parent-specific variants.	Sensitive to the annotation quality in the coordinate conversion step.	[113, 114, 116, 124, 125]
Allelic swap	Map reads on reference genome and re-process read overlapping heterozygous sites to test for unique mapping at the same location in both parental genomes.	Discard the reads not mapping to the reference in the initial step.	[126]

manuscript presented in section 2.4, I have used the personalised parental genomes strategy for its ability to perform well, even for high SNP density, which is the case in our study. Additionally, I designed a mask aimed at providing a strict control for remaining biases, including complex mapping bias events, such as ambiguous mapping (cf. section 2.2.4 below).

As presented in the section 2.2.3 of this chapter, the need to control for reference genome biases is crucial for the analysis of allele-specific data.

Using both genomic sequencing data and simulated reads, I generated a "mask" filtering the regions showing propensity for mappability bias. This mask comprises two parts : a genomic filter and a simulated read filter.

The genomic filter aims at masking regions with inherent weak mappability or indels. Indeed, if a deletion is present in only one of the two parental genotypes, it may create allelic imbalance because the reads bearing the deletion will not align at this position. To create this genomic filter, genomic DNA (gDNA) reads sequenced from the parental lines are processed following the same procedure described in section 2.2.3. Regions with no gDNA reads aligned are considered as not mappable and are included in the mask.

The simulated reads filter aims at masking region with inherent mappability biases between the two parental genotypes. For this purpose, the filter contrasts, for all genomic positions, the mappability of the reads coming from each parental genotypes. The filter contruction consists in generating transcriptomic and genomic reads spanning the whole genome and transcriptome at equal coverage, one read starting at each base pair position. Read lengths are designed to match the read length of our different data types. Regions are included in the filter if they present a difference in read coverage between the two

parental genotypes. This method also filters cases of ambiguous mapping, like the WASP algorithm (cf. allelic-swap strategy in Table 2.2). Indeed, mapping all possible read positions on parental genotype allows to detect parent-specific non-mappable regions. If these regions are uniquely mappable in the other parental genotype, the difference in simulated read coverage flags them as regions to mask in further analyses.

The final mappability mask is obtained by merging together the regions masked by the simulated read filter and the genomic filter. One advantage of this mask is its adaptability to each cross. We use a line-specific version of the filter for intra-individual comparisons of allelic-ratios. For inter-individual comparisons, the filtering method was adapted into a universal filter by combining the line-specific masks from all F1 lines.

2.2.5 Controlling for genotyping bias

In addition to mapping biases, incorrect annotation of heterozygous SNPs can lead to mis-assignment of reads to their alleles of origin [127].

In the event of a SNP annotated as heterozygous but in reality not segregating in the F1, reads will all be assigned to the same parental genotype (Fig. 2.4). As a result, the allelic ratio observed at that site will be completely imbalanced toward one of the parental alleles. It is therefore relevant to control for genotyping errors, in order to exclude cases of extreme imbalance caused by genotyping errors in the parental lines.

In the work presented in section 2.4, we have sequenced the genomic DNA (gDNA) of each F1 line (Table 2.1). I have then used this gDNA sequencing data to detect events of genotyping errors. Indeed, as gDNA data reflects the amount of DNA present in the F1 sample and is not impacted by transcriptional regulation, we expect to see a balanced allelic ratio at all heterozygous sites. As a result, after mapping the F1 gDNA reads with the same personalised parental genome strategy, I could test for statistical allelic imbalance in each SNP. For the SNPs annotated as heterozygous for a given F1 line, significant departure from allelic balance observed at the genomic DNA level was considered as evidence for genotyping error (cf. section 2.5.3). The SNP was therefore discarded from further analysis.

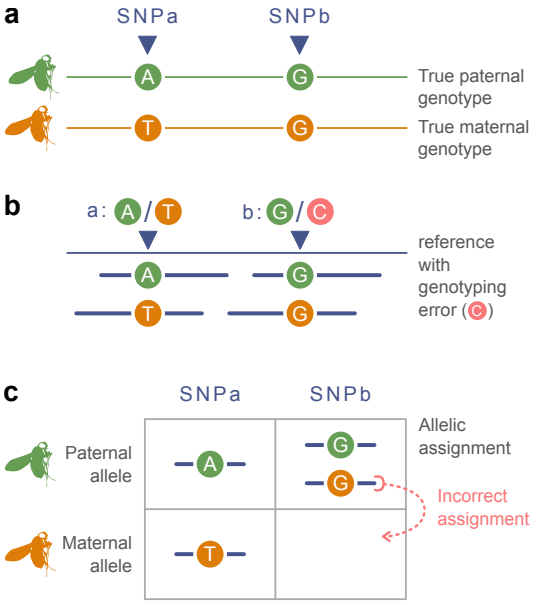

Figure 2.4: Genotyping error bias. Schematic of a case of genotyping error bias. A genotyping error arises when the true genotypes of the parental lines (green and orange circles in panel a) are erroneously annotated. Here in panel b, the genotype C (pink circle) is wrongly associating to the maternal line for SNPb, instead of the true genotype G (orange circle in panel a). In such cases, the reads sequenced from a F1 line (blue lines in panel b) are not properly assigned to their parent of origin (c). This genotyping error thus leads to a bias in allelic imbalance measures.

Having controlled for potential mapping and genotyping bias, allele-specific anal-

ysis offers a wide range of possible analyses to further explore transcriptional regulation.

2.3 Contribution to the published work

In collaboration with the Furlong lab (EMBL, Heidelberg), with the help of David Garfield and Emily Wong, my contributions focused on the bioinformatic analyses of the 248 pangenomic profiles generated by Bingqing Zhao, with the help of Rebecca Viales.

I have first diagnosed the potential mapping biases present in our dataset and further developed an adapted read mapping framework using as workflow manager Snakemake [128] to ensure reproductibility and automation. This work includes the development of masks to filter genome reference biases and genotyping errors.

I have then used the resulting read count data to integrate the signals from the different types of functional genomic assays (RNA-seq, ATAC-seq and ChIP-seq). I have used a partial correlation analysis framework to infer a network of direct interactions between regulatory layers.

Lastly, I participated in the drafting and rewriting of the Results and Methods sections, as well as in the design and generation of all figures.

2.4 The mechanisms and evolutionary relevance of regulatory variants in embryonic development

Swann Floc'hlay[1*], Emily Wong[2,3,4*], Bingqing Zhao[5*], Rebecca R Viales[5], Morgane Thomas-Chollier[1,6], Denis Thieffry[1], David A Garfield[5⋈] and Eileen EM Furlong[5⋈]

1. Institut de Biologie de l'École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France.
2. Victor Chang Cardiac Research Institute, Darlinghurst, New South Wales, Australia.
3. University of New South Wales, Sydney, Kensington, New South Wales, Australia.
4. St Vincent's Clinical School, UNSW Sydney, Kensington, New South Wales, Australia.
5. European Molecular Biology Laboratory (EMBL), Genome Biology Unit, D-69117, Heidelberg, Germany.
6. Institut Universitaire de France (IUF), 75005 Paris, France.

* equal contributions ; ⋈ corresponding authors

2.4.1 Abstract

Developmental gene expression patterns are driven by complex interactions between transcription factors, regulatory DNA sequence, and chromatin structure. As a result of these interactions, it can be complicated to predict the extent to which mutations affecting any of these regulatory 'layers' are propagated or buffered at the level of gene expression.

To better understand this, we quantified allele-specific changes in chromatin accessibility, histone modifications, and gene expression in F1 embryos generated from eight Drosophila crosses, at three embryonic stages, yielding a comprehensive dataset of 240 samples spanning multiple regulatory layers.

Genetic variation in *cis*-regulatory elements is common, highly heritable, and surprisingly consistent in its effects across embryonic stages. Much of this variation does not propagate to gene expression. When it does, it acts through two independent paths involving either changes in H3K4me3 or chromatin accessibility/H3K27ac signal. The magnitude and evolutionary impact of mutations is influenced by a genes' regulatory complexity (i.e. enhancer number), with developmental transcription factors being most robust to *cis*-acting variation.

While much of the variation affecting chromatin is based in *cis*, *trans*-acting variation dominates for gene expression. Our results suggest clear differences in evolutionary trajectory between regulatory layers as well as differences between functional gene classes.

2.4.2 Introduction

The development of a multicellular organism requires the tight regulation

of gene expression in both space and time to ensure that reproducible phenotypes are obtained across individuals and environmental conditions. Essential to this process are DNA regulatory elements (e.g. promoters and enhancers) whose sequence-specific interactions with transcription factors, polymerases, and other regulatory proteins encode the information needed to drive specific spatio-temporal expression patterns during development.

While a gene's expression pattern is typically quite precise, the DNA regulatory elements that control such expression states are replete with genetic variation (mutations) that can impact transcriptional regulation at multiple levels including transcription factor binding [1–3], chromatin state [4], transcriptional start site usage [5], gene expression levels [6, 7], and transcript isoform diversity [8]. Regulatory mutations also contribute greatly to variation in disease susceptibility among individuals [9, 10] and may contribute disproportionately to evolutionary change between species [11], demonstrating that genetically induced changes in transcriptional regulation can impact higher-level phenotypes.

Although regulatory mutations can have large effects, many behave effectively neutrally, and it is challenging to predict which mutations will have an impact. Part of the difficulty is that it is unclear on which regions of non-coding DNA actually have regulatory function. An additional challenge is the apparent robustness of gene regulatory networks. At least within a laboratory context, whole sections of regulatory DNA can be removed with little apparent impact on phenotype or fitness [12], and evolutionarily divergent regulatory sequences can often be swapped between species in transgenic assays with few detectable changes in gene expression profiles [13]. These studies demonstrate that developmental systems have the ability to compensate or "buffer" the effects of regulatory mutations, e.g. via compensation by other regulatory elements with partially overlapping activities [14–16].

The complex relationship between DNA sequences and regulatory function further complicates our understanding of how mutations can impact gene regulation. For example, mutations affecting TF binding motifs can have a large impact on gene expression [17]. But in some contexts/tissues, TF binding is driven by collective processes that can include protein-protein as well as protein-DNA interactions, such that mutations affecting a single TF motif may not substantially affect TF recruitment [18–21].

Moreover, many of the mutations affecting TF occupancy *in vivo* appear to lie outside of the TF's binding motif, and are likely due to variation affecting the binding of co-occurring TFs [1, 22, 23] or an overall change in DNA shape [24]. To make matters more complex, enhancer output is not a strict function of all TF's occupancy – enhancer often contain binding sites for multiple factors with redundant input, and in some cases, different combinations of TFs can produce the same expression output [21, 25, 26].

Even in cases in which an enhancer's activity is abolished by mutations, the gene's expression may still be robust, as genes may have many enhancers with partially overlapping activity, which can buffer the functional impact of genetic variation impacting a single enhancer [14–16]. With a few exceptions [27], this complex genotype-to-phenotype relationship cannot be modelled using regulatory sequence information alone, but rather must be evaluated empirically [21].

For mutations that do impact gene expression, there is also a question of evolutionary relevance. A substantial fraction of the regulatory mutations, particularly *cis*-acting mutations, appear to be inherited additively [28, 29], potentially making them direct targets for natural selection. In line with this, DNA regulatory elements often show evidence of directional [30] as well as balancing selection [31].

But while additive inheritance of gene expression variation is common, it is not universal [28, 32, 33], with individual genes showing variation in both the proportion of overall variation explained by additive effects and the extent to which genetic variation is influenced by *trans*-acting factors [34]. To our knowledge, there has been no attempt to explain differences in the heritability of gene expression with reference to regulatory architecture (e.g. enhancer number) or potential mechanisms for buffering expression against the impacts of mutations.

To better understand how mutations can impact developmental gene regulation, and to quantify the extent to which such mutations contribute to evolutionarily relevant variation, we made use of F1 Drosophila hybrid embryos and allele-specific quantification open chromatin (ATAC-Seq), enhancer and promoter activity (using H3K27ac or H3K4me3 H3K27ac ChIP-Seq as proxies, respectively), and gene expression (RNA-seq). Our design consists of a half-sibling panel in which F1 embryos were generated by crossing males from eight genetically distinct, wild-derived isogenic lines from the Drosophila Genetic Reference Panel (DGRP) [35] to females from a common, laboratory-derived isogenic reference strain.

In addition to having practical advantages for conducting large scale crosses, as described below, the use of a common female line allowed us to evaluate the impact of regulatory mutations while controlling for maternal effects, which can contribute disproportionately to variability in early developmental phenotypes [6, 36]. The design also encompasses an unusual scale of ecological and evolutionary differentiation: While the maternal and paternal lines are clearly the same species, the maternal lab strain was isolated in the laboratory more than 60 years ago [37–41], before the estimated invasion of the p-element in the North American population and before the widespread use of pesticides such as DDT and glycophosphates.

They thus represent an intermediate distance relative to the within-species crosses and between-species crosses typically used in allele-specific analyses. By collecting matched phenotypic measurements from two parental strains, we also estimated the heritability of *cis*-acting mutations and the relative magnitude of *trans*-acting genetic variation that contributes to phenotypic divergence.

Overall, we find allelic variation in chromatin accessibility and histone marks to be common and significantly correlated between regulatory layers, with the effects of regulatory mutations being more strongly coupled at promoters than enhancers. Specific classes of genes, such as developmental regulatory genes (e.g. transcription factors (TFs)) and genes with multiple regulatory elements, are in general more strongly buffered against the effects of this variation, with resulting impacts on the genetic architecture and heritability of gene expression variation.

We also observe multiple instances of selection having driven to near fixa-

tion, strong effect *cis*-regulatory mutations affecting genes involved in environmental response and pesticide resistance. Together, these measurements provide new insights into the functional impact of *cis*-regulatory DNA variation and how this is transmitted across different regulatory layers during embryogenesis, and how patterns of inheritance can influence the visibility of regulatory sequence variants to natural selection.

2.4.3 Results

Quantifying gene expression and regulatory element activity in hybrid embryos

We generated F1 hybrid embryos from mating eight genetically distinct inbred lines from the DGRP collection [35] to females from a common maternal line (Fig. 2.5a). The resulting F1 panel contains an average of 567,412 SNPs per cross, and a total of 1,455,988 unique SNPs covering a range of minor allele-frequencies and levels of conservation (phyloP scores) (Fig. 2.11a).

The F1 embryos were collected at three important stages of embryogenesis; 2-4 hours after egg laying, consisting primarily of pre-gastrulation, unspecified embryos (mainly stage 5), 6-8 hours (mainly stage 11), when major lineages within the three germ-layers are specified, and 10-12 hours (mainly stage 13), during terminal differentiation of tissue lineages (Fig. 2.5a).

For each developmental stage, we performed RNA-Seq, ATAC-Seq, and iChIP-seq for H3K27ac and H3K4me3 [43, 44], from the same collection of embryos (4 measurements x 3 stages x 8 genotypes=96 samples). In addition, we collected samples from the parents of one F1 genotype, forming a parent/offspring trio that allowed us to partition genetic differences between the parents into *cis* and *trans* for *cis* versus *trans* analysis [45].

All measurements were made in replicates from independent embryo collections to assess biological and technical variability, giving a total of 240 samples (192 F1 samples (96 x 2 replicates) + 48 parental (4 measurement x 3 stages x 2 genotypes x 2 replicates)). Read counts are highly correlated between biological replicates, with median correlation coefficients of 0.98 for RNA, ATAC and histone data (Fig. 2.11b, Methods).

To define non-coding features, ATAC-Seq and ChIP-Seq reads from each cross were mapped to each parental line independently and the significant peaks merged to produce a combined set of common peaks used in subsequent comparisons across all genotypes. In total, we identified 11,211 genes with detectable expression, 31,963 ATAC-Seq peaks, 19,769 H3K27ac peaks, and 6,648 H3K4me3 peaks, active at one or more stages of embryogenesis. Of these, 93.9%, 95.8%, 95.2%, and 96.9%, respectively, contained at least one SNP that distinguishes maternal and paternal haplotypes in at least one line.

The *CG12402* locus, a predicted ubiquitin-protein transferase, provides a good example of overall signal quality (Fig. 2.5a). The gene has dynamic expression, transitioning from very low to high expression from 2-4h to 10-12h. Accompanying this change are quantitative changes in chromatin accessibility, and to a lesser extent in histone modification levels, in its promoter-proximal region.

To examine the regulatory relationships between these different signals, we divided the data into promoter proximal

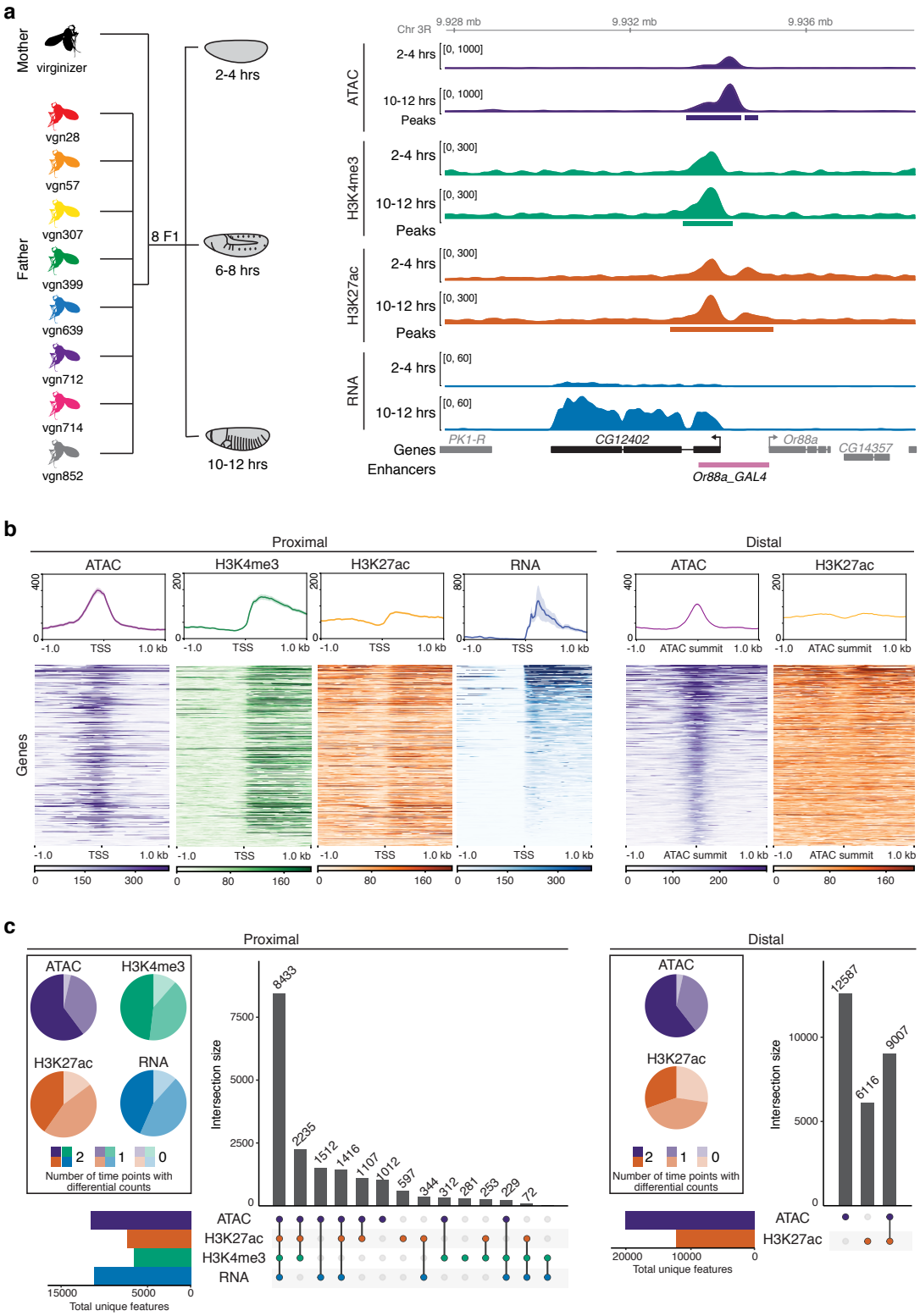

Figure 2.5: Quantifying gene expression and regulatory element activity in hybrid embryos. cf. legend on next page.

Figure 2.5: a. Left: Experimental design and data structure. RNA-seq, ATAC-seq and iChIP-seq of H3K4me and H3K27ac were performed on embryos of three developmental stages from 8 F1 hybrids with a common maternal line. Right: Genome browser overview for the *CG12402* gene locus showing all data for 2-4 hours and 10-12hrs for the genotype vgn28. Bottom track shows characterized enhancers [42]. b. Top panel shows density plots for read count signal from each data type for proximal and distal regions (left and right, respectively). Shaded regions indicate the 95% confidence intervals. Plots are centered at the TSS for promoter proximal regions, and ATAC summits for distal regions. Bottom panel shows a heatmap representation of the data type corresponding to the density plots shown above where rows are sorted by mean RNA-seq and mean ATAC-seq signal. c. Upset plots showing the colocalization of signal for proximal and distal regions (at peaks in regulatory regions and genes) for all four data types. Regions common between data types (filled circle) are joined by a vertical bar. Horizontal bar plots indicate the number of unique genes/features. Pie charts show the proportion of features with statistically different total read counts between time points (color indicates the number of times (0/1/2) the feature is differentially expressed).

(within +/- 500bp of an annotated transcriptional start site (TSS) or H3K4me3 peak) and distal (putative enhancers) elements. Looking globally at promoter proximal regions, all signals showed the expected enrichment and distribution around the TSS (Fig. 2.5b, proximal), demonstrating the quality of the data.

The ATAC-seq signal is highest directly at the promoter, representing occupancy of the basal transcriptional machinery, while H3K27ac and H3K4me4 signals are highest at the +1 nucleosome, reflecting the predominantly unidirectional nature of Drosophila promoters [46, 47]. Moreover, the levels of H3K27ac are higher at promoters compared to distal sites, as expected from *in vitro* studies [48, 49].

Interestingly, while all three regulatory signals (ATAC-seq, H3K27ac and H3K4me3) are highly correlated at the promoters of actively transcribed genes (8,433 promoters contain all 4 signals, Fig. 2.5c, left upset plot), 3,907 regions marked by H3K4me3 and overlapping peaks of ATAC-seq and/or H3K27ac show no detectable RNA-signal (Fig. 2.5c bar plots, 2.5b). Approximately 850 of these cases involve annotated transcripts of non-coding RNA (from Flybase) that lack a poly-A tail and were thus not selected in our Poly-A+ RNA-seq library. Taken together, this thereby suggests a surprising number of unannotated transcriptional events even within the well-annotated Drosophila genome.

The majority of H3K27ac (62.5%) and ATAC peaks (63.7%) are distal to an annotated promoter, representing likely enhancer elements. Of the distal ATAC peaks, 58% (12,587/21,594) have no H3K27ac signal and may represent inactive enhancers or other regulatory elements, e.g. insulators (Fig. 2.5c). The remaining 9,007 distal elements overlap H3K27ac signal (Fig. 2.5c, right), which is generally bimodally distributed around the ATAC-seq peak (Fig. 2.5b), suggesting they are active enhancers. The set of H3K27ac regions that do not overlap and ATAC-Seq peak show significantly lower signal than those that do have an overlap (Fig. 2.11c), and thus likely represent cases in which an ATAC-Seq peak was present, but below our threshold for detection.

Both gene expression (RNA-seq) and non-coding elements (based on ATAC-seq and chromatin signatures) show evidence of dynamic activity, with the majority (72%-96%) of features across all lines showing statistically significant changes in total counts between de-

velopmental time points across our F1 lines (Fig. 2.5c, pie charts; Methods), *CG12402* being one example (Fig. 2.5a). Taken together, this demonstrates both the quality and richness of the data and its usefulness to further annotate the regulatory landscape of the Drosophila genome at these important stages of embryogenesis.

Allele-specific variation is common across genotypes and regulatory layers

To test for allele-specific differences for each gene per line and time combination and each data type, we used an empirical Bayes framework to modeled allele-specific counts using a beta-binomial model (Fig. 2.12a). Most allelic ratio was centered at 50:50 across autosomes at both promoter proximal and distal elements (Fig. 2.6a), with a slight elevation in the magnitude of AI at distal sites (Fig. 2.12b). As expected, RNA allelic ratios were concordant with the direction of change of embryonic eQTL, previously quantified in these DGRP lines at the same stages of embryogenesis (Fig. 2.12c) [8].

To evaluate sex ratios in the embryo pools, and to set a reference point for evaluating allelic imbalance and dosage compensation on the X-chromosome [50], we performed genome DNA sequencing (gDNA) on each cross. This confirmed that our embryonic pools were relatively sex balanced, with the expected X-chromosome allelic ratio of ~ 0.67 observed across our gDNA dataset (Fig. 2.6f). Consistent with full dosage compensation on the maternally-derived male X chromosome [51], we observed a maternal:paternal ratio of 0.74 for mRNA (Fig. 2.6f; Methods).

Interestingly, a similar degree of up-regulation (dosage compensation) was not observed for chromatin data. For both chromatin accessibility and histone modifications, a ratio closer to 0.67 was observed at X chromosome sites (e.g. H3K27ac=0.688, H3K4me3=0.692), which is more similar to the genomic ratio (0.66) than the ratio of 0.75 expected under full dosage compensation (Fig. 2.6f).

The ratios showed no significant difference when comparing proximal to distal sites. Together, this argues against the hypothesis that the two-fold upregulation of gene expression on the male Drosophila X chromosome results from a two-fold increase in the loading of polymerase at its genes' promoters [52]. Our results rather indicate that whatever the mechanism of dosage compensation in Drosophila is, it does not lead to a linear increase in chromatin accessibility on the male X chromosome, though some increase in accessibility on the upregulated X is consistent with our measurements [53, 54]. Regardless of its cause, we used the empirically observed average ratio for X-chromosome features for each data type to form the null-hypothesis in subsequent beta-binomial tests for allelic imbalance.

Overall, statistically significant allelic imbalance is common, with 46% of genes and between 18-25% of non-coding features showing imbalance in at least one line at any time point (Fig. 2.6b, FDR <0.1). The magnitude of allelic imbalance is generally evenly distributed across SNPs with a range of minor allelic frequencies. Highly imbalanced peaks show a strong enrichment for extremely rare SNPs (including potentially *de-novo* mutations) found uniquely in the maternal line relative to the 205 lines of the full DGRP panel (Fig. 2.12d, chi2; $p<2.2e-16$), highlighting the disproportionate impact of rare and *de-novo* mu-

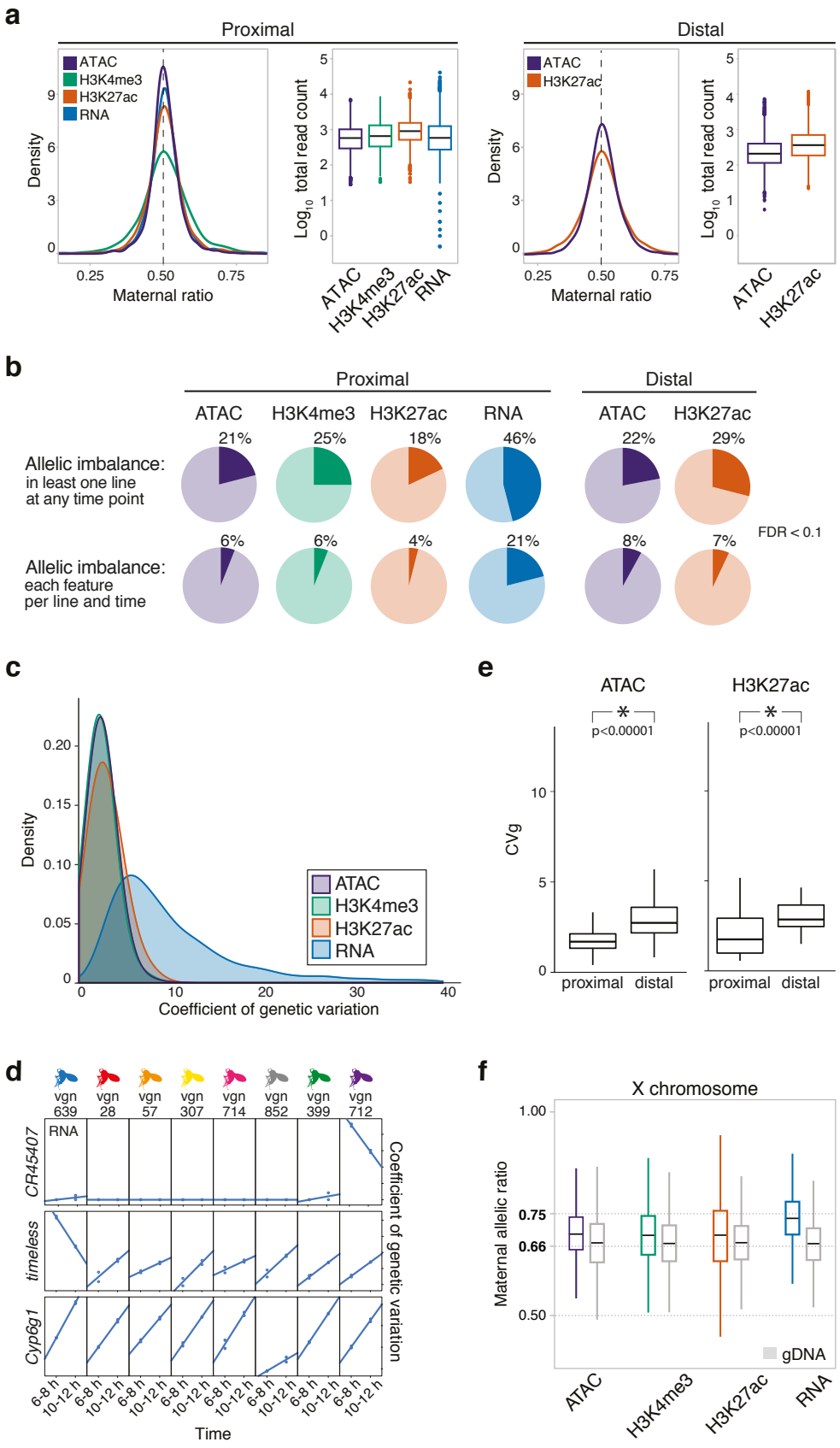

Figure 2.6: Allelic imbalance is common across regulatory data types. cf. legend on next page.

Figure 2.6: a. Density plot of allelic count distribution and matching boxplot showing total read count abundance (log10) in the autosomes at TSS proximal (left) and distal (right) regions for all data types assayed. b. Pie charts showing significantly allelic imbalance (AI) genes/features at promoter proximal (left, TSS +/-500 bp) and distal (right, 500-1500 bp +/- from TSS) regions for all four data types (FDR<0.1). Upper row shows AI events in at least one F1 line at any time point. Lower row shows AI events detected in all 8 F1 lines in all time points, on a per line and time basis. c. Smoothed histograms (left) show the distribution of coefficients of genetic variation for all features with statistically significant between-line variances within each regulatory layer. d. Line plots show three examples of individual lines having distinct expression profiles. Coefficients of genetic variation are typically larger for RNA than for non-coding features, an effect that often results from one or two lines having significantly altered expression relative to the panel as a whole. e. Box plots showing the distribution of the coefficient of genetic variation (CVg, y axis) for chromatin accessibility (left) and H3K27ac signal (right). Each panel compares the results for promoter-proximal and promoter-distal features. Genetic influences are more pronounced at distal than proximal regulatory elements in ATAC and H3K27ac. f. Box plot showing the distribution of the maternal allelic ratio of X chromosome in each data type. Each distribution is compared to the allelic ratio observed in genomic DNA for the same data type (genes/regulatory regions) in grey.

tations on phenotypes [8].

Allelic imbalance is more frequently observed for RNA than for other regulatory layers (Fig. 2.6b). In contrast to what is observed in mammals [55], promoter-proximal elements are slightly more polymorphic (pair-wise differences (pi)=0.132vs0.129, Wilcoxon-test p=1e-10) and evolve faster (phyloP=0.514vs0.560, Wilcoxon-test, p<2.2e-16) in Drosophila as compared to distal elements (putative enhancers). Despite this, distal peaks of open chromatin and H3K27ac show greater (Tukey's ASD, p<0.0001) and more frequent allelic imbalance (chi2; p<2.2e-16) than their proximal counterparts, highlighting the potential evolutionary relevance of distal regulatory mutations (Fig. 2.12b).

Although allelic imbalance is common, not all biological categories are equally affected. Imbalanced genes and associated regulatory features are enriched for fast-evolving and Drosophila-specific genes lacking clear categorical annotations [56, 57] and are depleted in TFs and their associated regulatory elements (Fig. 2.13; Methods), consistent with a previous eQTL study [8]. Allelic imbalance (AI) is also enriched for metabolic genes at the RNA level, although this is not observed for associated regions of open chromatin or histone modifications (Fig. 2.13).

The observed differences in AI among gene categories may reflect differential histories of selection; regulatory regions in the vicinity of TFs show a depletion of nucleotide diversity (pi, rank biserial correlation=-0.052, p<1e-4) and harbor more low-frequency SNPs (rank biserial correlation=-0.173, p=2.8e-3) compared to background. This AI could also be explained if different gene categories have different sensitivities to mutations (buffering), a point we explore below.

Directional imbalance suggests recent selection on environmental response genes

For most gene categories and regulatory elements, allelic imbalance is equally likely to favor the maternal or the paternal allele. A subset of categories, however, show consistent and often large, parent-specific biases (Methods). This trend is particularly striking for male-biased genes and regulatory elements

which show a strong over-representation of categories associated with immunity or insecticide resistance (Fig. 2.14a).

Exemplary of this trend is *Cyp6g1*, a gene that is not expressed in embryos of our maternal line, which is derived from a laboratory stock isolated before the widespread use of agricultural pesticides, or in embryos sequenced by the modENCODE project [58]. It is, however, strongly upregulated in every measured paternal haplotype from the wild, and its overexpression contributes to DDT resistance in multiple Drosophila species (Fig. 2.14b, [59, 60]. This result exemplifies how recent strong positive selection on *cis*-regulatory variation can shape embryonic gene expression even on relatively short time scales.

Because our F1 lines share a common maternal genotype, line effects are expected to be directly proportional to genetic effects/heritability [36] (Methods). This allows us to thus directly examine how allelic-imbalance, extreme or otherwise, relates to heritable variation in total counts. To make comparable the variances of genes and features with different mean read counts, we calculated the *coefficient of genetic variation* for each gene by scaling estimated 'between-line variances' by the variance stabilized mean read count of each feature such that line effects are expressed as a percentage deviation from a gene/features mean read counts [6, 61].

For chromatin features, the magnitude of genetic variation on measured signal is relatively modest, with the average peak varying by 5-10% of the mean phenotype among crosses (Fig. 2.6c,d). Heritability is higher at distal regulatory elements compared to their proximal counterparts (Fig. 2.6e, p<1e-5), consistent with the greater magnitude of AI at distal sites suggesting differences in the frequency of evolutionarily relevant mutations in these two site classes. For RNA, the magnitude of genetic effects are especially pronounced, with a coefficient of genetic variation of 9% and some genes showing coefficients of 40%, indicating that genetically encoded differences in expression can account for nearly half of some genes' mean expression levels.

For many genes, high coefficients of variation are driven by one or a few lines showing highly divergent patterns of expression (Fig. 2.6d). These highly variable genes, including several involved in response to environmental stressors and toxins (e.g. *Cyp6g1*), show a strong overlap with genes having extreme allelic imbalance (p-value<1e-6; Fig. 2.15a), suggesting that large differences in expression among individuals can be driven by large-effect *cis*-acting variants.

The impact of *cis*-acting genetic variation is largely consistent across development

We next evaluated if, and to what extent, allelic ratios change during development. Overall, we observed a surprising constancy of allelic imbalance between time points: Despite the temporal modularity of many *cis*-regulatory elements, imbalanced features at one time point have a 50% chance that it will be imbalanced in the subsequent timepoint (Fig. 2.7a, S2.16a).

To further quantify the potential impact of development on allelic ratios, we constructed a series of linear models comparing the effect sizes of genetics (genotype/line effect) vs developmental stages (time effect) on total counts and allelic ratios across our experimental design (Methods).

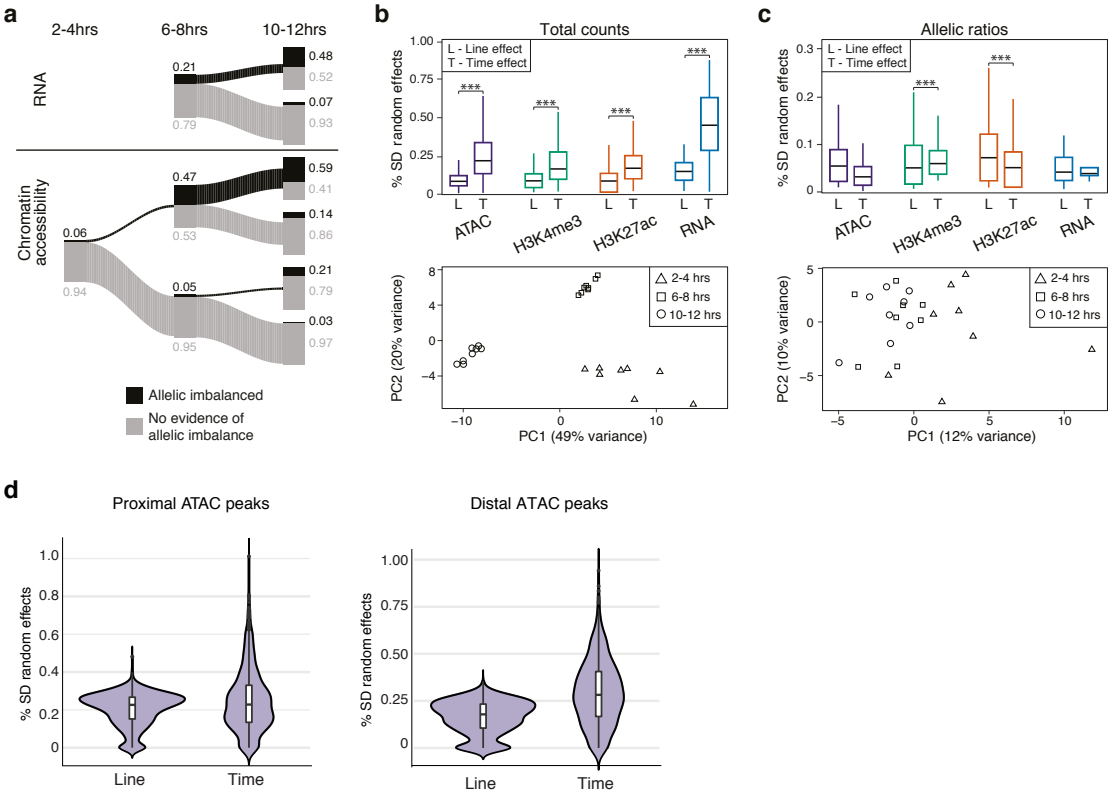

Figure 2.7: Allelic imbalance is generally not predictive of developmental times. a. The relationship of allelic imbalance across time points for RNA (upper panel) and chromatin accessibility (lower panel). Proportions of AI and non-AI features are shown in black and grey, respectively, and represented by the thickness of line. Exact proportions for each category are provided as numbers. Data for 2-4hr time point for RNA are not included due to presence of maternal transcripts at this stage. b. Top: Box plots showing the distribution of effect sizes obtained from mixed linear models, for total counts. For each type of data (gene/feature), the effect sizes of time (T) and line (L) effects are shown. Bottom: Principal component analysis of gene expression for total counts. c. Top: Box plots showing the distribution of effect sizes obtained from mixed linear models, for allelic ratios. For each type of data (gene/feature), the effect sizes of time (T) and line (L) effects are shown. Bottom: Principal component analysis of gene expression for allelic ratios. d. Results from mixed linear models examining the effect of developmental time versus line (genotype) between proximal and distal ATAC-seq peaks for total count data. Distal peaks show a larger time effect compared to genotype effect (Mann-Whitney U, p<2.2e-16), which was not observed at proximal peaks.

For total counts, developmental time was the greatest contributor to variation across all data types (Fig. 2.7b, upper panel), consistent with the clear time specific clustering by principal component analysis (Fig. 2.7b, lower panel, shown for RNA, Methods). Interestingly, this predominance of time is largely restricted to distal and not proximal sites for ATAC-Seq (Mann-Whitney U, p value=2e-61; Fig. 2.7d), likely reflecting the constitutive accessibility of promoters during Drosophila embryogenesis [62].

In contrast to the total counts, the impact of developmental time on allelic ratios is significantly reduced compared to genetic (line) effects (Fig. 2.7c, upper panel). Correspondingly, there is a lack of time-point specific clustering in PCA (Fig. 2.7c, lower panel), although there are some examples of allelic ratios that change over time in a coordinated

manner between regulatory layers (Fig. 2.16b).

Interactions between genetic and developmental effects can play an important role in gene regulation [63, 64]. We therefore looked for evidence of interaction effects in linear models fitted to total counts or allelic ratios containing only time, only genotype, time plus genotype (time + genotype), or interactions between the two (time x genotype).

Interaction effects occur frequently at the total count level and are particularly common for gene expression, making up nearly 30% of all analyzed models and highlighting a potentially important role for developmental stage by genetic (TxG) interactions in population-level variation during embryogenesis, as previously observed [8].

In contrast, there is little evidence for interaction effects for allelic ratios for gene expression or ATAC-seq peaks, consistent with both the relative stability of allelic ratios over time and the general additive heritability of *cis*-acting regulatory mutations observed in other studies [65, 66]. It is also consistent with the relatively small numbers of allelic ratios reported to show influences of gene by environment interactions across environmental conditions [67, 68].

In summary, allelic effects are often larger at distal sites, compared to promoter regions. Given that, and the dynamic nature of developmental enhancers, we were surprised to find, however, that allelic ratios at distal sites are surprisingly stable. Compared to total counts, allelic ratio is a poor predictor for developmental time. At the total count level, however, we observe more extensive genetic effects, with interaction (G x D) effects common for gene expression. This suggests that in addition to *cis*-acting mutations, *trans* effects may contribute importantly to genetic variation within populations.

Information flows in different directions across *cis*-regulatory layers

Although quantitative signals at chromatin features are highly correlated with gene expression, the relative causal relationships between chromatin accessibility, histone modifications, and gene expression remain unclear. To assess this, we made us of the relationships in allelic ratios between different regulatory layers to model the paths by which genetic variation influences regulatory phenotypes.

Allelic ratios in all pairs of datatypes are correlated, to varying extents (Fig. 2.8a), and in all cases we could reject the null hypothesis of independence (e.g. highest p-value between all comparisons=4.2e-17 for ATAC and H3K4me3). In parallel, we also tested for an enrichment/depletion of co-occurring, statistically significantly imbalanced (FDR<0.1) genes/features using an intersection-union test (Fig. 2.8b; Methods) and a distance of +/-1500 kb for assigning distal features to genes.

The co-occurrence of allelic imbalance is especially pronounced for chromatin features, in particular H3K4me3 and proximal H3K27ac with a log-odds >2.0 (Fig. 2.8b). Interestingly, for chromatin accessibility and H3K27ac, the co-occurrence of AI is more pronounced at promoters (proximal) than enhancers (distal) (Fig. 2.8b), despite allelic imbalance being slightly more frequent and of greater magnitude at distal sites (p<2.2e-16, Fig. 2.6b, Fig. 2.12b). This suggests that H3K27ac and chromatin accessibility are more functionally

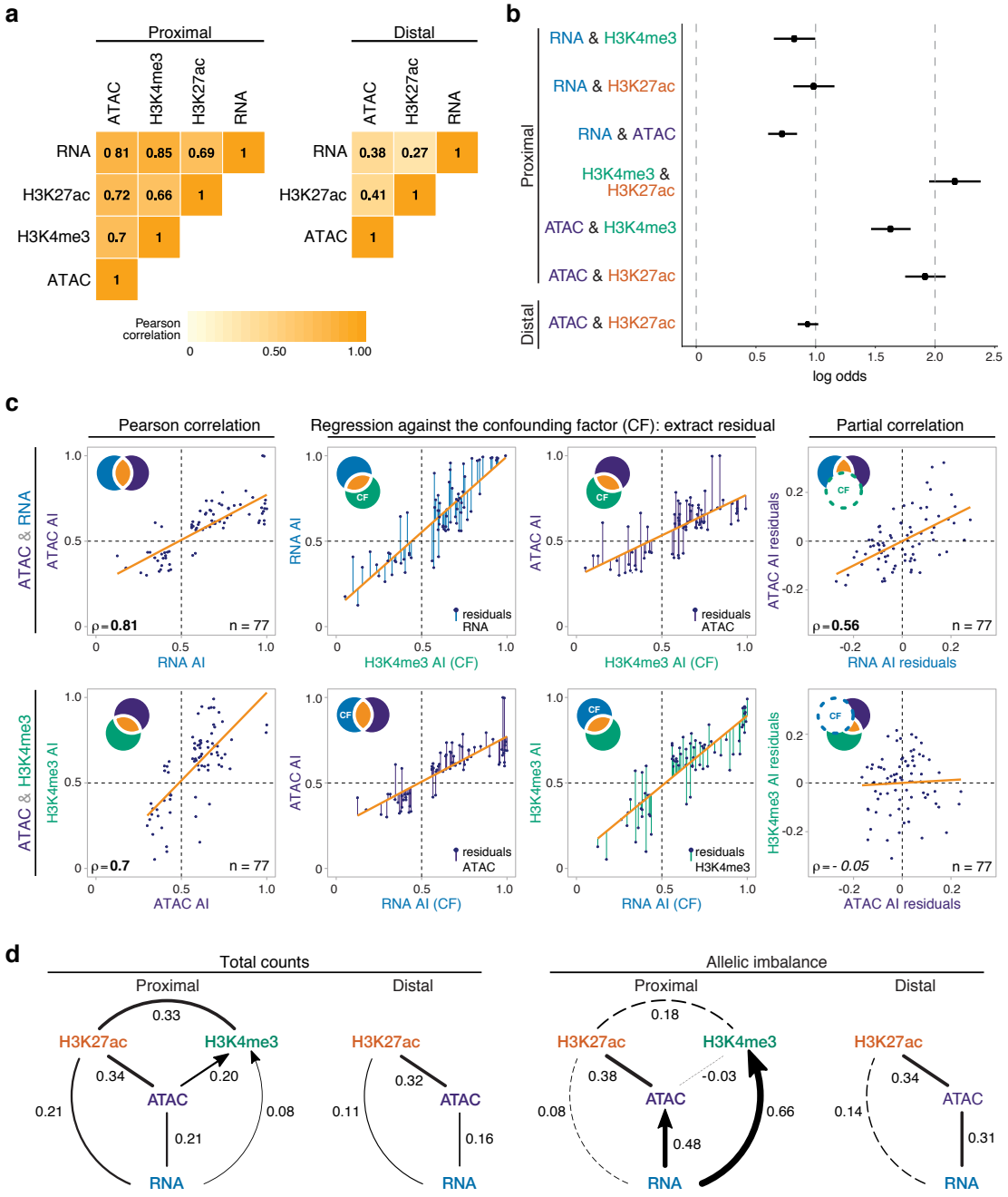

Figure 2.8: Allelic imbalance is propagated through regulatory layers via different epigenetic paths. cf. legend on next page.

Figure 2.8: a. Heatmap showing Pearson correlation coefficient of allelic ratios between each pair of data type for promoter proximal (left) and promoter distal (right) regions. Data restricted to 6-12hr and features/genes whose allelic ratio exceeds 0.5 +/- 0.06. b. Increased log odds of co-occurrence of allelic imbalance between two regulatory layers. X-axis shows the log odds based on intersection-union tests (Methods). 6-12hr data is shown. Bars stemming from dots are 95% confidence intervals. c. Stepwise example of a partial correlation analysis of allelic ratios for three variables (ATAC, RNA and H3K4me3). Partial correlation analysis between gene expression and chromatin accessibility is shown in the upper row, between proximal H3K4me3 signal and chromatin accessibility in the lower row. Venn diagram schematics, in top left, illustrate the variance of each variable and its shared proportion (orange), as measured by the linear regressions (orange lines). Left panels: Pearson correlations for the two comparisons are significant. Middle panels: regression of each initial variable against a third, confounding variable (upper row: H3K4me3 ; lower row: RNA). Residuals of the initial variables indicated as colored lines and represent the non-overlapping part of the circle of the same color in the schematic. Right panels: correlations of the residuals, which exclude the variance shared by the confounding factor (dashed circle in schematic). This resulting partial correlation is not significant in the bottom example, suggesting a lack of direct correlation within the pair H3K4me3-ATAC. d. Partial correlation and directional dependency regression for total counts (left) and allelic ratios (right). Significant partial correlations (solid lines) suggest dependencies among regulatory layers. For each significant edge ($p<0.01$), copula regression was used to assign directionality to the relationship (arrows, delta>0.01). Results are shown for promoter proximal and distal regions independently. Thickness of the lines indicates the value of partial correlations. Dashed lines indicate non-significance in partial correlation analysis.

coupled at promoters compared to enhancers, perhaps reflecting the fact that not all active enhancers seem to require H3K27ac [69, 70].

Due to the large amount of covariation between the different regulatory features (Fig. 2.8a), it is difficult to infer causal relationships from correlation data alone. To address this, we used partial correlation to identify independent, pairwise correlations between multiple co-varying variables beyond their global correlations after thresholding on allelic ratios to remove features/genes with low information content (Fig. 2.8c, 2.17a-b , Methods) [71, 72].

We first analyzed the total count data to evaluate the overall relationships among histone modifications and gene expression. Our results closely mirror those of Lasserre *et al.* in CD4+ and IMR90 cells [71], including finding a clear relationship between gene expression levels and the total abundance of H3K27ac that 'explains away' (at least in a statistical sense) much of the correlation between gene expression and promoter-proximal H3K4me3 (Fig. 2.8d, left). We also observed a statistically significant relationship between open chromatin and gene expression, though the strength of this partial correlation is reduced relative to standard Pearson correlation analyses (Fig. 2.17c), highlighting the value of histone modification data for predicting gene expression.

To assess the functional impact of *cis*-regulatory perturbations, we next applied the partial correlations analysis to allelic ratios (Fig. 2.8d, right). Relative to the total count data, we observed a much stronger relationship between open chromatin and gene expression for both proximal and distal regulatory elements unconditional on other regulatory layers, highlighting an important, possibly causal, link between mutations affecting accessibility (presumably TF occupancy) and gene expression (Fig. 2.8d).

A correlation is also observed between H3K27ac and open chromatin at pro-

moters, though interestingly, we see little evidence for a direct relationship between H3K27ac and gene expression itself (Fig. 2.8d, right). The latter is surprising as it differs from what is observed with total count data, and suggests that although promoter H3K27ac is highly correlated with, and even predictive of gene expression [73], they may not be mechanistically directly linked.

In contrast, allelic ratios for promoter proximal H3K4me3 show strong evidence of a direct correlation with gene expression that is independent of, at least in a statistical sense, allelic differences in chromatin accessibility or H3K27ac (Fig. 2.8d, right). Taken together, this analysis suggests two independent pathways by which selection on segregating mutations could influence gene expression, one affecting open chromatin and promoter-proximal H3K27ac, and the other influencing H3K4me3.

To explore these relationships further, we analyzed each edge identified by partial correlations using copula directional dependence analysis [74, 75], a statistical approach based on copula regression that evaluates the directionality of the pairwise relationships allowing for nonlinearities (Methods), to assign a direction to each edge for which there is clear evidence for greater explanatory weight in one direction.

For TSS-proximal regions, this placed RNA upstream of both H3K4me3 and open chromatin (Fig. 2.8d, arrow, right). Although counter intuitive at first glance, this suggests genetic variation causing changes to H3K4me3 is often asymmetrically coupled to changes to RNA, that is, gene expression is not highly sensitive to variations in H3K4me3 occupancy but changes to RNA is more predictive of H3K4me3 enrichment. This could reflect redundancy between regulatory elements, i.e. as multiple regulatory elements typically work together to regulate a gene's expression, changes in a single open chromatin region, as tested here, may not be sufficient to impact expression in an allele-specific manner.

Similarly, variation in allele-specific RNA counts better explain variation in chromatin accessibility compared to the reverse, hence, not all changes in open chromatin due to *cis*-acting variation lead to a corresponding change in gene expression (Fig. 2.8d, right panel). Our result is also consistent with the hypothesis that H3K4me3 is not required for gene expression, but may rather be deposited as a consequence, and be more involved in post-transcriptional events, as recently proposed [76].

In summary, *cis*-acting genetic variation shows greater covariance between open chromatin and H3K27ac enrichment at promoters compared to putative enhancers. By measuring informative dependencies to the effects of *cis*-acting genetic variation, we identified multiple epigenetic pathways affecting transcription. Specifically, genetic variation acts to change gene expression levels via interplay between at least two different promoter-proximal paths: open chromatin and H3K27ac, or H3K4me3. Moreover, the flow of information suggests that gene expression is often buffered against *cis*-acting mutations (presumably affecting TF binding) at associated regulatory elements.

Regulatory buffering varies depending on the gene classes and local chromatin architecture

Genes from different functional categories often have differences in the com-

plexity of their regulatory landscapes. Metabolic genes, for example, typically have relatively simple and more compact regulatory landscapes with fewer enhancers that are generally located close to the gene's promoter [77]. TFs, in contrast, have many enhancers often with partially overlapping spatial activity ("shadow enhancers") that are located at varying distances from the gene's promoter [78, 79].

This additional regulatory complexity is thought to make TFs more robust to deletions and mutations affecting their regulatory elements [16, 80–82]. To examine this, we used conditional probabilities with gene categories taken from the GLAD database for fly gene list annotation [83] to assess the extent to which allelic imbalance in the expression of different gene categories is independent of, or decoupled from, imbalance in their associated regulatory elements, treating gene categories with greater independence as being more 'buffered'.

Among all conditional comparisons, the expression of TFs, transmembrane genes, ancient genes (conserved bilaterian processes), genes of major signaling pathways, secreted genes, are most insensitive to imbalance in other regulatory layers (Fig. 2.9a). In contrast, genes and their regulatory elements associated with cytoskeletal function, glycoproteins, and, notably, metabolism show an increased sensitivity to allelic imbalance in other regulatory layers (Fig. 2.9a, Fig. 2.13). Taken together, our analyses suggest that in addition to purifying selection acting to remove genetic variation, regulatory buffering furthers the expression of TFs and other developmental regulatory factors from the effects of *cis*-acting mutations.

To more directly assess the relationship between buffering and regulatory complexity, we compared the number of ATAC-seq peaks in a gene's regulatory domain (+/- 1.5kb TSS) with the probability of imbalance in that gene's expression. Imbalanced genes have fewer associated ATAC-seq peaks genome-wide (Pearson's, r=0.1, p=1.7e-12; Fig. 2.9b). The trend is particularly striking for single-peak genes, which have significantly more AI than genes with multiple associated regulatory elements (Mann-Whitney U test, p-value=6.4e-6). Conversely , genes with more complex regulation (i.e. with a greater number of associated peaks) have more reproducible gene expression allelic ratios between biological replicates (Pearson's r=-0.05 Pearson's correlation, p-value=3.0e-7), indicating less variation in their expression.

Consistent with the observation of transcriptional robustness (a lack of AI) for genes with multiple regulatory elements, genes associated with partially redundant enhancers (or shadow enhancers) have a modest reduction in the frequency of allelic imbalance compared to genes without (Fig. 2.9c, chi2=5.3, p=0.02), based on a previously defined set of shadow enhancers for mesodermal genes [16]. Furthermore, genes associated with shadow enhancers show more evidence of buffering as evidenced by a greater degree of independence from allelic imbalance in associated gene regulatory elements ($p(RNAAI|ATACAI)$ = $0.04 vs 0.14$, respectively), supporting the conclusion that enhancers with partially overlapping activity can act to stabilize genetic variance in gene expression in the face of perturbations. We note, however, that this buffering is not absolute – even genes with multiple regulatory elements are more likely to be imbalanced when multiple associated peaks show unbalanced allelic ratios (gene ex-

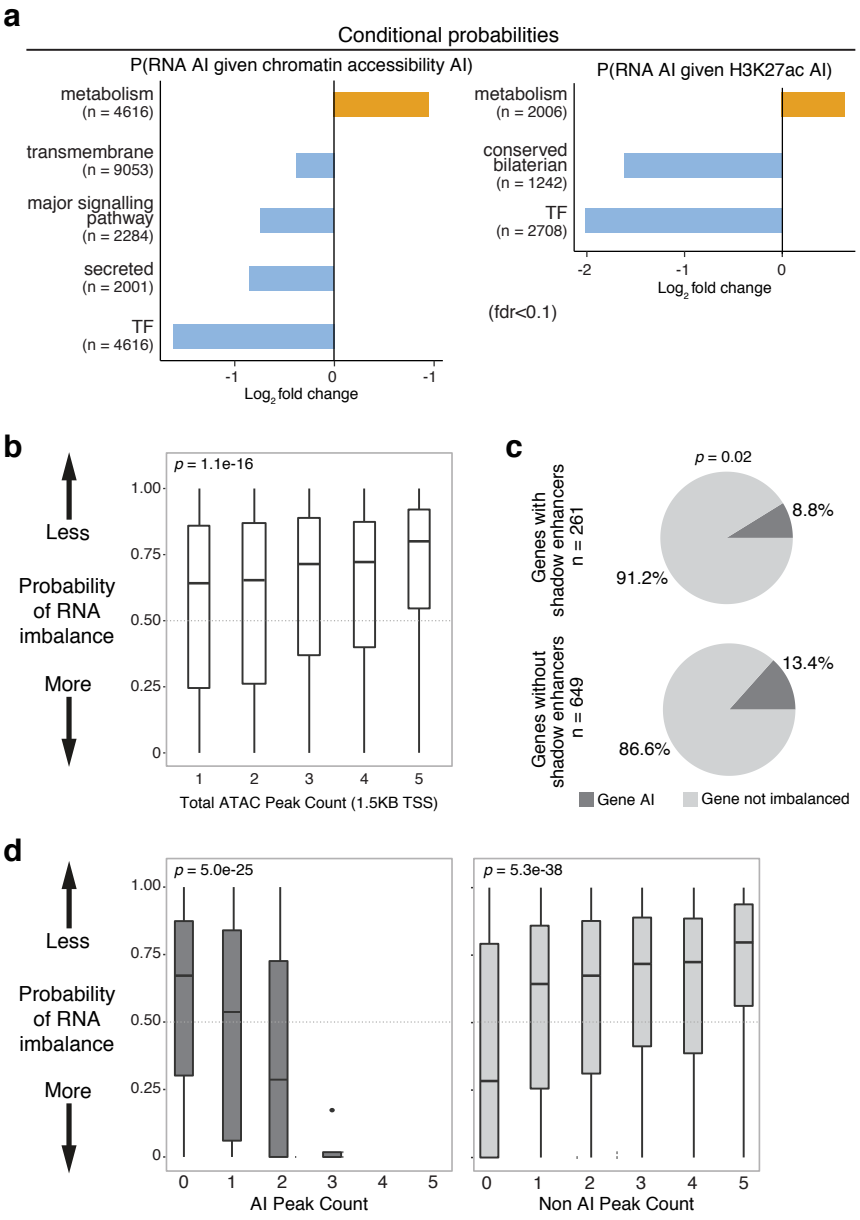

Figure 2.9: Regulatory buffering varies across gene categories and with local chromatin structure. a. Conditional probability of allelic imbalance in gene expression given allelic imbalance in associated chromatin peaks (left) and regions of H3K27ac (right) across gene categories. X-axis show log2 fold change where background is based on genome-wide expectation. Gene categories enriched (orange) or depleted (blue) for imbalance, relative to background, are indicated (FDR>0.1, Fisher's exact test). b. Box plots denote the probability of allelic imbalance in gene expression based on numbers of neighboring ATAC peaks (TSS<1.5kb). Genes associated to more ATAC peaks are more likely to show similar expression in both alleles compared to genes with fewer peaks. c. Pie charts displaying the proportion of genes with allelic imbalance in RNA associated to ATAC-seq peaks overlapping known partially redundant/shadow enhancers (top) or not (bottom). Genes associated with shadow enhancers are less likely to be allelic imbalanced compared to genes without (chi2=5.3, p=0.02). d. ATAC-seq peaks have a cumulative effect on gene expression. The probability of imbalance in gene expression (y-axis) is shown as a function of the number of ATAC-seq peaks that are allelic imbalanced (left) or not imbalanced (right).

pression AI p-value compared to proportion of imbalanced ATAC-seq peaks, Pearson's r=-0.1, p<2.4e-37; Fig. 2.9d).

In summary, there is a relationship between a gene's regulatory complexity and the degree to which its expression is influenced by functional mutations in its regulatory elements, with more regulatory elements providing a degree of buffering against genetic perturbations. Furthermore, allelic imbalance at multiple enhancers in the vicinity of a gene can have a cumulative influence on allele-specific gene expression.

Gene expression is less heritable than chromatin features

Gene expression phenotypes are influenced by linked, *cis*-acting, genetic variation, but also by *trans*-acting variation that is not directly captured using F1s alone. To estimate the relative impact of *trans*-acting variation, we collected iChIP-Seq, ATAC-Seq, and RNA-Seq information from a trio of lines consisting of one F1 line and stage-matched data from the maternal (vgn) and paternal (DGRP-399) lines.

Because F1 cells represent two haplotypes with a common nuclear environment, allelic ratios in the F1 can be seen as an estimate of the *cis*-based differences between the two parents. Differences in parental read counts not reflected in F1 allelic-ratios, in contrast, act as an estimate of the *trans*-acting contribution to between line divergence [65, 66, 84, 85].

Using a maximum likelihood framework, we classified features as *cis*, *trans*, *cis-trans*, or *conserved* and found a similar distribution among categories for all non-coding chromatin features, with *cis*-acting effects being more common than *trans* (59% vs. 41%, p<2.2e-16, chi2; Fig. 2.10a, S2.18a, Methods). This enrichment is particularly pronounced for histone modifications, with nearly twice as many *cis* influenced peaks compared to *trans* (Fig. 2.10a, S2.18a). Gene expression, in contrast, is more strongly influenced by *trans*-acting genetic variation (55% *trans* vs. 45% *cis*, p=0.0073, chi2; Fig. 2.10a). Moreover, a higher fraction of *cistrans* genes have more *trans* compared to *cis* variation that is the case for chromatin features (*trans* proportions 0.67vs.0.53, p=2.77e-05; Fig. 2.18b).

Previous studies suggest that, relative to *trans* influences, the effects of *cis*-acting mutations are more likely to be inherited in an *additive* manner [66, 86, 87]. This matters because while all heritable genetic differences can have evolutionary consequences, it is typically additive genetic variation, variation whose affects are common across all genetic backgrounds, that is most directly acted upon by natural selection [36].

By evaluating the extent to which the F1 signal (total read count) for each gene/feature departed from the parental average (a strictly additive model), we were able to make a similar evaluation within our data to better understand the differences in heritability among classes of genes and regulatory elements. For open chromatin, whether influenced by *cis* or *trans*, we could reject a non-additive model in fewer than 1% of cases (Fig. 2.18b), consistent with the finding that most variation affecting TF binding is inherited additively [66]. For gene expression, in contrast, the additive model could be rejected for 32% of genes, with *trans* influenced genes departing from an additive model far more often than *cis* (24% vs. 2%, p<1e-4; Fig. 2.10b).

To better understand the factors that

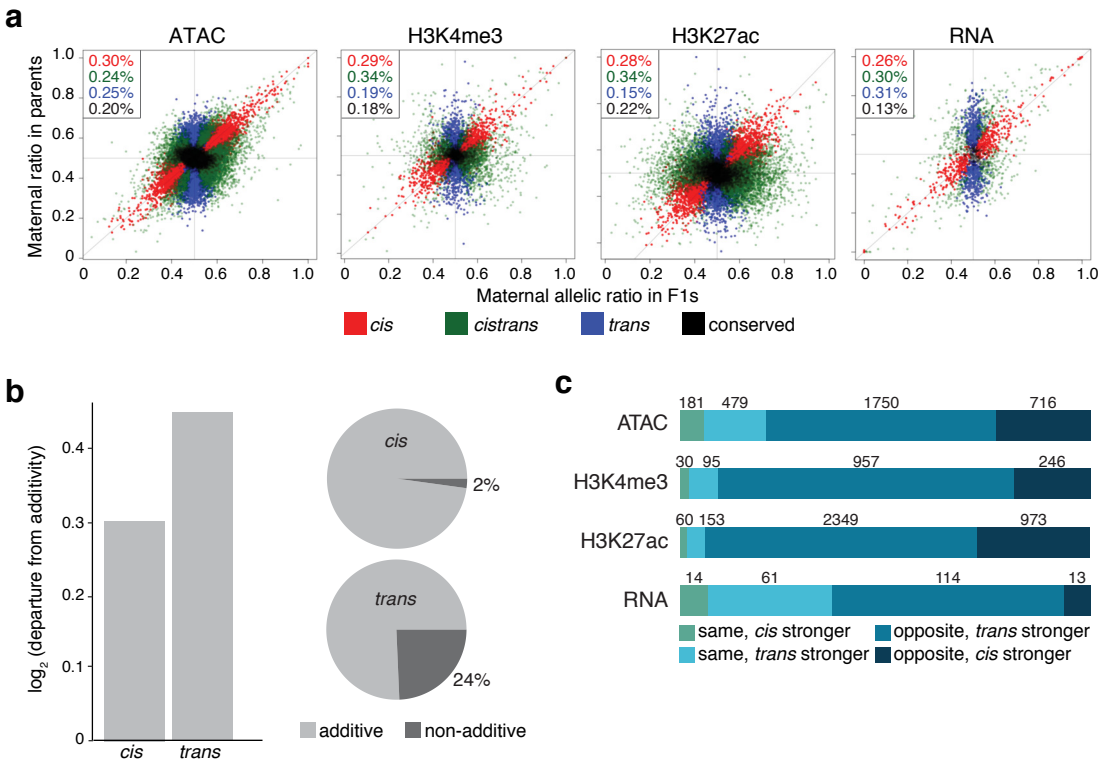

Figure 2.10: Chromatin features are more heritable than gene expression. a. Scatter plots of F1 allelic ratio (x-axis) against the maternal/paternal ratio observed in (normalized) parental, total count libraries. Genes/features along the diagonal are exclusively influenced by *cis*-acting variation, while vertically distributed genes/features show exclusively *trans*-influences. Colors indicate maximum likelihood classification into *cis*, *trans*, and *cistrans* (a mixture of *cis* and *trans*) or *conserved* genes/features. b. Left, bar plots shows the magnitude of deviation from additivity (parental mean) for features classified as *cis* vs. *trans* (BIC >= 2). Right, pie charts showing fraction of additive and non-additive genes for *cis* (upper) and *trans* (lower) classes. c. Classification of *cistrans* effects (BIC >= 2) for each regulatory layer into categories reflective of likely selective effects. Numbers and horizontal bars represent the size and relative proportions of each *cistrans* relative direction class in each data type. There is more directional selection (same directions, *cis* + *trans*>*cis*) than compensatory evolution (opposite directions, *cis* + *trans*<*cis*) in gene expression as compared to chromatin features.

contribute to the proportion of *trans*-acting variation (and, by association, non-additive heritability), we examined the contribution of regulatory complexity. Mirroring our buffering results, genes with more regulatory elements in their vicinity are more likely to be classified as *trans*-acting (*trans*=2.58 peaks per gene vs. 1.9, p=0.00094) and more likely to show non-additive inheritance (non-additively inherited genes=2.19 genes per peak vs. 1.82 genes per peak, p=1.4e-3).

Similarly, we see a significant, though modest, enrichment of *trans* influences among TFs and a depletion among metabolic genes, two categories that are strongly distinguished in the complexity of their associated regulatory landscapes. Correspondingly, among 80 tested gene categories (GLAD), DNA-binding transcription factors (p=3e-3) and, interestingly, mitochondrial associated genes (p=2e-6) stand out as the two gene categories with statistically elevated frequencies of non-additive inheritance (Fisher's exact test; Meth-

ods). Thus, while TFs generally show reduced genetic variation among lines (Fig. 2.15b-e) and reduced allelic imbalance in gene expression (Fig. 2.13), they are still affected by *trans*-acting variants whose non-additive inheritance reduces the efficacy by which selection can alter gene expression differences among different lines.

Genes influenced by both *cis* and *trans* acting variants (*cistrans*) provide an opportunity to understand patterns of recent selection: In the face of compensatory evolution, *cis* and *trans* acting influences are more likely to work in opposite directions, while directional selection will be more likely to reinforce *cis* and *trans* effects acting in the same direction.

Using the classification of Goncalves et al [65], we observed that *cis* and *trans* effects were much more likely to act in a compensatory manner as compared to gene expression: For chromatin accessibility and histone modifications, 13% of *cistrans* features were classified as *same* vs. 37% for RNA ($p<2.2e\text{-}16$ chi2). This suggests that for RNA there is either more frequent directional selection or less efficient selection against directional changes. Our finding is robust to the method used to classify *cis* + *trans* effects [84], with 63% of *cistrans* RNA features being classified as divergent for RNA vs. 22% for chromatin features ($p<2.2e\text{-}16$ chi2; Fig. 2.18d).

Taken together, these results suggest clear differences in evolutionary trajectory between regulatory layers which reflects population processes operating at different levels of organization, as well as differences between functional gene classes.

2.4.4 Discussion

We used genetic variation to better understand the impact of sequence variation in regulatory DNA on embryonic gene expression, and to shed light on how these effects are propagated or buffered through different layers of regulatory information during embryonic development. We generated allele-specific measurements of chromatin occupancy (ATAC-seq), chromatin activity state (using chromatin modifications) and gene expression (RNA-seq) in F1 embryos from eight different genotypes across multiple stages of embryogenesis. Our analysis of this extensive embryonic dataset led to several conclusions about the impact of regulatory mutations on transcriptional phenotypes.

First, although *cis*-acting genetic variation in gene expression and associated regulatory signals is fairly common in development, its effects are not equally distributed across the genome. Allelic variation is both more frequent and has greater magnitude at distal regulatory elements (putative enhancers) compared to promoters, despite genetic variation itself being more common at promoters. This may in part be due to differences in the relative importance of sequence content at promoters and enhancers – many promoters, particularly for broadly expressed genes, are remarkably tolerant to mutations [5].

Interestingly, while AI is more frequent and of greater magnitude at distal elements, it is less likely to be propagated to other regulatory layers (Fig. 2.7), suggesting that enhancer mutations are either more effectively buffered or of lower impact, a hypothesis that fits well with the observed robustness of gene expression to deletions that remove distal regulatory elements [14, 16]. But while

their impact is often lower, genetic variation affecting distal regulatory elements are not without evolutionary relevance - sufficiently large effect gene-by-gene or gene-by-environment interactions can theoretically serve to release this 'cryptic' genetic variation [88–90].

Whether such interactions are sufficiently common for regulatory traits is currently unknown, though we note here that the genetic contribution to total count gene expression varies considerably between time points, suggesting a substantial context-specificity to the mutations underlying gene expression variation.

Second, although all data types (open chromatin, histone modifications, RNA levels) are highly correlated, their explanatory values (potential causal relationships) as revealed by partial correlation analysis are far from equal. Using *cis*-acting mutations as perturbations to development, we observed a strong, potentially direct relationship between genetic variants affecting open chromatin (TF binding) at both proximal and distal sites and gene expression, as expected. The relationship between histone modifications and gene expression, however, proved more surprising – in contrast to total count data, both in this study and previously reported [71], we note a strong, potentially causal, link between allelic-imbalance in H3K4me3 signal and allelic imbalance in associated genes.

Although highly correlated with gene expression, the functional requirement of H3K4me3 is unclear. Our copula analysis placed RNA upstream of H3K4me3 (Fig. 2.7d), suggesting RNA levels are buffered against genetic variation affecting H3K4me3. Our results are also consistent with observations H3K4me3 may be deposited as a consequence of gene expression. This link, inferred from our statistical analysis of the impact of genetic variation on both properties, is supported by recent data from genetic ablation studies showing that RNA transcription does not require H3K4me3 [91–93]. In addition, we also observed a second, independent, pathway in which genetic mutations affecting H3K27ac impacted gene expression, but only when they were also associated with *cis*-influenced changes in chromatin accessibility.

Third, we observed that the impact on gene expression of *cis*-regulatory mutations is influenced by regulatory complexity, with genes that have more regulatory elements being less likely to show allelic imbalance (Fig. 2.9). In part, this may be due to selection against variation in regulatory elements associated with these genes. As observed in other studies [16], we found a clear reduction in allelic variation for elements associated with TFs and other developmental regulators as compared to elements associated with different gene categories.

However, selection is unlikely to be the whole story. Even when associated mutations are present, TFs and other complexly regulated genes show an unexpected degree of independence from allelic imbalance in associated regulatory layers, an active buffering process resulting fromt the presence of multiple regulatory inputs [94]. Notably, the buffering of genes with multiple regulatory elements is not absolute. On the contrary, as the number of allelic imbalanced open chromatin regions near a gene increases, as does the probability of a gene showing allelic imbalance.

On the basis of this, we propose that information averaging in a *cis*-regulatory context can be deployed to enhance the overall consistency of transcriptional re-

sponses, with clustered regulatory elements, including shadow enhancers, leading to a reduction in overall allelic imbalance.

Consistent with this, we observe via copula regression that *cis*-acting variation in gene expression could not be well predicted by variation, but rather the reverse – indicating a system in which allelic imbalance in gene expression is often associated with imbalanced regulatory elements, but that imbalance in chromatin accessibility does not imply an impact on gene expression. This in turn suggests a developmental regulatory landscape that is at once replete with functional mutations as well as mechanisms, active or passive, to buffer gene expression against the impacts of those mutation.

This does not mean that such mutations are evolutionarily irrelevant. On the contrary, large effect mutations can directly influence gene expression, with likely consequences for adaptive phenotypes, while small effect mutations, e.g. those affecting histone modifications or chromatin accessibility without affecting gene expression, may accumulate over time to have functionally relevant phenotypes.

Finally, we note that *trans*-acting variation is more common for RNA than for any other regulatory layer, with resulting consequences for the heritability and selectability of gene expression relative to chromatin features. Specifically, genes with complex regulatory landscapes, e.g. transcription factors, had a higher *trans* proportion of their overall genetic influences. This observation, which can be explained by the *cis* buffering effects of complex regulatory landscapes, has potentially counterintuitive evolutionary consequences, as predominantly *trans*-influenced genes are significantly more likely to show non-additive, and thus less selectable, patterns of inheritance. As a result, *trans*-acting variation affecting genes such as TFs may remain in populations even as negative selection and buffering act to reduce the influence of *cis*-acting mutations.

In summary, allelic variation in chromatin accessibility and histone modifications at regulatory elements is prevalent in the genome and capable of propagating across regulatory layers. Information flow depends on the type of regulatory element and appears mitigated at developmental factors. Notably, these *cis*-regulatory changes to individual genes do not have an appreciable effect on overall developmental programs.

2.4.5 Methods

Detailed methods are provided in the supplementary methods section 2.4.7.

Fly husbandry, crosses and embryo collection

F1 hybrid embryos were generated by crossing males from eight genetically distinct inbred lines from the Drosophila Genetic Reference Panel (DGRP) collection [35] to females from a common maternal "virginizer" line. The virginizer line contains a heat-shock inducible pro-apoptotic gene (hid) on the Y chromosome [95] of a laboratory reference strain (w1118). We made the virginizer line isogenic by backcrossing for over 20 generations [96]. Placing embryos from the virginizer stock at 37°C kills all male embryos, thereby facilitating the collection of a large population of isogenic virgin females, which we mated to males of

different DGRP lines (DGRP lines are listed in Fig. 2.5a). In addition, we collected samples from the parents of one genotype (399) for cistrans analysis (see below).

RNA-seq, ATAC-seq and iChIP-seq

For three developmental stages (2-4hr, 6-8hr and 10-12hr after egg-laying), we performed RNA-Seq, ATAC-Seq, and iChIP-seq for H3K27ac and H3K4me3 for pooled embryos of each F1 strain. All experiments were made in replicates from independent embryo collections. iChIP-seq experiments were performed as described in Lara-Astiaso *et al.* [44]. ATAC-seq libraries were 125bp PE, RNA-Seq 118bp PE, and iChIP-seq 75bp PE. In addition, gDNA from 100 embryos per F1 cross was extracted and 75bp SE libraries constructed. All libraries were run on a Bioanalyzer chip, multiplexed and sequenced with Illumina machines.

Sequencing reads processing

Strain-specific genomes and liftOver chain files were constructed for each DGRP paternal line using custom scripts to insert SNPs and indels into the Drosophila dm3 assembly (version 5 from FlyBase). To annotate these parental genomes, we used pslMap [97] to shift reference annotations r5.57 to the parental genomes. ATAC-seq and ChIP-seq reads were mapped using BWA [98], while RNA-seq reads mapped using STAR [99]. In all cases, overlapping read pairs were trimmed so each base was covered only once by the higher quality read. The resulting alignments against both parental genome mappings were merged into a single alignment file. To generate allele-specific counts, the resulting reads were scored for their overlap with known, cross-specific SNPs. Discordant reads (those overlapping alleles assigned to different parents) were discarded. Genomic DNA was generated for each of the F1 lines to filter potentially miscalled variants and simulated reads from each parental genome were used to assess and filter out regions with likely mappability errors. Peak calling was then performed using Macs2 (-broad) for iChIP-seq reads and Hotspot for ATAC-seq reads [100, 101]. To compare between lines and times, we constructed merged peak coordinates across samples (Supplementary methods).

Test for allele-specific imbalance

Two sources of maternally deposited transcripts can bias our estimation of allelic imbalance. (i) Those originating from unfertilized eggs (ii) those still present in fertilized eggs. First, maternally deposited transcripts were identified using RNA-seq data of unfertilised eggs from the same developmental time windows as the F1 samples. These transcripts were filtered out across all time points. Second, based on experiments by 6 hours post egg-laying no maternally deposited transcripts in fertilized eggs were detected but not before. Hence, 2-4hr RNA-seq data, where genes with maternal deposition constituted the bulk of detected genes, were excluded from allele-specific downstream analyses.

To test for allelic imbalance, we used an empirical Bayesian method to test the null hypothesis that there is no difference in read counts between F1 alleles for each feature of each data set (RNA-seq, ATAC-seq, H3K4me3, H3K27ac).

Individual tests were performed for each line and for each time point. Total F1 counts ($n_g^{s,i,t}$) can be modeled on an allele-specific basis ($z_g^{s,i,t}$) using a beta-binomial distribution. Specifically, $z_g^{s,i,t}$ denotes the number of reads from the maternal allele mapped to feature f for pool of individuals i, of paternal strain s, at time t. $n_g^{s,i,t}$ denotes the total number of reads mapping to genes for pool of individuals i of strain s, at time t.

$$z_f^{s,i,t} \; Bi(n_f^{s,i,t}, p_f^{s,i,t}) \tag{2.1a}$$

$$p_g^{s,i,t} \; Be(\alpha, \beta) \tag{2.1b}$$

where α, β are the shape parameters of the beta distribution. We tested the following scenarios by maximum likelihood estimation:

$$No\ imbalance : \alpha = \beta \tag{2.2a}$$

$$Allelic\ imbalance : \alpha \neq \beta \tag{2.2b}$$

Due to limited replicates per condition, we borrowed information across features to reduce the uncertainty of estimates and improve testing power by assuming that all features have the same mean-variance relationship [102, 103]. We use empirical data to estimate the over-dispersion parameter (ρ) for each data type based on the beta-binomial distribution. Maximum likelihood estimation used to obtain α and β for each feature of time t and strain s. ρ is calculated as follows:

$$\rho = \frac{1}{\alpha + \beta + 1} \tag{2.3}$$

The mean over-dispersion value for all features was used as the shrinkage term. Likelihood ratio tests (df=1) were used to obtain a p-value, which was adjusted using the false discovery rate (FDR) procedure [104]. Autosomes were tested separately to sex chromosomes; features on Chromosome X were tested using a background allelic ratio of no imbalance that is centered upon the averaged ratio of maternal versus paternal alleles across the data set being compared (i.e. RNA, ATAC, H3K4me3, H3K27ac). Autosomal features were tested using a null distribution of 0.5

Allele-specific changes across lines and developmental time

We use a linear mixed-effects model where random effect components were incorporated to estimate variability between pools of individuals, time points and lines.

$$y_f^{d,s,r,t} = \mu_f + \delta_f^t + \omega_f^s + (\delta\omega)_f^t \tag{2.4a}$$

$$\omega_f^s \sim N(0, \sigma_f^2) \tag{2.4b}$$

μ_f is the intercept term. δ_f^t is a random effect term denoting time. ω_f^s is a random effect based on strain and $(\delta\omega)_f^t$ is a interaction term for time by strain.

To infer the significance of time or strain dependent allele bias, we restrict the values that the parameters can take. Library size differences were corrected for at the allele-combined count level using the TMM method in the R package 'edgeR' [102] prior to analysis. Count data was filtered for reads with more than 20 allele-combined counts. Each autosomal feature was tested using read counts at SNPs common to all lines. Not all features contained enough information for statistical testing. Analyses were limited to features with at least six samples in each of the three time points in at least four genetic strain.

Allele-specific changes across regulatory layers

Intersection-union tests were used to test for the pairwise co-occurrence of allelic imbalance in overlapping genes/features, limited to autosomes, based on rejecting the null hypothesis if a significant outcome with respect to the feature compared at the same time point exists for both data types [105].

To infer pairwise relationships between regulatory data types while reducing indirect relations, we performed partial correlation analysis using the R package 'GeneNet' [106] for both allelic ratios and total count data. Directional dependence modeling was done in a regression framework using copulas to describe the bivariate distribution between our pairwise datasets [107]. Copula regression was used to infer the flow of information for pairwise relationships that showed a significant relationship in partial correlation analyses.

Conditional probabilities for the probability of allelic imbalance given imbalance in a different regulatory data type were calculated by the definition:

$$P(A|B) = \frac{P(A \cup B)}{P(B)} \qquad (2.5)$$

where A and B are the probabilities of allelic imbalance in each data type.

Cis/*trans* analysis

For one F1 line (vgn x 399) and its parental lines, we use maximum likelihood estimation (MLE) to compare parental and offspring ratios simultaneously to determine whether gene expression, chromatin accessibility, H3K4me3 and H3K27ac enrichments are influenced by *cis*-, *trans*-, *conserved* or both *cis*- and *trans*- acting by modeling read counts. For parents, we modeled the data using negative binomial distributions and modeled allelic differences in F1 alleles using beta-binomial distribution (Supplemental Methods). We constrained parameter estimation for each model based on four different regulatory scenarios and derived maximum likelihood values for each hypothetical case on a site-by-site basis. In the presentation of the proportions of features assigned to each category (Fig. 2.18d), we presented the maximum likelihood assignment. In subsequent analyses, we limited analyses to features that showed a BIC difference>=2.

Test for compensatory mutation

Following the procedure of Goncalves *et al.* [65], for all genes classified as having both *cis*- and *trans*-acting influences, we asked if the *cis* and *trans* contributes act to reinforce one another (same direction) or if they operated in opposite directions. Formally, for the *ith* gene, we define the average log2 fold change for the parental lines as x_i and the average log2 allelic ratio from the F1 data as y_i. We then classified:

Opposite–*cis*:$(0<y_i<x_i)$or$(0>y_i>x_i)$

Opposite–*trans*:$(x_i<0<y_i)$or$(y_i<0<x_i)$

Same–*cis*:$(0<x_i<y_i<2x_i)$or$(0>x_i>y_i>2x_i)$

Same–*trans*:$(0<2x_i<y_i)$or$(0>2x_i>y_i)$

A complementary analysis following Landry *et al.* [84] can be found in the supplemental methods.

Measuring additive vs. non-additive heritability

In the case of additively inherited gene expression (or read counts for any of our measured features), we expect that the signal observed in the F1 should be equal to the midpoint (average) of the two parents, while non-additively inherited genes/features should show a significant departure from that midpoint. To formally test for non-additivity, we made use of the standard workflow in DESeq2 with two modifications. First, we set the 'betaPrior' option equal to TRUE. After setting the reference genotype to the F1 (vgn x 399) using the 'relevel' function, we then extracted the results using the 'results' function and the contrast vector c(0,1,-.5, -.5) to contrast the full value of the F1 genotype with 1/2 (vgn + 399). Features with an FDR$<$0.1 were considered as "non-additive".

2.4.6 Supplementary figures

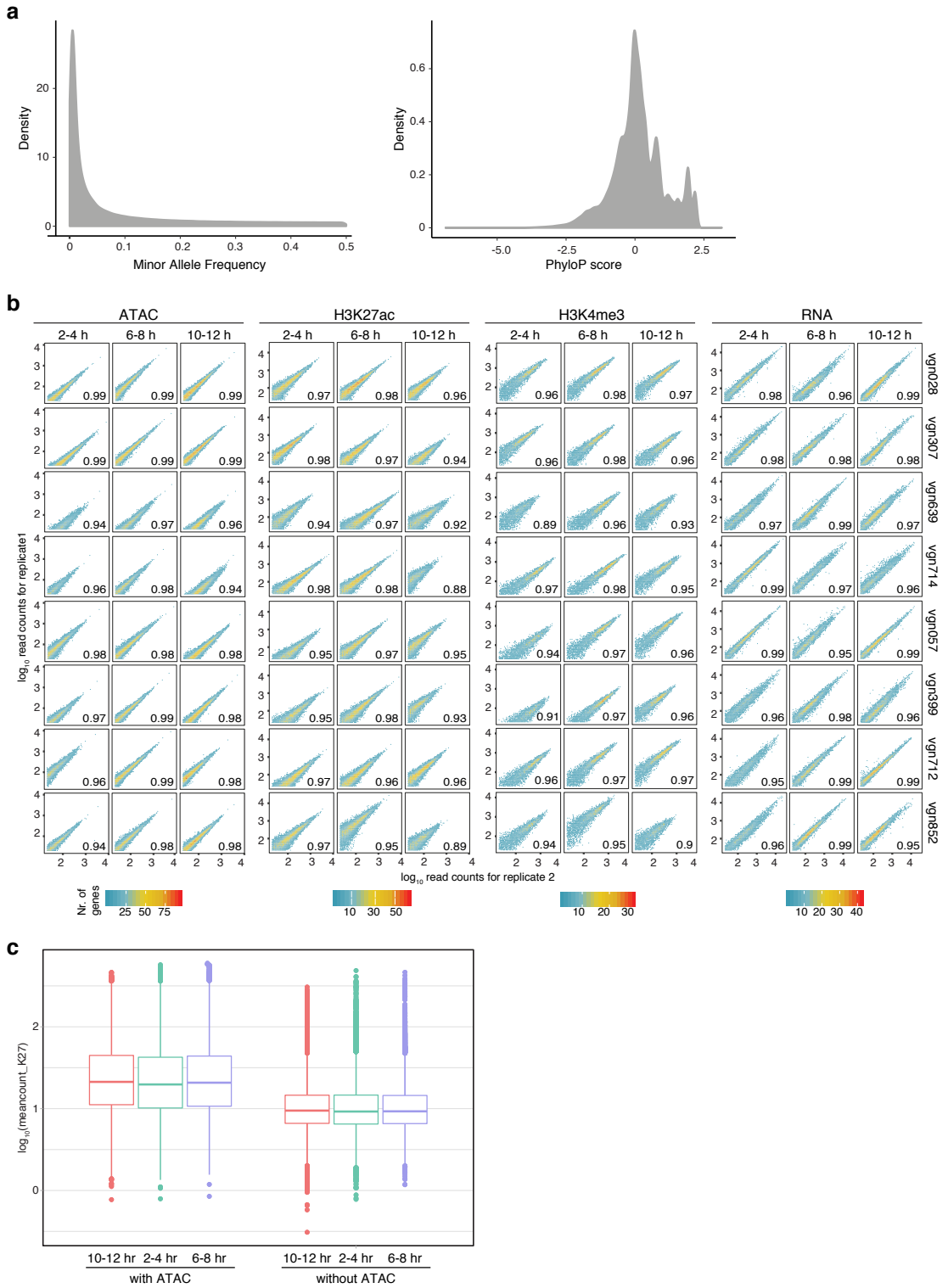

Figure 2.11: General properties of the data – distribution of SNPs and reproducibility. a. Smoothed histograms show the distribution of minor allele frequencies of SNPs in regulatory elements (left) and regulatory element phyloP scores (right). Regulatory elements are defined here as peaks of ATAC-Seq, H3K27ac, or H3K4me3. MAF are taken from the full 205 lines of release 2 of the Drosophila Genetic Reference Panel. b. Gene expression levels and read counts from accessible chromatin/ChIP-Seq show consistently high levels of correlation between replicates. Pearson correlation coefficients indicated. c. Coverage plots comparing distal peaks of H3K27ac that do (left) and do not (right) overlap annotated ATAC-Seq peaks show an overall lower read count for the second category.

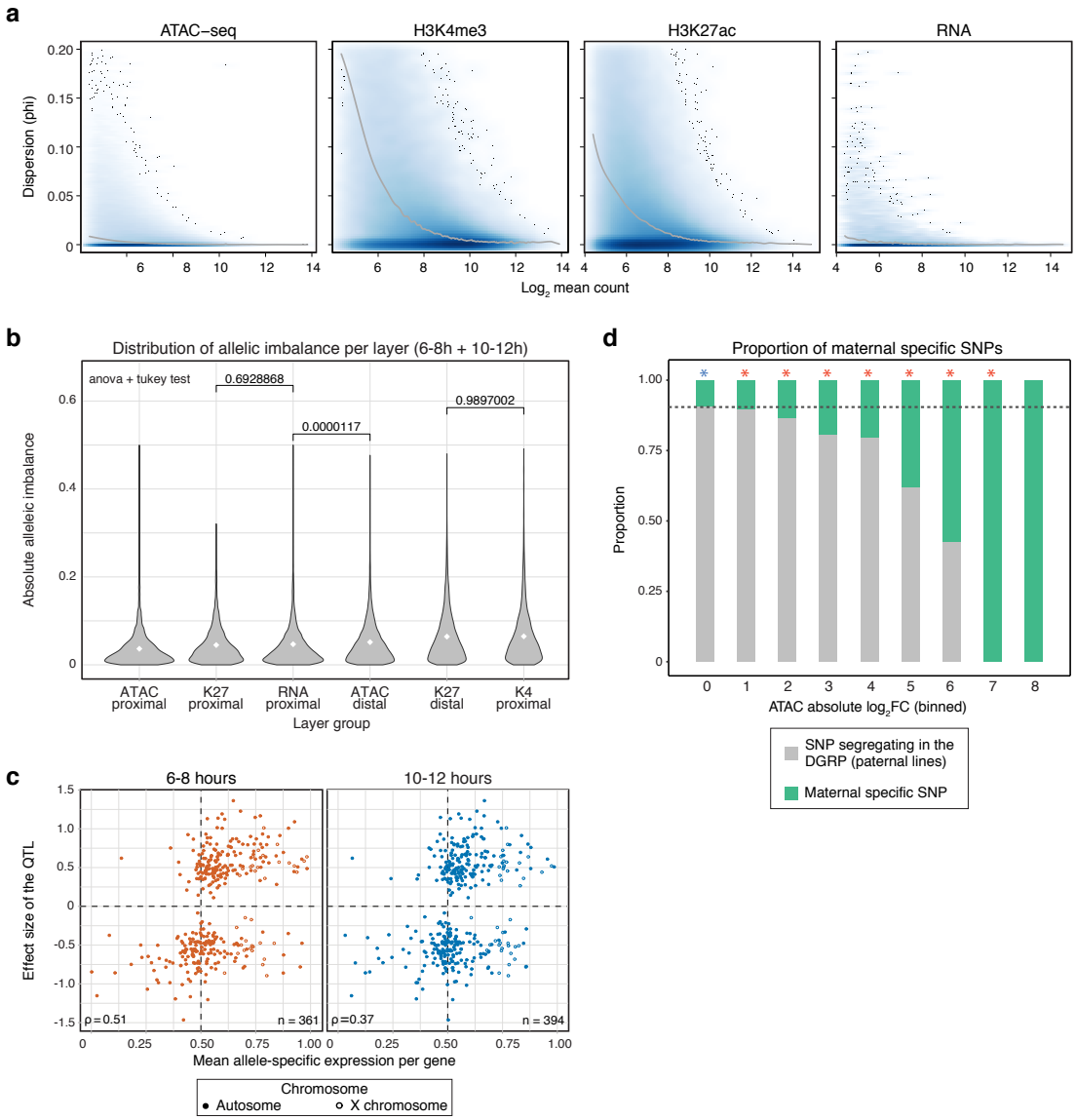

Figure 2.12: Proportion and dispersion of SNPs and allelic ratios. a. Dispersion is largely constant as a function of read count across all data types. Density plots show the beta-binomial dispersion parameter estimated across our pooled replicates for each feature per time point and per genotype (y-axis) plotted against the log2 averaged (arithmetic mean) total count across replicates (x-axis). b. Distribution of the absolute departure from allelic imbalance for each data type. For the same data type, distal features show a significantly higher amplitude of imbalance compared to proximal features (tukey test, p<0.0001, three highest p-value shown). c. Spearman correlations (ρ) between F1 allelic ratios in autosomes and the effect sizes of associated eQTL identified in Cannavo *et al.* [8], show consistence concordance across effect sizes. Allele-specific expression predicts the direction of eQTL in 69% of cases for autosomes. RNA-seq data from 2-4 hours are excluded from the analysis, as the presence of maternal transcripts may bias the allelic ratio measures toward the maternal side (right quadrants). d. The proportion of SNPs unique to the maternal line compared to any of the DGRP lines (green bars) is greater in highly imbalanced ATAC peaks at 6-8 and 10-12 hours, indicating a correlation between the presence of rare (potentially *de novo*) mutations and large effect sizes. Dotted line shows the average proportion across all the allelic imbalance (AI) values. Green and red asterisks indicate a statistically significant depletion and enrichment, respectively, of maternal specific SNPs for each bin of AI values.

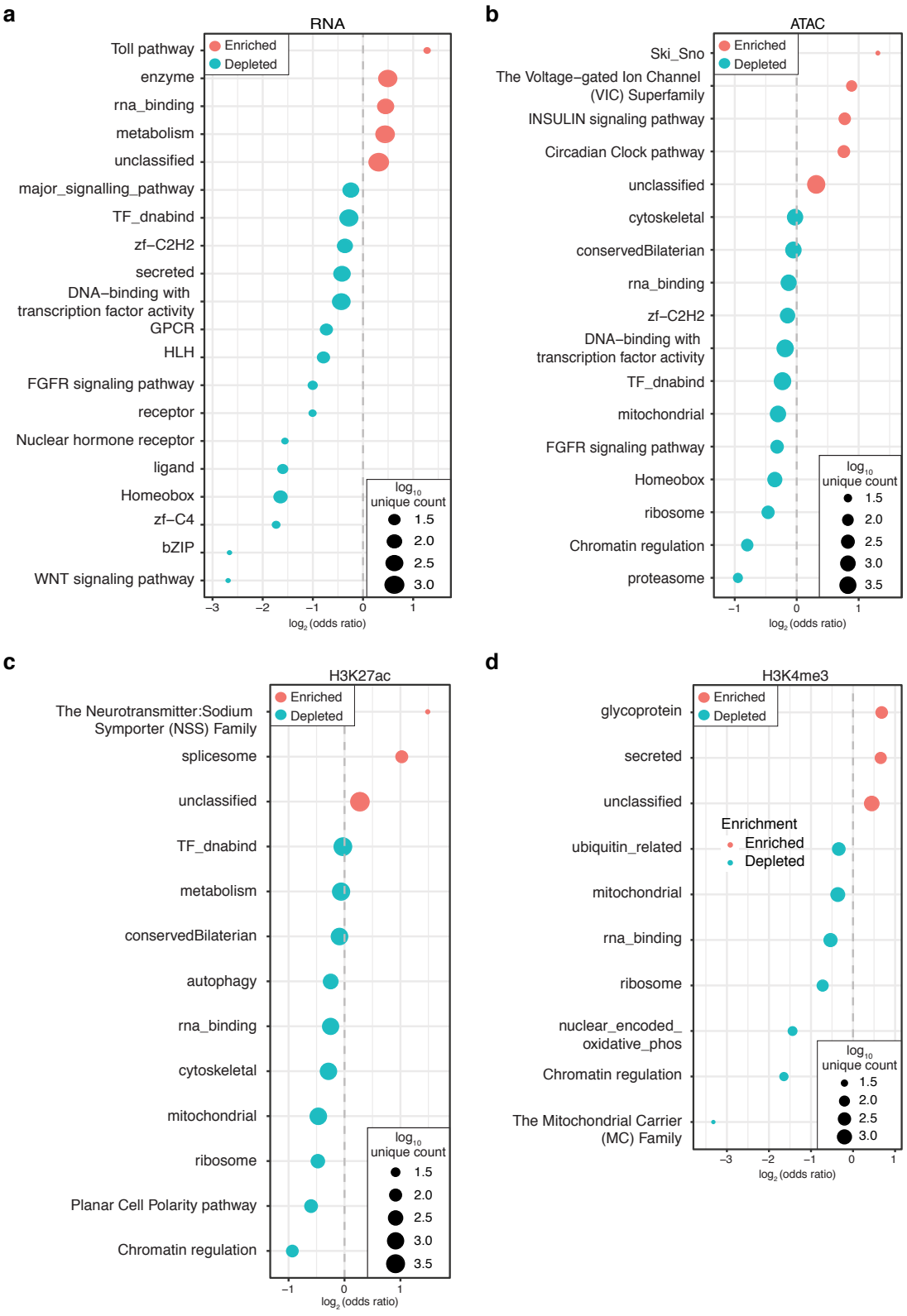

Figure 2.13: Allelic variation is consistently depleted for transcriptional regulation and enriched for the expression of metabolic genes. GO term enrichments for all GLAD categories with statistically significant ($p<0.01$, Fisher's-Exact Test) enrichment or depletion of significant allelic imbalance for (a) RNA, (b) ATAC-Seq or (c-d) histone modifications. Bubble size represents the number of genes/features per category. For regulatory elements, category assignments were made on the basis of closest gene annotations.

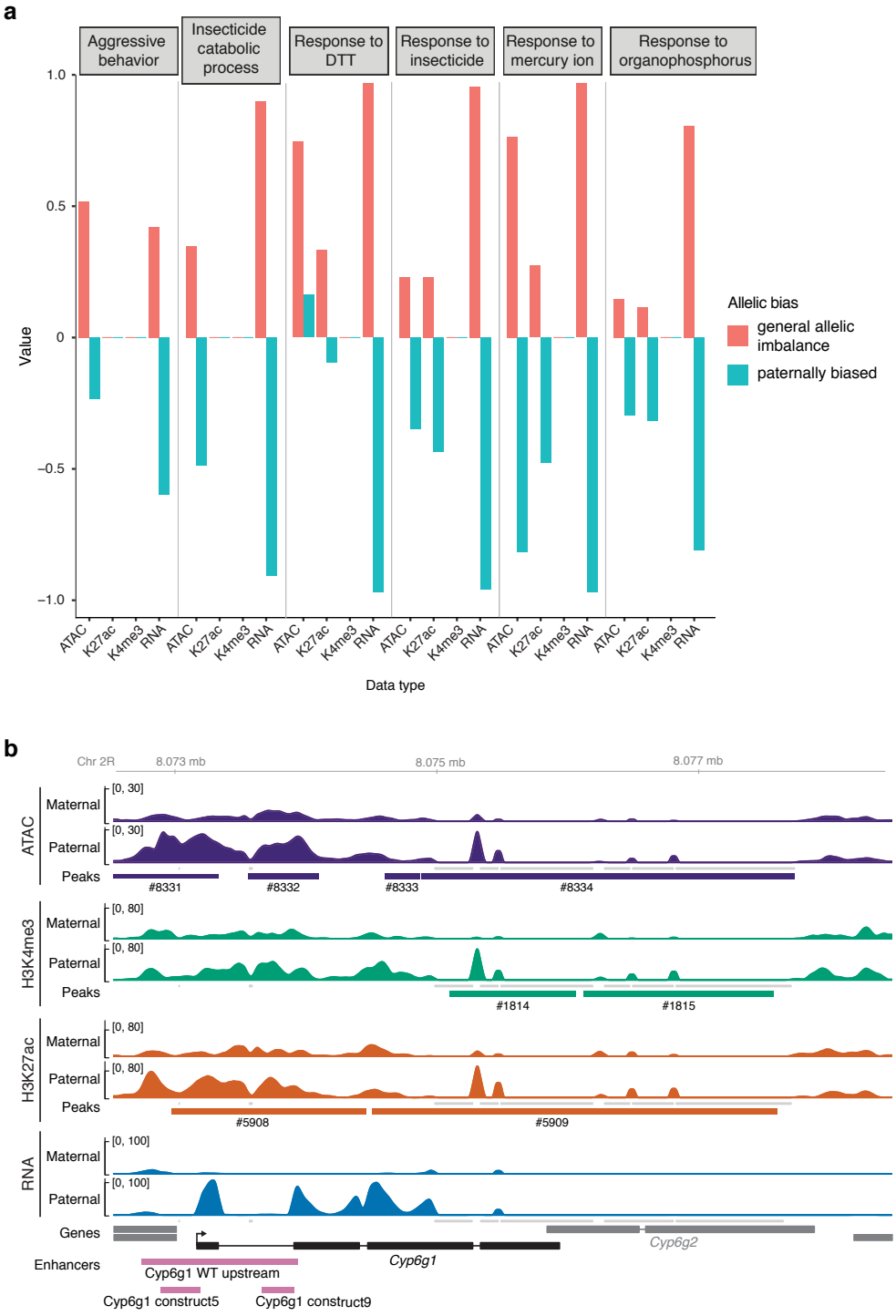

Figure 2.14: Regulatory changes on paternal haplotypes are enriched for genes related to pesticide resistance and environmental response. a. Categories of genes involved in environmental response and immunity show consistent biases involving the paternal lines. Correlation coefficient from point-biserial correlation analyses: red bars indicate correlation with the absolute value of allelic imbalance, blue bars with negative values indicate correlation with allelic imbalance with the paternal allele being more highly expressed. b. *Cyp6g1*, a DTT resistant gene, is upregulated in every regulatory layer and in gene expression in the paternal allele, while the maternal allele shows no evidence of expression (as recorded in FlyBase). Grey bars indicate locations of non-uniform mappability across lines.

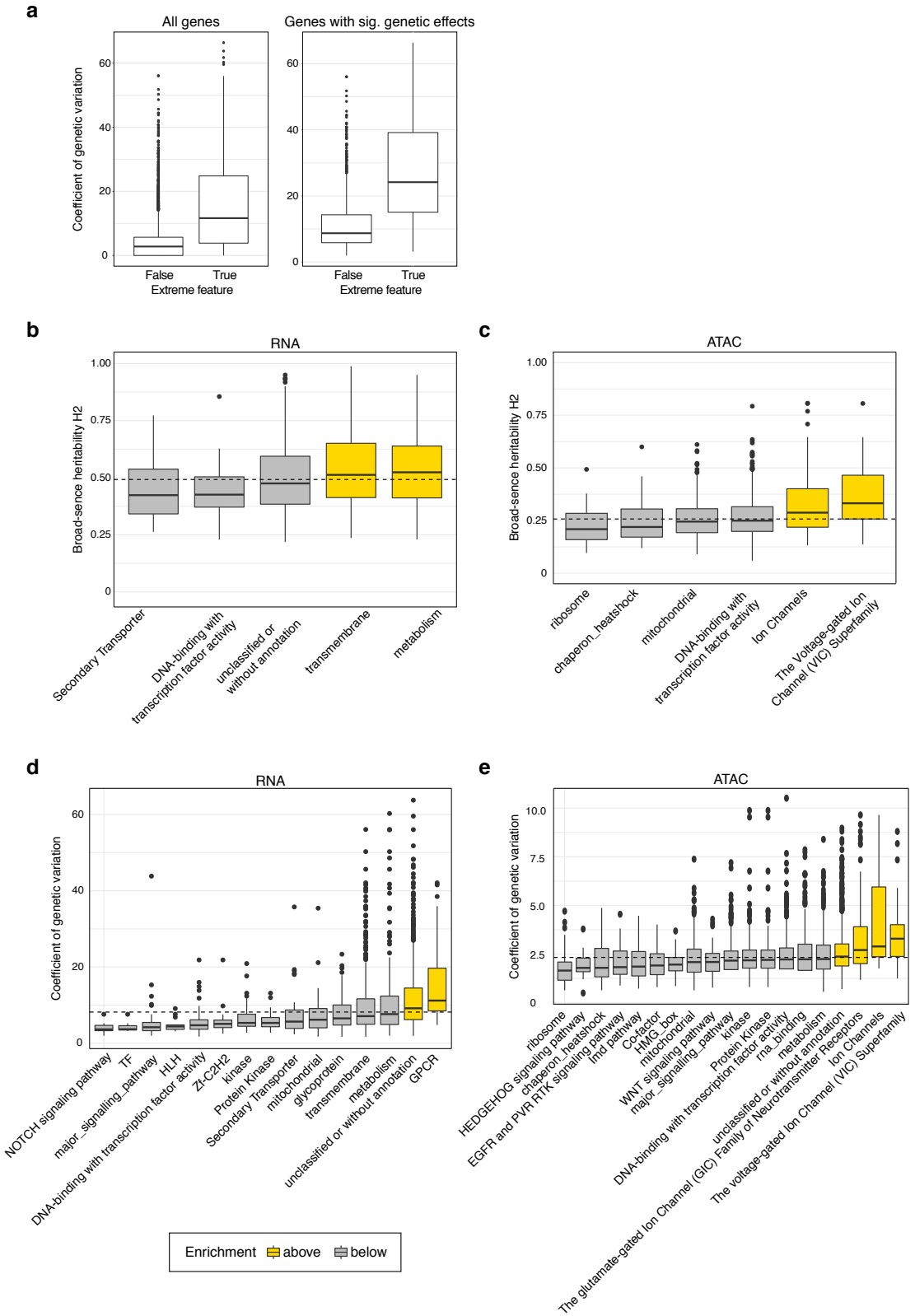

Figure 2.15: Transcriptional regulators show reduced Heritability and smaller genetic effect sizes. cf. legend on next page.

Figure 2.15: a. Comparing the coefficient of genetic variation for highly imbalanced genes vs. all others (left) or only those with significant line effects (right). b-e. Box plots show the distribution of broad sense heritability (b,c) or coefficient of genetic variation (d,e) for categories of genes (b, d) or associated ATAC-Seq peaks (c, e) with statistically significant ($p < 0.01$) enrichment or depletion of genetic variation in a rank-biserial analysis. Transcription factors and related categories show a consistent depletion of genetic variation in all contrasts.

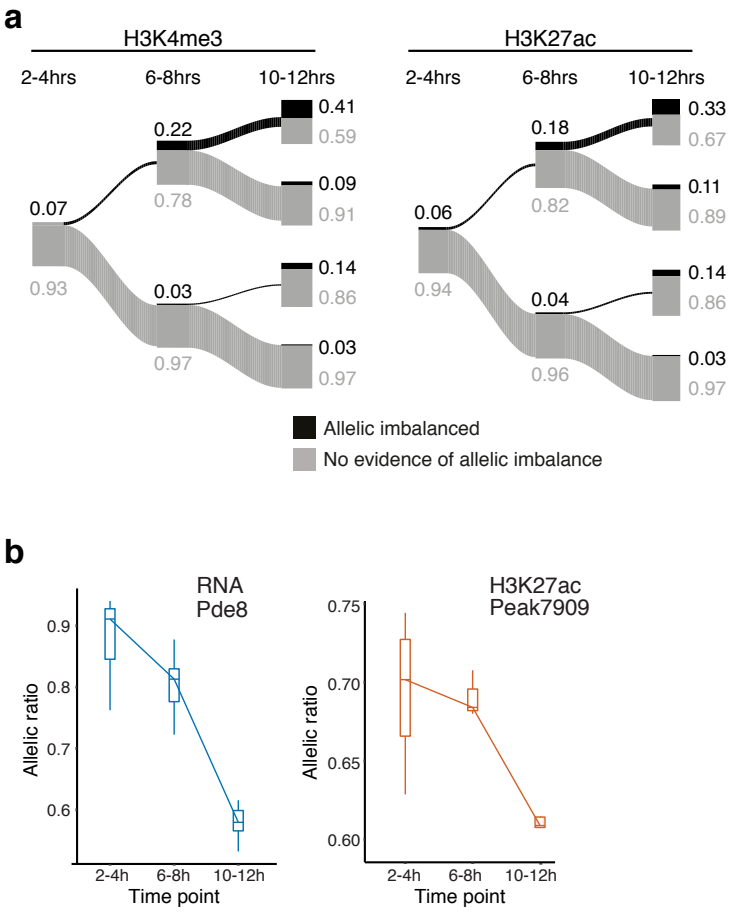

Figure 2.16: Time and genotype effect over development. a. Flow diagram showing dynamics of allelic imbalance (AI) in chromatin marks across developmental time. Proportions of AI and non-AI features are shown in black and grey, respectively, and represented by the line thickness. Exact proportions for each category are provided as numbers. b. Histograms showing allelic imbalance for the *pde8* gene. Allelic imbalance changes across time for gene expression level and associated H3K27ac enrichment.

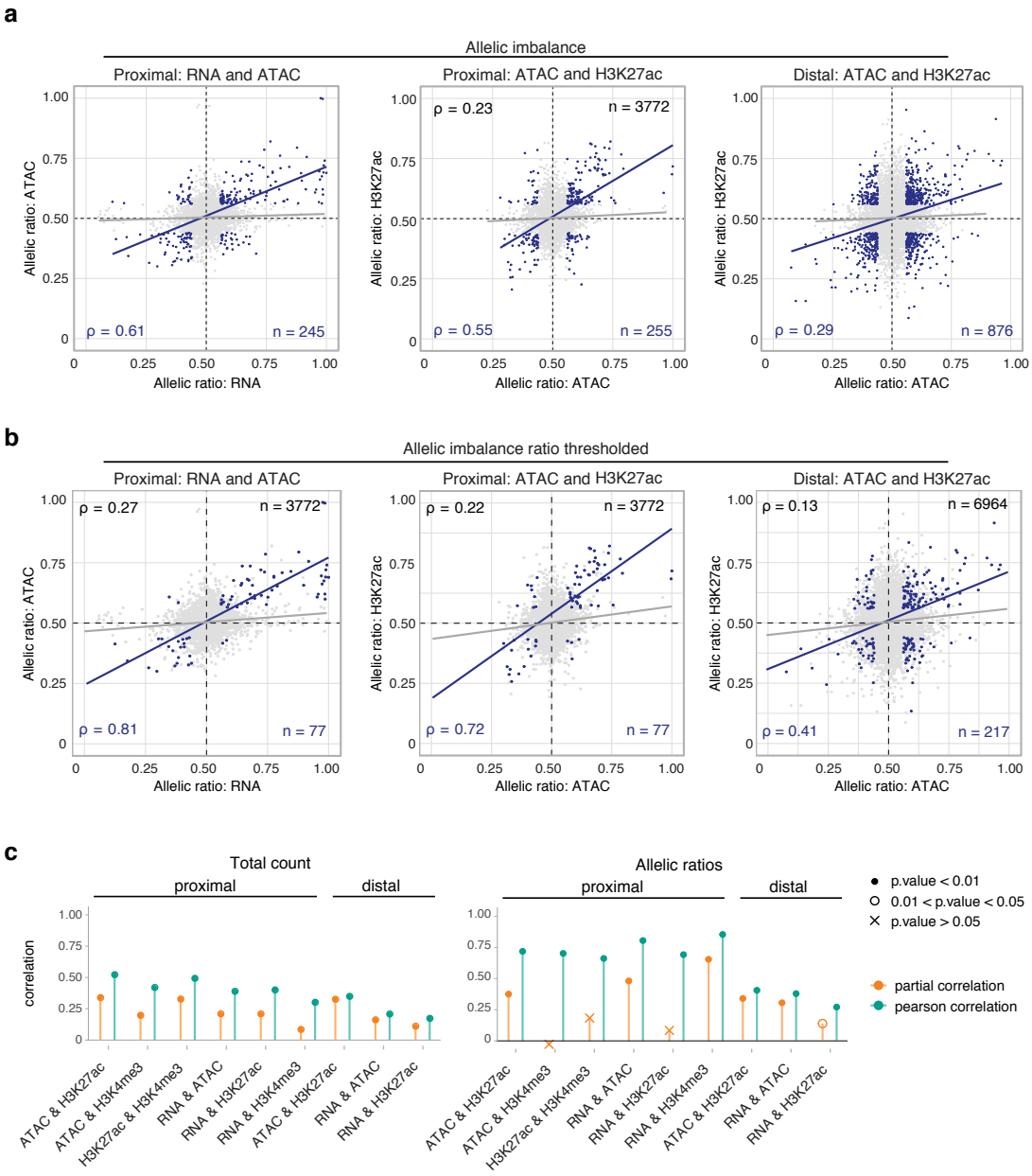

Figure 2.17: Partial correlation analysis reveals potentially causal relationships among regulatory layers. a, b. Overall Pearson correlations in allelic ratios (black values, grey points and regression lines) between regulatory layers. Correlations are generally modest with little correspondence in allelic ratios between overlapping non-coding features and associated genes. Correlations in allelic ratios for more imbalanced features (AI 0.5 +/- 0.06 for both regulatory layers being compared; blue values, points and regression line) are stronger. a. AI events present in the two data sets being compared. b. AI events present in all four data sets (blue points). c. Comparison between the partial (orange) and the Pearson (blue) correlation for total count (left) and allelic ratios (right). The decrease in partial correlation denotes a lack of direct relationship within the overall correlations.

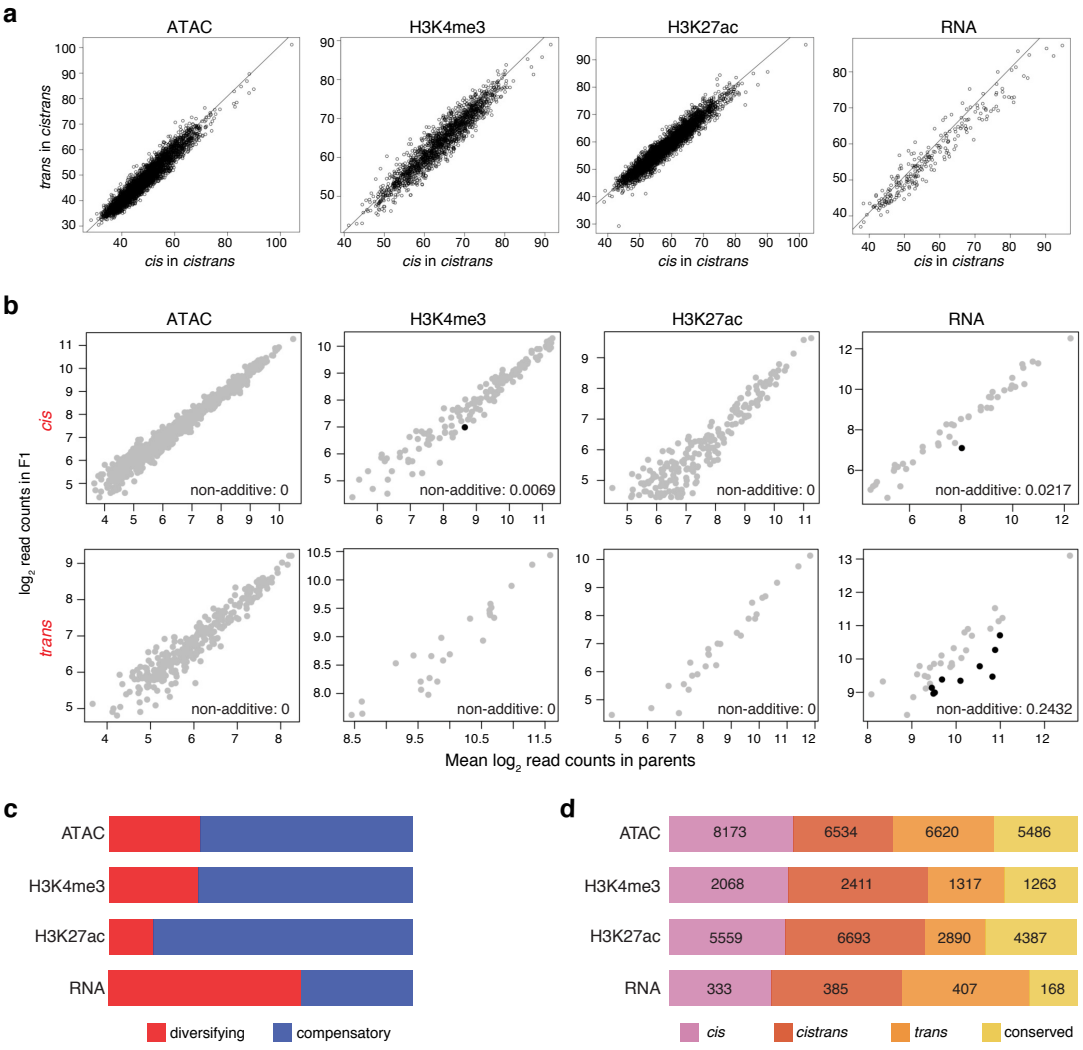

Figure 2.18: Differences in the frequency of *cis* and *trans* acting genetic variation among regulatory layers influences the heritability of regulatory phenotypes. a. For each regulatory layer, *cistrans* classified genes/features were assessed to evaluate the relative contributes of *cis* and *trans* to the *cis*/*trans* signal. Only RNA has evidence for an unequal contribution (more *trans* than *cis*). b. Scatterplots showing total read counts in the F1 lines vs. the mean of the two parents for genes/features classified as *cis* (top) or *trans* (bottom). Shown in black are genes/features with an F1 total read count significantly different (FDR <= 0.1) than the parental mean, indicating non-additive heritability. Only *trans*-influenced RNA has a frequency of non-additive heritability meaningfully distinct from 0. c. Using an alternative classification scheme [84], we assessed the frequency of diversifying and compensatory evolution for genes and features. For regulatory features, the concordance of *cis* and *trans* effects suggests predominantly compensatory evolution, while RNA shows a pattern more consistent with diversifying selection. d. Maximum likelihood *cis*-trans classification composition for each data type (genes and regulatory layers). Numbers and horizontal bars represent the size and relative proportions of each *cis*-*trans* class in each of the four regulatory layers. Genetic variation affects regulatory layers and gene expression with *cis*, *trans*, a mixture of *cis* and *trans* or *conserved* effects.

2.4.7 Supplementary methods

Personal genome construction

As a starting place for all personal genomes, we began with Drosophila reference assembly dm3 as downloaded from the UCSC genome browser (version 5 from FlyBase) along with reference annotations r5.57 from FlyBase. To generate reference genomes for our paternal lines, we downloaded variant calls for the full 205 lines of the DGRP (dgrp2.gnets.ncsu.edu) made against the dm3/v5 D. melanogaster reference genome.

For each paternal line from the DGRP, we used a custom script to insert into the reference genome SNPs and indels from the VCF file. Changes in coordinates were recorded in a liftOver chain file for subsequent steps. Heterozygous SNPs were replaced with the appropriate ambiguity code with missing data ('.') recorded as an N. Heterozygous indels were inserted as a string of N equal to the length of the longer haplotype. In the case of the Virginizer line, we made use of a VCF file generated for reference genome dm6/v6 [96] and converted the coordinates to dm3 using pslMap (genome.ucsc.edu, v5).

The same steps for reference generation was used for the DGRP paternal lines. For all parents, a genotype-specific set of annotations was created using liftOver in conjunction with custom software for translating the r5.57 reference GFF3 file into the coordinate frame of the custom parental genome.

Read mapping

Read were trimmed for adaptor sequences and sequencing quality using skewer (v0.2.2) and seqtk (v1.0) respectively, with default parameters. In order to avoid mappability bias, we used the parental genome mapping strategy (see mappability filter section, below) and mapped the reads on both personalised parental genomes [108].

Reads from ATAC-seq and ChIP-seq were mapped using BWA (v0.7.12) [98], reads from RNA-seq were mapped using STAR (v2.5.1b) [99] and FlyBase gene annotation version 5.57. Aligned reads were clipped when overlapping their read pairs using clipOverlap (v1.0.14). Using the appropriate chain files and pslMap (genome.ucsc.edu, v5), alignment coordinates were converted into the reference Drosophila melanogaster r5.57 genome coordinate space, also used in the DGRP project for variant calling.

Resulting alignments from both parental genome mappings were merged into a single alignment file, where reads aligned in both cases were reported only once (selecting the alignment with the highest mapping score).

Mappability filter

To ensure equal mappability across the two parental genomes for a given F1 line, we made use of two approaches. First, genomic DNA sequencing data from all the parental lines were mapped using STAR on their personalized genome and converted into the r5.57 reference using pslMap. Coverage data were produced using pslToBed from the UCSC genome browser utilities. For each of our F1 crosses, genomic regions showing a null coverage in either the mother or the father line were discarded from the analysis by trimming the portion of the aligned reads overlapping such regions.

Second, for each of the parental genotypes, we simulated transcriptomic and genomic reads spanning all the genome with equal coverage (one read starting at each base pair). The resulting reads were mapped on the corresponding parental genome and converted into the reference genome coordinates in the same manner as the RNA-seq, ATAC-seq and ChIP-seq experiments. For each of the F1 lines, regions showing a different coverage between the paternal and maternal synthetic reads mapping were not considered when calling allele-specific measures.

This step captured mappability issues caused by ambiguous bases and Ns introduced during the construction of the parental reference genomes. In order to compare total coverage measures across samples, a universal mappability filter encompassing all the line-specific filtered regions was applied to trim the reads before further analysis.

Quality control

We evaluated the quality of our sequencing data in three ways. First, we looked at pairwise correlations between replicates, observing a Spearman correlation of at least 0.95 in all cases (Fig. 2.11b). Second, we performed a principal component analysis at the level of total counts to look for evidence of issues for specific samples (e.g. failure to cluster with a replicate or clustering with the wrong time point). Third, in the case of RNA, we looked for correlations between our samples and the modENCODE time series of development (Fig. 2.19, below) [109].

Through these last two steps, we realized significant issues with the 6-8hr time point for the parental line 399 – while the replicate correlations is high, these samples appear closer to 10-12hr than they do 6-8hr. We thus removed these two samples from all analyses, thus reducing our *cis*/*trans* analyses for RNA to only the 10-12hr time point. No similar staging issues are apparent in the open chromatin or histone modification data.

Peak calling

For ChIP-seq experiments targeting H3K4me3 and H3K27ac marks, peak calling was performed for each sample on each parental genome using Macs2 (v2.1.1) [100] with the broad option and default parameters. After converting all peak calls to the dm3 coordinates, we merged peaks using the bedtools merge function to produce a single peak set used across all lines and all developmental time points. For ATAC-seq experiments, regions of chromatin accessibility were defined as the merge of peak summits called by Hotspot (v4.0.0) [101] with a score higher than 5 in at least one of the F1 samples after extending the summits by 200bp in both directions.

Total signal quantification

For each individual sample, total coverage signal was evaluated at the feature level (genes or peaks) using custom python scripts built around the pysam package. Each read mapping to at least one of the two parental genomes and not filtered by the mappability filter was assigned to its overlapping feature. Reads not overlapping a SNP were also included in this process, as this measure is not allele-specific.

To quantify changes in total read counts between time points, we imported the

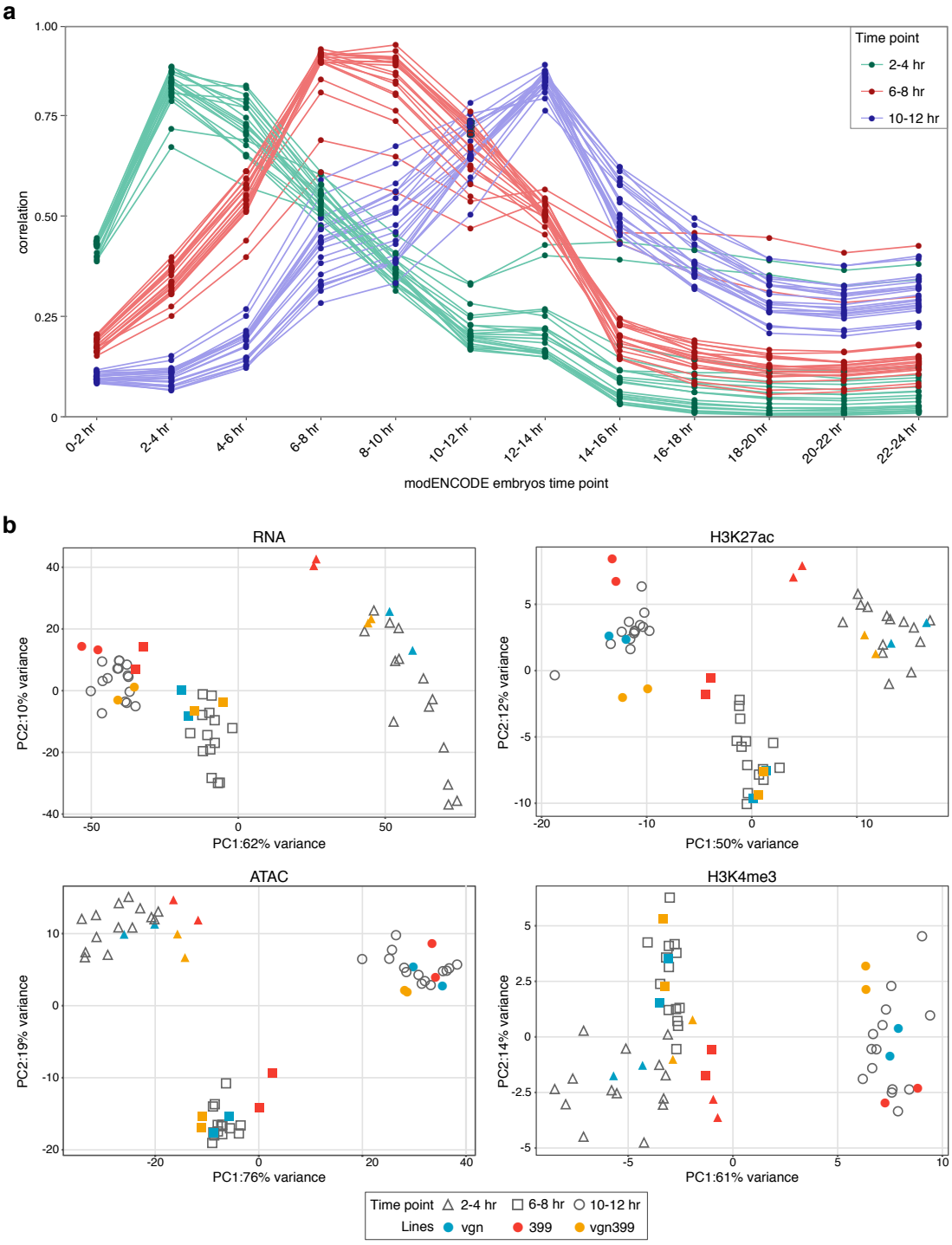

Figure 2.19: Total count data cluster with similar time points. a. Correlation of gene expression total count data with modENCODE time series. As expected, the highest correlation correspond to the comparison with equivalent time points. b. Principal component analysis of total count data for each data type. Parent-offspring trio is shown in color. For gene expression data (upper, left), the two replicates of parental line 399 at 6-8hr cluster together with the 10-12hr samples. We removed these samples from further analysis.

total count data into DESeq2 and fit a model consisting of line + time. For any gene with a significant (FDR<0.1) time effect, we subsequently looked for evidence of changes between neighboring time points using the contrasts option in DESeq2.

One concern in such an analysis is that some genes may have expression levels that are simply too low to provide statistical power. To avoid this issue, we presented results only from genes whose mean expression across all three time points was equal or greater to the read count threshold identified via DESeq2's implementation of independent filtering [110].

Upon visual inspection of the total count distribution across data types, we set the minimum of 20 reads per feature as the threshold for detecting expressed genes and non-coding peaks. Total counts were library scaled and TMM normalized using EdgeR [102]. Values are expressed in log10(Counts Per Million).

Allele-specific signal quantification

For each dataset, allele-specific counts were performed at the feature level, i.e. per genes for RNA-seq and per peak for ATAC-seq and ChIP-seq. Based on their genotype at the SNP location, reads overlapping a feature were assigned to the maternal or paternal haplotypes. Reads not overlapping a segregating SNP or reads with disagreeing assignment between SNPs were ignored in the measurement.

Cases of genotyping errors can potentially lead to incorrect allelic imbalance measure if a SNP is wrongly called as segregating in a given cross. In order to correct for these events, we performed a genomic DNA-seq experiment for each of the F1 lines and processed it in the same manner as the ATAC-seq. Using a two-sided binomial test with false discovery rate correction, we tested for each SNP whether the number of reads assigned to the maternal and the paternal haplotypes were significantly different from an expected 50:50 ratio in the autosomes.

In chrX, the expected ratio was empirically measured from the 1000 SNPs with the highest coverage in chrX. Only SNPs with a minimum coverage of 15 reads for autosomes and 10 reads for chrX were tested. SNPs considered as significantly imbalanced ($p<0.05$) for the genomic DNA data were formatted as missing data (N) for alignment and were ignored when performing the allele-specific measures. In order to evaluate the sex ratio of our pool of embryos, allele-specific counts of gDNA reads were also performed at the feature level.

Due to the presence of maternally deposited transcripts, a portion of the genes has an allelic imbalance biased toward the mother in the RNA-seq data. In order to detect them, we used RNA-seq data of unfertilised eggs from the same developmental time windows as the F1 samples. To identify genes with maternal deposition, we plotted the log10 read count of each gene as measured in freshly laid eggs (the first time window), using the bimodality of this distribution to set a threshold for "expressed".

The majority of these genes were excluded from subsequent analyses. However, as previously noted [111], we observed a population of these transcripts that decayed over time, becoming not detected by 6-8 hours. Formally, we quantified this population as those transcripts showing significant evidence of

decay (using DESeq2) between 2-4h and 10-12h (Fig. 2.20). As this population of transcripts shows a 50:50 autosomal ratio in the 6-8h and 10-12h datasets, we included them in our analyses.

In addition to maternal transcript removal, because we used a poly-A selection step in the construction of our RNA-Seq libraries, we removed from our analyses categories of genes and transcripts that largely or entirely lack polyadenylation signal (e.g. ncRNA, snRNA, rRNA).

Allele-specific changes across lines and developmental time

In the analysis of linear mixed-effects model, to infer the significance of time or strain dependent allele bias, we restrict the values that the parameters can take:

$$M_0 : \mu_f \neq 0, \delta_f^t \neq 0, \omega_f^s \neq 0 \; ; \; (\delta\omega)_f^t \neq 0$$

$$M_1 : \mu_f \neq 0, \delta_f^t = 0, \omega_f^s \neq 0 \; ; \; (\delta\omega)_f^t = 0$$

$$M_2 : \mu_f \neq 0, \delta_f^t \neq 0, \omega_f^s = 0 \; ; \; (\delta\omega)_f^t = 0$$

where M_0 is the full model that controls for effects due to time, genotype as well as the time by genotype effect. M_1 is a model where we assume no allelic balance between time points after controlling for strain effects. Conversely, M_2 accounts for allelic imbalance changes by time points while controlling for strain effects. Each model is fitted to the data in turn by maximizing the likelihood using the R library 'lme4' **[RM55]**.

In order to identify regions that show significantly different allelic ratio due to time, we used a likelihood ratio test based on the chi-square distribution with two degrees of freedom to compare between M_0 and M_1. Similarly, we compare between M_0 and M_2 to assess for differences due to genotype. We adjusted p-values following multiple testing using FDR correction [104] and considered q-values below 0.1 as denoting a significant allelic imbalance due to a time or genotype effect.

Gene category enrichment of allelic imbalance

To better understand the biological functions affected by *cis*-regulatory variation, we looked for the enrichment/depletion of allelic imbalance in functional categories using a Fisher's exact test. We focus here on the gene-centric GLAD categories [83], which are broadly representative of the trends observed using other ontologies. In these analyses, chromatin features (ATAC and histone mark peaks) were assigned to the closest gene, though similar results were obtained if we limit our analysis only to promoter-proximal elements (<500bp from the assigned TSS).

With this analysis, we observe an enrichment of allelic imbalance in a set of genes and associated non-coding features that could not be assigned to any known GLAD category. Such set collectively represents fast-evolving and Drosophila-specific genes, referred as the 'unclassified' category [112, 113] (Fig. 2.13). Enrichment analyses themselves were carried out using Fisher's exact test (for binarized data) or a point-biserial correlation, effectively a Pearson's correlation coefficient for circumstances in which one variable is continuous and the other categorical (to ensure robust results, point-biserial correlations were also compared to non-parametric rank-biserial correlation analyses). In all enrichment analyses, features/genes from both the X chromosome and autosomes

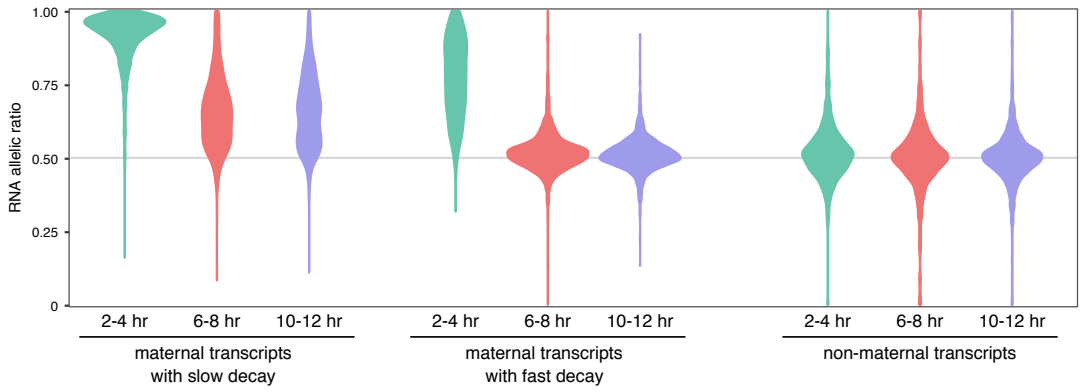

Figure 2.20: Maternally deposited transcripts show different rates of decay. Distribution of allelic ratios for gene expression data for line vgn28 at each time point. The genes are separated into three categories, based on their expression in freshly laid eggs. Left: genes detected in unfertilized eggs (maternal transcripts) showing no evidence of decay at 10-12h. Middle: genes detected in the eggs and showing significant decay between 2-4hr and 10-12h. Right: genes not detected in eggs, only zygotically expressed. Zygotic genes and maternally deposited genes showing evidence of fast decay are included in the analysis.

were included. Similar analyses were performed to understand enrichment of H2 and the coefficient of genetic variation (Fig. 2.15) and evolutionary rate data.

Allele-specific changes across regulatory layers

Each set of non-coding features were split into TSS-distal and TSS-proximal subsets. Features are considered as part of the TSS-proximal set if their nearest TSS is not further than 500bp away or if they overlap a region called as a H3K4me3 peak. For each subset, we defined regions of overlap between the regulatory layers as the overlap portion of two or more non-coding features. In the case of the overlap of three features, at least one base pair must be shared with all the layers to create an overlap.

For the proximal subset, genes features are assigned to a given overlap region if the distance between the overlap boundaries and the TSS is smaller or equal to 500bp. For the distal subset, overlaps are associated with the gene having the closest TSS. To avoid mis-assignment of TSS to proximal *cis*-regulatory overlaps, we excluded TSS positioned in the 600bp upstream region of other TSS.

As we noticed a clear drop in the correlation between the regulatory layers when the distance between the nearest TSS and the non-coding features exceed 1500bp (Fig. 2.21a), we restricted the TSS-distal set to overlaps with a maximum distance to TSS of 1500bp.

Partial correlation analysis was performed using the R package 'GeneNet' [106] for both allelic ratios and total count data and for TSS-proximal and TSS-distal sets (excluding chromosome X). We used features with no missing data in any of the regulatory layers (Fig. 2.17a). For allelic ratio data, we observed a distinct increase in Pearson correlations between layers as the AI fold change increased, suggesting a threshold below which allelic imbalance was effectively "noise".

To establish a filtering threshold for "noisy" allelic imbalance, we separated

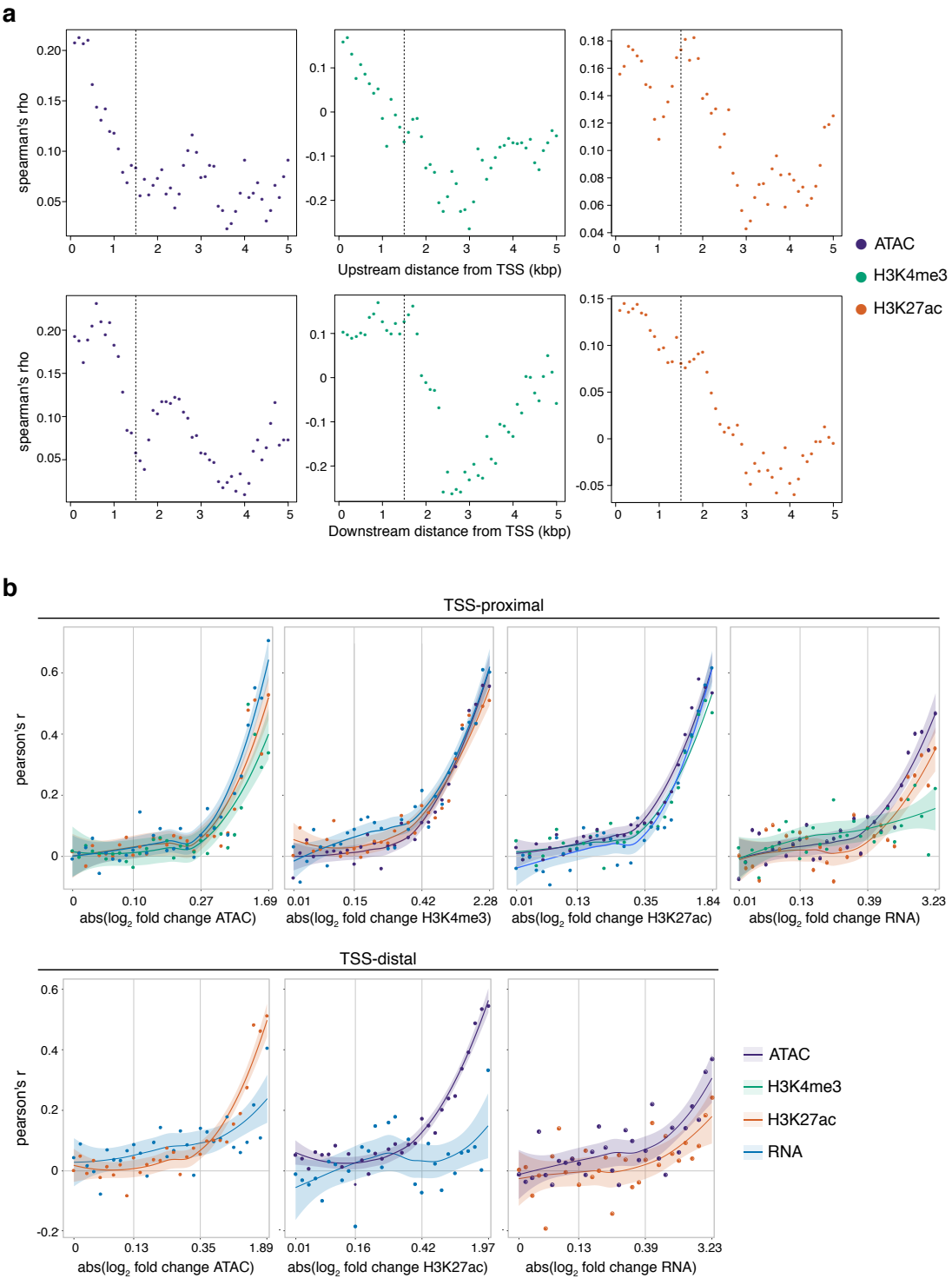

Figure 2.21: AI and TSS distance thresholding increases correlation between regulatory layers. cf. legend on next page.

Figure 2.21: a. Spearman correlation of allelic ratios between genes expression and the other regulatory layers, as a function of their distance from TSS. Correlations are shown for non-coding features located from 0 to 5kb upstream (upper row) or 0 to 5kb downstream (lower row) of the TSS. As the correlation values show a clear drop for distances further than 1.5kb, we removed features greater than 1.5kb from TSS (upstream and downstream) from the TSS-distal overlap set. b. Pearson correlation of allelic ratios between regulatory layers for TSS-proximal (upper row) and TSS-distal (lower row) features at 10-12hr. Correlations are binned into 30 quantiles based on the absolute log2 fold change of allelic ratio. Values on x-axis show the mean log2 fold change in the 1st, 10th, 20th and 30th quantiles from left to right. Shaded regions indicate the 90% confidence intervals. In most cases, we see an inflexion point around the 18th quantile, which was used to set the AI threshold of 0.5 +/- 0.06 in further analysis.

each dataset into 30 bins of total allelic imbalance and plotted the average correlation in allelic fold change between datasets for each bin (Fig. 2.21b). In each case, there is a general inflection point in the correlation at an allelic ratio of 0.5 +/- 0.06. We filtered loci that fell below this threshold to improve the covariance signal of AI datasets for partial correlation analysis (Fig. 2.8a, Fig. 2.17b). For both TSS-proximal and TSS-distal analysis, the partial correlation results are largely consistent between time points. Time points were thus pooled for this analysis. Because the different ratios for autosomes and X chromosomes would lead to an artificially inflated correlation, X chromosome genes/features were removed for this analysis.

Directional dependence modeling was done in a regression framework using copulas to describe the bivariate distribution between our pairwise datasets. A copula is a multivariate distribution where the marginal distributions are uniform [**RM58**, 114]. Any multivariate function, $F(x, y)$, can be represented in a copula as a function of its marginals, $C(F_X(x), F_Y(y))$. Hence, given two random variables X and Y, the copula $C(u, v)$ reflects this dependency, and $U = F_X(x), V = F_Y(y)$ are the marginal variables with uniform distributions.

Given an asymmetric copula, it is possible to infer directional dependence based on the proportion of total variance of V that can be explained by the copula regression $r_{(V|U)}(u)$ compared to the proportion of total variance of U that can be explained by the copula regression $r_{(U|V)}(v)$ [107, 115]. We use the method of Lee and Kim [107] to infer the flow of information for pairwise relationships that showed a significant relationship in partial correlation analyses. Analyses were performed for allele-specific and total counts and at different developmental time points (excluding 2-4hr and sex chromosomes). Chromosome X data was kept for both analyses and it removal did not change the direction of findings but did increase effect size.

For the locus-specific test of gene expression with numbers of open chromatin regions, ATAC-seq peaks were associated to genes at +/-1500bp from the TSS and the relationship between the regulatory datasets was tested on a locus-specific basis. We note that the results obtained from this analysis showing less imbalance in expression with more local peaks are unlikely to be due to differences in statistical power, as genes linked to shadow enhancers and multiple peaks are typically more highly expressed (wilcox test, p<1.5e-69). Hence, there is greater power to detect allelic imbalance.

Cis-trans analysis

For one F1 line (vgn x 399), we use maximum likelihood estimation (MLE) to compare parental and offspring ratios simultaneously to determine whether gene expression, chromatin accessibility, H3K4me and H3K27ac enrichments are influenced by *cis*-, *trans*, *conserved* or both *cis*- and *trans*- acting by modeling read counts in parents using negative binomial distributions and the F1 alleles using beta-binomial distribution. We then find the most likely model for each gene.

For each gene, F0 counts from each strain can be modeled as a negative binomial marginal distribution, while F1 counts were modeled using a beta-binomial distribution where the parameters of the beta distribution modeled the proportional contribution from each allele. For each data type, there were 2 replicates (i) for each F0 strain and 2 replicates (j) for F1 samples. F0 counts for each strain (x_i,and y_i) were assumed to follow negative binomial distributions while F1 counts (n_j), were modeled on an allele-specific basis (z_j) using a beta-binomial distribution:

$$x_i \sim Po(\mu_i), y_i \sim Po(v_i), z_j \sim Bi(n_j, p_j)$$

$$\mu_i \sim Ga(r, \frac{p_\mu}{1-p_\mu}), v_i$$

$$\mu_i \sim Ga(r, \frac{p_v}{1-p_v}), p_j$$

$$\mu_i \sim Be(\alpha, \beta)$$

where x_i is formally defined as the count of the variant in the ith vgn F0 line, y_i is the binding intensity of the variant in the ith dgrp399 F0 line, n_j is the number of reads mapping across both allelic variants in the jth F1 hybrid and z_j is the number of reads mapping to the vgn allele in the jth F1 hybrid.

We estimate the dispersion parameter r for F0 samples using the 'estimateDispersions' function within DESeq2 with local regression fit. r was used as the reciprocal of the fitted dispersion value from DESeq2.

We constrained parameter estimation for each distribution based on four different regulatory scenarios and derived maximum likelihood values for each hypothetical case on a site-by-site basis. The four models are described below:

$$Conserved : p_\mu = p_v \; and \; \alpha = \beta$$

$$Cis : p_\mu \neq p_v \; and \; \frac{\alpha}{\alpha+\beta} = \frac{\frac{p_\mu}{1-p_\mu}}{\frac{p_\mu}{1-p_\mu} + \frac{p_v}{1-p_v}}$$

$$Trans : p_\mu \neq p_v \; and \; \alpha = \beta$$

$$Cistrans : p_\mu \neq p_v \; and \; \alpha \neq \beta$$

In our presentation of the proportions of features assigned to each category (Fig. 2.18d), we presented the maximum likelihood assignment. In subsequent analyses, however, we limited our analyses to features that showed a BIC difference $>= 2$. For the *cis* and *trans* assignments, we focused only on autosomal features. This was in part due to the complications of calculating *cis*/*trans* components for two different sets of expected ratios, but also because the difference in expected ratio between the X chromosome and the autosomes can influence the power to detect allelic imbalance and, thus, influencing the assignments of *cis* vs. *trans*, with resulting complications for downstream analyses (e.g. categorical enrichment)

A challenge in assessing *cis* and *trans* proportions during development is that differences in staging between samples

can induce differences in read counts that will not be reflected in allelic ratios. If these differences stem from genetically based differences in developmental rates, then the classification would reflect genuine *trans* differences. Environmental variation or differences in sample handling, however, can also lead to developmental shifts.

To evaluate this possibility, we looked first to see if genes and features classified for *trans* frequently showed evidence of an increase in log2 fold-changes between time points. For all regulatory layers, we observed a significant increase in log2 fold-changes between time points for genes and features classified as *trans* ($p<2.2e-16$). However, we see no evidence for a coordinated shift in the parental ratios used to calculate *trans* relative to log2 fold-changes between time points - genes and feature counts that increase during development are equally likely to show higher expression in either the maternal line (vgn) or paternal line (DGRP-399).

We thus conclude that while a portion of our observed *trans* effects may result from differences in developmental timing, they are likely genetic in origin, as global shifts in developmental staging (e.g. from handling errors or differences in collection temperature) would induce clear correlations between log2 fold-change over development and expression bias towards the more developmentally advanced parent.

To avoid potential interaction effects with 'time', we fit a separate model for each time point (all three time points in the case of the chromatin data, and excluding 2-4h in the case of RNA). For each gene/feature, we used the 'lmer' function from lme4 to estimate a random effect for line after applying the 'vst' function in DESeq2 to bring the trait values more closely in line with normality. To evaluate the significance of the resulting fit, the model was compared to a null model consisting only of an intercept using the anova function. FDR values were calculated from the resulting vector of p-values using the 'qvalue' function in R. Estimated line variances and residual variances were extracted from the model using the 'VarCorr' function.

Line variances were treated as proportional to broad-sense heritability (H^2). We calculated the coefficient of genetic variation (CVg) by scaling our estimated 'between-line variances' by the variance stabilized mean read count of each feature such that genetic variation is presented as a percentage deviation from the average of the population [6, 116]. We used the formula below:

$$CV_g = 100 \times \frac{\sqrt{line\ variance}}{trait\ mean}$$

In our analysis, H^2 and coefficient of genetic variation were considered meaningful as long as the line-model was significant with an FDR<0.1. The result of the above process generated one value per time point for each feature. An alternative is to fit a similar model to the above for all times (excluding 2-4h in the case of RNA) and including a term for 'time'. The resulting distributions for H^2 and coefficient of genetic variation were qualitatively similar. The enrichment calculations were carried out as described above.

2.4.8 References

1. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328,** 232–5 (2010).
2. Spivakov, M. *et al.* Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biol* **13,** R49 (2012).
3. Behera, V. *et al.* Exploiting genetic variation to uncover rules of transcription factor binding and chromatin accessibility. *Nat Commun* **9,** 782 (2018).
4. Waszak, S. M. *et al.* Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell* **162,** 1039–50 (2015).
5. Schor, I. E. *et al.* Promoter shape varies across populations and affects promoter evolution and expression noise. *Nat Genet* **49,** 550–558 (2017).
6. Garfield, D. A. *et al.* The impact of gene expression variation on the robustness and evolvability of a developmental gene regulatory network. *PLoS Biol* **11,** e1001696 (2013).
7. Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347,** 664–7 (2015).
8. Cannavo, E. *et al.* Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature* **541,** 402–406 (2017).
9. Epstein, D. J. Cis-regulatory mutations in human disease. *Brief Funct Genomic Proteomic* **8,** 310–6 (2009).
10. Lowe W. L., J. & Reddy, T. E. Genomic approaches for understanding the genetics of complex disease. *Genome Res* **25,** 1432–41 (2015).
11. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13,** 59–69 (2011).
12. Ahituv, N. *et al.* Deletion of ultraconserved elements yields viable mice. *PLoS Biol* **5,** e234 (2007).
13. Borok, M. J., Tran, D. A., Ho, M. C. & Drewell, R. A. Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila. *Development* **137,** 5–13 (2010).
14. Hong, J. W., Hendrix, D. A. & Levine, M. S. Shadow enhancers as a source of evolutionary novelty. *Science* **321,** 1314 (2008).
15. Frankel, N. *et al.* Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466,** 490–3 (2010).
16. Cannavo, E. *et al.* Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Curr Biol* **26,** 38–51 (2016).
17. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* **10,** 3583 (2019).
18. Junion, G. *et al.* A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* **148,** 473–86 (2012).
19. Uhl, J. D., Zandvakili, A. & Gebelein, B. A Hox Transcription Factor Collective Binds a Highly Conserved Distal-less cis-Regulatory Module to Generate Robust Transcriptional Outcomes. *PLoS Genet* **12,** e1005981 (2016).

20. Doitsidou, M. *et al.* A combinatorial regulatory signature controls terminal differentiation of the dopaminergic nervous system in C. elegans. *Genes Dev* **27,** 1391–405 (2013).

21. Khoueiry, P. *et al.* Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. *Elife* **6** (2017).

22. Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M. & Snyder, M. Genetic analysis of variation in transcription factor binding in yeast. *Nature* **464,** 1187–91 (2010).

23. Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res* **22,** 860–9 (2012).

24. Lu, R. & Rogan, P. K. Transcription factor binding site clusters identify target genes with similar tissue-wide expression and buffer against mutations. *F1000Res* **7,** 1933 (2018).

25. Brown, C. D., Johnson, D. S. & Sidow, A. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* **317,** 1557–60 (2007).

26. Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E. E. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462,** 65–70 (2009).

27. Bullaughey, K. Changes in selective effects over time facilitate turnover of enhancer sequences. *Genetics* **187,** 567–82 (2011).

28. Powell, J. E. *et al.* Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. *PLoS Genet* **9,** e1003502 (2013).

29. Albert, F. W., Bloom, J. S., Siegel, J., Day, L. & Kruglyak, L. Genetics of trans-regulatory variation in gene expression. *Elife* **7** (2018).

30. He, B. Z., Holloway, A. K., Maerkl, S. J. & Kreitman, M. Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. *PLoS Genet* **7,** e1002053 (2011).

31. Garfield, D., Haygood, R., Nielsen, W. J. & Wray, G. A. Population genetics of cis-regulatory sequences that operate during embryonic development in the sea urchin Strongylocentrotus purpuratus. *Evol Dev* **14,** 152–67 (2012).

32. Yang, S. *et al.* Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals. *BMC Genomics* **15,** 13 (2014).

33. Lloyd-Jones, L. R. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *Am J Hum Genet* **100,** 371 (2017).

34. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45,** 1238–1243 (2013).

35. Mackay, T. F. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* **482,** 173–8 (2012).

36. Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits* xvi, 980 p. (Sinauer, Sunderland, Mass., 1998).

37. Levis, R. & Rubin, G. M. The unstable wDZL mutation of Drosophila is caused by a 13 kilobase insertion that is imprecisely excised in phenotypic revertants. *Cell* **30,** 543–50 (1982).

38. Pimpinelli, S. *et al.* Transposable elements are stable structural components of Drosophila melanogaster heterochromatin. *Proc Natl Acad Sci U S A* **92,** 3804–8 (1995).

39. Sgrò Carla M.; Partridge, L. Evolutionary Responses of the Life History of Wild-Caught Drosophila melanogaster to Two Standard Methods of Laboratory Culture. *The American Naturalist* **156,** 13 (2000).

40. Hoffmann, A. A., Hallas, R., Sinclair, C. & Partridge, L. Rapid loss of stress resistance in Drosophila melanogaster under adaptation to laboratory culture. *Evolution* **55,** 436–8 (2001).

41. Orozco-terWengel, P. *et al.* Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Mol Ecol* **21,** 4931–41 (2012).

42. Kvon, E. Z. *et al.* Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512,** 91–5 (2014).

43. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10,** 1213–8 (2013).

44. Lara-Astiaso, D. *et al.* Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345,** 943–9 (2014).

45. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. *Nature* **430,** 85–8 (2004).

46. Core, L. J. *et al.* Defining the status of RNA polymerase at promoters. *Cell Rep* **2,** 1025–35 (2012).

47. Mikhaylichenko, O. *et al.* The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes Dev* **32,** 42–57 (2018).

48. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23,** 800–11 (2013).

49. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **24,** 1595–602 (2014).

50. Lucchesi, J. C. & Kuroda, M. I. Dosage compensation in Drosophila. *Cold Spring Harb Perspect Biol* **7** (2015).

51. Georgiev, P., Chlamydas, S. & Akhtar, A. Drosophila dosage compensation: males are from Mars, females are from Venus. *Fly (Austin)* **5,** 147–54 (2011).

52. Conrad, T., Cavalli, F. M., Vaquerizas, J. M., Luscombe, N. M. & Akhtar, A. Drosophila dosage compensation involves enhanced Pol II recruitment to male X-linked promoters. *Science* **337,** 742–6 (2012).

53. Urban, J. *et al.* Enhanced chromatin accessibility of the dosage compensated Drosophila male X-chromosome requires the CLAMP zinc finger protein. *PLoS One* **12,** e0186855 (2017).

54. Pal, K. *et al.* Global chromatin conformation differences in the Drosophila dosage compensated chromosome X. *Nat Commun* **10,** 5355 (2019).

55. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160,** 554–66 (2015).

56. Mi, H. *et al.* Assessment of genome-wide protein function classification for Drosophila melanogaster. *Genome Res* **13,** 2118–28 (2003).

57. Turner, L. M., Chuong, E. B. & Hoekstra, H. E. Comparative analysis of testis protein evolution in rodents. *Genetics* **179,** 2075–89 (2008).

58. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

59. Daborn, P., Boundy, S., Yen, J., Pittendrigh, B. & ffrench-Constant, R. DDT resistance in Drosophila correlates with Cyp6g1 over-expression and confers cross-resistance to the neonicotinoid imidacloprid. *Mol Genet Genomics* **266,** 556–63 (2001).

60. Battlay, P., Schmidt, J. M., Fournier-Level, A. & Robin, C. Genomic and Transcriptomic Associations Identify a New Insecticide Resistance Phenotype for the Selective Sweep at the Cyp6g1 Locus of Drosophila melanogaster. *G3 (Bethesda)* **6,** 2573–81 (2016).

61. Berney, D., Petcher, T. J., Schmutz, J., Weber, H. P. & White, T. G. Conformations and biological properties of apomorphine and its phenanthro(10,1-b,c)azepine homologue. *Experientia* **31,** 1327–8 (1975).

62. Cusanovich, D. A. *et al.* The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555,** 538–542 (2018).

63. Paaby, A. B. & Gibson, G. Cryptic Genetic Variation in Evolutionary Developmental Genetics. *Biology (Basel)* **5** (2016).

64. Yadav, A., Dhole, K. & Sinha, H. Differential Regulation of Cryptic Genetic Variation Shapes the Genetic Interactome Underlying Complex Traits. *Genome Biol Evol* **8,** 3559–3573 (2016).

65. Goncalves, A. *et al.* Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res* **22,** 2376–84 (2012).

66. Wong, E. S. *et al.* Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. *Nat Commun* **8,** 1092 (2017).

67. Moyerbrailean, G. A. *et al.* High-throughput allele-specific expression across 250 environmental conditions. *Genome Research* **26,** 1627–1638 (2016).

68. Knowles, D. A. *et al.* Allele-specific expression reveals interactions between genetic variation and environment. *Nature Methods* **14,** 699–702 (2017).

69. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat Genet* **44,** 148–56 (2012).

70. Pradeepa, M. M. *et al.* Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat Genet* **48,** 681–6 (2016).

71. Lasserre, J., Chung, H. R. & Vingron, M. Finding associations among histone modifications using sparse partial correlation networks. *PLoS Comput Biol* **9,** e1003168 (2013).

72. Pai, A. A., Pritchard, J. K. & Gilad, Y. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet* **11,** e1004857 (2015).

73. Karlic, R., Chung, H. R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107,** 2926–31 (2010).

74. Kim, J. M. *et al.* A copula method for modeling directional dependence of genes. *BMC Bioinformatics* **9,** 225 (2008).

75. Lee, N. & Kim, J. M. Copula directional dependence for inference and statistical analysis of whole-brain connectivity from fMRI data. *Brain Behav* **9,** e01191 (2019).

76. Howe, F. S., Fischl, H., Murray, S. C. & Mellor, J. Is H3K4me3 instructive for transcription activation? *Bioessays* **39,** 1–12 (2017).

77. Zabidi, M. A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518,** 556–9 (2015).

78. Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13,** 613–26 (2012).

79. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167,** 1170–1187 (2016).

80. Xiong, N., Kang, C. & Raulet, D. H. Redundant and unique roles of two enhancer elements in the TCRgamma locus in gene regulation and gammadelta T cell development. *Immunity* **16,** 453–63 (2002).

81. Cretekos, C. J. *et al.* Regulatory divergence modifies limb length between mammals. *Genes Dev* **22,** 141–51 (2008).

82. Montavon, T. *et al.* A regulatory archipelago controls Hox genes transcription in digits. *Cell* **147,** 1132–45 (2011).

83. Hu, Y., Comjean, A., Perkins, L. A., Perrimon, N. & Mohr, S. E. GLAD: an Online Database of Gene List Annotation for Drosophila. *J Genomics* **3,** 75–81 (2015).

84. Landry, C. R. *et al.* Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. *Genetics* **171,** 1813–22 (2005).

85. Tirosh, I., Reikhav, S., Levy, A. A. & Barkai, N. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324,** 659–62 (2009).

86. Lemos, B., Araripe, L. O., Fontanillas, P. & Hartl, D. L. Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. *Proc Natl Acad Sci U S A* **105,** 14471–6 (2008).

87. Meiklejohn, C. D., Coolon, J. D., Hartl, D. L. & Wittkopp, P. J. The roles of cis- and trans-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression. *Genome Res* **24,** 84–95 (2014).

88. Gibson, G. & Dworkin, I. Uncovering cryptic genetic variation. *Nat Rev Genet* **5,** 681–90 (2004).

89. Schneider, R. F. & Meyer, A. How plasticity, genetic assimilation and cryptic genetic variation may contribute to adaptive radiations. *Mol Ecol* **26,** 330–350 (2017).

90. Zheng, J., Payne, J. L. & Wagner, A. Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks. *Science* **365,** 347–353 (2019).

91. Clouaire, T. *et al.* Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev* **26,** 1714–28 (2012).

92. Margaritis, T. *et al.* Two distinct repressive mechanisms for histone 3 lysine 4 methylation through promoting 3'-end antisense transcription. *PLoS Genet* **8,** e1002952 (2012).

93. Clouaire, T., Webb, S. & Bird, A. Cfp1 is required for gene expression-dependent H3K4 trimethylation and H3K9 acetylation in embryonic stem cells. *Genome Biol* **15,** 451 (2014).

94. Waymack, R., Fletcher, A., Enciso, G. & Wunderlich, Z. Shadow enhancers suppress input transcription factor noise through distinct regulatory logic. *bioRxiv,* 778092 (2019).

95. Starz-Gaiano, M., Cho, N. K., Forbes, A. & Lehmann, R. Spatially restricted activity of a Drosophila lipid phosphatase guides migrating germ cells. *Development* **128,** 983–91 (2001).

96. Ghavi-Helm, Y. *et al.* Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat Genet* (2019).

97. Zhu, J. *et al.* Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput Biol* **3,** e247 (2007).

98. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–95 (2010).

99. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

100. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9,** R137 (2008).

101. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* **43,** 264–8 (2011).

102. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–40 (2010).

103. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15,** 550 (2014).

104. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57,** 12 (1995).

105. Berger, R. L. & Hsu, J. C. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* **11,** 37 (1996).

106. Opgen-Rhein, R. & Strimmer, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst Biol* **1,** 37 (2007).

107. Lee, N. & Kim, J. M. Copula directional dependence for inference and statistical analysis of whole-brain connectivity from fMRI data. *Brain Behav* **9,** e01191 (2019).

108. Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res* **21,** 1728–37 (2011).

109. Graveley, B. R. *et al.* The developmental transcriptome of Drosophila melanogaster. *Nature* **471,** 473–9 (2011).

110. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci U S A* **107,** 9546–51 (2010).

111. Thomsen, S., Anders, S., Janga, S. C., Huber, W. & Alonso, C. R. Genome-wide analysis of mRNA decay patterns during early Drosophila development. *Genome Biol* **11,** R93 (2010).

112. Mi, H. *et al.* Assessment of genome-wide protein function classification for Drosophila melanogaster. *Genome Res* **13,** 2118–28 (2003).

113. Turner, L. M., Chuong, E. B. & Hoekstra, H. E. Comparative analysis of testis protein evolution in rodents. *Genetics* **179,** 2075–89 (2008).

114. Sklar, A. Random variables, joint distribution functions, and copulas. *Kybernetika* **9,** 12 (1973).

115. Sungur, E. A. A Note on Directional Dependence in Regression Setting. *Communications in Statistics - Theory and Methods* **34,** 9 (2005).

116. Berney, D., Petcher, T. J., Schmutz, J., Weber, H. P. & White, T. G. Conformations and biological properties of apomorphine and its phenanthro(10,1-b,c)azepine homologue. *Experientia* **31,** 1327–8 (1975).

2.5 Complementary results

2.5.1 Construction of the mappability mask

The construction of the mappability mask consists in a workflow of bioinformatics tools and personalised scripts (Fig. 2.22a-b), generating two distinct sets of filtered regions : the genomic mask and the synthetic mask.

Firstly, the generation of personalised annotations for each parental genome is necessary to properly process transcriptomic data. In this respect, I designed a Snakemake pipeline (see section 2.5.8) in order to convert the coordinates of each annotated element from the reference genome (dm3, GFF format) into a new genome coordinate space (Fig. 2.22a).

The conversion is performed using standard file formats, such as chain and PSL (Pattern Space Layout) files, which allow a base-wise coordinate conversion, in combination with the UCSC tool pslMap. The format conversion was notably involving the conversion between 1-based fully closed (GFF) and 0-based half-open (BED, BAM, PSL) coordinate systems.

Secondly, starting from the parental genome sequence (FASTA files) and personalised genome annotation, I have designed a second pipeline to simulate genomic and transcriptomic sequencing reads (Fig. 2.22b).

Simulated genomic reads cover the whole genome homogeneously, with one read starting at each base pair position. Genomic reads have been simulated using two different read length (75bp and 100bp), in order to best match ATAC-seq and ChIP-seq data (Fig. 2.22c).

Similarly, simulated transcriptomic reads are 100bp long and cover all the annotated coding regions with no change in depth of coverage (Fig. 2.22c), except in the case of overlapping exons.

Simulated reads are mapped to their respective personalized genome and translated into the reference coordinate space (Fig. 2.22b). Regions showing a difference in coverage between the two parental genomes, presumably stemming from insertions, deletions, heterozygous sites or inherent mappability issues, are included in the final synthetic mask (Fig. 2.22c).

In complement to the synthetic mask, I have also generated a genomic mask, based on genomic DNA sequencing data from the parental lines. The genomic reads are mapped to both parental genomes and translated into the coordinate space of the reference genome, similarly to the simulated read processing (Fig. 2.22b). A region displaying a null coverage in at least one of the parent is included in the mask. The genomic filter thus aims at discerning the regions of the genome presenting low mappability, heterozygocity or genotyping errors.

The visual inspection of the read alignments permitted to distinct a trend related to the use of a standard "local" alignment strategy. Indeed, by clipping the reads presenting mismatches at their extremities, the "local" alignment tends to discard numerous read extremities comprising SNPs. This phenomenon intensifies when a splice junction is also present in the vicinity. This alignment strategy may therefore be too conservative and discard unbiased regions.

As a result, we chose to use the "end to end" alignment strategy for all our mapping steps. However, forbidding

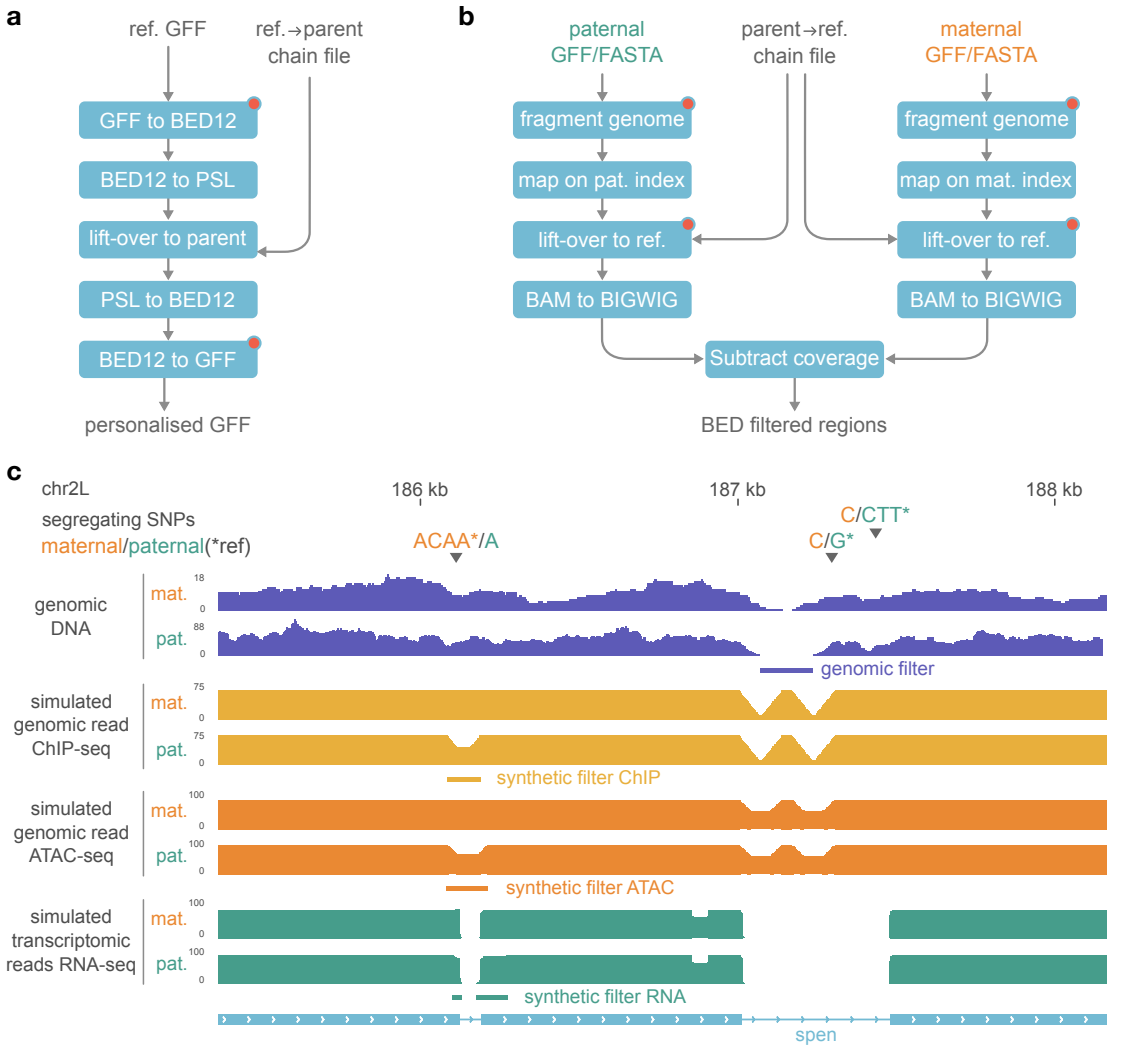

Figure 2.22: Mappability mask construction. a,b: Schematic of the workflows used for the construction of personalised parental genome annotations (a) and mappability masks based on simulated reads (b). The lift-over step consists in converting the read alignments coordinates from a given genome assembly into a new coordinate space (eg. reference). Steps labelled with a red dot are done using personalised scripts, chain files were provided by the Furlong laboratory. c: Example browser snapshot of the mappability mask construction for the F1 line vgn852. An exonic deletion present in the paternal genotype (ACAA/A) prevents a fraction of the paternal simulated reads from correctly mapping on the reference ; the regions with a difference in coverage between the two genotypes are included in the synthetic mask. A low mappability intronic region prevents both genomic and simulated reads from mapping ; the regions with a null genomic coverage in any of the parental lines are included in the genomic mask. ref.: reference.

read clipping requires high quality reads, in order to avoid incorrect alignment. Consequently, the sequenced reads were trimmed based on sequencing quality score prior to mapping.

2.5.2 Impact of the synthetic mask

The complete mask, combining genomic and synthetic filters for all lines, represents 25Mbp, including 8Mbp in exonic and UTR regions (Fig. 2.23a). The mappability filter correction does not show clear change in allelic ratio

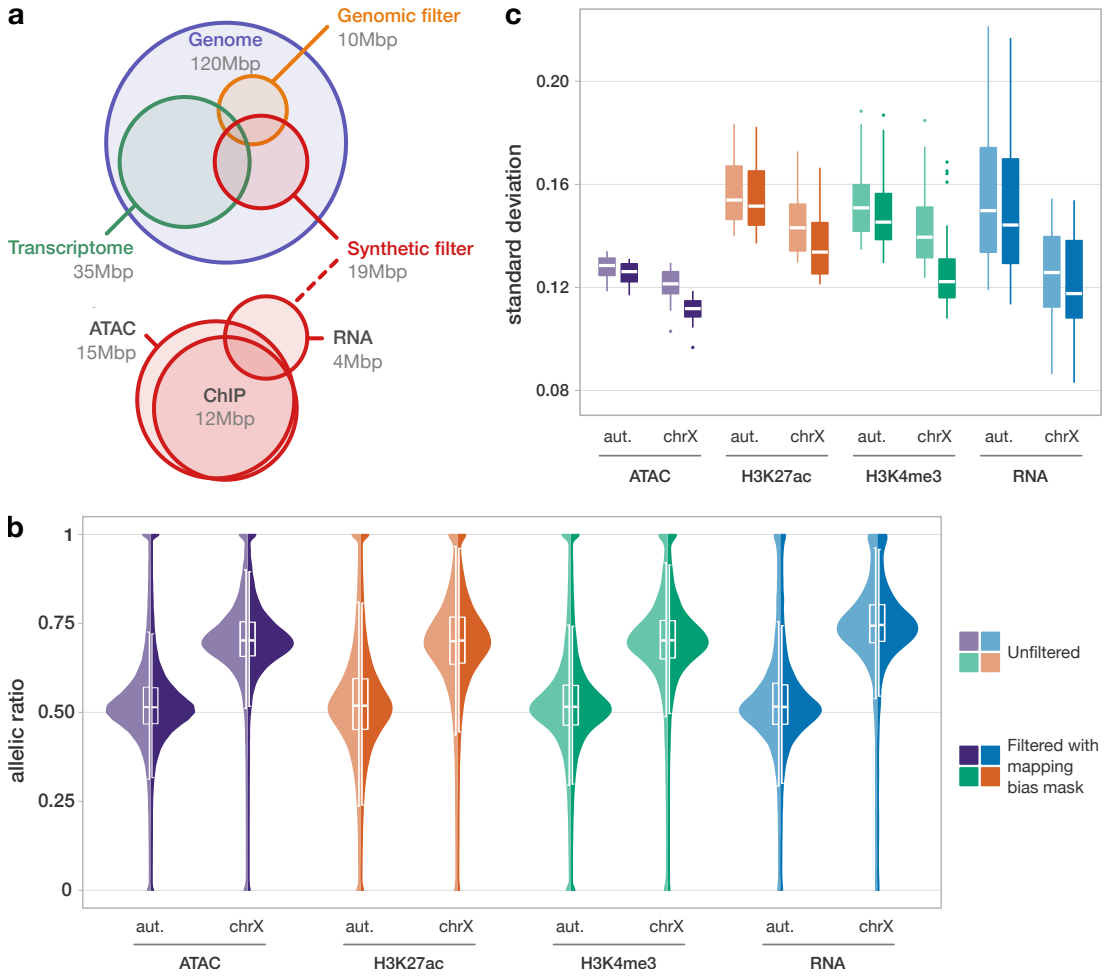

Figure 2.23: Building a mappability mask. a: Proportional Venn diagram showing the composition and size of the universal mappability mask. b: Distribution of allelic ratios for mappability filtered data (dark colors) and unfiltered data (light colors) for all F1 samples (maternal genes removed, c.f. section 2.5.4). c: Distributions of the standard deviations of allelic ratios per sample, in each date type. The observed small decrease in standard deviation for corrected data (dark colors), compared to non-corrected data (light colors) suggests that the mappability mask mostly discards regions with elevated allelic imbalance. aut.: autosomes.

mean values (Fig. 2.23b). However, the correction tends to decrease the standard deviation of the distribution (Fig. 2.23c). This trend is in agreement with the expected effect of discarding regions with differential mappability levels, as they should correspond to regions with large allelic imbalance.

2.5.3 Impact of using F1 genomic data to discard genotyping errors

Following the mappability filtering step, we observed a small enrichment of extreme imbalanced features at both ends of the allelic ratio distribution (Fig. 2.23b). These features have the totality of their overlapping reads assigned solely to one parental allele.

To assess whether this measure reflects genuine imbalances or genotyping errors, we used genomic DNA (gDNA) se-

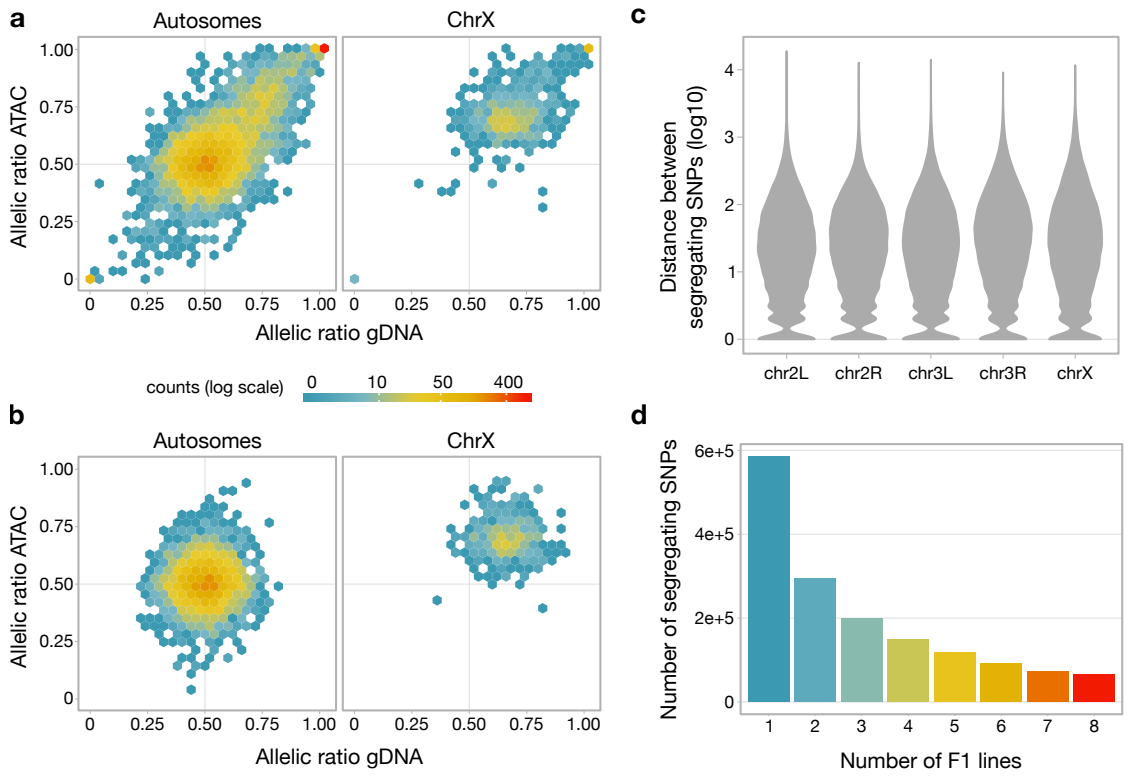

Figure 2.24: Testing for allelic imbalance in genomic DNA (gDNA) data. a,b: Heatmaps showing the allelic ratio correlation between ATAC-seq and gDNA-seq data in line vgn852 at 10-12hr, for autosomes and chrX, before (a) and after (b) correction of genotyping errors. In absence of correction (a), we note an enrichment for extreme imbalanced features in both cases, suggesting the presence of genotyping errors. After correction (b), these extreme features are filtered. c: Violin plots showing the distribution of distances between SNPs segregating in the F1 line vgn852, for each chromosome, after gDNA correction. d: Bar plots showing the proportion of SNPs shared between the F1 lines after gDNA correction. The large majority of the SNPs are found segregating in only one of the F1 lines.

quencing data from the F1 lines (cf. section 2.2.5).

Most of the measured SNPs were effectively relatively balanced between the two alleles, yet a small fraction was showing extreme imbalance (Fig. 2.24a). The extreme imbalance at genomic DNA level was also correlated with extreme imbalance in other datatypes, such as ATAC-seq, in the same measured features (Fig. 2.24a).

To test departure from a balanced allelic ratio in gDNA data, I performed a two-sided binomial test for each SNP following the formula :

$$m_s^{cl} \sim Bi(n_s^l, p^{cl}), \tag{2.10}$$

where m represents the number of maternal reads overlapping the SNP s within chromosome c in F1 line l, Bi denotes the binomial distribution, the parameter n_s^l refers to the total number of reads mapped on SNP s in F1 line l, and the parameter p^{cl} represents the expected allelic ratio under H0 conditions (allelic balance) in chromosome c and F1 line l.

As we work with samples of pooled female and male embryos, it is expected that allelic ratio in chromosome X (chrX) will not be centred at a 50:50

ratio. Indeed, as half of our embryo pool has two copies of chrX (females) and the other only one (males), the allelic ratio is expected to lie near $\frac{1}{3}$. As a result, to test the chrX SNPs, we use as H0 hypothesis the average allelic ratio from the 1000 SNPs with highest coverage in chrX. For autosomes, the H0 hypothesis reflect a balanced state ($p = 0.5$).

In each F1 line, we found on average 62,303 SNPs ($\pm 5,726$ standard deviation) with significant allelic imbalance (FDR$<$5%), which represents approximately 10% of SNPs segregating in the cross. After discarding the SNPs showing significant imbalance, the resulting distribution of allelic ratios is improved. Indeed, the enrichment for extreme imbalance at the tail of distribution is no longer observed (Fig. 2.24b).

Additionally, we noticed that the genomic DNA (gDNA)-based correction also brings the mean allelic ratios closer to their expected average (0.5 for autosomes and 0.66 for chrX). For example, following gDNA correction in ATAC-seq data for line vgn852 at 10-12hr, the average allelic ratio drops from 0.59 to 0.50 in autosomes and from 0.75 to 0.70 in chrX (Fig. 2.24c). After correction, we still conserve a high number of heterozygous SNPs, with a majority of them being line-specific (Fig. 2.24d). The average distance between these SNPs is approximately 80bp. This high density is sufficient to measure allele-specific expression in more than 94% of all detected coding and non-coding features.

2.5.4 Impact of using egg data to discard maternal transcripts

After correcting for mappability biases and genotyping errors, we still observe a shift in the allelic ratio toward the maternal allele for RNA-seq data at early time point. The genes highly expressed and highly imbalanced at 2-4hrs are suspected to have their transcripts maternally deposited in the egg.

The presence of unfertilised eggs and the maternal deposition of mRNAs in embryos result in a maternal shift in the allelic ratio distribution. In order to avoid discarding all the maternally-deposited genes from our analysis, I tested several methods aiming at correcting their allelic ratios.

Using RNA-seq data of unfertilised eggs as a model group for gene expression, I tried to estimate the fraction of maternally deposited reads within the maternal read pool of the F1. I tested four deconvolution methods, including three published tools : PERT [129], ISOpure ([130] and unmix from DESeq2 R package [70]. I designed a fourth, more naive method, which uses the slope of the linear regression between F1 maternal counts and egg counts as an estimate of the egg fraction in maternal signal.

None of these deconvolution methods gave satisfactory results. Most of the estimated egg fractions were over-corrective for the samples with small imbalance bias and under-corrective for highly biased samples (Fig. 2.25). This is likely due to the relative large differences in gene expression between the model data of unfertilised eggs and the tested maternal-only reads of F1 data. Additionally, *trans*-acting paternal signal in F1 data may also have an effect on gene expression.

I further tested an expectation maximisation approach. This method adjusts the estimation of reads coming from unfertilised eggs, until the mode of the allelic distribution is centred at 0.5. Although this method seemed to give good

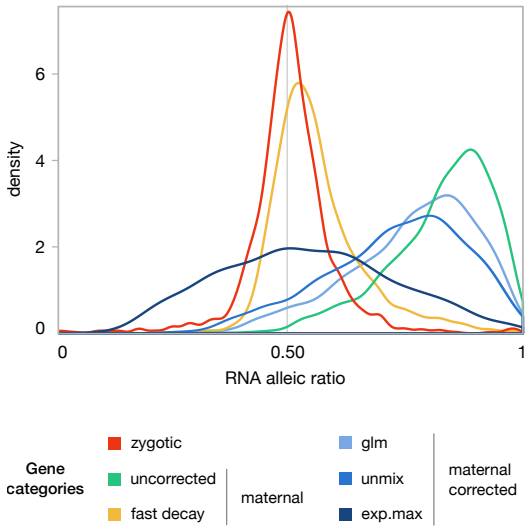

Figure 2.25: Maternal transcript correction. Distribution of RNA-seq allelic ratios for sample vgn57 at 10-12hr. Genes with exclusive zygotic expression (red) or fast-decaying maternal expression (yellow) are both centered at a balanced 0.5 ratio. Genes with slow maternal decay (green) are shifted toward the maternal side due to maternally-deposited transcripts, still present in the embryos. The three methods tested to infer and remove the proportion of deposited transcripts in the maternal counts failed to provide satisfactory correction (blue hues).

results for low bias samples, the corrected distribution of samples with large allelic bias had a very large variance compared to the uncorrected distribution, and could not be reliably used for allele-specific analysis (Fig. 2.25).

As we did not find an efficient method to correct the allelic bias in maternally-deposited genes, we decided to discard the RNA-seq data for 2-4hrs time point. In 6-8hr time point, we still observe a smaller maternal bias originating from maternally deposited transcripts that did not fully decay by that time.

To correct for this remaining bias, we used DESeq2 R package to perform a differential gene expression analysis in egg data between 2-4hrs and 10-12hrs, as described in the Methods of the manuscript. Genes showing a significant decrease in gene expression between 10-12hr and 2-4hr are considered to have most of their maternally-deposited transcripts decayed by 6-8hr. Thus, this set of genes with fast decaying maternal transcripts are rescued for further analyses.

In conclusion, allele-specific data analysis requires careful pre-processing steps, taking into consideration the potential biases affecting the allelic ratios. Such biases can arise from the reference genome during the mapping step, genotyping errors during the read assignment step, or from maternally deposited transcripts. Making use of simulated reads, genomic DNA-seq data and RNA-seq data of unfertilised eggs, my analyses address and correct all these three kinds of biases.

2.5.5 Delineation of genomic regions with overlapping signals

One challenge of this study is to integrate heterogeneous data types coming from multiple samples. To integrate signal from genes (RNA-seq) and chiefly intergenic regions (ATAC and histone marks peaks), we generated a common set of genomic ranges, based on a combination of peak-calling done in each F1 lines.

Comparing results between data types requires to link different features based on their genomic location proximity (e.g. ATAC peaks and H3K27ac peaks). This becomes complex when simultaneously comparing multiple layers of both genic and intergenic signal. Indeed, each data type has different region characteristics in term of number, location and length. Hence, associating regions of different types is not always intuitive.

Several tools exist to make non-coding

Table 2.3: Non-coding genome annotation tools.

Algorithm Association type	Assignment strategy	Limitations	Ref.
GREAT peak-genes	Assign non-coding peaks to genes based on TSS-proximal regions and nearest TSS.	No association between non-coding peaks.	[131]
ChIPseeker peak-peak	Compute co-occurrence likelihood from overlaps of non-coding peaks. Use shuffled peaks as control.	Pairwise assignment only, no association with genes.	[132]
ChromHMM peak-peak	Train a Hidden Markov Model on multiple binarised non-coding signals in bins of 200bp.	No association with genes.	[133]

region associations (Table 2.3). However, none of them fully address the challenge of a simultaneous assignment of non-coding regions to other non-coding regions and genes.

To address this issue, I have designed an algorithm based on the tools presented in table 2.3 to build a list of overlaps across the four data types.

I first applied a binarisation strategy, similar to chromHMM, in order to integrate non-coding regions together. I used the peak-calling results to define the genomic regions with either presence or absence of signal for each type of data (Fig. 2.26).

Secondly, I computed the overlaps between each of these genomic ranges using the R package GRanges [134], similarly to ChIPseeker. Regions needed to overlap by at least one base pair to be linked together. In case where three regions were overlapping but no base pairs were shared by all of them, I considered two distinct overlap events, with one feature being present twice.

Thirdly, I assigned the non-coding overlaps obtained from the first step to the genes, using a strategy similar to GREAT (Fig. 2.26). On the one hand, overlaps closer than 500bp from a TSS were directly assigned to the associated gene, constituting the promoter-proximal set. On the other hand, overlaps further apart than 500bp from a TSS were assigned to the gene with the nearest TSS, forming the promoter-distal set.

Several adjustments were made. Indeed, promoter-proximal overlaps are frequently present between two TSS in opposite direction. In this "head-to-head" conformation, where upstream regions overlap, it is difficult to properly assign the non-coding overlap to its target gene.

In order to avoid mis-assignments, TSS having another TSS present in their 600 bp upstream region were discarded in the analyses involving correlations between data types.

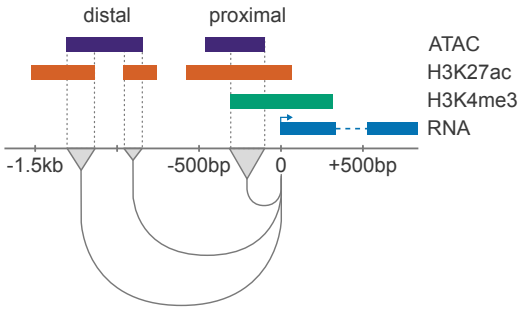

Figure 2.26: Defining genomic region overlaps. Schematic summarising the method to assign non-coding region to genes. Non-coding regions from different assays (ATAC-seq and ChIP-seq) are linked together if they all share at least one base pair (grey triangles). Genes are associated to each non-coding region overlapping the 500bp region upstream of their TSS (proximal overlap). Distal non-coding regions are assigned to the gene with the nearest upstream or downstream TSS (distal overlap).

Additionally, we observed a clear drop in the correlation between gene expression (RNA-seq) and enhancer activity (ATAC-seq, ChIP-seq) when the distance between the non-coding overlap and the TSS was larger than 1.5kb (cf. Method within the manuscript in section 2.4). As a result, we discarded from the overlapping regions located over 1.5 kb from the nearest TSS.

2.5.6 Probing direct interactions with partial correlation

By applying Pearson correlation on the proximal and distal overlap sets described in section 2.5.5, I noted that allelic ratios are highly correlated between all the features (genes, peaks, ...) overlapping the same regions. This result suggests that all regulatory layers are indirectly affecting each other and behaving in a synchronized way.

Consequently, I performed a partial correlation analysis in order to extract the fraction of direct correlation from these highly co-varying features. The partial correlation method uses the residuals, obtained after regressing confounding variable(s) against variables of interest, to perform Pearson correlation measures.

Although this method is efficient in discriminating direct from indirect correlation, it requires to have complete and independent measures. I therefore discarded, prior to the analysis, overlaps with missing data and/or involving features already assigned to another overlap.

I first applied partial correlation for each time point, and further validated the resulting correlation values using bootstrapping (Fig. 2.27). As I obtained similar results in 6-8 hrs and 10-12 hrs time points, I applied the analysis again on the combined 6-12hr data in order to increase the statistical power.

This final result is presented and described in the manuscript (section 2.4).

2.5.7 Exploring allelic imbalance at the SNP level

In 2018, Jacobs *et al.* published an analysis of ATAC-seq data from 23 inbred Drosophila strains from the DGRP [53]. Using a multivariate regression analysis, they detected 4,289 SNPs (1.5% of their tested set) significantly co-segregating with the local chromatin accessibility levels of enhancers (caQTL). They identified Grainyhead as the pioneer transcription factor having a binding motif impacted by the presence of a SNP.

I performed the same type of analysis using ATAC-seq total count (TC) and allelic ratio (AR) data from our project.

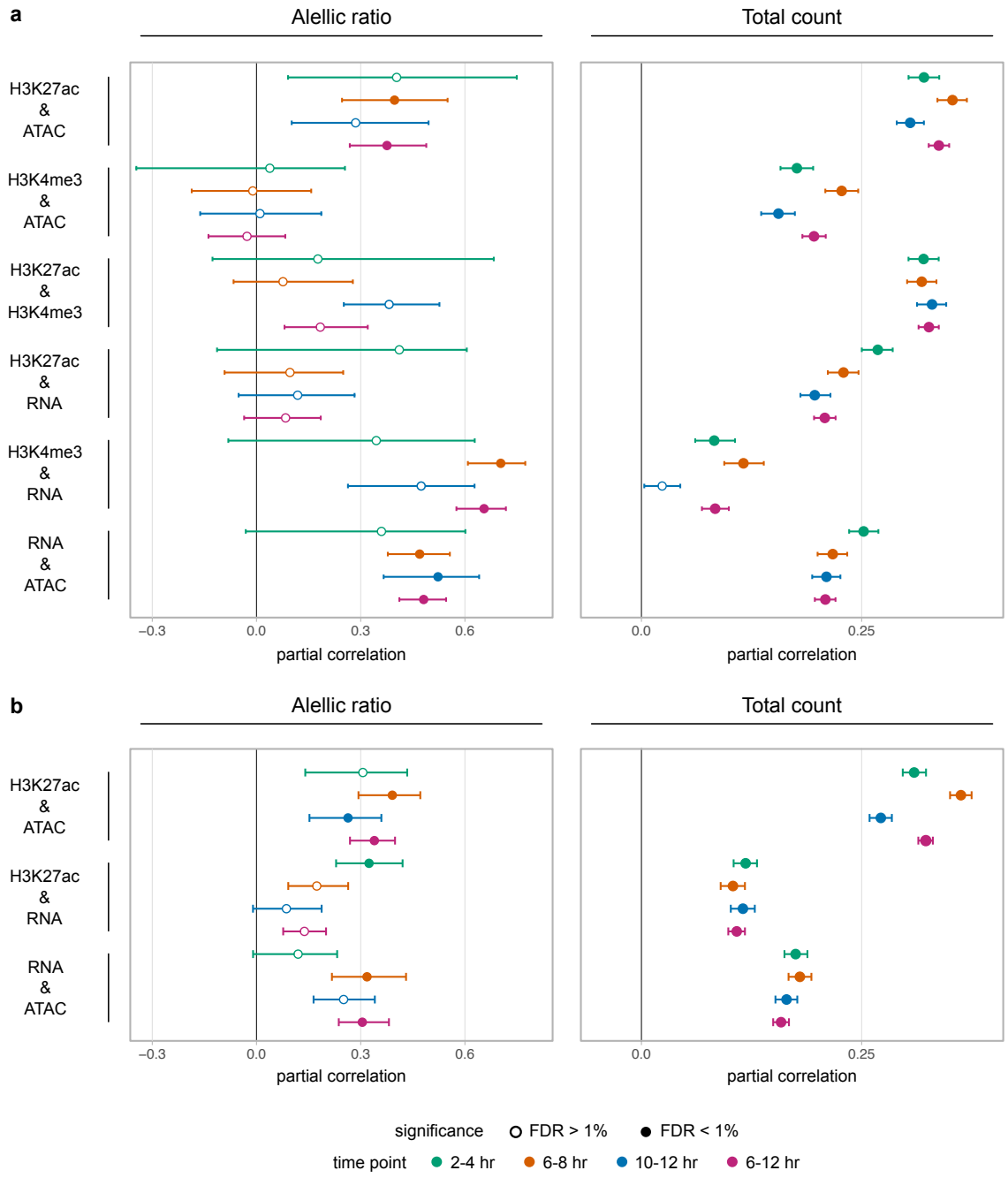

Figure 2.27: Assessing direct relationship using partial correlation. Partial correlation results obtained for each pairwise data type comparison, at each time point, for promoter-proximal (a) and promoter-distal (b) features. Correlation results are shown for allelic ratio data (left) and normalised total count data (right), in autosomes. Wiskers represent the 95% confidence interval obtained with bootstrap analysis (80% sub-sampling, 2000 iterations). Circles position depict the partial correlation values obtained using the full dataset, significant interactions (FDR<1%) are shown with filled circles. Although most of the correlations are significant in total count data, partial correlations in allelic ratios are more variable and significant in only a small subset of the tested interactions, suggesting a smaller number of direct relationships.

I fitted a linear regression model for each SNP overlapping a promoter-distal overlap (putative enhancer), using the following formulas :

$$|log_2(AR_s^l)| \sim \alpha_s + \beta_s \times snp_s^l + \epsilon, \quad (2.11a)$$

$$TC_s^l \sim \alpha_s + \beta_s \times snp_s^l + \epsilon, \quad (2.11b)$$

where AR_s^l and TC_s^l respectively denote the ATAC-seq allelic ratio and total count for line l at SNP position s, α is the intercept term, β is the estimated slope coefficient, snp_s^l is a Boolean variable depicting the status of SNP s in line l (1 if heterozygous, 0 if homozygous), and ε is the residual variable.

I obtained 1,031 SNP significantly co-segregating with ATAC allelic imbalance (0.8% of tested set, FDR<0.1) (Fig. 2.28) and one SNP co-segregating with global accessibility level.

I tested whether these SNPs were impacting a specific motif binding site using Var-tools from the RSAT suite [135]. This tool computes the difference δ in motif matches between two haplotypes for all motifs from the JASPAR insect database (approximately 1,300 non-redundant motifs). I compared the obtained δ with a control set of SNPs not co-segregating with ATAC-seq signal. I did not find motifs differentially impacted between segregating and non-segregating SNPs sets.

This negative result might be due to two main aspects of the design of the study. Firstly, the data were collected from whole embryos, which can limit the detection of enriched sequences, as transcription factors usually act in a tissue-specific manner. Secondly, eight different lines might not grant sufficient power to detect the effect of a SNP.

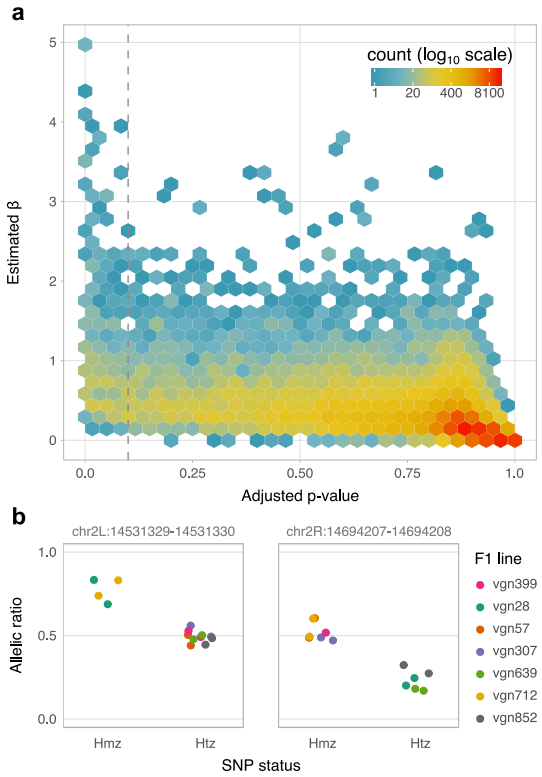

Figure 2.28: Detecting SNP co-segregating with allelic imbalance. a: Heatmap showing the distribution of the estimated β and the associated adjusted p-values, obtained from the linear regression analysis described in equation 2.11a. Vertical dashed line represents the threshold used to select the SNP considered as significantly co-segregating with ATAC-seq allelic imbalance (p.adj<10%). b: Scatter plot showing two examples of significant SNPs. Left panel shows a SNP tending to co-vary with the presence of paternal allelic imbalance. Right panel shows a SNP which tends to increase allelic imbalance when present in the cross (heterozygous state).

2.5.8 Script availability

All analyses reported in the presented manuscript have been performed using the programming languages Python (https://www.python.org/) and R (http://www.R-project.org), and the version control system Git (https://git-scm.com). The scripts are available in a GitHub repository and the data have been deposited in the ArrayExpress database

a

```
IDS, = glob_wildcards("./fileA_{id}")

rule all:
    input:
        expand("res_{id}", id = IDS)

rule merge_files:
    input:
        A = "fileA_{id}",
        B = "fileB_{id}"
    output:
        "temp_{id}.txt"
    shell:
        "cat {input.A} {input.B} > {output}"

rule apply_cowsay:
    input:
        "temp_{id}"
    output:
        "res_{id}"
    shell:
        "cat {input} | cowsay > {output}"
```

b

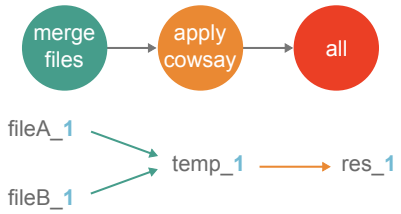

Figure 2.29: Snakefile example workflow. a: Example of a master Snakefile requiring specific inputs targeted using an "id" wildcard. Here, the Snakefile includes rules for merging two input files and applying the "cowsay" command. b: Rule graph generated from the Snakefile (top) and schematic of the different files generated in the case of a wildcard with value "1" (bottom).

(www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-8877, E-MTAB-8878, E-MTAB-8879 and E-MTAB-8880.

The mappability masks were built using the Python-based workflow management tool Snakemake [128]. This system is based on a master script (Snakefile) (Fig. 2.29a) recapitulating the architecture of the workflow using wildcard names. It generates a rule graph (Fig. 2.29b) setting up the rule dependencies and automatically submits jobs to the cluster, following the graph order.

This software has the advantage to combine multiple processing steps into a single job submission of the master script. In addition, a Snakefile can be easily shared and reused, which ensure a higher consistency in data storage and script usage, especially for complex analyses. Lastly, the wildcard system makes the workflow easily adaptable and scalable to different datasets.

Chapter 3

Depicting *trans*-regulation using logical modelling

3.1 Study summary . . . 107

3.2 Methodological background . . . 107
3.2.1 The regulatory role of TGF-β signalling . . . 107
3.2.2 Draw me a TGF-β map . . . 108
3.2.3 Feedback circuits . . . 109
3.2.4 Logical formalism . . . 110
3.2.5 Dynamical simulation . . . 111

3.3 Contribution to the published work . . . 113

3.4 Deciphering and modelling the TGF-β signalling interplays specifying the dorsal-ventral axis of the sea urchin embryo . . . 114
3.4.1 Abstract . . . 114
3.4.2 Introduction . . . 114
3.4.3 Results . . . 117
3.4.4 Discussion . . . 128
3.4.5 Materials and methods . . . 131
3.4.6 Acknowledgements . . . 135
3.4.7 References . . . 136

3.5 Complementary results . . . 139
3.5.1 Script availability . . . 139

We shall gradually approach the correct view -or, to put it more modestly, the one that I propose as the correct one.

Erwin Schrödinger, What is life?, 1944

Deciphering and modelling the TGF-β signalling interplays specifying the dorsal-ventral axis of the sea urchin embryo

Swann Floc'hlay[1], Maria Dolores Molina[2*], Céline Hernandez[1*], Emmanuel Haillot[2], Morgane Thomas-Chollier[1,3], Thierry Lepage[2⋈] and Denis Thieffry[1⋈]

* equal contributions ; ⋈ corresponding authors

3.1 Study summary

In the presented manuscript (currently under review in the journal Development), we aimed at better understanding the gene regulatory network governing the onset of dorsal-ventral axis specification in the sea urchin *Paracentrotus Lividus* (Lamarck, 1816). Using a logical formalism, implemented in the software GINsim [136–139], we delineated the main regulatory interactions involved in this process. We further analysed the network dynamics using the stochastic simulation software MaBoSS [95, 140], and characterised diffusion signalling using the mulicellular framework implemented in EpiLog [141]. Together, these analyses have highlighted the crucial role of the cross-inhibition between the two TGF-β pathways Nodal and BMP2/4. Additionally, we noticed that the network structure inherently provides an advantage for the dorsal cascade activation.

3.2 Methodological background

3.2.1 The regulatory role of TGF-β signalling

In collaboration with the Lepage lab (Institut Valrose, Nice), we chose to study the gene regulatory network controlling dorsal-ventral axis specification in the sea urchin for several reasons.

Firstly, the specification of the main axes of body plan is crucial for embryogenesis. In chordate, dorsal-ventral patterning is defined through the positioning of a master organiser, at the dorsal side of the embryo, equivalent to the Spemann's organiser observed in amphibians [3].

However, the evolutionary origin of this organiser, and the precise mechanisms governing its establishment are not fully understood. As part of the deuterostome phyla, the sea urchin is an interesting system to study the evolutionary divergence between the chordate and the echinoderm organisers [142].

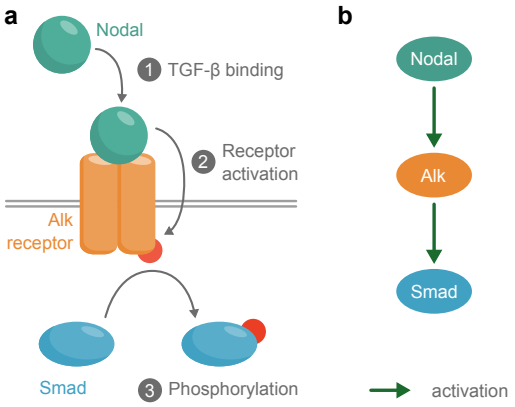

Figure 3.1: TGF-β regulatory map. Schematic of a TGF-β signalling pathway. a: In the Nodal TGF-β signalling pathway, the Nodal ligand binds to Alk receptors type I and II. This triggers the activation of a kinase activity and the phosphorylation of Smad proteins. b: This signalling pathway can be decribed in a regulatory map as a directed acyclic graph of two sequencial activations.

Secondly, it is known that the dorsal-ventral axis specification is governed by Transforming Growth Factor β (TGF-β) morphogen gradients (Fig. 3.1a). TGF-β signaling is particularly interesting to study ; it involves a kinase activity triggered from the ALK membrane receptors. The resulting phosphorylation cascade can act on a large number of downstream signalling components, including the Smads proteins [143].

TGBF-β are of particular interest due to their known implication in cell proliferation [143], as well as the emergence of multi-cellularity [143]. A better understanding of the regulatory interactions involved in TGF-β signalling would benefit multiple domains, including cancer research [144].

3.2.2 Draw me a TGF-β map

Proper mapping of the TFG-β regulatory network constitutes the first abstraction step to further explore its behaviour [82]. Indeed, molecular interaction maps integrate multiple regulatory processes, enabling the delineation of the key regulatory circuits embedded within the network.

The notions framing the design of regulatory maps take roots in the domain of graph theory. The vertices (or nodes) represent the molecular compounds, and the edges represent their interactions (Fig. 3.1b). The vertices mainly refer to genes, although they may encompass other elements (eg. proteins) or molecular mechanisms (eg. apoptosis). In a gene regulatory map, interactions are chiefly directed, from an active gene toward a targeted element downstream of the regulatory cascade. Consequently, the edges of the regulatory graphs can be represented as oriented arcs, with the possibility to assign them a specific sign and shape, reflecting the type of interactions. By convention, activation are positive green arrowheads and repression are negative red hammerheads (Fig. 3.2b). Today, multiple bioinformatic tools exist to design a regulatory map, such as CellDesigner [145], GINsim [139] and BioTapestry [146] (cf. Systems Biology Graphical Notation project, https://sbgn.github.io/). Although each of these tools have specific modelling capacities (eg. description of activity flow, process and entity relationship), they all adopt a common standard for model encoding: the Systems Biology Mark-up Language (SBML, http://www.sbml.org) [147]. This compliant format is used as an exchange format to promote the sharing of information between different scientific communities and software.

This set of tool is used to integrate into a formal mathematical framework the documented genes and regulatory interactions compiled from the existing literature and/or novel experiments (cf. section 1.2.4). As a result, we obtain a reg-

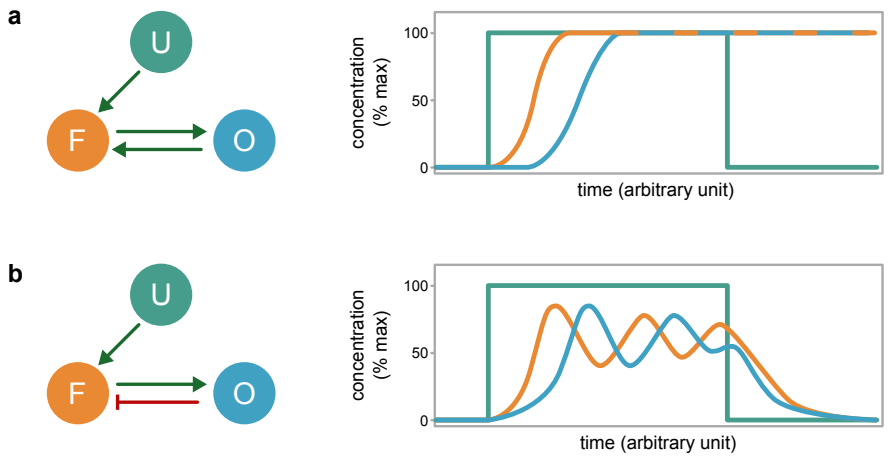

Figure 3.2: Feedback circuits dynamic. Characteristics of a minimal feedback circuit. a: In a positive feedback circuit (left panel), gene O, once activated by an upstream gene F, has a positive action (green arrow) on the expression level of gene F. This reciprocal activation module allows for the maintenance of the expression of O and F (right panel, blue and orange lines respectively), even if the activity of the upstream activator of gene F, gene U, stops (green line). b: In a negative feedback circuit (left panel), upon sufficient activation from gene F, gene O will be activated and in turn trigger an inhibition action (red arrows) on gene F. This asymmetrical module will create oscillations alternating between rises and drops in gene expression levels of genes O and F (right panel), that will further stabilise at an intermediary level. However, a loss of activation from the upstream activator of gene F, gene U, will completely shutdown the module.

ulatory graph of interconnected nodes from which we can already derive interesting properties. For example, morphogens and master regulators, such as Nodal (Fig. 3.1b) and BMP, are usually present at the top of the network, having few upstream components and a high number of interactions targeting downstream components.

3.2.3 Feedback circuits

A the onset of embryogenesis, the establishment and maintenance of the expression of key regulatory genes is crucial to properly specify the main presumptive territories. The fine tuning of the expression of these genes involves multiple regulations and feedback, forming regulatory circuits [148]. Regulatory circuits (also often called feedback loops or feedback circuits) are defined as simple circular sequences of regulatory interactions (Fig. 3.2). In any such defined circuit, each component exerts an indirect effect on itself, with a sign simply depending on the product of the signs of all the regulatory interactions taking part in this circuit. Hence, regulatory circuits can be classified into positive versus negative circuits, depending on the parity of the number of negative interactions involved.

Positive feedback circuits (Fig. 3.2a) trigger their own activation pathway. They are defined as a cycle comprising an even number of negative interactions [149]. Zero being an even number, a cycle with only positive interactions is also a positive circuit.

Such positive circuits have several key characteristics. Firstly, they can provide a high signal sensitivity. Indeed, their structure enable to amplify the activation signal, even in the case of low initial levels [148]. Secondly, by self-promoting the induction of their target gene, positive circuits have a steep response time, similar to a switch behav-

ior [148, 150]. Last but not least, these circuits confer multi-stationary properties to the network [148, 151]. On the one hand, upon an initial transient activation, the reciprocal activation of two genes may self-perpetuate the activation signal, even long after the upstream activation has ceased. On the other hand, in absence of initial activation the circuit is maintained in a complete inactivated state. This switch-like property of transient signal memory is necessary for the establishment of different stable states (cf. section). On a biological perspective, it reflects the potential of a cell to commit into a specific differentiated state, subsequently to a transient specification signal.

In negative feedback circuits (Fig. 3.2b), downstream components inhibit the activation of their own upstream activator. These circuits are defined by a set of negative interactions present in odd number [149] within the cycle. Like positive feedback circuits, they carry interesting characteristics for the tight control of gene regulation. Firstly, they tend to stabilise the expression target gene at an intermediate level, corresponding to an homeostatic state [148]. Indeed, such circuits function like thermostats ; they push the system back and forth toward an intermediate expression level [149]. Secondly, negative feedback circuits behavior greatly varies depending on the response time. In the case of a fast response time following activation, these circuits reach a steady-state transcript concentration. They even tend to reach it more rapidly than an unregulated, open circuit [152]. Conversely, if the response time is delayed, negative feedback circuits may trigger a sustained oscillatory behavior [148], where the target gene expression level is alternating between high and low concentrations.

The capacity for negative circuits to either maintain homeostasis or oscillatory behaviors is of particular use in transcription regulation. Indeed, it is likely due to their capacity to maintain an homeostatic state that these circuits are commonly found within the *Escherichia coli* regulatory network [153]. Additionally, their potential to sustain oscillatory behavior is exploited by networks controlling cellular cycles and circadian rhythms [94, 154].

As a result, regulatory circuits can convey robustness and precision to signalling and regulatory networks. They enable the readjustment of gene activity level following perturbation (e.g. change or loss of upstream activation level). However, in the presence of multiple intertwined circuit modules, it becomes difficult to grasp the corresponding dynamical properties using the sole intuition. Hence, one then needs to use a more formal approach to model and simulate such networks.

3.2.4 Logical formalism

A large variety of both quantitative and qualitative approach have been applied to formalise regulatory networks, including the Boolean approach [21].

According to the Boolean formalisation, the activity of each gene is approximated by a binary variable, taking the value 0 when the activity is negligible, and the value 1 when it displays a functional activity. The state of a node is dictated by a Boolean function, taking as input the binary states of the upstream nodes and returning a binary solution. The different input states are combined using the logical operators AND, OR and NOT, which respectively correspond to the logical sum (inclusive OR), product and negation (Fig. 3.3.

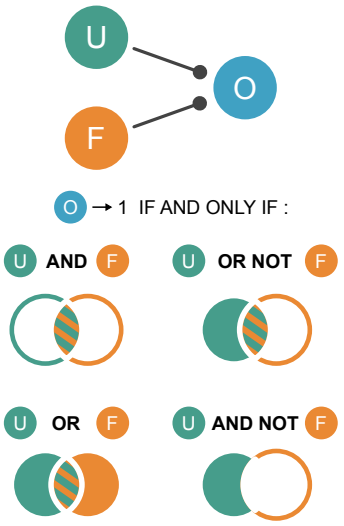

Figure 3.3: Boolean function construction. Four Boolean function examples in a simple circuit of two upstream genes (top schematic). For each Boolean function, the conditions meeting its requirements to activate gene O are shown as colored area in the Venn diagrams. If U and F are activators, they may activate O either if they are both required (U AND F) or if only one of them is necessary (U OR F). If U is activator and F is inhibitor, gene O can be activated when F is active, solely if gene U is a dominant activator (U OR NOT F). Conversely, the activation of a dominant inhibitor F fully excludes the possibility to activate O (U AND NOT F).

The logical formalism is well-suited to model gene regulatory networks. Indeed, we often lack the quantitative information of exact gene dosage and chiefly refer to a presence/absence of gene activity. Consequently, the activation state of a gene can be easily abstracted as either active (ON) or inactive (OFF). Additionally, the combination of inputs with logical operators enable to model inter-dependencies and context-sensitivity, for example the requirements for multiple activators and/or the exclusion of repressors.

In some cases, for example considering the action of morphogens, the Boolean approximation is too crude. However, functional differences associated to different concentration or activity ranges can be modelled using a generalized logical framework, for example the extension of Boolean logics to multilevel variables proposed by R. Thomas [96, 149, 161]. In the context of this multilevel extension, a gene can be associated with multiple discrete activation levels (usually up to three) if relevant biological information justifies it. A logical rule is then assigned to each of these activation level, thereby defining in which regulatory contexts this level may be attained or maintained.

With the logical formalism, one can precisely infer the activity of a gene as a function of its associated Boolean function and the states of the upstream components. Given this information, it becomes then possible to study the network dynamic behavior with adapted simulation tools.

3.2.5 Dynamical simulation

For logical model of limited size, it is possible to list all the possible states. Each model state corresponds to one specific combination of active and inactive levels for all the components (or node) of the model. Computing the total number of states in a model can be achieved with the following formula :

$$\prod_{i=1}^{m} i^{x_i} \tag{3.1}$$

where i represents the number of logical levels in a given node (e.g. 2 for a binary node), m the highest possible number of logical levels reached within the network. i is raised to a power x_i, where x_i represents the total number of nodes able to reach level i.

Given the logical rules and the current level of each component at a given state,

Table 3.1: Updating strategies of logical rules.

updating strategy	principle	properties	tools
synchronous [155]	All components are simultaneously updated.	Unique attractor (stable state or simple cycle)	BoolNet [156], BoolSim [157], GeNeTool [158, 159], GINsim [136–139], The Cell Collective [160]
fully asynchronous	Defined by R. Thomas [161]. Transition only concerns one component at a time, and all the possible orders of component updating are considered.	Multiple possible attractors (stable states and/or simple and more complex cycles	BoolNet [156], GINsim [136–139]

one can compute all the possible component changes. Based on these trends, two main updating strategies are used: the synchronous and fully asynchronous strategies (cf. Table 3.1).

Using one of these updating strategies, one can compute a State Transition Graph (STG), where nodes represent states of the model, whiles arcs represent transitions enabled by the logical rules (Fig. 3.4). This graph is of particular interest to computationally explore the attractors of the model, in which the system may be trapped with no possibility to escape. Attrators can take the form of a single stable state, with no further possible transitions. It can also be formed by cycling transitions, either within a simple loop, with each node having a single successor, or complex loop, comprising multiple embedded simple loops [162]. These attractors are biologically relevant, as they usually coincide with cellular differentiated states (stable states) or periodical behavior like cell cycle and circadian clock (simple and complex cycles) [94, 154].

Under the synchronous updating assumption, the system will necessarily reach and remain trapped in either a stable state (Fig. 3.4b) or a simple cycle (Fig. 3.4a). Indeed, synchronous updating is a deterministic strategy and therefore only permits at most one single transition in each state. In contrast, the fully asynchronous updating is in general non deterministic and can lead to more complex dynamics. Consequently, it may potentially lead to alternative stable states (Fig. 3.4a) or more attractors, including complex (potentially transient) cyclic behaviours [162].

In the manuscript presented in the next section, I chose to use the asynchronous updating strategy, much more realistic from a biological point of view, in the absence of synchronicity constraints. Interestingly, asynchronous simulations can be refined by considering time delays [161] or probabilistic transition rates [95]. To build my model, I used the software GINsim [136–139], developed in my host team, which ease the encoding and the in-depth analysis of the dy-

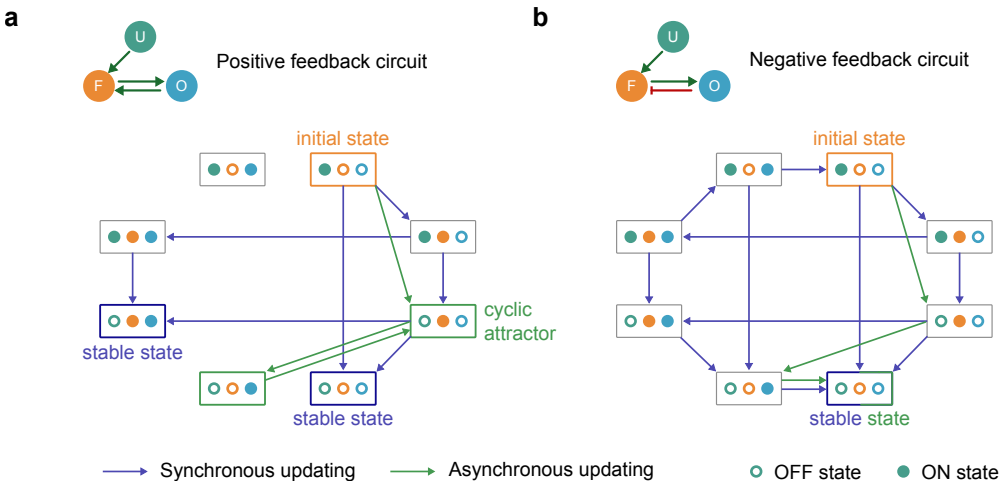

Figure 3.4: State transition graphs. State transition graphs obtained for minimal positive (a) and negative (b) feedback circuits, starting with a temporary activation of gene U as initial state (orange box). Each state of the model is represented as a box (node), comprising the state of each individual component (filled and empty circle for ON and OFF state, respectively). The circle colours correspond to the color of the U (green), F (orange) and O (blue) genes shown in the top schematic. Synchronous and asynchronous updating are shown as blue and green arrows, respectively. In the positive feedback circuit (a), synchronous updating can reach two possible stable states, including one with both genes F and O activated. Conversely, synchronous updating only reaches a single cyclic attractor, transiting between the activation of F and O. In the negative feedback circuit (a), both updating lead to a single stable state where all the nodes are turned OFF, although asynchronous updating reaches a much wider range of different states.

namics and structures of regulatory networks. Its memory-efficient implementation and its palette of tools allow for the analysis of large networks, often limited by the issue of combinatorial explosion (cf. equation 3.1).

3.3 Contribution to the published work

In collaboration with the Lepage lab (iBV, Nice), and with Céline Hernandez and Aurélien Naldi in my team, my contributions focused on the delineation of the Gene Regulatory Network, on its dynamical modelling, and on the analysis of the simulation results.

First,using the software GINsim and relying on an extensive analysis of the literature, as well as on recent *in-situ* experimental data generated by Maria Dolores Molina, I built an extensive regulatory network of the main TGF-β signalling pathways and regulatory components driving early embryonic dorsal-ventral specification in the sea urchin *Paracentrotus lividus*.

Next, I used the stochastic simulation software MaBoSS to further explore the differences in likelihoods to reach alternative expression patterns.

In a third step, I used the multi-cellular logical modelling software EpiLog to simulate the changes in gene expression and signalling activities simultaneously in the different territories of the ectoderm, explicitly taking into account inter-cellular interactions.

Lastly, I participated in the drafting and rewriting of all manuscript sections, as well as in the design and the generation of all the figures.

3.4 Deciphering and modelling the TGF-β signalling interplays specifying the dorsal-ventral axis of the sea urchin embryo

Swann Floc'hlay[1], Maria Dolores Molina[2*], Céline Hernandez[1*], Emmanuel Haillot[2], Morgane Thomas-Chollier[1,3], Thierry Lepage[2⋈] and Denis Thieffry[1⋈]

1. Institut de Biologie de l'École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France.
2. Institut Biologie Valrose, Université Côte d'Azur, Nice, France.
3. Institut Universitaire de France (IUF), 75005 Paris, France.

* equal contributions ; ⋈ corresponding authors

3.4.1 Abstract

During sea urchin development, secretion of Nodal and BMP2/4 ligands and their antagonists Lefty and Chordin from a ventral organizer region specifies the ventral and dorsal territories. This process relies on a complex interplay between the Nodal and BMP pathways through numerous regulatory circuits. To decipher the interplay between these pathways, we used a combination of treatments with recombinant Nodal and BMP2/4 proteins and a computational modelling approach. We further developed a logical model focusing on cell responses to signalling inputs along the dorsal-ventral axis, which was extended to cover ligand diffusion and enable multicellular simulations.

Our model simulations accurately recapitulate gene expression in wild type embryos, accounting for the specification of the three main ectodermal regions, namely ventral ectoderm, ciliary band and dorsal ectoderm. Our model further recapitulates various mutant phenotypes. Temporal analysis revealed the dominance of the BMP pathway over the Nodal pathway, and suggested that the rate of Smad activation governs D/V patterning of the embryo. These results indicate that a mutual antagonism between the Nodal and BMP2/4 pathways is the fundamental mechanism driving early dorsal-ventral patterning.

3.4.2 Introduction

During embryonic development, cell fate is specified by transcription factors activated in response to instructive signals. Regulatory interactions between signalling molecules and their target genes form networks, called Gene Regulatory Networks (GRN) [1].

Deciphering such GRNs is a key for developmental biologists to understand how information encoded in the genome is translated into cell fates, then into tissues and organs, and how morphological form and body plan can emerge from the linear DNA sequence of the chromosomes [2]. Noteworthy, the gene regulatory network that orchestrates morphogenesis of the ectoderm along the dorsal-ventral axis of the embryo of the model sea urchin *Paracentrotus lividus* has started to be uncovered in great de-

tail [3–10].

The ectoderm of the sea urchin larva is constituted of two opposite ventral and dorsal territories, separated by a central ciliary band (Fig. 3.5a):

The ventral ectoderm is the territory at the centre of which the mouth will be formed. Specification of the ventral ectoderm critically relies on signalling by Nodal, a secreted growth factor of the TGF-β family. *nodal* expression is turned on by maternal factors, while Nodal stimulates and maintains its own expression through a positive feedback circuit. Nodal is zygotically expressed and is thought to dimerise with another TGF-β ligand maternally expressed called Univin [8]; the Nodal-Univin heterodimer promotes Alk4/5/7 signalling and the activation of Smad2/3 together with Smad4. The ventral ectoderm boundary is thought to be positioned by the activity of the product of the Nodal target gene *lefty*, which prevents the expansion of *nodal* expression out of the ventral ectoderm region via a diffusion-repression mechanism [11–14].

The ciliary band ectoderm is a proneural territory located between the ventral and dorsal ectoderm [15]. The ciliary band is made of prototypical cuboidal epithelial cells and runs along the arms of the pluteus larva. Unlike specification of the ventral and the dorsal ectoderm, which actively requires TGF-β signalling, specification of the ciliary band tissue does not rely on Nodal or BMP signalling, and this tissue develops as a "default" state of the ectoderm in the absence of these signals.

The dorsal ectoderm is the territory that will differentiate into the apex of the pluteus larva. Its specification relies on the diffusion of the ventrally synthesized protein BMP2/4, which promotes dorsal fates by activating phosphorylation of Smad1/5/8 via the activation of the BMP type I receptors Alk1/2 and Alk3/6. The inhibition of BMP signalling on the ventral side and the translocation of BMP2/4 to the dorsal side requires the product of the *chordin* gene, which is activated in the ventral ectoderm downstream of Nodal signalling [16]. Glypican5 is expressed downstream of BMP2/4 signalling and contributes to stabilize BMP signalling on the dorsal side by a positive feedback circuit. In addition, the BMP ligands Admp1 and Admp2 provide robustness to signalling fluctuations of BMP through an expansion-repression mechanism and autoregulation [4, 17–23]. This mechanism, which relies on the transcriptional repression of *admp1* expression by BMP2/4/ADMP2 signalling, allows *admp1* expression to increase and ADMP1 protein to be shuttled to the dorsal side by Chordin when BMP signalling decreases. Thus, an increase in *admp1* expression compensates for the reduction of the intensity of BMP signalling.

One prominent feature of the D/V onset specification is that it relies extensively on the maternal inputs Panda and Univin, which respectively represses and promotes ventral fate. Univin is a TGF-β related to Vg1 and GDF1/3 signalling through the Nodal/Activin receptors. Panda is a secreted factor structurally related to members of the TGF-β superfamily presumed to repress ventral fate by a still unidentified mechanism. Finally, the transcriptional repressor Yan/Tel acts as a negative regulator of *nodal* expression, whose function is required downstream of Panda to restrict *nodal* expression to the ventral side [24].

Previous studies have shown that Nodal produced by the ventral ectoderm is

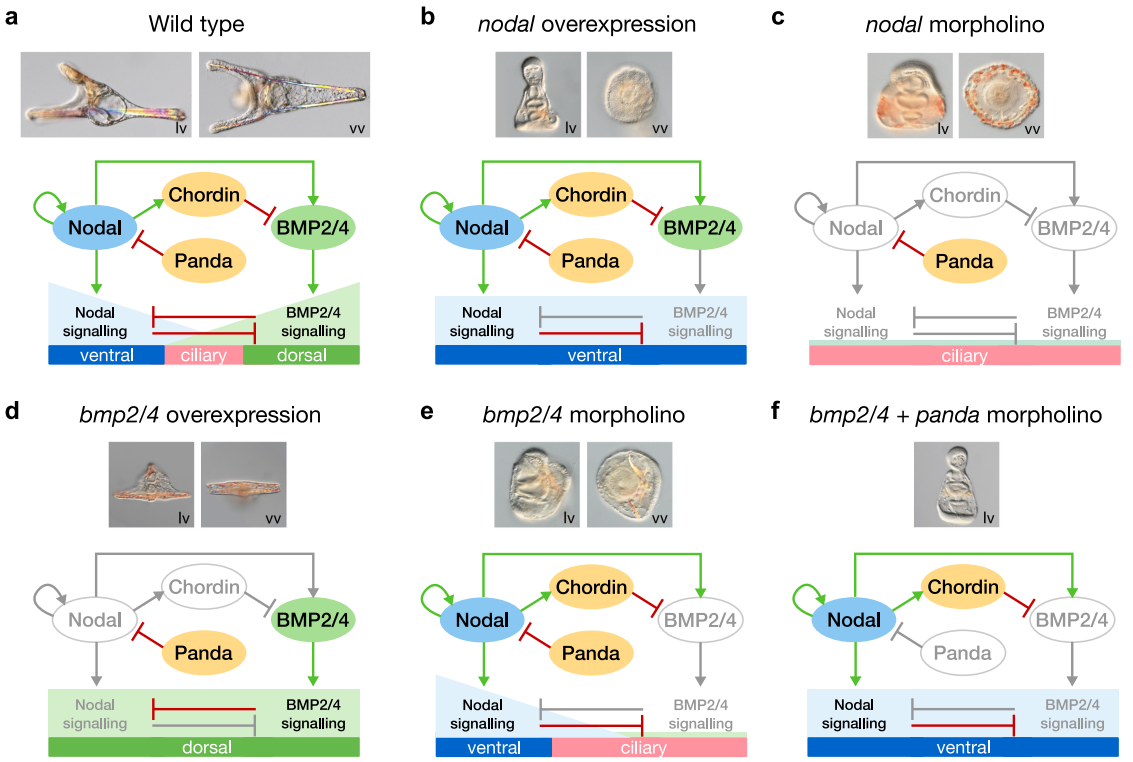

Figure 3.5: Panda, Nodal and BMP2/4 signalling directs patterning of the Dorsal/Ventral axis of the sea urchin embryo. Summary of the morphological phenotypes and identity of the ectodermal territories of wild-type (control) embryos and embryos following perturbations of Nodal or BMP signalling. a: In control embryos, the balance between Nodal and BMP signalling patterns the ectoderm in three main territories: Nodal signalling specifies the ventral ectoderm, BMP2/4 signalling specifies the dorsal ectoderm, while a ciliary band develops at the interface between them. b: The whole ectoderm differentiates into ventral territory when nodal is overexpressed. c: Both Nodal and BMP2/4 signalling are absent in Nodal morphants, which gives rise to an expanded large ciliary band. d: Following BMP2/4 overexpression, all the ectoderm acquires dorsal identity. e: In contrast, after BMP2/4 inhibition, ventral territories are not perturbed but an ectopic ciliary band develops in place of the presumptive dorsal ectoderm. f: Simultaneous perturbation of both the TGF-β related factor Panda and BMP2/4 signalling allows the expansion of Nodal signalling to the whole territory resulting in the ventralisation of the ectoderm. The genes, proteins or interactions that are inactive following each perturbation are denoted in light grey. Activation and inhibition interactions are respectively shown by green and red arrows. lv, lateral view. vv, vegetal view.

a strong ventralising signal. Overexpression of *nodal* causes all ectodermal cells to adopt a ventral fate [4, 9, 25] (Fig. 3.5b), whereas a loss of Nodal function prevents specification of both ventral and dorsal fates and causes the ectoderm to differentiate as a ciliary band (Fig. 3.5c). Conversely, the activity of BMP2/4 protein promotes dorsalisation, and overexpression of BMP2/4 forces all ectoderm cells to adopt a dorsal fate [3, 4, 16] (Fig. 3.5d). In contrast, removing the function of the BMP2/4 ligand from fertilization on prevents specification of dorsal fates, leading to formation of an ectopic ciliary band territory in the dorsal region (Fig. 3.5e). Additionally, knocking down *panda* expression in this BMP2/4 loss-of-function experiment enables *nodal* to be expressed through the dorsal side of the ectoderm and promotes ventral fates in all ectodermal cells (Fig. 3.5f). Conversely, a local *panda* overexpression specifies the D/V axis by promoting dorsal fates, suggesting that *panda* is suffi-

cient to break the radial symmetry of the embryo and necessary to specify the D/V axis. The BMP and Nodal ligands thus show strongly antagonistic activities. However, the mechanism underlying this antagonism and the resulting cell fate decision still awaits clarification.

Due to the largely non-cell autonomous nature of the D/V GRN and to the many events of protein diffusion and feedback circuits involved, an intuitive understanding of the logic of the network is hard to obtain. For example, Nodal and BMP2/4 are co-expressed in the ventral territory, but active signalling pathways are located at opposite poles of the D/V axis. In this context, a model of the D/V gene regulatory network is very useful to formalize the complex regulatory interactions at stake [26].

A Gene Regulatory Network can be modelled as a static regulatory graph with standardized annotations to represent molecular interactions between key regulator components [1]. This regulatory graph can be supplemented with threshold levels and regulatory rules to obtain a predictive, dynamical model [27–30].

In the present study, we built a logical model (i.e. using Boolean algebra for the regulatory rules) of the sea urchin dorsal-ventral specification GRN to (i) assess its accuracy, (ii) compare the model predictions for different perturbations with the observed gene expression patterns, (iii) explore the dynamics of the system, and (iv) develop a multicellular framework to test the ability of the model to generate the spatial patterns observed in wild type or perturbed embryos.

3.4.3 Results

Model building

We constructed a model of the GRN driving the D/V patterning of the sea urchin ectoderm. We started by compiling experimental data to identify the key genes and regulatory interactions (Fig. 3.6). The raw data that provided the spatial and temporal expression information to build the model were derived from high resolution in situ hybridization analysis, Northern blot experiments and systematic perturbations experiments. Loss-of-function experiments via morpholino injections are particularly important for GRN reconstruction since they allow to test if a gene is required for activation of another gene. Gain-of-function experiments via mRNA injection were also used in many instances to test for the ability of a given gene to induce another gene when overexpressed.

Based on these data, we first built a regulatory graph representing the D/V GRN in a single cell, using the software GINsim (ginsim.org) [31]. This construction is an iterative process: the model is subjected to simulations that are confronted to experimental data, and the network and regulatory rules are progressively refined until the simulations qualitatively recapitulate the experimental observations (Fig. 3.6).

We present here the final model and the simulations (cf. section below) that recapitulate the patterns observed experimentally. The model encompasses a total of 31 nodes, linked by 25 activations and 16 inhibitions (Fig. 3.7). The nodes included into the model correspond to signalling and regulatory components, while the signed arcs denote regulatory effects between these components. Sig-

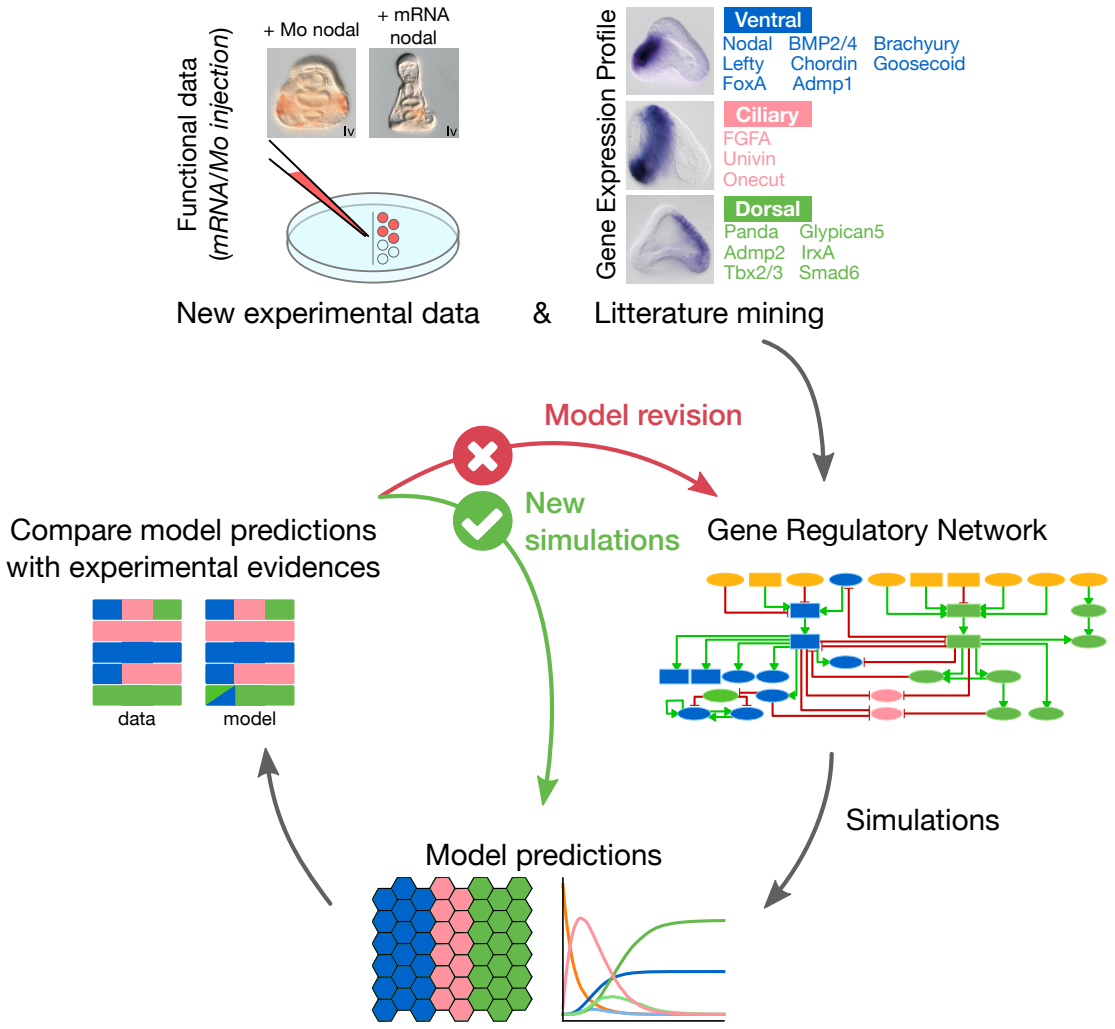

Figure 3.6: Iterative integration of biological data into the GRN model. The GRN model has been built through an iterative process. A first version based on literature curation and experimental evidence was set and then simulated in wild-type and perturbation conditions. The results of wild-type and mutant simulations were systematically compared with experimental results. In case of discrepancy, the regulatory graph and the logical rules were refined, and the behaviour of the model was then re-examined through the same process.

nalling factors are modelled as input nodes (in yellow in Fig. 3.7), providing activating or inhibiting signals through the corresponding membrane receptors.

Each non-input node is classified as ventral (eleven nodes shown in blue in Fig. 3.7), ciliary band (two nodes shown in pink in Fig. 3.7) or dorsal (eight nodes shown in green in Fig. 3.7), according to the reported location and time of activation in the presumptive ectodermal regions. For example, *goosecoid* is activated by the Nodal cascade in the ventral region in wild-type condition, and is thus considered as a ventral gene. The model encompasses the main regulatory components of TGF-β signalling pathways, including the ligands, negative regulators such as the proteins that trigger receptor degradation, downstream transcription factors, and antagonists. Each node of the model is annotated with textual explanations and database links (in particular to PubMed) documenting our modelling assumptions (main references include the GRN diagram published for

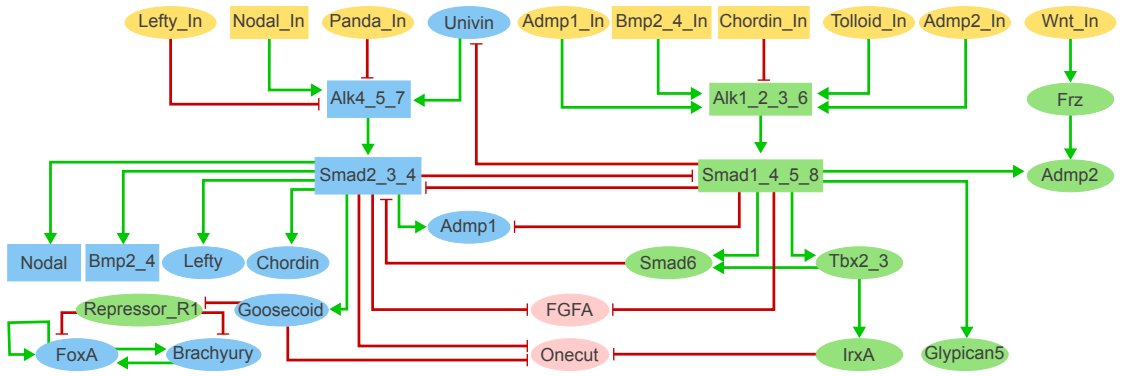

Figure 3.7: Logical model integrating the main signalling pathways controlling specification the dorsal-ventral axis in the embryo of the sea urchin *Paracentrotus Lividus*. Relying on a logical formalism, this model was defined and analysed using the software GINsim. Green and red arrows represent activations and repressions, respectively. Ellipsoid and rectangular components represent Boolean and multivalued nodes, respectively. The components in yellow correspond to model inputs. Internal components are coloured according to their domain of expression along the dorsal-ventral axis, i.e. dorsal (green), ventral (blue) or ciliary (pink) regions.

Paracentrotus lividus in Haillot *et al.* [3], Lapraz *et al.* [4, 32], Range *et al.* [8] and Saudemont *et al.* [9]).

Among the 31 components of the model, 22 are associated with Boolean variables (ellipsoid nodes in Fig. 3.7, taking the values 0 or 1 depending on their activation state), while the remaining components are associated with multilevel variables (rectangular nodes in Fig. 3.7, associated with three or four integer levels, from zero to 2 or 3, see below). The nine input nodes (shown in yellow in Fig. 3.7) define 2,304 possible input value configurations. Using the Java library bioLQM [33], we identified 1,258 stable states, which can be split into three main patterns based on the active nodes: 456 ventral, 450 ciliary and 352 dorsal patterns (cf. Jupyter notebook).

An antagonism between the Nodal and BMP2/4 pathways drives allocation of cell fates along the dorsal-ventral axis

A key feature of the D/V GRN is the strong antagonistic activities of BMP2/4 and Nodal. To correctly account for this aspect in the model, additional experiments were conducted to better characterize the underlying mechanisms. We first tested whether difference in relative intensity between both pathways could favour the establishment of one cell fate over the other.

Treatment with an intermediate dose of BMP2/4 protein resulted in embryos developing with a straight archenteron, no mouth, and covered with a ciliary band like ectoderm (Fig. 3.8a), which is a prototypical Nodal loss-of-function phenotype. Similarly, at intermediate doses of Nodal, the embryos developed with a reduced apex, a phenotype resembling the BMP2/4 loss-of-function phenotype.

These observations suggest that ectodermal cells receive both antagonistic ventralising Nodal and dorsalising BMP2/4 signals, and integrate them even at intermediary doses at the level of the *cis*-regulatory sequences of their target genes. However, since these treatments were performed soon after fertilisation, it was not clear whether the outcome was reflecting an antagonism between Nodal and BMP occurring during cell

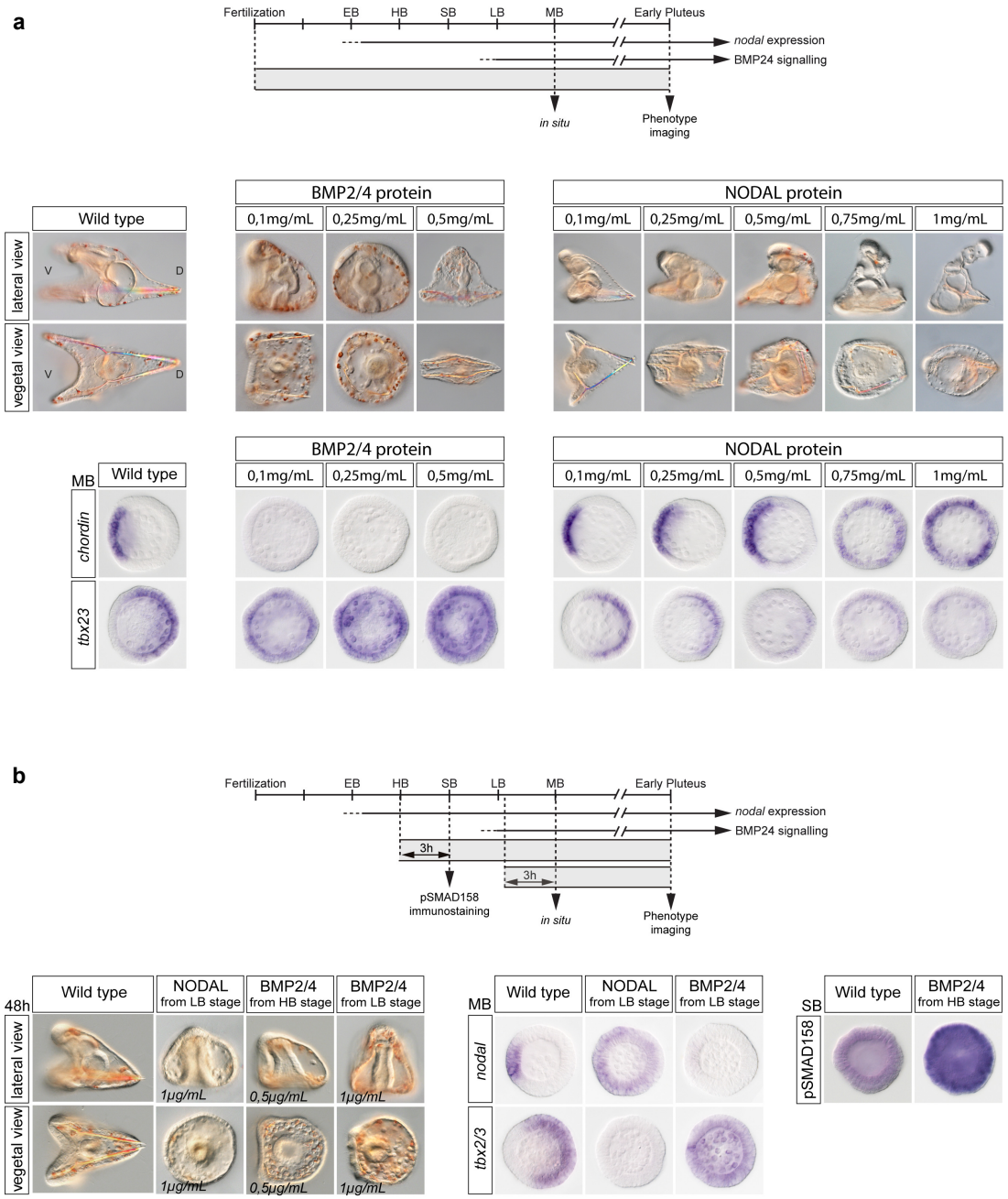

Figure 3.8: BMP2/4 and Nodal signalling antagonize each other to pattern the D/V axis of the sea urchin embryo. a: Continuous treatments at increasing concentrations with Nodal or BMP2/4 protein from the fertilized egg stage. Treatments with increasing concentrations of BMP2/4 protein progressively dorsalise the embryo at the expense of the ventral territories, as reflected by the expansion of the expression of the dorsal marker *tbx2/3* and the repression of the expression of the ventral marker *chordin*. On the other hand, treatments with increasing concentrations of Nodal protein gradually ventralises the embryo at the expense of the dorsal territories as reflected by the gradual expansion of the expression of the ventral marker *chordin* and the repression of the expression of the dorsal marker *tbx2/3*. a: Three-hour Nodal or BMP2/4 protein treatments at late blastula and hatching blastula stages are sufficient to cross-antagonise each other signalling pathway. Three-hour Nodal protein treatment at late blastula stage results in rounded-shaped embryos partially ventralised that overexpress the ventral marker *nodal* at the expense of the dorsal marker *tbx2/3*. Complementary, three-hour BMP2/4 protein treatment at hatching or late blastula stages promotes massive pSMAD1/5/8 signalling and results in partially dorsalised embryos that overexpress the dorsal marker *tbx2/3* at the expense of the ventral marker *nodal*. EB, Early Blastula; HB, Hatching Blastula; SB, Swimming Blastula; LB, Late Blastula; MB, Mesenchyme Blastula; V, Ventral; D, Dorsal.

fates allocation, or if they were only the consequence of one pathway being activated early and dominantly in all cells of the embryo following the injection of mRNA into the egg.

To address this issue, we repeated the Nodal and BMP2/4 treatment at late blastula or early mesenchyme blastula stage. Treatments with Nodal or with BMP2/4 proteins at late blastula radialised the embryos by respectively inducing ventral or dorsal fate in all ectodermal cells (Fig. 3.8b). These results confirm that the Nodal and BMP2/4 pathway act antagonistically also during the fate specification phase and that a competition based on dosage rather than on time of activation is driving the regulation. In order to take into account these results , we paid a particular attention to the encoding of this antagonism in our model.

First, we associated nine nodes of the model with ternary or quaternary variables (rectangular nodes in Fig. 3.7, taking values from 0 to 2 or from 0 to 3, respectively). These multivalued nodes allow for a more fine-grained encoding of the activation states of key morphogens and downstream components whose effects are dose-sensitive (nodes Nodal_In, Chordin_In, BMP2_4_In, Alk4_5_7, Alk2_3_6, Smad2_3_4, Smad1_4_5_8 in Fig. 3.7).

Second, the antagonism between the Nodal and BMP pathways is encoded in the model in the form of a double reciprocal inhibition between Smad2_3_4 and Smad1_5_8_4 (Fig. 3.7), which implements the competition of these signalling complexes for the shared molecule Smad4. Note that each of these two inhibitory interactions can be counteracted by an increased activity of the other antagonistic pathway, following the dose-dependent competition hypothesis (this is encoded in the corresponding logical rules, see Table 3.2).

Model stable states match experimental wild-type and morphants phenotypes

To test our model, we ran simulations and compared the resulting stable states with the wild type phenotypes observed experimentally. We applied different sets of values for the inputs nodes, each set corresponding to a specific territory of the ectoderm (Table 3.2 and Materials and methods). Using proper combinations of active inputs, the model returns stable state(s), which are then compared with the list of marker genes expected to be expressed in the corresponding territory, based on *in-situ* hybridization experiments.

We first ran simulations using initial states corresponding to early 32-cell stage embryos signalling (Fig. 3.9a), preceding the later blastula stage. This stage corresponds to the onset of specification of the ventral organiser, which forms at the opposite side of the gradient of Panda mRNA [3]. This pattern is correctly recapitulated by the stable states obtained from the wild-type simulations (Fig. 3.9b-c).

The resulting stable states were then used to specify the initial conditions reflecting a later blastula stage of embryogenesis, after diffusion and shuttling of maternal factors have taken place (Fig. 3.9b). Indeed, as multiple diffusion events occur, some model inputs expressed in one territory are active in a broader region for blastula simulations (Fig. 3.10a).

After some iterative refinements of the rules, model simulations qualitatively recapitulated the expression patterns expected for each individual territory

a - Input nodes	Value ventral	Value ciliary	Value dorsal
Nodal_In	2	1	1
Lefty_In	1	1	1
Panda_In	0	0	0
Admp1_In	1	1	1
Bmp2_4_In	1	1	1
Chordin_In	1	1	1
Tolloid_In	0	0	1
Wnt_In	0	0	0
Admp2_In	0	0	0

b - Internal nodes	Value	Logical rule	Expression	Initial state
Univin	1	!Smad2_3_4:2	Ventral	1 (basal)
Alk4_5_7	1	(Nodal_In:1 & !Lefty_In & Univin & !Panda) | (Nodal_In:2 & Univin & !Panda)	Ventral	0
Alk4_5_7	2	Nodal_In:3 & Univin	Ventral	0
Smad_2_3_4	1	Alk4_5_7:1 & !Smad1_4_5_8 & !Smad6	Ventral	0
Smad_2_3_4	2	Alk4_5_7:2	Ventral	0
Nodal	2	Smad2_3_4	Ventral	0
Bmp2_4	1	Smad2_3_4:1	Ventral	0
Bmp2_4	2	Smad2_3_4:2	Ventral	0
Lefty	1	Smad2_3_4	Ventral	0
Chordin	1	Smad2_3_4	Ventral	0
Goosecoid	1	Smad2_3_4	Ventral	0
Repressor_R1	1	!Goosecoid	Ventral	0
FoxA	1	(FoxA | Brachyury) & !Repressor_R1	Ventral	0
Brachyury	1	!Repressor_R1 | FoxA	Ventral	0
Alk1_2_3_6	1	((Admp2_Trans & !Chordin_In) | (Admp2_Trans & Tolloid_In & !Chordin_In:2)) & !Bmp2_4_In:2	Dorsal	0
Alk1_2_3_6	1	((Bmp2_4_In:1 & Admp1_In & !Chordin_In) | (Bmp2_4_In:1 & Admp1_In & Tolloid_In & !Chordin_In:2)) & !Bmp2_4_In:2	Dorsal	0
Alk1_2_3_6	2	Bmp2_4_In:2 & !Chordin_In:2	Dorsal	0
Smad1_4_5_8	1	Alk1_2_3_6:2 & Smad2_3_4:2	Dorsal	0
Smad1_4_5_8	2	(Alk1_2_3_6:2 & !Smad2_3_4:2) | (Alk1_2_3_6:1 & !Smad2_3_4)	Dorsal	0
Tbx2_3	1	Smad1_4_5_8:2	Dorsal	0
IrxA	1	Tbx2_3	Dorsal	0
Smad6	1	Tbx2_3 | Smad1_4_5_8:2	Dorsal	0
Glypican5	1	Smad1_4_5_8:2	Dorsal	0
Frz	1	Wnt_In	Dorsal	0
Admp2	1	Frz | Smad1_4_5_8:2	Dorsal	0
FGFA	1	!Smad2_3_4 & !Smad1_4_5_8:2	Ciliary	0
Onecut	1	!(IrxA | Goosecoid | Smad2_3_4)	Ciliary	0

Table 3.2: Logical rules of the unicellular model. Logical rules are used to define the behaviour of each node, for each territory, relative to its direct upstream regulatory nodes. Input nodes (a) do not have any assigned rule, as they are set to a given fixed value specific for each territory when performing simulations. Internal nodes (b) have a logical rule assigned for each possible level they can converge to and an initial state. All nodes have a basal level of 0 (inactive), except Univin (basal level 1), as it tends to be ubiquitously active without the need for activators. The logical rules combines literals, each representing the activity of one node, with the Boolean operators OR ("|"), AND ("") and NOT ("!").

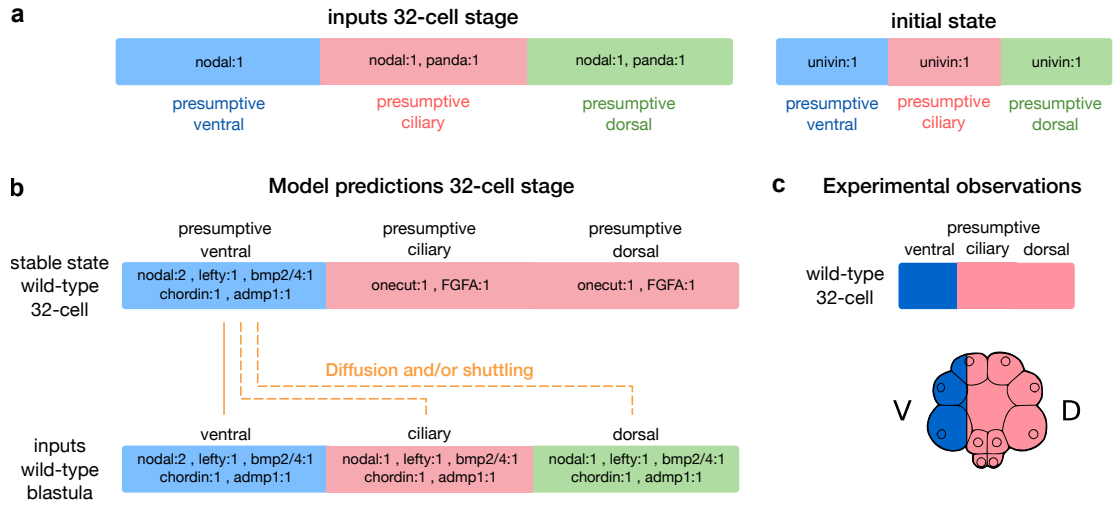

Figure 3.9: Simulation of early 32-cell stage and specification of later stage inputs. By restricting the active input nodes to combinations of Nodal and Panda (a), our unicellular model recapitulates the patterns observed in the 32-cell stage embryos (a upper part, c). In wild-type condition, panda is expressed in the presumptive dorsal region and restricts nodal expression to the presumptive ventral region [3]. The stable states resulting from our ventral wild-type simulation (b upper part) were then used to define the input node values for the simulation of later developmental stages, taking into account the diffusion and shuttling events known to occur from the ventral region to further dorsal territories in this developmental time window (b lower part). V, ventral; D, dorsal.

(ventral, ciliary, dorsal) (Fig. 3.10b). Hence, we can conclude that, the regulatory graph shown in Fig. 3.7 supplemented by the logical rules of Table 3.2 are sufficient to specify the three main ectodermal D/V patterns of the sea urchin embryo.

To further validate and explore the properties of our model, we simulated loss- or gain-of-function experiments (mRNA or morpholino injection) by restricting the range of reachable levels for one or multiple node(s), e.g. to zero for a loss-of-function, or to a higher value for a gain-of-function (cf. Material and methods). As in the wild-type conditions, we assessed the relevance of our model by comparing the resulting stable states with the patterns observed in the corresponding *in-situ* experiments following mRNA or Morpholino injection in the embryo at early stage. For seven of the eight mutants simulated, the model returned a unique single stable state in each region, which qualitatively matched experimental observations (Fig. 3.10b-d).

In the cases of *nodal* Morphants and of *lefty* mRNA overexpression, the ventral cascade fails to be established, leading to the absence of both Nodal and BMP2/4 pathways, and the presence of a default ciliary state in all the ectodermal cells. Following *nodal* mRNA overexpression, competition between Smad2/3 and Smad1/5/8 for Smad4 creates an advantage for the ventral cascade producing a fully ventralised embryo. The same pattern is obtained for the *lefty* Morpholino, because this perturbation impacts the diffusion-repression mechanism controlled by Lefty [34], enabling *nodal* expression to propagate without restriction [13].

In the case of overexpression of the dorsal fate repressor *chordin* or in the case of BMP morphants, the absence of BMP2/4 signalling fosters a ciliary band state in the presumptive dorsal territory.

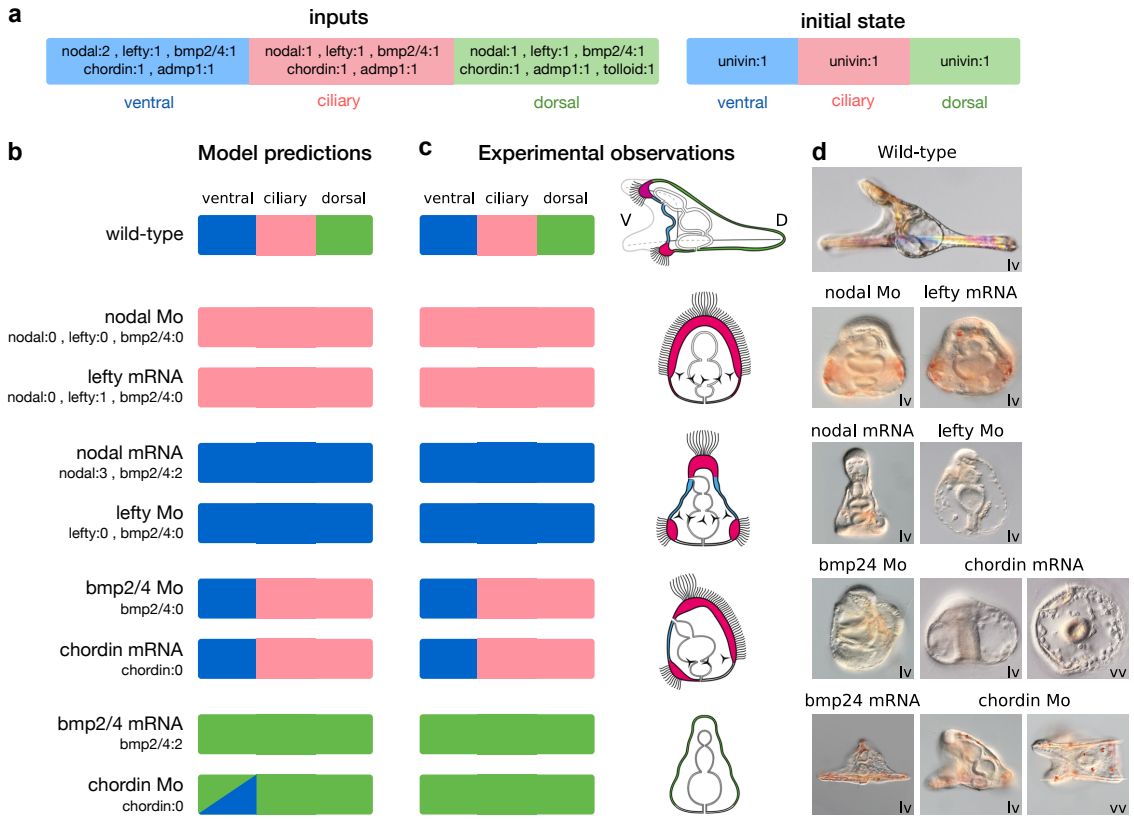

Figure 3.10: Comparison of blastula model simulations and experimental results. With proper logical rules (see Table 3.2), inputs and initial state conditions (a), our model gives rise to different stable patterns (b), which qualitatively match experimental results (c-d). Note that, in the simulation of the chordin morpholino injection, the model predicts two possible stable patterns in the ventral region, whereas experiments point to a mild ventral territory present in these embryos. In the middle area (c), the organisation of dorsal (green), ventral (blue) and ciliary (pink) territories are schematized based on images of wild-type and morphant embryos (d). lv, lateral view; vv, ventral view.

However, as BMP2/4 is not necessary for the expression of *nodal*, the ventral cascade maintains a wild-type expression pattern in these morphants. Finally, following *bmp2/4* overexpression, the competition for the common Smad is driven toward the activation of the dorsal cascade, giving a fully dorsalised ectoderm state.

Interestingly, in the case of the *chordin* morpholino, the model returned two stable states (denoted by the green and blue triangles at the bottom of Fig. 3.10b) in the presumptive ventral region, corresponding to ventral and dorsal fates, respectively. This situation is hereafter further investigated using a probabilistic framework.

Stochastic logical simulation of the *chordin* Morphants

Using stochastic simulations, one can unfold the temporal dynamics of the regulatory network for given initial conditions and estimate the prevalence of any reachable stable state. In the case of the ventral region in *chordin* morphant conditions, we have seen that our model can reach two different stable states, corresponding to ventral and dorsal expression patterns.

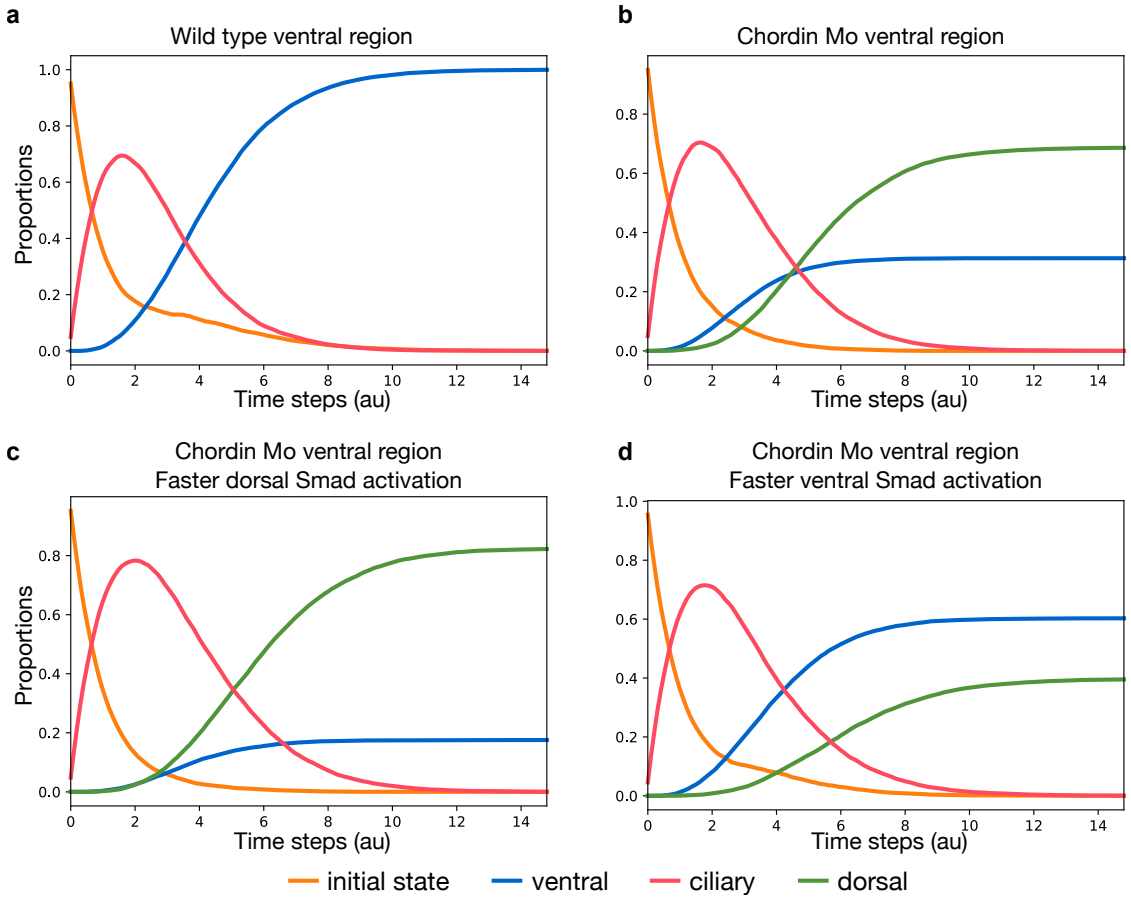

Figure 3.11: Probabilistic time-course simulations of the unicellular model starting with the ventral initial state. Temporal evolutions of the mean activation levels of Goosecoid, Iroquois and Onecut, representing ventral (blue), dorsal (green) and ciliary (pink) phenotypes, respectively. All simulations start from a ventral initial state (orange). The first plot (a) corresponds to the wild type, while the three other ones (b-d) correspond to chordin knock down conditions. Simulations (a-b) were performed with equal up and down state transition rates. Further simulations were performed using rates favouring the dorsal cascade (c), or favouring the ventral cascade (d) (see methods for details).

Using the software MaBoss (maboss.curie.fr) [35], we performed stochastic temporal simulations of our model to generate mean time plots and estimate the probability to reach each of these stable states. In the absence of precise kinetic information, we first used equal rates for all (up or down) state transitions. In the wild type ventral region, as expected, all stochastic simulations gave rise to a ventral expression pattern (Fig. 3.11a).

In contrast, for the *chordin* morphant, the dorsal state is reached about twice as often as the ventral state (Fig. 3.11b). In other words, the dorsal pathway is more likely to win the competition for Smad4. This partial dominance of the dorsal pathway matches the experimental observations of weak dorsal patterns for *chordin* morphants (Fig. 3.10d), presumably resulting from the co-activation of the two antagonistic pathways.

In the *chordin* morphants, BMP signalling goes unrestricted in the ventral ectoderm, promoting dorsal fates and repressing the ventral cascade. However, since Nodal is critically required for *bmp2/4* activation, *nodal* down-regulation in turn leads to the repres-

sion of BMP2/4 signalling. Therefore, in the absence of Chordin, both the ventral and dorsal cascades are activated through feedback circuits. This conclusion is supported by experimental observation of transient Smad1/5/8 signalling and *tbx2/3* expression in the ventral ectoderm [16].

To further assess whether this imbalance in favour of the dorsal pathway activation is sensitive to kinetic (transition) rates, we ran stochastic simulations with different Smad activation rates for the two pathways. An imbalance in favour of the dorsal Smad activation increased the gap between the final proportion of dorsal and ventral stable state compared to wild type (Fig. 3.11c). On the contrary, an imbalance in favour of the ventral Smad activation inverted the relationship, with a higher fraction of ventrally specified states than dorsally specified states, almost mirroring the ratios obtained with equiprobable transition rates (Fig. 3.11d).

In conclusion, the outcome of the competition between the two pathways is strongly sensitive to the kinetic rates for the activation of the different Smads. The pathway specific Smad firstly activated immediately represses the other one by pre-empting Smad4 and thereby fosters the corresponding state stable.

Multicellular simulations emphasize the crucial role of long-range signal diffusion

In the preceding section, we simulated the behaviour of cells of the three different presumptive territories by selecting appropriate combinations of signalling input levels, which were considered as fixed for the whole duration of the simulations.

To model more precisely the production and diffusion of signalling molecules across the ectoderm, we used the software EpiLog (epilog-tool.org) [36], which supports simulations of an epithelium encompassing multiple cells connected through signalling of diffusing elements. The behaviour of each cell of the epithelium is modelled by the same cellular logical model, but levels of input signal directly depend on the signal values output by neighbouring cells. The input signals perceived by a given cell are integrated into logical diffusion rules, and updated synchronously (see Materials and methods and Table 3.3) (Fig. 3.12a). Hence, in general, input levels change over time.

To simulate the wild type condition, we initialised the model with a small concentration of Nodal and Smad2/3/4 in the ventral territory (corresponding to the three left-most range of cells) (Fig. 3.12b). This simulation correctly recapitulated the expected contiguous ventral, ciliary and dorsal ectoderm territories (Fig. 3.12c). This result suggests that the relatively simple diffusion rules properly account for the dynamics of the proteins governing the dorsal-ventral patterning. In addition, this result highlights the crucial role of Nodal to direct specification of the ectoderm along the whole dorsal-ventral axis.

As in the case of the unicellular model, we can apply specific perturbations to assess their impact at the tissue level. As shown in Fig. 3.12c, our multicellular simulations accurately recapitulated the phenotypes of the different morphant patterns observed experimentally. Note that the chordin morpholino has been discarded from these simulations, as it gave rise to two different stable states, which cannot be covered with the Epilog deterministic input updating approach.

Interestingly, in the course of *lefty* mor-

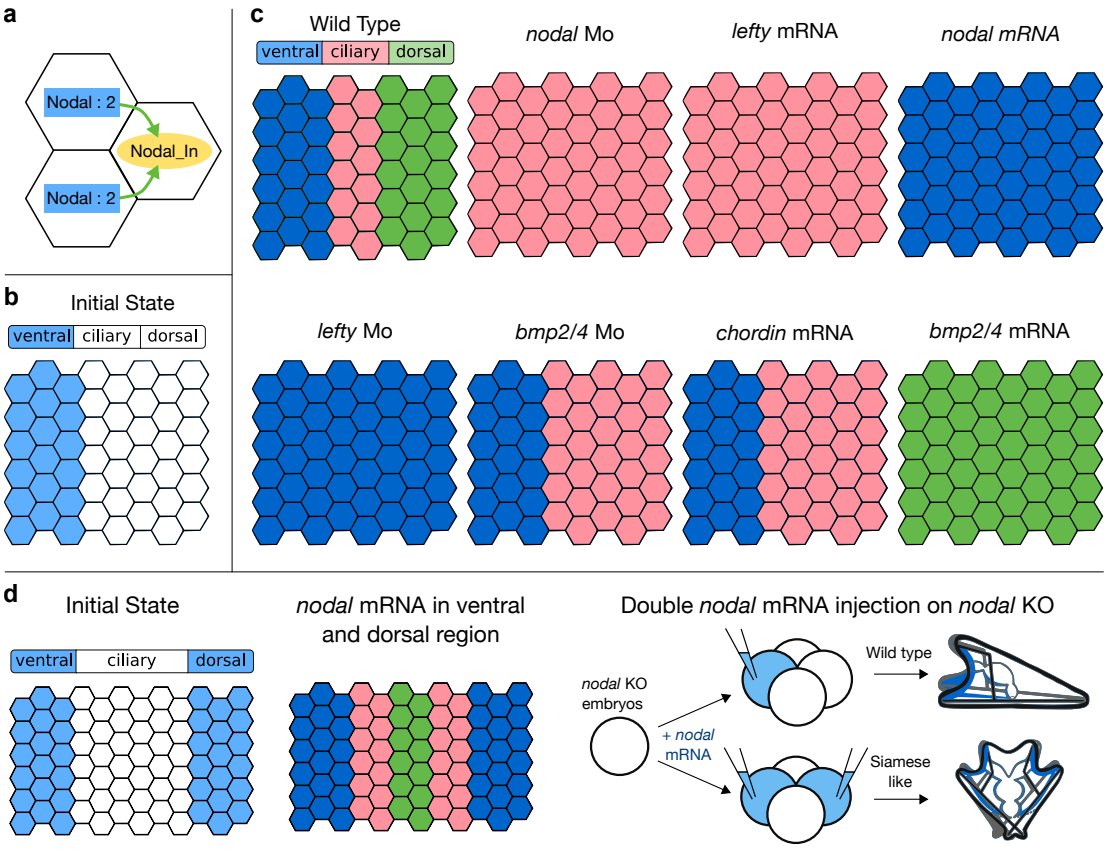

Figure 3.12: Multicellular logical simulations for wild-type and mutant conditions, using the software Epilog. Across the multicellular epithelium, specific logical rules have been defined to model the diffusion of signalling components (a) (see Table 3.3 for diffusion rules). At the initial state, only Nodal is activated in the presumptive ventral territory (b). Multicellular simulation results for the wild-type and morphant conditions (c) qualitatively match our experimental results. Considering a larger epithelium, we further simulated the injection of nodal mRNA in two opposite blastomeres of a four-cell embryo (d), which resulted in an embryo displaying a mirror symmetric pattern of ventral, ciliary and dorsal territories along the dorsal-ventral axis, as observed by Lapraz *et al.* [4].

pholino simulation, we could clearly see a shift in the ventral-ciliary frontier, which progressively moved toward the dorsal side, as Nodal diffused in the absence of Lefty repression, until it reached the fully ventralised stable state (see Fig. 3.13).

Using Epilog, it is further possible to perform local perturbations by modifying the initial levels of one or several signalling molecules at specific epithelium locations. Using this feature, we could recapitulate in silico the results of an experiment reported by [4], who injected *nodal* mRNA into two opposite cells of a 4-cell stage *nodal* knock-down embryo (i.e. following a *nodal* Morpholino injection in the egg).

This experiment triggered the formation of an ectopic, inverted D/V axis and resulted in the development of siamese pluteus larvae with two ventral sides, two ciliary bands and a central dorsal territory. Using Epilog and imposing *nodal* and *smad2/3/4* activity at the initial state in both the ventral and the dorsal side of the epithelium, our spatial logical simulation qualitatively reproduced the siamese pattern observed experimentally (Fig. 3.12d).

Input nodes	Value	Logical diffusion rule	Interpretation
Admp1_In	1	{Admp1[0:],min = 1}	Admp1 takes level 1 if a least one cell expresses Admp1 with no distance restriction
Bmp2_4_In	1	{Bmp2_4[0:], min = 1} & !{Bmp2_4:2[0], min = 1}	Bmp2_4 takes level 1 if at least one cell expresses Bmp2_4 with no distance restriction and the cell does not already express Bmp2_4 at level 2
Bmp2_4_In	2	{Bmp2_4:2[0], min = 1}	Bmp2_4 takes level 2 if the cell already expresses Bmp2_4 at level 2
Chordin_In	1	{Chordin[0:2], min = 1}	Chordin takes level 1 if at least one cell expresses Chordin among the cell itself and its neighbors at a distance equal or lower than two cells
Chordin_In	2	No function	Chordin cannot take level 2 by diffusion
Lefty_In	1	{Nodal:2[0:1], min = 1, max = 2}	Lefty takes level 1 if one or two cell(s) express Nodal at level 2 among the cell itself and its direct neighbors
Nodal_In	2	{Nodal:2[0:1], min = 3}	Nodal takes level 2 if at least three cells express Nodal et level 2 among the cell itself and its direct neighbors
Nodal_In	3	No function	Nodal cannot take level 3 by diffusion
Admp2_In	1	{Admp2[:0],min = 1}	Admp2 takes level 1 if the cell already expresses Admp2
Tolloid_In	*	No function	No Tolloid diffusion
Panda_In	*	No function	No Panda diffusion
Wnt_In	*	No function	No Wnt diffusion

Table 3.3: Logical diffusion rules used in Epilog. Logical rules are used to define the diffusion dynamics perceived by the inputs nodes, depending on the values of the output nodes in the neighbouring cells. Diffusion rules are defined in the format "N:L[D],S", with N as the node emitting diffusing signal, L its required activation level, D the distance range to perceive diffusion and S the minimum and/or maximum number of cell required in this state. For example, the seventh row in the table specifies that cells will have their Nodal input node value converging toward the value 2 if at least three cells are expressing Nodal at a value 2 at a maximum distance of one cell (i.e. among the target cell itself and its direct neighbours).

3.4.4 Discussion

Gene regulatory networks integrate documented interactions between transcription factors, signalling components, and their target genes, which ultimately translate the information encoded into the genome into morphological form and body plan. However, as our delineation of developmental systems progresses, we are facing increasingly large and complex networks, which cannot be fully and rigorously understood without proper formalisation. This is, for example, clearly the case for the GRN governing D/V patterning of the sea urchin embryo, which relies on numerous signalling and regulatory factors, involved in multiple positive and negative feedback circuits.

In our modelling study, several key choices had to be made. As little is known regarding detailed mechanisms and kinetic parameters, we opted for a qualitative, logical formalism. However, to properly model morphogen diffusion and dose-dependent effects, we considered a multilevel extension of the classical Boolean framework. Importantly, in the course of its conception, the model was systematically tested through ex-

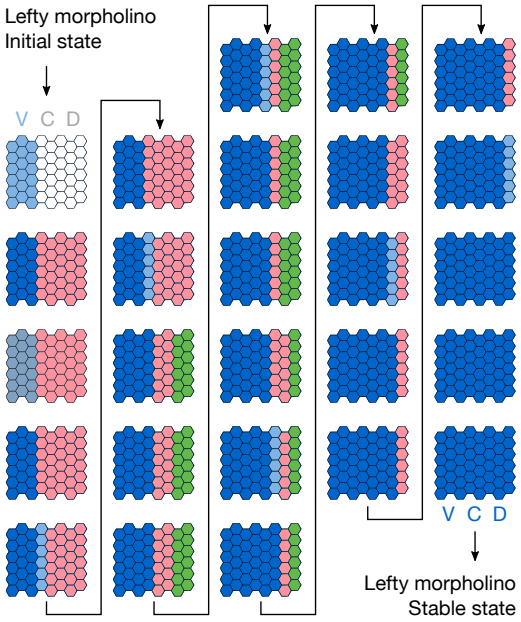

Figure 3.13: EpiLog simulation of the Lefty morpholino condition. Starting from the initial state with the Nodal pathway initiated in the ventral region, the Epilog simulation of the *lefty* morpholino condition ultimately results in a fully ventralized stable state. The activity dynamics of the three markers genes *goosecoid* (blue, ventral), *onecut* (pink, ciliary) and *irxa* (green, dorsal) denotes the progressive shift of the ciliary boundary toward the dorsal side, as the loss of Lefty enables Nodal to diffuse freely outside of the ventral region. V, ventral; C, cilliary; D, dorsal.

tensive simulations of wild-type and perturbed conditions. In wild-type conditions, our unicellular model fully recapitulated each territory pattern independently. We further took advantage of a recent multicellular extension of the logical framework to explicitly simulate spatial pattern formation, whose results can be more easily compared directly with the phenotypes of wild-type and mutant embryos.

A key step in our study was to model the interplay between the Nodal and BMP pathways. In this respect, we were guided by our experiments dealing with the treatment of embryos with recombinant Nodal or BMP2/4 proteins at blastula stage (i.e. after the initial specification of the ventral and dorsal territories). These experiments, which demonstrated that over-activation of one of these pathways is sufficient to abrogate signalling from the other pathway, highlighted the strong antagonism between Nodal and BMP2/4 signalling and suggested that this antagonism resulted from a direct competition between the two pathways activated by these TGF-βs ligands rather than from their timing of activation.

The competition between Nodal and BMP2/4 played a key role in understanding the regulatory dynamics within the *chordin* knock-down experiments, which was the only morphant not fully recapitulated by our model for all three territories. In the case of the *chordin* knock-down, our logical model predicted that both the ventral and dorsal steady states were possible in the presumptive ventral region. Accordingly, in the *chordin* morphant, both the Nodal and the BMP2/4 pathways are activated, antagonising each other. Consequently to this ectopic activation of BMP signalling, the ventral territory in *chordin* morphants displays a transient dorsalisation, before reversing towards a ventral ectoderm fate during gastrulation, as shown by the presence of a mouth opening.

However, this brief dominance of the dorsal fate over the ventral fate in the conditions of Nodal and BMP2/4 co-activation is not well understood. To further explore the underlying regulatory mechanism of this dorsal fate dominance, we performed a stochastic logical simulation of the unicellular model in *chordin* knock-down condition. This analysis resulted in a higher proportion of active dorsal fates over ventral fates, in agreement with the experiments. This result suggests that the transient dorsal dominance is encoded

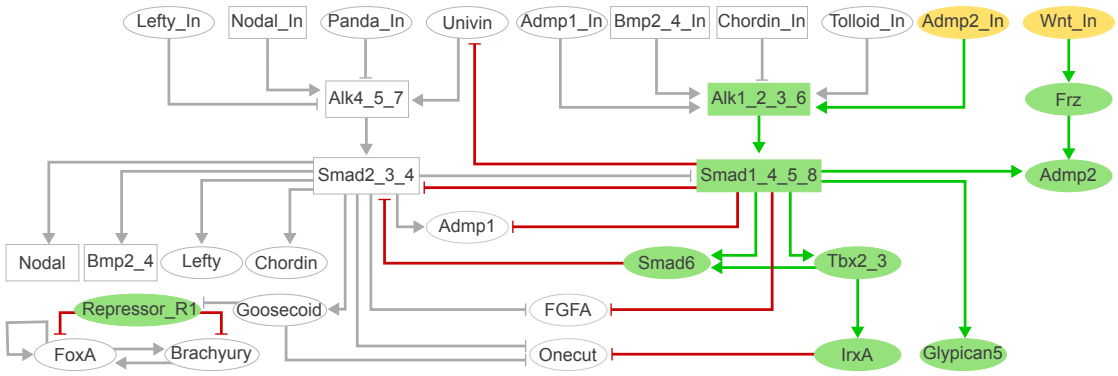

Figure 3.14: Simulation of the boundary ectoderm in the unicellular model. Stable state obtained with the unicellular model when considering Admp2 and Wnt inputs active. Active nodes (yellow for inputs and green for dorsal nodes) and edges (green for activation and red for inhibition) are shown in colour, inactive ones are shown in grey. This stable state corresponds exactly to the dorsal stable state shown in Figure 3.10.

in the structure of the GRN. Indeed, even in the case of the *chordin* morphants, the model accurately recapitulates the conflict caused by the coactivation of the Nodal and BMP2/4 pathways in the ventral ectoderm. However, by modulating the rates associated with the different Smads and performing additional simulation, we showed that the outcome of the competition between the two pathways is strongly sensitive to these rates.

At this point, our model remains limited to the major early dorsal-ventral patterning events occurring in the sea urchin embryo. However, in the future, this model could be tentatively extended to integrate novel data and to explore more refined specification and differentiation events. For example, it could be extended to investigate the specification of the boundary ectoderm region, located at the interface between the ectoderm and endomesoderm, which plays a central role in positioning the clusters of primary mesenchyme cells and spicules patterning [9, 37–41]. This process is known to depend on Wnt signalling, presumably in conjunction with Nodal, BMP2/4 and ADMP2 signalling [4, 9, 41, 42]. With the current unicellular model, the simulation with the input levels corresponding to the boundary ectoderm (i.e. Admp2 and Wnt active) results in a dorsal stable state (Fig. 3.14).

Another possible addition to the model would be the integration of the negative feedback of Smad6 on BMP2/4 pathway [37–41]. Indeed, Smad6 is activated by the dorsal signalling downstream of BMP2/4 ; it buffers the activation of the dorsal pathway by acting as an inhibitor on BMP2/4 signalling. Such a negative feedback circuit typically generates an oscillatory behaviour. In the frame of the Boolean logic, the consideration of this negative feedback circuit would result in a cyclic attractor with alternation of active and inactive BMP and Smad6 activities, which are more difficult to interpret than stable states.

Our logical model focuses on the blastula and gastrula developmental stages of sea urchin embryogenesis. One possible extension would be to further explore the regulatory interactions taking place at earlier stages. In the case of the 32-cell stage, our model correctly recapitulates wild-type pattern mainly driven by Panda expression. Furthermore, the simulation results of Panda

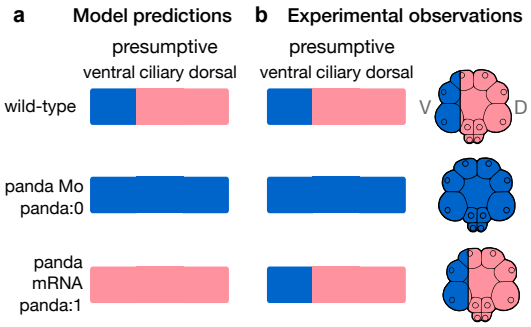

Figure 3.15: Simulation of *panda* perturbations at the 32-cell stage. Starting with a restricted combination of inputs with Nodal and Panda, we simulated the 32-cell stage D/V patterning using the unicellular model. In wild-type condition, model predictions (a) correctly recapitulates experimental observations (b). It is also the case for the simulation of Panda loss-of-function, mirroring the fully ventralised phenotype observed upon *panda* Mo injection. Simulations of *panda* overexpression fully abrogate ventral specification and result in a global ciliary band phenotype (a), whereas experimental evidences show no impact of global overexpression of *panda* on the onset of D/V patterning [3] (b). V, ventral; D, dorsal.

loss-of-function in 32-cell stage conditions mirror the fully ventralised phenotype obtained experimentally [3] (Fig. 3.15).

However, the simulations of Panda overexpression show discrepancies relative to the experimental observations. In this case, our model predicts the ventral region become dorsalized, whereas global injection of *panda* mRNA does not impact the wild-type pattern. Indeed, current models suppose that an asymmetry of *panda* mRNA provides the spatial cue that in turn controls the polarised activation of downstream genes. Therefore, an asymmetry of *panda* mRNA or of Panda protein constitutes the main driving signal to allocate cell fates, rather than a change in overall Panda concentration.

This signalling based on multicellular gradient cannot be currently recapitulated by our unicellular model, as it requires to integrate inputs from multiple surrounding cells and also to rely on relative differences in concentration instead of absolute levels. Further development of EpiLog features could enable us to build logical rules accounting for the specificities of such multicellular gradient signalling.

To conclude, we have shown that logical modelling can capture several salient dynamical features of the GRN governing early dorsal-ventral patterning of sea urchin embryos, including the key role played by intercellular interactions. Such models should therefore be useful to further explore the complex interplay between maternal factors and zygotic genes, which orchestrates patterning of the ectoderm of the sea urchin embryo downstream of intercellular signals. To this end, we provide our models and the Jupyter notebook implementing all our analyses within the CoLoMoTo docker environment (see supplementary materials).

3.4.5 Materials and methods

Animals, embryos and treatments

Adult sea urchins (*Paracentrotus lividus*) were collected in the bay of Villefranche-sur-Mer. Embryos were cultured as described in [43, 44]. Fertilization envelopes were removed by adding 1mM 3-amino-1,2,4 triazole (ATA) 1min before insemination to prevent hardening of this envelope followed by filtration through a 75 µm nylon net. Treatments with recombinant BMP2/4 or Nodal (RD) proteins were performed at the time indicated in the schemes by adding the recombinant protein diluted from stocks in 1mM HCl in 24 well plates containing about 1000 embryos

in 2ml of artificial sea water [16].

Anti-phospho-Smad1/5/8 Immunostaining

Embryos were fixed with 4% formaldehyde for 15 min at swimming blastula stage (3 hours after adding BMP2/4 protein) then briefly permeabilized with methanol. Anti-Phospho-Smad1 (Ser463/465) / Smad5 (Ser463/465) / Smad9 (Ser465/467) from Cell Signalling (D5B10 Ref. 13820) was used at 1/400. Embryos were imaged with an Axio Imager.M2 microscope.

In situ hybridization

In situ hybridization was performed using standard methods [45] with DIG-labelled RNA probes and developed with NBT/BCIP reagent. The *nodal*, *chordin* and *tbx2/3* probes have been described previously [16, 25]. Control and experimental embryos were developed for the same time in the same experiments. Embryos were imaged with an Axio Imager M2 microscope.

Overexpression of mRNAs and morpholino injections

For overexpression studies, *nodal*, *lefty*, *chordin* and *bmp2/4* capped mRNAs were synthesized from NotI-linearized templates using mMessage mMachine kit (Ambion). After synthesis, capped RNAs were purified on Sephadex G50 columns and quantitated by spectrophotometry. *Nodal* , *lefty*, *chordin* and *bmp2/4* mRNAs were injected at $400\,\mu g\,ml^{-1}$, $200\,\mu g\,ml^{-1}$, $1000\,\mu g\,ml^{-1}$ and $500\,\mu g\,ml^{-1}$, respectively. Capped mRNA were injected mixed with Tetramethylrhodamine Dextran (10000MW) at $5\,mg\,ml^{-1}$ [25]. Morpholino oligonucleotides were dissolved in sterile water and injected at the one-cell stage together with Tetramethylrhodamine Dextran (10000MW) at $5\,mg\,ml^{-1}$ as already described [3, 25].

Logical formalism

We built our model using the multilevel logical formalism introduced by R. Thomas [30]. This qualitative approach relies on graph-based representations of the network and of its dynamics. The network is formalized as a *regulatory graph*, where nodes denote molecular species (e.g. proteins), whereas signed arcs denote regulatory interactions, positive or negative.

The nodes can take a limited number of integer values, only two (0 or 1) in the simplest, Boolean case, but potentially more when biologically justified, for example in the case of morphogens with clearly distinct activity ranges. Hence, each regulatory arc is associated with an integer threshold, always 1 in the Boolean case, but potentially higher in the case of a multilevel regulator.

Logical rules further specify how each node reacts to the presence or absence of the corresponding incoming interactions. Specific (non-overlapping) Boolean rules are defined for each value of each node. Boolean rules are built by combining literals (i.e. valued component) with the logic operators AND (denoted "&"), OR (denoted "|") and NOT (denoted "!").

Table 3.2 lists the formula associated with the different components of our model. Note that the formula associated with zero values are omitted, as they can be computed directly as the complement of the formulae defined for the other values for a given node.

For example, the formula of the node FoxA :

$$FoxA \rightarrow 1 \equiv (FoxA|Bry)\,\&\,!R1 \quad (3.2)$$

can be translated into "FoxA node tends toward the value 1 if and only if FoxA or Brachyury (Bry) are active and Repressor_R1 ($R1$) is not active". In this example, the regulatory actions from Brachyury to FoxA and from Repressor_R1 to FoxA correspond respectively to an activation and an inhibition.

The levels of the input (unregulated) nodes are defined at the start of simulations. Using the Boolean rules of Table 3.2, we can simulate the behaviour of the system for different input value combinations. In this respect, we use the asynchronous updating approach proposed by R. Thomas [46], which consists in following all the different possible single unitary changes of values induced by the rules.

The dynamical behaviour of the model is generally represented in terms of a *state transition graph*, where each node represents a discrete state of the model (i.e. a vector listing the values of the different components of the model), whereas each arc represents a state transition.

In this work, we took advantage of the implementation of this logical formalism into the user-friendly Java software suite GINsim (version 3.0, see ginsim.org [47]). In our analyses, we particularly focused on stable states (see e.g. Fig. 3.9b), which typically denote cell differentiation states. These can be directly computed (i.e. without unfolding the state transition graph) using a very efficient algorithm implemented in GINsim [48].

Wild type simulation

We simulated the behaviour of each dorsal-ventral region independently, considering different sets of values for the input nodes in the ventral, ciliary and dorsal presumptive territories. These sets of input values were defined based on previously published results (see Results section).

As we simulate each territory individually, the unicellular model cannot directly take into account the diffusion of morphogens, which are therefore specified as input levels (e.g. the presence of Lefty is considered as an active input in the ciliary regions, although it is known that it diffuses from the ventral region). For each simulation, we extract the resulting stable state(s) and classify them as ventral, ciliary or dorsal pattern depending on the set of output node levels.

For example, the initial conditions for the simulation of the ventral ectoderm territory considered as active inputs Nodal (level 2), Lefty, Chordin and BMP2/4. This combination produced a stable state in which all the ventral nodes were active and the dorsal nodes inactive. In contrast, when the initial conditions were set as Nodal (level 1), Lefty. BMP2/4, Chordin and Tolloid being active, the resulting stable state corresponded to the dorsal regulatory state.

Mutant simulations

Genetic perturbations are defined in GINsim by constraining one or sometimes several nodes of the model. To simulate a *knock down* mutant (e.g. injection of a morpholino), the level of the corresponding node is set and maintained to 0. In the case of an ectopic

expression (e.g. injection of a mRNA), the level of the corresponding node is set and maintained to its maximal value, which can be 1 or higher in case of a multilevel node.

Morphogen diffusion is taken into account through the specifications of proper input values, which thus need to be adjusted for each mutant. For example, the ectopic activation of Nodal is known to induce the activation of its downstream target BMP2/4 very early on; hence, the corresponding input variables must be set at their highest levels for the simulation of ectopic nodal expression.

Stochastic modelling using MaBoss

When several stable states can be reached (as in the case of *chordin* morpholino), we have performed probabilistic simulations to evaluate the probability to reach each of these stable states from the specified initial conditions. In this respect we used the software MaBoss (maboss.curie.fr), a C++ software enabling the simulation of continuous/discrete time Markov processes, applied to Boolean networks.

The original unicellular model is converted into the MaBoSS compliant format using a specific export functionality of GINsim, which involves the replacement of multilevel nodes by sets of Boolean variables, without affecting the model dynamic [35]. Per default, all up and down rates are considered equal, but these can be modified at will.

In this study, we used MaBoSS to simulate the *chordin* morpholino perturbation (comparing it with the wild-type situation), which resulted in two possible stable states in the unicellular model. The inputs were fixed as for the ventral configuration (Nodal, Lefty, BMP2/4 and Admp1 active) in the presence or inactivation of Chordin. We then modified the propensity to activate the ventral or the dorsal cascade by adjusting the ratios of the rates assigned to the Alk receptors corresponding to each of the two cascades: 0.5/0.5 (equiprobable rates), 0.75/0.25 (ratio favouring the dorsal Alk), 0.25/0.75 (ratio favouring the ventral Alk).

Multicellular simulation using EpiLog

We took advantage of recent software Epilog (epilog-tool.org, v1.1.1.) [36] to perform multicellular simulations. The use of Epilog implies the definition of additional logical rules for the diffusion of signals, e.g. of the values of input nodes depending on the output nodes active in neighbouring cells, taking into account their distance from the target cell. For example, the rule :

$$Nodal : 2[0:1],\ min\ =\ 3 \qquad (3.3)$$

states that a cell will have its Nodal input node value converging toward the value 2 if at least 3 cells are expressing Nodal at a value 2 at a maximum distance of one cell (i.e. the target cell itself and its direct neighbours).

Our epithelium model is eight cells wide and six cells long, made of hexagonal shaped cells, each one being in direct contact with at most six different neighbouring cells. The top and bottom part of the epithelium are wrapped together to allow diffusion of signalling molecules through these two sides.

Each cell behaves according to the model described in our unicellular simulations. In contrast with our previous unicellular simulations, the inputs are dynamically updated based on the signals perceived in each cell, depending on

the activation levels of the output nodes of neighbouring cells.

The rules integrating the extracellular signals are identical for all cells of the epithelium. In our epithelium simulations, input nodes of all cells are updated in a synchronous manner. Hence, each epithelium simulation gives rise to a deterministic trajectory ending in a single attractor at the level of the whole tissue (a stable state for the simulations reported here).

Multicellular wild type and mutant simulations

For our epithelium simulations, we define the initial state by selecting the nodes that will be active in a specific set of cells at the start of the simulation. During simulations, the values of these nodes can change depending on the model state and on paracrine signalling.

To simulate a wild-type embryo, we set the model to an initial state where the ventral cascade is starting to be activated in the ventral region of the epithelium (3 leftmost cell columns of the epithelium), with the initial and transient activation of Smad1_4_5_8 and Nodal output nodes. Univin is also ubiquitously present at initial state. As in the unicellular simulation, for simulating perturbations, the target nodes are set and maintained at a fixed value.

For the siamese simulation, we use the wild-type logical model with a larger epithelium, with an initial state accounting for a ventral expression of Smad1_4_5_8 and Nodal on the ventral side, but also on the dorsal side of the epithelium (3 rightmost cell columns of the epithelium), i.e. a symmetrical activity pattern.

Model and code availability

The unicellular and multicellular models, together with the Jupyter notebook encoding all the simulations performed with GINsim and MaBoss, are available in a GINsim model repository and a GitHub repository. The Jupyter notebook uses the colomoto-docker image (github.com/colomoto/colomoto-docker, v2020-01-24) [49]. The models can be uploaded in zginml and peps format, to be open with GINsim (v3.0.0b) and EpiLog (v1.1.1), respectively. The unicellular model has been further deposited in SBML qual format in the BioModels database (ID MODEL2002190001).

3.4.6 Acknowledgements

We thank Aline Chessel for excellent technical help. We are indebted to Guillaume Lavisse for insightful comments. We thank Aurélie Martres for taking care of the sea urchins. We thank Aurelien Naldi for continuous help and support on the development of the Colomoto Jupyter Notebook. We thank Claudine Chaouiya and Pedro Monteiro for their support regarding the use of EpiLog. We thank Mathurin Dorel for his help in defining a preliminary version of the cellular model.

3.4.7 References

1. Arnone, M. I. & Davidson, E. H. The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124,** 1851–1864 (1997).
2. Levine, M. & Davidson, E. H. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences* **102,** 4936–4942 (2005).
3. Haillot, E., Molina, M. D., Lapraz, F. & Lepage, T. The Maternal Maverick/GDF15-like TGF-β Ligand Panda Directs Dorsal-Ventral Axis Formation by Restricting Nodal Expression in the Sea Urchin Embryo. *PLOS Biology* **13,** e1002247 (2015).
4. Lapraz, F., Haillot, E. & Lepage, T. A deuterostome origin of the Spemann organiser suggested by Nodal and ADMPs functions in Echinoderms. *Nature Communications* **6,** 8434 (2015).
5. Li, E., Materna, S. C. & Davidson, E. H. Direct and indirect control of oral ectoderm regulatory gene expression by Nodal signaling in the sea urchin embryo. *Developmental Biology* **369,** 377–385 (2012).
6. Li, E., Materna, S. C. & Davidson, E. H. New regulatory circuit controlling spatial and temporal gene expression in the sea urchin embryo oral ectoderm GRN. *Developmental Biology* **382,** 268–279 (2013).
7. Li, E., Cui, M., Peter, I. S. & Davidson, E. H. Encoding regulatory state boundaries in the pregastrular oral ectoderm of the sea urchin embryo. *Proceedings of the National Academy of Sciences* **111,** E906–E913 (2014).
8. Range, R. *et al.* Cis-regulatory analysis of nodal and maternal control of dorsal-ventral axis formation by Univin, a TGF-β related to Vg1. *Development* **134,** 3649–3664 (2007).
9. Saudemont, A. *et al.* Ancestral Regulatory Circuits Governing Ectoderm Patterning Downstream of Nodal and BMP2/4 Revealed by Gene Regulatory Network Analysis in an Echinoderm. *PLOS Genetics* **6,** e1001259 (2010).
10. Su, Y.-H. *et al.* A perturbation model of the gene regulatory network for oral and aboral ectoderm specification in the sea urchin embryo. *Developmental Biology* **329,** 410–421 (2009).
11. Chen, D., Zhao, M. & Mundy, G. R. Bone Morphogenetic Proteins. *Growth Factors* **22,** 233–241 (2004).
12. Cheng, S. K., Olale, F., Brivanlou, A. H. & Schier, A. F. Lefty Blocks a Subset of TGFβ Signals by Antagonizing EGF-CFC Coreceptors. *PLOS Biology* **2,** e30 (2004).
13. Duboc, V., Lapraz, F., Besnardeau, L. & Lepage, T. Lefty acts as an essential modulator of Nodal activity during sea urchin oral–aboral axis formation. *Developmental Biology* **320,** 49–59 (2008).
14. Sakuma, R. *et al.* Inhibition of Nodal signalling by Lefty mediated through interaction with common receptors and efficient diffusion. *Genes to Cells* **7,** 401–412 (2002).
15. Angerer, L. M., Yaguchi, S., Angerer, R. C. & Burke, R. D. The evolution of nervous system patterning: insights from sea urchin development. *Development* **138,** 3613–3623 (2011).
16. Lapraz, F., Besnardeau, L. & Lepage, T. Patterning of the Dorsal-Ventral Axis in Echinoderms: Insights into the Evolution of the BMP-Chordin Signaling Network. *PLOS Biology* **7,** e1000248 (2009).

17. Ben-Zvi, D., Shilo, B.-Z., Fainsod, A. & Barkai, N. Scaling of the BMP activation gradient in Xenopus embryos. *Nature* **453,** 1205–1211 (2008).

18. De Robertis, E. M. Spemann's organizer and the self-regulation of embryonic fields. *Mechanisms of Development* **126,** 925–941 (2009).

19. Joubin, K. & Stern, C. D. Formation and maintenance of the organizer among the vertebrates. *The International Journal of Developmental Biology* **45,** 165–175 (2001).

20. Kimelman, D. & Pyati, U. J. Bmp signaling: turning a half into a whole. *Cell* **123,** 982–984 (2005).

21. Lele, Z., Nowak, M. & Hammerschmidt, M. Zebrafish admp is required to restrict the size of the organizer and to promote posterior and ventral development. *Developmental Dynamics* **222,** 681–687 (2001).

22. Reversade, B. & De Robertis, E. M. Regulation of ADMP and BMP2/4/7 at opposite embryonic poles generates a self-regulating morphogenetic field. *Cell* **123,** 1147–1160 (2005).

23. Willot, V. *et al.* Cooperative Action of ADMP- and BMP-Mediated Pathways in Regulating Cell Fates in the Zebrafish Gastrula. *Developmental Biology* **241,** 59–78 (2002).

24. Molina, M. D. *et al.* MAPK and GSK3/ß-TRCP-mediated degradation of the maternal Ets domain transcriptional repressor Yan/Tel controls the spatial expression of nodal in the sea urchin embryo. *PLoS genetics* **14,** e1007621 (2018).

25. Duboc, V., Röttinger, E., Besnardeau, L. & Lepage, T. Nodal and BMP2/4 signaling organizes the oral-aboral axis of the sea urchin embryo. *Developmental Cell* **6,** 397–410 (2004).

26. Wilczynski, B. & Furlong, E. E. M. Challenges for modeling global gene regulatory networks during development: Insights from Drosophila. *Developmental Biology. Special Section: Gene Regulatory Networks for Development* **340,** 161–169 (2010).

27. Mbodj, A. *et al.* Qualitative Dynamical Modelling Can Formally Explain Mesoderm Specification and Predict Novel Developmental Phenotypes. *PLOS Computational Biology* **12,** e1005073 (2016).

28. Peter, I. S. in *Methods in Cell Biology* (eds Hamdoun, A. & Foltz, K. R.) 89–113 (Academic Press, 2019).

29. Peter, I. S., Faure, E. & Davidson, E. H. Predictive computation of genomic logic processing functions in embryonic development. *Proceedings of the National Academy of Sciences* **109,** 16434–16442 (2012).

30. Thomas, R. & D'Ari, R. *Biological feedback* (CRC Press, Boca Raton, FL etc., 1990).

31. Chaouiya, C., Naldi, A. & Thieffry, D. Logical modelling of gene regulatory networks with GINsim. *Methods in Molecular Biology* **804,** 463–479 (2012).

32. Lapraz, F. *et al.* RTK and TGF-beta signaling pathways genes in the sea urchin genome. *Developmental Biology* **300,** 132–152 (2006).

33. Naldi, A. BioLQM: A Java Toolkit for the Manipulation and Conversion of Logical Qualitative Models of Biological Networks. *Frontiers in Physiology* **9** (2018).

34. Juan, H. & Hamada, H. Roles of nodal-lefty regulatory loops in embryonic patterning of vertebrates. *Genes to Cells: Devoted to Molecular & Cellular Mechanisms* **6,** 923–930 (2001).

35. Stoll, G. *et al.* MaBoSS 2.0: an environment for stochastic Boolean modeling. *Bioinformatics* **33,** 2226–2228 (2017).

36. Varela, P. L., Ramos, C. V., Monteiro, P. T. & Chaouiya, C. EpiLog: A software for the logical modelling of epithelial dynamics. *F1000Research* **7,** 1145 (2019).

37. Armstrong, N. & McClay, D. R. Skeletal pattern is specified autonomously by the primary mesenchyme cells in sea urchin embryos. *Developmental Biology* **162,** 329–338 (1994).

38. Armstrong, N., Hardin, J. & McClay, D. R. Cell-cell interactions regulate skeleton formation in the sea urchin embryo. *Development* **119,** 833–840 (1993).

39. Duloquin, L., Lhomond, G. & Gache, C. Localized VEGF signaling from ectoderm to mesenchyme cells controls morphogenesis of the sea urchin embryo skeleton. *Development* **134,** 2293–2302 (2007).

40. Hardin, J., Coffman, J. A., Black, S. D. & McClay, D. R. Commitment along the dorsoventral axis of the sea urchin embryo is altered in response to NiCl2. *Development* **116,** 671–685 (1992).

41. Röttinger, E. *et al.* FGF signals guide migration of mesenchymal cells, control skeletal morphogenesis and regulate gastrulation during sea urchin development. *Development* **135,** 353–365 (2008).

42. McIntyre, D. C., Seay, N. W., Croce, J. C. & McClay, D. R. Short-range Wnt5 signaling initiates specification of sea urchin posterior ectoderm. *Development* **140,** 4881–4889 (2013).

43. Lepage, T. & Gache, C. Purification and characterization of the sea urchin embryo hatching enzyme. *The Journal of Biological Chemistry* **264,** 4787–4793 (1989).

44. Lepage, T. & Gache, C. Early expression of a collagenase-like hatching enzyme gene in the sea urchin embryo. *The EMBO journal* **9,** 3003–3012 (1990).

45. Harland, R. M. In situ hybridization: an improved whole-mount method for Xenopus embryos. *Methods in Cell Biology* **36,** 685–695 (1991).

46. Thomas, R. Regulatory networks seen as asynchronous automata: A logical description. *Journal of Theoretical Biology* **153,** 1–23 (1991).

47. Naldi, A. *et al.* Logical Modeling and Analysis of Cellular Regulatory Networks With GINsim 3.0. *Frontiers in Physiology* **9,** 646 (2018).

48. Naldi, A., Thieffry, D. & Chaouiya, C. *Decision Diagrams for the Representation and Analysis of Logical Models of Genetic Networks* in *Computational Methods in Systems Biology* (eds Calder, M. & Gilmore, S.) (Springer, Berlin, Heidelberg, 2007), 233–247.

49. Naldi, A. *et al.* The CoLoMoTo Interactive Notebook: Accessible and Reproducible Computational Analyses for Qualitative Biological Networks. *Frontiers in Physiology* **9** (2018).

3.5 Complementary results

3.5.1 Script availability

The analyses presented in the manuscript have been implemented using the Python programming language (python.org) and the Jupyter Notebook interface (jupyter.org). This framework has the advantage of embedding annotation in Markdown format (daringfireball.net/projects/markdown) together with blocks of code, making the script easy to share and reuse (Fig. 3.16).

The script is based a notebook from the Consortium for Logical Models and Tools (CoLoMoTo, colomoto.org) [163]. This notebook gathers a large panel of softwares specifically adapted to qualitative modelling, provided as python modules. In my analysis, I made use of the packages GINsim [136], bioLQM [164], MaBoSS [140] and Pint [165].

In order to avoid conflicts due to dependencies distribution requirements, the script uses the CoLoMoTo Docker image (github.com/colomoto/colomoto-docker). This image is constructed using Docker (docker.com) and works as a software container. This embedding guarantees the portability and reproducibility of the script, irrespective of the user working environment.

Lastly, the model is made available as an SMBL-qual file [166]. This format is an extension of the Systems Biology Markup Language (SMBL) [147] (cf. section 3.2.2) ; it provides a standardised formatting, adapted for qualitative logical models. If further allows for an easy transfer of the models between different tools, such as GINsim and EpiLog in this study.

The jupyter notebook and the models (unicellular and multicellular) are available in a GitHub repository and a GINsim repository.

In [3]:
```
# Compute the total number of stable states
fps = biolqm.fixpoints(lqm2)
print(len(fps)), "fixpoints found")

# Show one example stable state
ginsim.show(lrg, fps[1257])
```

```
1258 fixpoints found
```

Out[3]:

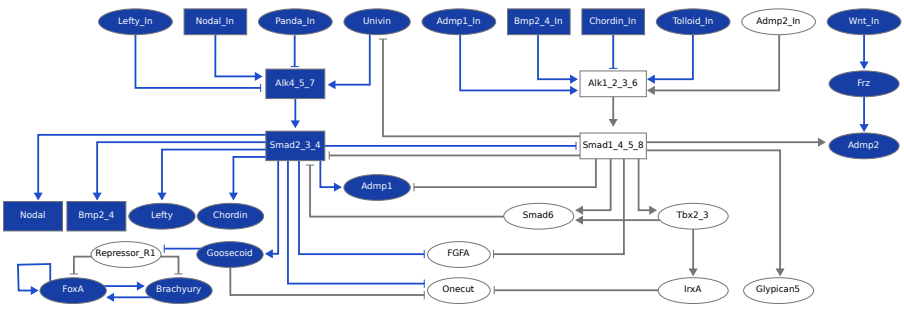

Figure 3.16: Jupyter notebook. Example of code block (grey box) and output graphic from the Jupyter notebook. In this example, the python packages bioLQM and GINsim are used to respectively compute the total number of stable states of the model and make a visual representation of one of them.

Chapter 4

Conclusions and perspectives

> *Having the same set of evidence, each of them will make a choice based on their belief and assumptions, yet the wisest position seems to keeping the doubt from your capacity to know the truth without bias.*
>
> **Twelve Angry Men, Reginald Rose, 1954**

In light of the previous chapters, we have demonstrated how mechanistic modelling and probabilistic inference methods can help to improve our view of transcription regulation and support the search for novel *cis*- and *trans*-acting regulatory interactions. Furthermore, these methods show complementary assets and have the potentials to be combined into a common network modelling framework.

This closing chapter is structured as follows :

⋈ I will first outline the biological insights gained from this work. On the one hand, I will contrast the roles of the histone marks H3K27ac and H3K3me3 in transcription regulation (chapter 2). On the other hand, I will discuss the network circuits governing TGF-β cross-inhibition (chapter 3).

⋈ Secondly, I will give a brief overview of the methodological assets and pitfalls of the strategies used in this thesis, namely the allele-specific data analysis (chapter 2) and the logical modelling (chapter 3). Additionally, I will present preliminary results of a method I am currently implementing, which aims at refining TF motif discovery in individual ChIP-seq histone data.

⋈ Lastly, I will outline the different prospects of this work, with a specific focus on the intergation potential of the mechanistic and probabilistic network modelling methods, in order to aim for a system-wide dynamical GRN.

4.1 Biological aspects

4.1.1 Coupling between epigenetic and transcriptional regulatory mechanisms

In the analyses presented in Chapter 2, I took advantage of a comprehensive dataset of multi-omics measures reflecting the transcriptional activity of heterozygous F1 embryos. Relying on the genetic variation present in each sample, I measured the relative signal coming from each allele and inferred an allelic ratio measure. This led me to delineate specific interactions between regulatory layers based on partial correlation analysis of allelic ratio and total read signal.

The resulting networks displayed inter-

esting characteristics. First, the network constructed from total read signal assigned similar correlation values between each pair of regulatory layers, except for a weak direct correlation between RNA and H3K4me3 (ie. independent from ATAC and/or H3K27ac covariation).

In contrast, the network constructed from allelic ratio displayed a more asymmetrical structure, with only three strongly supported interactions : between RNA and H3K4me3, between RNA and ATAC, and between ATAC and H3K27ac. In contrast with total read signal, the analysis of allelic ratios has the advantage of excluding *trans*-mediated correlation. This result therefore suggests that two *cis*-interactions are directly linked to gene expression (ATAC and H3K4me3), while H3K27ac only directly interact in *cis* with chromatin accessibility (ATAC). However, as correlation does not imply causation, these inferred *cis*-interactions are not directed. Indeed, a correlation solely reflects the level of co-variation between two components, but does not provide information regarding the direction of the interaction.

Recent studies have hinted to the potential causal role of H3K27ac and H3K4me3 in the activation of gene expression. The role of the H3K4me3 has been especially challenging to interpret. This histone mark is associated with active promoter [33], but several studies suggest that H3K4me3 is not required for transcription activation [167]. Additionally, this mark tends to be long-lived and persists even after the shutdown of transcription induction [168]. A recent study demonstrated that H3K4me3 broad marks tend to be enriched for cell-type specific genes [169]. These observations are consistent with the hypotheses from Benayoun et al. [170], who propose that H3K4me3 bears the role of a persistent marker for transcriptional memory, used to flag genes requiring consistent activation or fast re-induction [168].

These hypotheses are consistent with our result, showing a direct *cis*-effect between RNA and H3K4me3, but no correlation between H3K4me3 and the active enhancer signatures H3K27ac and ATAC. In light of this network, we may hypothesise that transcriptional induction induces the deposition of H3K4me3 at the promoters of genes requiring maintained activity over time.

The role of H3K27ac has also been explored at length. It is associated with active enhancer and promoter regions [33]. Contrary to H3K4me3, H3K27ac has been characterised to be a highly dynamic and short-live mark, with the ability to decrease DNA affinity for nucleosome [171]. Its presence induce the binding of bromodomain-bearing proteins, often associated with a transcriptional activation function. As Barnes et al. [171] nicely summarise it, H3K27ac can be pictured both as a crowbar to remove histone from DNA, and as a post-it for short-term CRMs labelling. However, some studies also suggest that H3K27ac signalling is more complex and dynamically adjusting through time, as the presence of histone deacetylases (HDAC) seem to be necessary for the establishment of transcriptional activity [172]. This hypothesis also supports our analysis results, more precisely the presence of a direct interaction between H3K27ac and chromatin accessibility (ATAC) (ie. independent from RNA and/or H3K4me3). As we did not find any evidence for a direct link between H3K27ac and RNA, our study further suggest that H3K27ac does not directly impact gene expression but rather acts via the regulation of chromatin accessibility.

Taken together, our result are consistent with a linear view of transcriptional regulation, starting from H3K27ac deposition, followed by the chromatin opening (ATAC), the activation of the transcription (RNA), and ending with H3K4me3 labelling. However, the causal role of H3K27ac on chromatin opening is still debated [173, 174]. Furthermore, we still lack a mechanistical understanding of a large part of this process. To further explore this regulation dynamic, several tracks can be followed.

On the one hand, assaying a higher number of histone modifications (eg. H3K4me1, H3K27me3, H3K9me3) could lead to refinements of our conclusions. Indeed, a growing number of post-translational histone modifications can be characterised at a pangenomic scale. Furthermore, the very recent ChromID technique [175] could be used to characterise the set of factors and cross-interactions involved in chromatin remodelling.

On the other hand, it should be possible to test some of the interactions predicted from our statistical analysis by performing local allelic perturbations, for example using the Crispr-Cas9 system [25]. The perturbation of a CRM sequence or the deposition of histone mark on only one of the two alleles could help to assess whether the perturbation is propagated on one or two alleles, depicting a causal *cis*-effect or *trans*-effect respectively.

4.1.2 The mechanims of TGF-β cross-inhibition

In the Chapter 3, I integrated the existing data on dorsal-ventral axis specification in sea urchin embryo into a predictive, mechanistic, mathematical model.

Based on the resulting logical model, I simulated the wild-type and various mutant backgrounds and recapitulated the documented embryonic patterns. In the case of the Chordin loss-of-function, we observed one discrepancy pointing to a pending question regarding the dominance of time-driven versus concentration-driven cross-inhibitory competition between TGF-β pathways. This led us to design novel experiments, presented in the manuscript, whose results supported the hypothesis of the concentration-driven hypothesis. According to this result, TGF-β signalling can be switched off by the antagonist TFG-β signal when present in higher concentration, irrespective of the time of activation.

To further explore the model dynamics, I used a probabilistic extension of the logical formalism to compare the likelihood of alternative fates starting form specific initial conditions. My results indicate that the network structure provides an inherent advantage to the dorsal BMP pathway. This result is consistent with the weak dorsal embryo patterning observed in Chordin loss-of-function experiments. This result further implies that dorsal regulation is dominant over the ventral regulation when equal transition rates are used.

Interestingly, our model correctly recapitulates the expected specification patterning of the ectoderm, although we still only have a partial mechanistic understanding. For example, the mechanism governing the repression of Nodal by Panda is still not completely understood [176]. Novel experimental results could potentially help to refine the model. For example, ATAC-seq and ChIP-seq experiments could help to delineate the active CRMs and enable to formally integrate the enhancers within the network, together with regulatory

rules reflecting TF binding, including cooperative and antagonist effects.

4.2 Methodological aspects

4.2.1 Allele-specific measurements can contrast *cis*- versus *trans*- effects

In the analysis presented in Chapter 2, we take advantage of the F1 cross design to extract allele-specific measures. One advantage of using allele-specific data to study *cis*-regulatory variation is that it considerably lowers the noise coming from *trans*-acting mechanisms (Fig. 4.1) [103, 104]. Indeed, as both alleles are present in the same nucleus, they share the exact same cellular environment and the same embryonic developmental timing. Hence, a variation acting in *trans* from one of the two alleles will not lead to allelic imbalance outcomes, as it will equally affect the cellular environment of the two alleles.

A valuable consequence of this experimental design is that it offers a framework to study heritability and genomic imprinting. Indeed, one can compare, for a given gene or non-coding region, the allelic imbalance observed in the F1s against the imbalance observed when directly comparing the two parents. If allelic imbalance is present between the two parents but not in the F1, it suggest a *trans*-acting mechanism. On the contrary, if allelic imbalance is maintained both between the parents and between the alleles of the F1, it implies that a *cis*-acting mechanism is taking place. Going further, the F1 experimental design offers a framework to test additivity model of heritability, by comparing total expression levels [108].

With this characteristic, the allele-specific framework could be especially relevant in GWAS. Indeed, QTL target *cis*-acting variant, but in most cases the discrimination between *cis*- and *trans*-QTL solely rely on genomic vicinity, which may lead to mis-assignment [177]. Using F1 rather than isogenic lines, GWAS could leverage their capacity to contrast cis- from trans QTL. Recently, the Deplancke laboratory has used this strategy by performing F1 crosses from DGRP lines and concluded that only 10% of the QTLs could be attributed to a *cis*-effect [178]. The use of allele-specific data is therefore a powerful tool to better infer causal genetic variations. More specifically, it could be an efficient alternative to GWAS for detecting rare deleterious variants. Consequently, allele-specific studies show a great potential for biomedical application in the case of genetic diseases. Additionally, the need for isogenic crosses can now be avoided for human application with the recent methods of *de novo* haplotype-resolved geonome assembly.

Although allele-specific studies offer great advantages, they also come with challenges, in particularly reference mapping bias. Multiple tools have emerged to tackle this problem. Several publications now call for the end of haploid reference genomes. Indeed, reference genomes commonly stem from random wild-type individuals and do not bear the expected characteristic of a "gold-standard", such as the equitable representation of the population genetic diversity [179]. Now that sequencing costs and speed enable to sequence a large number of samples, the design of a better consensus reference genome or systematic *de-novo* assembly might become more efficient. Indeed, these strategies can increase the power to detect variants and to help to understand complex biological mechanisms, such as transvection and muta-

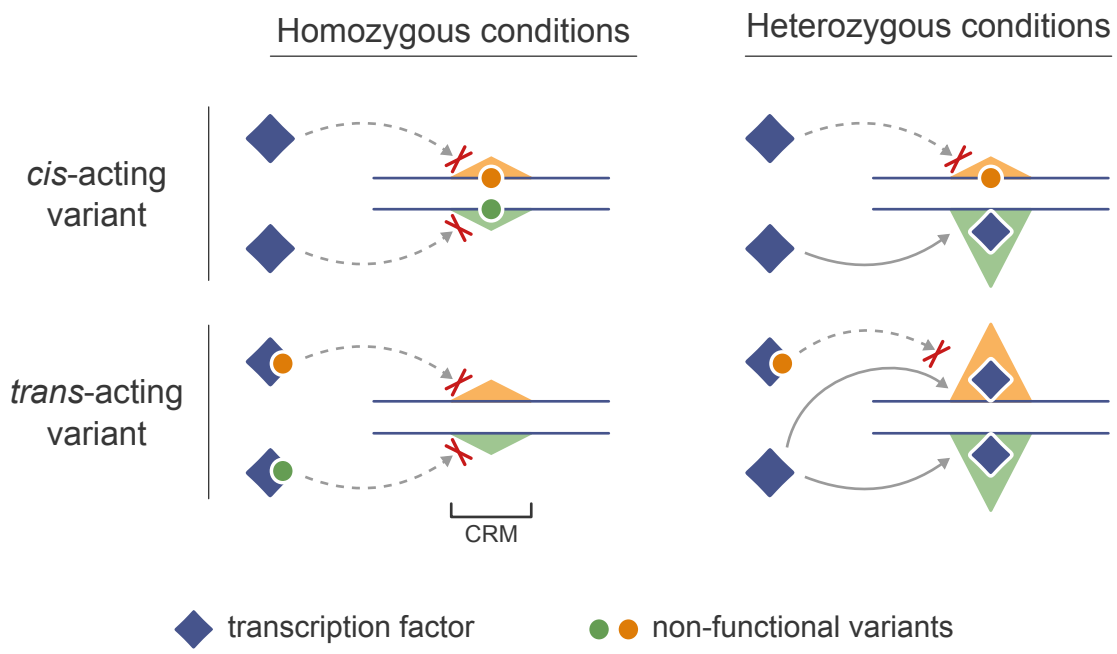

Figure 4.1: Distinguishing *cis* from *trans* acting variants. Schematic of the different impacts of *cis* and *trans* regulatory variants on a *cis*-regulatory module (CRM). In homozygous conditions (left), mutations (orange/green circles) affecting either the CRM sequence (*cis*, top) or the transcription factor (TF) structure (*trans*, bottom) prevent regulatory activity in both alleles (orange/green small triangles). In heterozygous conditions (right), the functional allele, either in *cis* or *trans*, shows CRM activity (large green triangles). The allele with non-functional *trans*-variant may have its CRM activity rescued by the binding of TFs produced from the other functional allele. The common cellular environment in heterozygous diploid conditions therefore minimises the confounding impact of *trans*-acting variants.

tion penetrance [180].

For example, Garrison et al. [181] propose the use of *variation graphs* as reference. In this formalisation, both DNA strands are represented together with segregating genetic variations, enabling the alignment on both parental genotypes at the same time. Additionally, new tools based on read pseudo-alignments may completely circumvent the pitfalls of reference mapping biases [182].

4.2.2 DNA binding motifs from ChIP-seq targeting histone marks

In the analysis presented in Chapter 2, I designed an *ad hoc* method to integrate multiple types of data with varying genomic spans and genomic locations. Although the analysis of ATAC-seq, ChIP-seq targeting TFs (ChIP-seq TF) and RNA-seq data usually lead to well-defined genomic regions, the analysis of ChIP-seq targeting histone marks (ChIP-seq histone) data usually yields fuzzier and larger genomic regions.

The combined analysis of multiple ChIP-seq histone datasets can yield genomic regions more precisely defined, for example using ChromHMM [133]. However, the analysis of individual histone mark datasets remains hindered by the large size of the broad peak signal (sometimes dozens of kb). Yet, more advanced approaches can take advantage of the intrinsic properties of ChIP-seq histone signal. Indeed, the landscape of H3K27ac histone mark signal typically follows a U-shape or peak-valley-peak (PVP) pattern, mir-

roring the alternation of nucleosome-bound and nucleosome-free DNA [78]. Furthermore, the local minima (valleys) depicting a local depletion of nucleosome are likely to reflect the position of a *cis*-regulatory region, bound by TF. Consequently, this property could be exploited to detect enhancer regions.

DNA motif analyses of ChIP-seq histone peaks often yield poor results because the large peaks implies a low signal/noise ratio. By narrowing the search space onto the putative CRM region, similar to a ChIP-seq TF, we could enrich this ratio and improve the detection of specific TF binding motifs.

One challenge to address in order to narrow peak regions is the background noise, which can lead to the generation of a large number of false positive valleys, for example when extracting them based on changes in signal slope sign. Several bioinformatic tools are available to detect valley patterns in noisy biological signals (Table 4.1). The most common strategy to remove background noise consists in fitting the signal to a smoothing function.

Two of these methods have already been applied to ChIP-seq histone data: EpiSafari and PARE. The valleys obtained with EpiSafari are shown to overlap well with ChIP-seq TF and DNase signal [78], although EpiSafari does not specifically reduce the size of the search space compared to a standard peak calling procedure.

In order to further explore this strategy of search space reduction, I adapted the algorithm from Meers et al. (EcHo [185]) for the analysis of ChIP-seq histone signal (Fig. 4.2). I chose this method because it starts from an initial set of predetected peaks, rather than performing a genome-wide search. Consequently, it is better suited to refine signal from a pre-existing search space.

Within each broad histone peak region, the signal was smoothed by a Loess regression with 20 different window sizes, spanning between 5% and 100% of the peak length (Fig. 4.2a). The best smoothing window was selected based on the normalised standard error of the loess curve and the normalised standard deviation of its first derivative. These two measures give a quantitative score for the over-fitting and under-fitting level of each window size (Fig. 4.2b). After smoothing each peak with the best-scoring window, the valley regions were defined as the local minima of the loess curve. For each peak, I tested the statistical significance of each valley by comparing the average number of reads in the 200bp of the valley regions against the two flanking local maxima regions (one-sided Poisson test). The 200bp region size was chosen to include at least one individual nucleosome region, which is known to coil 146bp of DNA [6].

I tested the capacity of two methods to detect significantly enriched motifs compared to a standard enrichment from MACS2 peak regions: EpiSafari, and my implemented algorithm detecting valleys within the MACS2 peaks. I used the H3K27ac ChIP-seq dataset from Sebastiaan Meijsing lab on U2OS cell lines treated with glucocorticoids [187]. The motif enrichment analysis was performed using the 'peak-motif' tool from the RSAT suite [73–75], with identical parameters for all analyses (oligo-analysis, -nmotifs 10 -minol 6 -maxol 8) and sequences from MACS2 peak regions as control.

Compared to the sequences of MACS2 full peak regions, the sequences obtained from EpiSafari regions were significantly

Table 4.1: Existing tools for signal smoothing and valley detection.

Tool name	Smoothing method	Type of biological signal	Ref.
PARE	Gaussian fitting	ChIP-seq histone	[183]
EpiSafari	Spline fitting	ChIP-seq histone	[78]
MSR	Gaussian fitting	ChIP-seq Polymerase II	[184]
EChO	Loess	Cut&Run fragment size	[185]
LastWave	Wavelet fitting	DNA replication timing	[186]

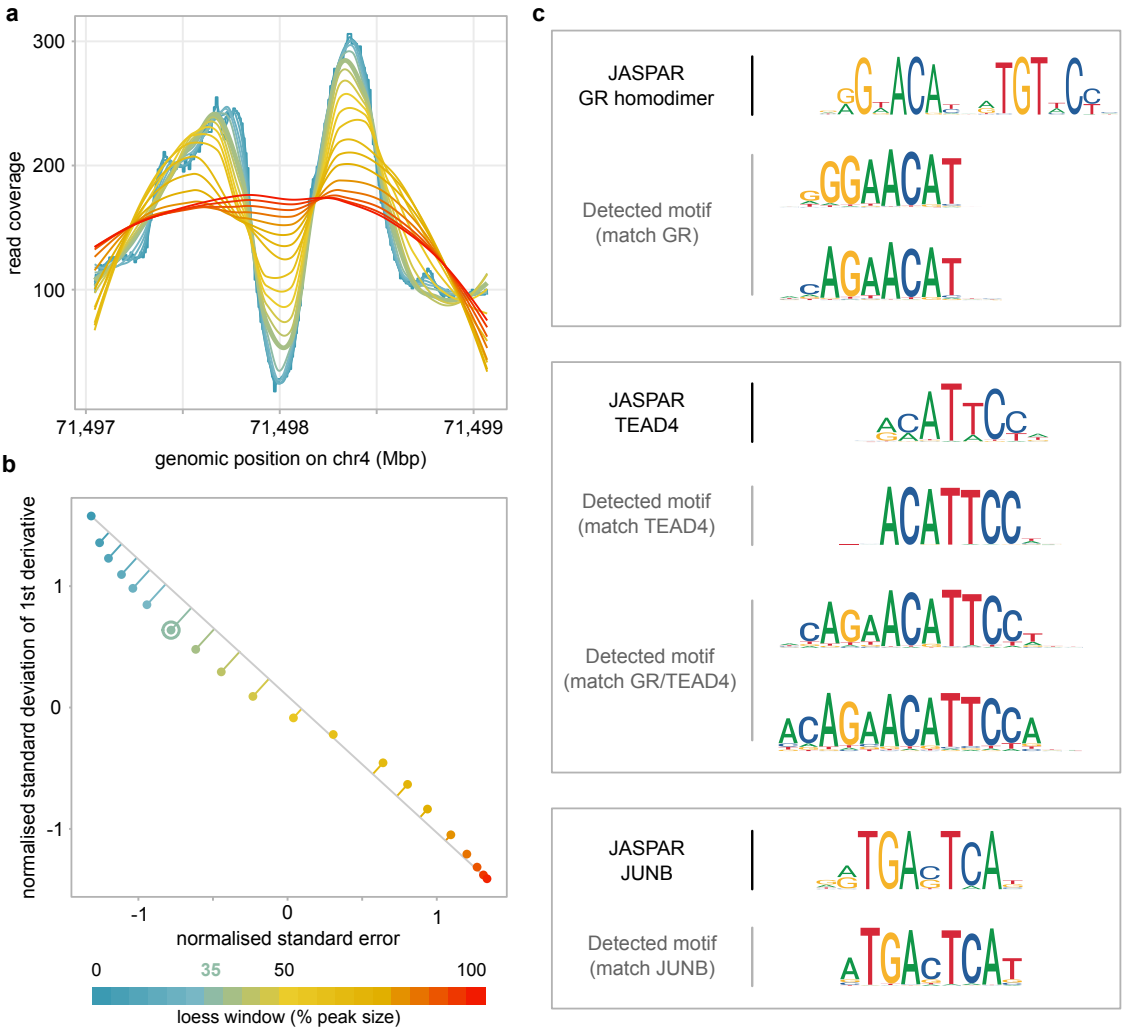

Figure 4.2: Valley detection in H3K27ac signal. a. The raw H3K27ac signal obtained within a MACS2 peak is shown in blue. The other curves correspond to the smoothing of this signal using increasing window sizes for Loess regression. This peak present the typical peak-valley-peak pattern suggesting a CRM region at the center. b. Scatter plot used to choose the best window size (circled point, 35% peak size) based on the over-fitting (y-axis) and under-fitting score (x-axis). The best window size corresponds to the point with the largest orthogonal distance from the two most extreme points. c. Motifs significantly enriched within the valleys detected by loess method, as compared to the full MACS2 peak regions. Detected motifs match glucocorticoid receptor (GR), TEAD and JUNB motifs from the JASPAR PWM database and further suggest the additional binding of a heterodimer complex GR/TEAD.

enriched in mostly AT-rich motifs, including several matches for FOX and GATA binding profiles. The presence of such motifs agrees with the hypothesis of Starick *et al.* [188], which suggests that FOXA1 and GATA may help tethering glucocorticoid receptor (GR) to DNA.

Contrary to EpiSafari, the valley regions obtained from my implemented algorithm were chiefly enriched in GC-rich sequences, notably similar to the SP1 binding profile. The enriched motifs detected by 'peak-motif' recapitulated the expected glucocorticoid receptor (GR) motif monomer (Fig. 4.2c). The detection of GR motif suggests that my method is capable of reducing the signal/noise ratio for motif detection in ChIP-seq histone signal, compared to a standard search based on the full peak regions, which did not detect GR motif.

The motif search in the valleys, generated by my algorithm, also detected a significant enrichment for members of the AP1 family (FOS, JUNB). This result is consistent with the study of Biddie *et al.* [189], which suggests that AP1 binding is required to maintain chromatin accessibility for GR binding in a cell-type specific manner. Additionally, 'peak-motif' detects a significant enrichment for the TEAD4 motif in the sequences obtained from my algorithm compared to the full peak sequences. Furthermore, another potentially combined motif, resembling adjacent GR and TEAD4 motifs, suggests the binding of a heterodimer GR/TEAD4 (Fig. 4.2c).

Previous experiments from the Meijsing lab further support this GR/TEAD4 heterodimer hypothesis. Indeed, relying on base-pair resolution binding profiles (ChIP-exo) and STARR-seq of different dimer combinations (eg. GR/GR, GR/TEAD), Schöne *et al.* [190] and Starick *et al.* [188] demonstrated that (i) footprint signal from ChIP-exo targeting GR matches a heterodimer GR/TEAD profile and (ii) synergistic effects of GR/TEAD tend to increase transcriptional expression while reducing signal noise. TEAD4 might therefore act as a regulator of GR activity.

To contrast this result, I performed a reverse analysis to detect motifs with sequences significantly enriched within MACS2 peaks, compared to the valley regions obtained from my algorithm. This analysis chiefly yielded AT-rich sequences, consistent with the expected result that valleys targeting CRMs retain most of the GC-rich regions. These results suggest that my algorithm, based on a search-space reduction strategy, performs better at defining regions of interest within the histone signal, likely to comprise CRMs. However, this preliminary result needs to be complemented with additional analyses, performed on other cell lines and histone marks.

4.2.3 The iterative process of network modelling

In the analysis presented in Chapter 3, I took a knowledge-based approach to build a mechanistic dynamical model.

The logical formalism used is particularly valuable for studying network for which only or mainly qualitative prior information is available. Indeed, logical modelling discretises protein levels and reaction rates, reducing the need for precise quantitative data. In addition, it is relatively straightforward to explore the asymptotic behaviour of logical models. For example, Traynard *et al.* used refined model-checking techniques to explore the mammalian cell cycle from Fauré *et al.* [191]. They spec-

ified novel dynamical properties, identified the attractors, and could ultimately validate several novel components and interactions, along with refined logical rules.

In contrast with the mammalian cell cycle, our model gives raise to several stable states corresponding to the different presumptive embryonic tissue. We first used an efficient algorithm to identify these stable states and then compared them with the documented expression patterns of marker genes to validate our logical rules. Next, we used a probabilistic extension of logical modelling to further characterise the dynamics of the network. This way, we highlighted novel properties of the model and validated our logical rules.

However, having a model with a coherent dynamics is not an end goal. Rather, modelling involves constant cross-talks between experimentalist and modellers. On the one hand, novel simulations can help to delineate experimental tests with the potential to generate interesting insights. For example, the obtention of two stable states for the Chordin KO led me to suggest an experiment consisting in injecting Nodal and BMP proteins at different concentrations and different developmental stages. This led to evidences supporting the hypothesis that Chordin regulation was governed by a concentration difference rather than by a difference in activation timing. The model rules were then adapted accordingly.

On the other hand, novel (potentially independent) experimental results can be used to improve a pre-existing model. Hence, a model will never perfectly reflect the reality and can always be refined based on new discoveries. One testimony of this iterative process can be found in Davidson's work [43], with the production of various refined GRN version over several decades.

Currently, the construction and refinement of a logical model is often manual. As a result, the abstraction work from experimental evidence to a mathematical formulation of the logical rules can be error-prone or lead to model overfitting. A potential improvement to automate this step it to take advantage of the existing databases of curated interactions [192]. Furthermore, several software tools are tackling the considerable challenge to infer regulatory rules from quantitative data, such as the Inferelator [193] and CaSQ (A. Niarakis, unpublished).

4.3 Prospects

Using probabilistic and mechanistic modelling approaches, we have explored different regulatory mechanisms involved in early embryonic development. The two approaches yielded complementary results: on one hand, the probabilistic network inference highlighted general patterns of transcriptional regulation. On the other hand, the mechanistic network modelling delineated precise regulatory dynamics in a specific signalling pathway.

4.3.1 Aiming at a system-wide model

Probabilistic modelling, based on quantitative data, is efficient to infer novel interactions. In contrast, mechanistic modelling, capable of flexible adjustment and simulation, are more efficient at driving the interpretation of a regulatory network. As a result, these two strategies offer complementary benefits for regulatory network construction and

can supplement each other. The recent mathematical and experimental advances regarding these two approaches render possible the integration of genetic variation, molecular components and regulatory process into dynamical, system-level, predictive model.

One the one hand, probabilistic modelling method take advantage of the continuous growth of high-throughput techniques to refine the characterisation of molecular actors, and to increase its power to infer regulatory interactions. On the other hand, mechanistic modelling method benefits from this gain in network resolution to guide logical rule refinement and the integration of novel interactions.

These new approaches allow to depict the mechanistic dynamic of transcription regulation at an unrivalled depth, and therefore open exciting possibilities for the reconstruction of dynamic, predictive, system-wide GRN models. Furthermore, this level of precision is a consequent leverage for the analysis of enhancer logic. Indeed, a more precise characterisation of the TF binding dynamic (eg. by integrating multiple ChIP-seq TF assays) can help to better understand how cooperative and antagonistic binding may adjust enhancer activity.

4.3.2 Towards the inference of regulatory networks

Multiple advances have recently emerged in the field of high-throughput sequencing to better characterise the molecular actors of enhancer regulation and their associated interactions [25]. More specifically, several methods aim at inferring the interactions between TFs and their target CRMs on a genome-wide scale.

The growing mine of available high-throughput data and TF motifs, notably through the elaboration of international consortium (ENCODE [79], Roadmap [80]) and databases (JASPAR [194]), enables the inference of probabilistic networks from the combined analysis of independent datasets. For example, i-cisTarget [195] combines information from more than a thousand ChIP-seq TF datasets to infer CRM regions associated to a set of co-expressed genes. Additionally, i-cisTarget exploits PWM databases to infer enriched TF motifs within the detected CRMs, resulting in a regulon network, associating a candidate master TF with its targeted downstream genes. Another tool, TFregulomeR [77], relies on a high-dimensional integration of ChIP-seq datasets to deconvolve context-specific motifs of homodimers and heterodimer TFs from a general mixture TF motif.

The emergence of single-cell techniques opens novel prospects for the inference of direct interaction between TF and target gene (ie. regulon network) [61]. Indeed, single-cell techniques have the considerable advantage of recapitulating cell-type specific clusters of cells. For example, the SCENIC workflow [196] infer sets of co-expressed genes from scRNA-seq (using tree-based ensemble method) and infer their associated regulon network using i-cisTarget. The activity of these regulons can then be explored across the whole single-cell space and used to infer the different cell types.

Interestingly, very recent technical developments make possible to synchronously probe multiple molecular identities within the same cell (eg. sc-CAT [197]). With this new dimension, the development of tools to properly reconstruct heterogeneous multi-layered

network (HMLN) brings promising possibilities for network inference (cf. Lee *et al.* [89] and Hawe *et al.* [86] for an overview of the existing techniques). In particular, spatial reconstruction [198] and matrix factorisation methods [197] show promising scalability and flexibility characteristics to explore large heterogeneous single-cell data. Together with the advances of high resolution microscopy for genome visualisation [62], these novel techniques can be used to generate precise snapshots of genome unfolding in time and space at a cellular resolution.

The integration of multi-omics data within the same network opens exciting possibilities for the study of cell specification and transcription regulation. Yet, several challenges await to be solved. Benchmarking effort should be carefully undertaken to assess the reproducibility and robustness of the inferred regulons. Similar to the idea of a reference genome, the construction of consensus cell atlases and reference molecular maps could facilitate the development of predictive models. Efforts have already been overtaken in that direction, such as the Fly Cell Atlas (flycellatlas.org), FlyBase (flybase.org) and REDfly (redfly.ccr.buffalo.ed) databases, which aim at gathering the Drosophila community around a comprehensive atlas of annotated cell types and CRMs that would serve as a reference "backbone" for future analyses.

Lastly, the contribution of machine learning methods such as Deep Learning approach may be extremely valuable. Indeed, convolution and latent spaces visualisation have already greatly contributed to network inference and proved to be efficient at delineating patterns from quantitative data [61]. Similarly, Deep Learning approaches may be well suited to dissect regulatory information from enhancer sequences and derive a precise description of how the sequences dictate TF binding and encode the signal into a regulatory command [199].

4.3.3 Qualitative inference of network dynamics

With the increasing number of characterised cell types and the parallel increase in size of the related networks, it becomes evident that graph-theory and model-based approaches will have a role to play in the inference of system-level GRN.

Indeed, the high number of samples ease the detection of interactions by avoiding the "large p (variables), small n (samples)" limitation. However, this approach may still lack prediction and simulation capacities for cell specification trajectories, as well as proper method to explore the global system dynamic. Indeed, current omics methods tend to drive a descriptive, component-focused research, while sometime loosing the scope of the dynamic, biological processes, linking these components together. As we continuously increase our knowledge on specific molecular identities and TF-gene interactions, we are lacking a causal understanding of their mechanisms and their integration within molecular processes and within cellular tissues [200]. In that sense, mechanistic modelling approaches might be well suited to answer this challenge.

Indeed, mechanistic modelling methods are more efficient to explore dynamical behaviors and are therefore more suited to address the current challenge of causality inference in cell biology [200]. For example, Collombet *et al.*

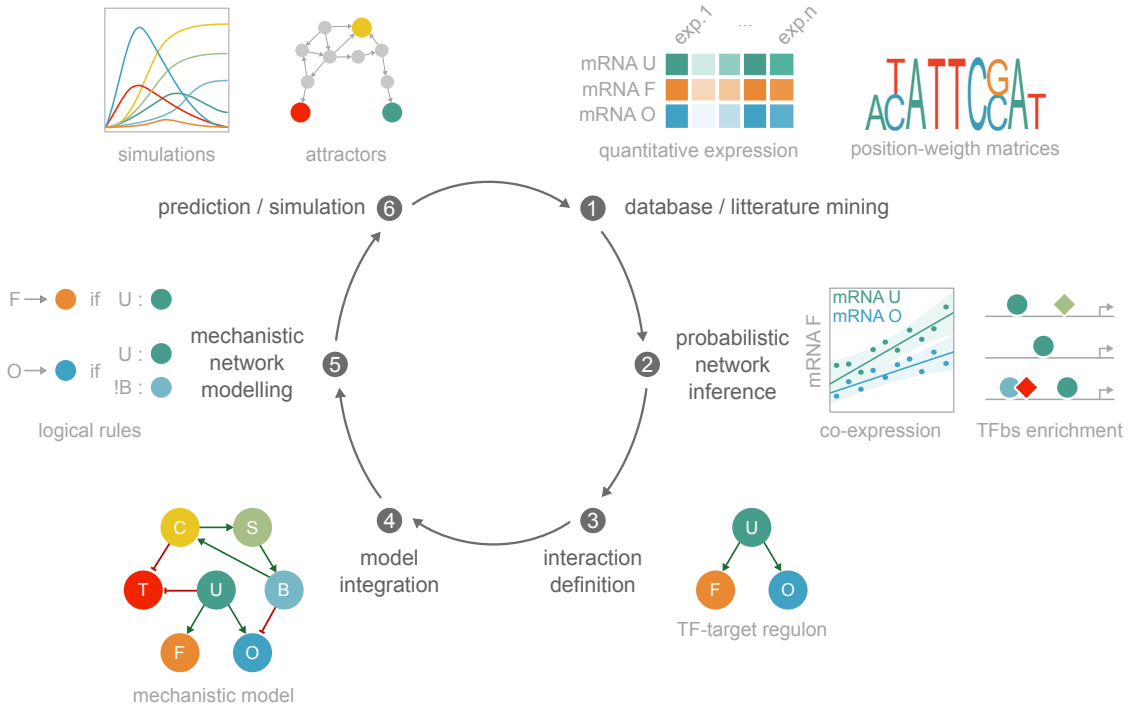

Figure 4.3: An integrative view of mechanistic and probabilistic networks. General idea of a workflow combining two network construction approaches. Starting from chiefly quantitative data (1), probabilistic inference methods (2) enable to characterise new TF-target interactions (3). These interactions are integrated into a pre-existing mechanistic model (4), together with refined rules adapting the network dynamic (5). The simulations and predictions resulting from the refined model are compared with experimental observations (6). A new iteration of this workflow is started to either (i) correct the model in case of disagreement with experimental data (ii) further refine the model in case of agreement.

could assess the potential role of candidate regulations, inferred from ChIP-seq data meta-analyses, by integrating them within the logical regulatory network of hematopoietic cell specification and assessing their impact on the simulations [201].

However, adapted tools are essential to scale up the current model-checking and regulatory rules inference methods toward a system-level GRN. Indeed, as model size increases, its dynamic complexity may become challenging to perform model checking efficiently. A potential improvement is the unit-testing approach (Hernandez, Naldi, *et al.* unpublished), where a network can be subdivided into modules that could be individually tested in depth.

In summary, combining the assets of probabilistic network inference and mechanistic GRN modelling offers exciting possibilities to better understand how enhancers implements the regulatory logic stemming from TF binding into a robust transcriptional signal (Fig. 4.3). With this perspective in mind, it is clear that proper combinations of biology and mathematics have a great potential to unveil the remaining secrets of the DNA.

List of Figures

1.1 Model organisms considered in this study. . . . 12
1.2 Contrasting views of the operon model. . . . 14
1.3 Enhancer-promoter looping. . . . 18
1.4 Drosophila in-situ 19
1.5 ChIP-seq TF processing. . . . 23
1.6 Two network modelling strategies 27

2.1 Linking genotypes and phenotypes 35
2.2 SNP-based read assignment 35
2.3 Cases of mapping bias in read allelic assignment 36
2.4 Genotyping error bias 38
2.5 Quantifying gene expression and regulatory element activity in hybrid embryos 44
2.6 Allelic imbalance is common across regulatory data types 47
2.7 Allelic imbalance is generally not predictive of developmental time . . 50
2.8 Allelic imbalance is propagated through regulatory layers via different epigenetic paths 52
2.9 Regulatory buffering varies across gene categories and with local chromatin structure 56
2.10 Chromatin features are more heritable than gene expression 58
2.11 General properties of the data – distribution of SNPs and reproducibility 67
2.12 Proportion and dispersion of SNPs and allelic ratios 68
2.13 Allelic variation is consistently depleted for transcriptional regulation and enriched for the expression of metabolic genes 69
2.14 Regulatory changes on paternal haplotypes are enriched for genes related to pesticide resistance and environmental response 70
2.15 Transcriptional regulators show reduced Heritability and smaller genetic effect sizes 71
2.16 Time and genotype effect over development 72
2.17 Partial correlation analysis reveals potentially causal relationships among regulatory layers 73
2.18 Differences in the frequency of *cis* and *trans* acting genetic variation among regulatory layers influences the heritability of regulatory phenotypes 74
2.19 Total count data cluster with similar time points 77
2.20 Maternally deposited transcripts show different rates of decay 80
2.21 AI and TSS distance thresholding increases correlation between regulatory layers 81
2.22 Mappability mask construction 93
2.23 Building a mappability mask 94

2.24 Testing for allelic imbalance in genomic DNA (gDNA) data 95
2.25 Maternal transcript correction 97
2.26 Defining genomic region overlaps 99
2.27 Assessing direct relationship using partial correlation 100
2.28 Detecting SNP co-segregating with allelic imbalance 101
2.29 Snakefile example workflow 102

3.1 TGF-β regulatory map. 108
3.2 Feedback circuits. 109
3.3 Logical rule construction. 111
3.4 State transition graphs. 113
3.5 Panda, Nodal and BMP2/4 signalling directs patterning of the Dorsal/Ventral axis of the sea urchin embryo 116
3.6 Iterative integration of biological data into the GRN model 118
3.7 Logical model integrating the main signalling pathways controlling specification the dorsal-ventral axis in the embryo of the sea urchin *Paracentrotus Lividus* 119
3.8 BMP2/4 and Nodal signalling antagonize each other to pattern the D/V axis of the sea urchin embryo 120
3.9 Simulation of early 32-cell stage and specification of later stage inputs 123
3.10 Comparison of blastula model simulations and experimental results . 124
3.11 Probabilistic time-course simulations of the unicellular model starting with the ventral initial state 125
3.12 Multicellular logical simulations for wild-type and mutant conditions, using the software Epilog 127
3.13 EpiLog simulation of the Lefty morpholino condition 129
3.14 Simulation of the boundary ectoderm in the unicellular model 130
3.15 Simulation of *panda* perturbations at the 32-cell stage 131
3.16 Jupyter notebook. 139

4.1 Distinguishing *cis*- from *trans*-acting variants. 145
4.2 Valley detection in H3K27ac signal. 147
4.3 An integrative view of mechanistic and probabilistic networks. 152

Bibliography

1. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362,** 709–715 (1993).
2. Spielberg, S. *JURASSIC PARK* 1993.
3. Gilbert, S. F. *Developmental Biology* 6th (Sinauer Associates, 2000).
4. Laubichler, M. D. & Davidson, E. H. Boveri's long experiment: Sea urchin merogones and the establishment of the role of nuclear chromosomes in development. *Developmental biology* **314,** 1–11 (2008).
5. Morgan, T. H. The fertilization of non-nucleated fragments of Echinoderm-eggs. *Archiv für Mikroskopische Anatomie* **2,** 268–280 (1895).
6. Alberts, B. *et al. Molecular Biology of the Cell* 4th (Garland Science, 2002).
7. Brenner, S., Dove, W., Herskowitz, I. & Thomas, R. Genes and development: molecular and logical themes. *Genetics* **126,** 479–486 (1990).
8. Soyfer, V. N. The consequences of political dictatorship for Russian science. *Nature Reviews. Genetics* **2,** 723–729 (2001).
9. Maddox, B. The double helix and the 'wronged heroine'. *Nature* **421,** 407–408 (2003).
10. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3,** 318–356 (1961).
11. Peluffo, A. E. The "Genetic Program": Behind the Genesis of an Influential Metaphor. *Genetics* **200,** 685–696 (2015).
12. Yaniv, M. The 50th anniversary of the publication of the operon theory in the Journal of Molecular Biology: past, present and future. *Journal of Molecular Biology* **409,** 1–6 (2011).
13. Monod, J. & Jacob, F. General Conclusions: Teleonomic Mechanisms in Cellular Metabolism, Growth, and Differentiation. *Cold Spring Harbor Symposia on Quantitative Biology* **26,** 389–401 (1961).
14. Ptashne, M. Specific binding of the lambda phage repressor to lambda DNA. *Nature* **214,** 232–234 (1967).
15. Morange, M. Quelle place pour l'épigénétique ? *médecine/sciences* **21,** 367–369 (2005).
16. Waddington, C. H. *The Strategy Of The Genes* (1957).
17. Arnone, M. I. & Davidson, E. H. The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124,** 1851–1864 (1997).

18. Britten, R. J. & Davidson, E. H. Gene Regulation for Higher Cells: A Theory. *Science* **165,** 349–357 (1969).

19. Wieschaus, E. & Nüsslein-Volhard, C. The Heidelberg Screen for Pattern Mutants of *Drosophila* : A Personal Account. *Annual Review of Cell and Developmental Biology* **32,** 1–46 (2016).

20. Sugita, M. Functional analysis of chemical systems in vivo using a logical circuit equivalent. II. The idea of a molecular automaton. *Journal of Theoretical Biology* **4,** 179–192 (1963).

21. De Jong, H. Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *Journal of Computational Biology* **9,** 67–103 (2002).

22. Thomas, R. Boolean formalization of genetic control circuits. *Journal of Theoretical Biology* **42,** 563–585 (1973).

23. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* **13,** 613–626 (2012).

24. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics,* 1–17 (2019).

25. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics,* 1–19 (2020).

26. Mikhaylichenko, O. *et al.* The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes & Development* **32,** 42–57 (2018).

27. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics* **12,** 7–18 (2011).

28. Rivera, C. M. & Ren, B. Mapping Human Epigenomes. *Cell* **155,** 39–55 (2013).

29. Furlong, E. E. M. & Levine, M. Developmental enhancers and chromosome topology. *Science* **361,** 1341–1345 (2018).

30. Wang, X. *et al.* Analysis of Genetic Variation Indicates DNA Shape Involvement in Purifying Selection. *Molecular Biology and Evolution* **35,** 1958–1967 (2018).

31. Olins, D. E. & Olins, A. L. Chromatin history: our view from the bridge. *Nature Reviews. Molecular Cell Biology* **4,** 809–814 (2003).

32. Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease. *Nature* **571,** 489–499 (2019).

33. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128,** 693–705 (2007).

34. Reinberg, D. & Vales, L. D. Chromatin domains rich in inheritance. *Science* **361,** 33–34 (2018).

35. Steensel, B. v. & Furlong, E. E. M. The role of transcription in shaping the spatial organization of the genome. *Nature Reviews Molecular Cell Biology* **20,** 327–337 (2019).

36. Bolt, C. C. & Duboule, D. The regulatory landscapes of developmental genes. *Development* **147** (2020).

37. Magnani, L., Eeckhoute, J. & Lupien, M. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends in Genetics* **27,** 465–474 (2011).

38. Bozek, M. & Gompel, N. Developmental Transcriptional Enhancers: A Subtle Interplay between Accessibility and Activity. *BioEssays* **42,** 1900188 (2020).

39. Kvon, E. Z. *et al.* Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512,** 91–95 (2014).

40. Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E. E. M. Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature* **462,** 65–70 (2009).

41. Saudemont, A. *et al.* Ancestral Regulatory Circuits Governing Ectoderm Patterning Downstream of Nodal and BMP2/4 Revealed by Gene Regulatory Network Analysis in an Echinoderm. *PLOS Genetics* **6,** e1001259 (2010).

42. Davidson, E. H. *et al.* A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. *Developmental Biology* **246,** 162–190 (2002).

43. Peter, I. S. & Davidson, E. H. A gene regulatory network controlling the embryonic specification of endoderm. *Nature* **474,** 635–639 (2011).

44. Rivera, J., Keränen, S. V. E., Gallo, S. M. & Halfon, M. S. REDfly: the transcriptional regulatory element database for Drosophila. *Nucleic Acids Research* **47,** D828–D834 (2019).

45. Kudtarkar, P. & Cameron, R. A. Echinobase: an expanding resource for echinoderm genomic information. *Database: The Journal of Biological Databases and Curation* **2017** (2017).

46. Elkon, R. & Agami, R. Characterization of noncoding regulatory DNA in the human genome. *Nature Biotechnology* **35,** 732–746 (2017).

47. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107,** 1–8 (2016).

48. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* **21,** 22 (2020).

49. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* **22,** 1813–1831 (2012).

50. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).

51. Van den Berge, K. *et al.* RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. *Annual Review of Biomedical Data Science* **2,** 139–173 (2019).

52. Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews. Genetics* **15,** 709–721 (2014).

53. Jacobs, J. *et al.* The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. *Nature Genetics* **50,** 1011 (2018).

54. Bozek, M. *et al.* ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the Drosophila blastoderm. *Genome Research* **29,** 771–783 (2019).

55. He, C. & Bonasio, R. A cut above. *eLife* **6,** e25000 (2017).

56. He, Q., Johnston, J. & Zeitlinger, J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology* **33,** 395–401 (2015).

57. Madrigal, P. On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions. *Frontiers in Bioengineering and Biotechnology* **3** (2015).

58. Baranello, L., Kouzine, F., Sanford, S. & Levens, D. ChIP bias as a function of cross-linking time. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* **24,** 175–181 (2016).

59. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics* **19,** 329–346 (2018).

60. Ludwig, C. H. & Bintu, L. Mapping chromatin modifications at the single cell level. *Development* **146,** dev170217 (2019).

61. Sagar & Grün, D. Deciphering Cell Fate Decision by Integrated Single-Cell Sequencing Analysis. *Annual Review of Biomedical Data Science* **3,** – (2020).

62. Lakadamyali, M. & Cosma, M. P. Visualizing the genome in high resolution challenges our textbook understanding. *Nature Methods,* 1–9 (2020).

63. Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics* **14,** 671–683 (2013).

64. Bailey, T. *et al.* Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS computational biology* **9,** e1003326 (2013).

65. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11,** R25 (2010).

66. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).

67. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

68. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).

69. Feng, J., Liu, T. & Zhang, Y. Using MACS to Identify Peaks from ChIP-Seq Data. *Current Protocols in Bioinformatics* **34,** 2.14.1–2.14.14 (2011).

70. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550 (2014).

71. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

72. Lun, A. T. & Smyth, G. K. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research* **44,** e45 (2016).

73. Thomas-Chollier, M. *et al.* RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research* **40,** e31–e31 (2012).

74. Nguyen, N. T. T. *et al.* RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Research* **46,** W209–W214 (2018).

75. Medina-Rivera, A. *et al.* RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Research* **43,** W50–W56 (2015).

76. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16,** 16–23 (2000).

77. Lin, Q. X. X., Thieffry, D., Jha, S. & Benoukraf, T. TFregulomeR reveals transcription factors' context-specific features and functions. *Nucleic Acids Research* **48,** e10–e10 (2020).

78. Harmanci, A., Harmanci, A. S., Swaminathan, J. & Gopalakrishnan, V. EpiSAFARI: sensitive detection of valleys in epigenetic signals for enhancing annotations of functional elements. *Bioinformatics* **36,** 1014–1021 (2020).

79. Ecker, J. R. *et al.* ENCODE explained. *Nature* **489,** 52–54 (2012).

80. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology* **28,** 1045–1048 (2010).

81. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE* **5,** e12776 (2010).

82. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology* **9,** 770–780 (2008).

83. Mbodj, A. *et al.* Qualitative Dynamical Modelling Can Formally Explain Mesoderm Specification and Predict Novel Developmental Phenotypes. *PLOS Computational Biology* **12,** e1005073 (2016).

84. Wilczynski, B. & Furlong, E. E. M. Challenges for modeling global gene regulatory networks during development: Insights from Drosophila. *Developmental Biology. Special Section: Gene Regulatory Networks for Development* **340,** 161–169 (2010).

85. Penfold, C. A. & Wild, D. L. How to infer gene networks from expression profiles, revisited. *Interface Focus* **1,** 857–870 (2011).

86. Hawe, J. S., Theis, F. J. & Heinig, M. Inferring Interaction Networks From Multi-Omics Data. *Frontiers in Genetics* **10** (2019).

87. Huynh-Thu, V. A. & Sanguinetti, G. Gene regulatory network inference: an introductory survey. *arXiv,* 1801.04087 (2018).

88. Opgen-Rhein, R. & Strimmer, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology* **1,** 37 (2007).

89. Lee, B., Zhang, S., Poleksic, A. & Xie, L. Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis. *Frontiers in Genetics* **10,** 1381 (2019).

90. Hüebschmann, D. *et al.* Deciphering programs of transcriptional regulation by combined deconvolution of multiple omics layers. *bioRxiv,* 199547 (2017).

91. Andrieu, C., de Freitas, N., Doucet, A. & Jordan, M. I. An Introduction to MCMC for Machine Learning. *Machine Learning* **50,** 5–43 (2003).

92. Huynh-Thu, V. A. & Geurts, P. dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. *Scientific Reports* **8,** 1–12 (2018).

93. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics* **22,** 403–434 (1976).

94. Traynard, P., Fauré, A., Fages, F. & Thieffry, D. Logical model specification aided by model-checking techniques: application to the mammalian cell cycle regulation. *Bioinformatics* **32,** i772–i780 (2016).

95. Stoll, G., Viara, E., Barillot, E. & Calzone, L. Continuous time Boolean modeling for biological signaling: application of Gillespie algorithm. *BMC systems biology* **6,** 116 (2012).

96. Thomas, R. & D'Ari, R. *Biological feedback* (CRC Press, Boca Raton, FL etc., 1990).

97. Ghavi-Helm, Y. *et al.* Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature Genetics* **51,** 1272–1282 (2019).

98. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287,** 2185–2195 (2000).

99. Mackay, T. F. C. *et al.* The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482,** 173–178 (2012).

100. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20,** 467–484 (2019).

101. Cannavò, E. *et al.* Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature* **541,** 402–406 (2017).

102. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS biology* **8,** e1000294 (2010).

103. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics* **11,** 533–538 (2010).

104. Signor, S. A. & Nuzhdin, S. V. The Evolution of Gene Expression in cis and trans. *Trends in Genetics* **34,** 532–544 (2018).

105. Bao, Y. *et al.* Unraveling cis and trans regulatory evolution during cotton domestication. *Nature Communications* **10,** 1–12 (2019).

106. Connelly, C. F., Wakefield, J. & Akey, J. M. Evolution and Genetic Architecture of Chromatin Accessibility and Function in Yeast. *PLOS Genetics* **10,** e1004427 (2014).

107. Tirosh, I., Reikhav, S., Levy, A. A. & Barkai, N. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324,** 659–662 (2009).

108. Wong, E. S. *et al.* Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. *Nature Communications* **8,** 1092 (2017).

109. Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503,** 487–492 (2013).

110. Goncalves, A. *et al.* Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Research* **22,** 2376–2384 (2012).

111. Fear, J. M. *et al.* Buffering of Genetic Regulatory Networks in Drosophila melanogaster. *Genetics* **203,** 1177–1190 (2016).

112. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Regulatory changes underlying expression differences within and between Drosophila species. *Nature Genetics* **40,** 346–350 (2008).

113. McManus, C. J. *et al.* Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Research* **20,** 816–825 (2010).

114. Coolon, J. D., McManus, C. J., Stevenson, K. R., Graveley, B. R. & Wittkopp, P. J. Tempo and mode of regulatory evolution in Drosophila. *Genome Research* **24,** 797–808 (2014).

115. Osada, N., Miyagi, R. & Takahashi, A. Cis- and Trans-regulatory Effects on Gene Expression in a Natural Population of Drosophila melanogaster. *Genetics* **206,** 2139–2148 (2017).

116. Quinn, A., Juneja, P. & Jiggins, F. M. Estimates of allele-specific expression in Drosophila with a single genome sequence and RNA-seq data. *Bioinformatics* **30,** 2603–2610 (2014).

117. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic Variation in Human Gene Expression. *Science* **297,** 1143–1143 (2002).

118. Moyerbrailean, G. A. *et al.* High-throughput allele-specific expression across 250 environmental conditions. *Genome Research* **26,** 1627–1638 (2016).

119. Knowles, D. A. *et al.* Allele-specific expression reveals interactions between genetic variation and environment. *Nature Methods* **14,** 699–702 (2017).

120. Kilpinen, H. *et al.* Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription. *Science* **342,** 744–747 (2013).

121. Waszak, S. M. *et al.* Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell* **162,** 1039–1050 (2015).

122. Stevenson, K. R., Coolon, J. D. & Wittkopp, P. J. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics* **14,** 536 (2013).

123. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25,** 3207–3212 (2009).

124. Yuan, S. *et al.* One Size Doesn't Fit All - RefEditor: Building Personalized Diploid Reference Genome to Improve Read Mapping and Genotype Calling in Next Generation Sequencing Studies. *PLOS Computational Biology* **11,** e1004448 (2015).

125. Coolon, J. D., Stevenson, K. R., McManus, C. J., Graveley, B. R. & Wittkopp, P. J. Genomic imprinting absent in Drosophila melanogaster adult females. *Cell Reports* **2,** 69–75 (2012).

126. Geijn, B. v. d., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods* **12,** 1061–1063 (2015).

127. Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics* **6,** 847–859 (2005).

128. Koster, J. & Rahmann, S. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).

129. Qiao, W. *et al.* PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. *PLoS Computational Biology* **8** (2012).

130. Anghel, C. V. *et al.* ISOpureR: an R implementation of a computational purification algorithm of mixed tumour profiles. *BMC Bioinformatics* **16,** 156 (2015).

131. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28,** 495–501 (2010).

132. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31,** 2382–2383 (2015).

133. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols* **12,** 2478–2492 (2017).

134. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology* **9,** e1003118 (2013).

135. Santana-Garcia, W. *et al.* RSAT Var-tools: an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding. *bioRxiv,* 623090 (2019).

136. Naldi, A. *et al.* Logical Modeling and Analysis of Cellular Regulatory Networks With GINsim 3.0. *Frontiers in Physiology* **9,** 646 (2018).

137. Chaouiya, C., Naldi, A. & Thieffry, D. Logical modelling of gene regulatory networks with GINsim. *Methods in Molecular Biology* **804,** 463–479 (2012).

138. Naldi, A. *et al.* Logical modelling of regulatory networks with GINsim 2.3. *Bio Systems* **97,** 134–139 (2009).

139. Gonzalez, A. G., Naldi, A., Sánchez, L., Thieffry, D. & Chaouiya, C. GINsim: a software suite for the qualitative modelling, simulation and analysis of regulatory networks. *Bio Systems* **84,** 91–100 (2006).

140. Stoll, G. *et al.* MaBoSS 2.0: an environment for stochastic Boolean modeling. *Bioinformatics* **33,** 2226–2228 (2017).

141. Varela, P. L., Ramos, C. V., Monteiro, P. T. & Chaouiya, C. EpiLog: A software for the logical modelling of epithelial dynamics. *F1000Research* **7,** 1145 (2019).

142. Lapraz, F., Haillot, E. & Lepage, T. A deuterostome origin of the Spemann organiser suggested by Nodal and ADMPs functions in Echinoderms. *Nature Communications* **6,** 8434 (2015).

143. Wu, M. Y. & Hill, C. S. TGF-β Superfamily Signaling in Embryonic Development and Homeostasis. *Developmental Cell* **16,** 329–343 (2009).

144. Batlle, E. & Massagué, J. Transforming Growth Factor-β Signaling in Immunity and Cancer. *Immunity* **50,** 924–940 (2019).

145. Kitano, H., Funahashi, A., Matsuoka, Y. & Oda, K. Using process diagrams for the graphical representation of biological networks. *Nature Biotechnology* **23,** 961–966 (2005).

146. Longabaugh, W. J., Davidson, E. H. & Bolouri, H. Visualization, documentation, analysis, and communication of large scale gene regulatory networks. *Biochimica et biophysica acta* **1789,** 363–374 (2009).

147. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19,** 524–531 (2003).

148. Thieffry, D. Dynamical roles of biological regulatory circuits. *Briefings in Bioinformatics* **8,** 220–225 (2007).

149. Thomas, R., Thieffry, D. & Kaufman, M. Dynamical behaviour of biological regulatory networks–I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bulletin of Mathematical Biology* **57,** 247–276 (1995).

150. Hermsen, R., Erickson, D. W. & Hwa, T. Speed, sensitivity, and bistability in autoactivating signaling circuits. *PLoS computational biology* **7,** e1002265 (2011).

151. Ferrell, J. E. Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Current Opinion in Cell Biology* **14,** 140–148 (2002).

152. Rosenfeld, N., Elowitz, M. B. & Alon, U. Negative autoregulation speeds the response times of transcription networks. *Journal of Molecular Biology* **323,** 785–793 (2002).

153. Thieffry, D., Huerta, A. M., Pérez-Rueda, E. & Collado-Vides, J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *BioEssays* **20,** 433–440 (1998).

154. Traynard, P., Feillet, C., Soliman, S., Delaunay, F. & Fages, F. Model-based investigation of the circadian clock and cell cycle coupling in mouse embryonic fibroblasts: Prediction of RevErb-α up-regulation during mitosis. *Bio Systems* **149,** 59–69 (2016).

155. Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology* **22,** 437–467 (1969).

156. Müssel, C., Hopfensitz, M. & Kestler, H. A. BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics* **26,** 1378–1380 (2010).

157. Di Cara, A., Garg, A., De Micheli, G., Xenarios, I. & Mendoza, L. Dynamic simulation of regulatory networks using SQUAD. *BMC bioinformatics* **8,** 462 (2007).

158. Faure, E., Peter, I. S. & Davidson, E. H. A new software package for predictive gene regulatory network modeling and redesign. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* **20,** 419–423 (2013).

159. Peter, I. S., Faure, E. & Davidson, E. H. Predictive computation of genomic logic processing functions in embryonic development. *Proceedings of the National Academy of Sciences* **109,** 16434–16442 (2012).

160. Helikar, T. *et al.* The Cell Collective: Toward an open and collaborative approach to systems biology. *BMC Systems Biology* **6,** 96 (2012).

161. Thomas, R. Regulatory networks seen as asynchronous automata: A logical description. *Journal of Theoretical Biology* **153,** 1–23 (1991).

162. Garg, A., Di Cara, A., Xenarios, I., Mendoza, L. & De Micheli, G. Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics (Oxford, England)* **24,** 1917–1925 (2008).

163. Naldi, A. *et al.* The CoLoMoTo Interactive Notebook: Accessible and Reproducible Computational Analyses for Qualitative Biological Networks. *Frontiers in Physiology* **9** (2018).

164. Naldi, A. BioLQM: A Java Toolkit for the Manipulation and Conversion of Logical Qualitative Models of Biological Networks. *Frontiers in Physiology* **9** (2018).

165. Pauleve, L. Reduction of Qualitative Models of Biological Networks for Transient Dynamics Analysis. *IEEE/ACM transactions on computational biology and bioinformatics* **15,** 1167–1179 (2018).

166. Chaouiya, C. *et al.* SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC systems biology* **7,** 135 (2013).

167. Hödl, M. & Basler, K. Transcription in the absence of histone H3.3. *Current biology: CB* **19,** 1221–1226 (2009).

168. Howe, F. S., Fischl, H., Murray, S. C. & Mellor, J. Is H3K4me3 instructive for transcription activation? *BioEssays* **39,** e201600095 (2017).

169. Park, S., Kim, G. W., Kwon, S. H. & Lee, J.-S. Broad domains of histone H3 lysine 4 trimethylation in transcriptional regulation and disease. *The FEBS journal* (2020).

170. Benayoun, B. A. *et al.* H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **158,** 673–688 (2014).

171. Barnes, C. E., English, D. M. & Cowley, S. M. Acetylation & Co: an expanding repertoire of histone acylations regulates chromatin and transcription. *Essays in Biochemistry* **63,** 97–107 (2019).

172. Greer, C. B. *et al.* Histone Deacetylases Positively Regulate Transcription through the Elongation Machinery. *Cell Reports* **13,** 1444–1455 (2015).

173. Zhang, T., Zhang, Z., Dong, Q., Xiong, J. & Zhu, B. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biology* **21** (2020).

174. Raisner, R. *et al.* Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation. *Cell Reports* **24,** 1722–1729 (2018).

175. Villaseñor, R. *et al.* ChromID identifies the protein interactome at chromatin marks. *Nature Biotechnology,* 1–9 (2020).

176. Haillot, E., Molina, M. D., Lapraz, F. & Lepage, T. The Maternal Maverick/GDF15-like TGF-β Ligand Panda Directs Dorsal-Ventral Axis Formation by Restricting Nodal Expression in the Sea Urchin Embryo. *PLOS Biology* **13,** e1002247 (2015).

177. Hasin-Brumshtein, Y. *et al.* Allele-specific expression and eQTL analysis in mouse adipose tissue. *BMC genomics* **15,** 471 (2014).

178. Frochaux, M. V. *et al.* cis-regulatory variation modulates susceptibility to enteric infection in the Drosophila genetic reference panel. *Genome Biology* **21,** 6 (2020).

179. Ballouz, S., Dobin, A. & Gillis, J. A. Is it time to change the reference genome? *Genome Biology* **20,** 159 (2019).

180. Aleman, F. The Necessity of Diploid Genome Sequencing to Unravel the Genetic Component of Complex Phenotypes. *Frontiers in Genetics* **8** (2017).

181. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology* **36,** 875–879 (2018).

182. Audoux, J. *et al.* DE-kupl: Exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. *Genome Biology* **18,** 1–15 (2017).

183. Pundhir, S., Bagger, F. O., Lauridsen, F. B., Rapin, N. & Porse, B. T. Peak-valley-peak pattern of histone modifications delineates active regulatory elements and their directionality. *Nucleic Acids Research* **44,** 4037–4051 (2016).

184. Knijnenburg, T. A. *et al.* Multiscale representation of genomic signals. *Nature Methods* **11,** 689–694 (2014).

185. Meers, M. P., Janssens, D. H. & Henikoff, S. Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded. *Molecular Cell* **75,** 562–575.e5 (2019).

186. Audit, B. *et al.* Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm. *Nature Protocols* **8,** 98–110 (2013).

187. Thormann, V. *et al.* Expanding the repertoire of glucocorticoid receptor target genes by engineering genomic response elements. *Life Science Alliance* **2** (2019).

188. Starick, S. R. *et al.* ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome Research* **25,** 825–835 (2015).

189. Biddie, S. C. *et al.* Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Molecular Cell* **43,** 145–155 (2011).

190. Schöne, S. *et al.* Synthetic STARR-seq reveals how DNA shape and sequence modulate transcriptional output and noise. *PLOS Genetics* **14,** e1007793 (2018).

191. Fauré, A., Naldi, A., Chaouiya, C. & Thieffry, D. Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinformatics* **22,** e124–e131 (2006).

192. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods* **13,** 966–967 (2016).

193. Bonneau, R. *et al.* The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology* **7,** R36 (2006).

194. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research* **46,** D260–D266 (2018).

195. Herrmann, C., Van de Sande, B., Potier, D. & Aerts, S. i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Research* **40,** e114–e114 (2012).

196. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nature Methods* **14,** 1083–1086 (2017).

197. Liu, L. *et al.* Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nature Communications* **10,** 1–10 (2019).

198. González-Blas, C. B. *et al.* Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. *bioRxiv,* 2019.12.19.882381 (2019).

199. Atak, Z. K. *et al.* Prioritization of enhancer mutations by combining allele-specific chromatin accessibility with deep learning. *bioRxiv,* 2019.12.21.885806 (2019).

200. Bizzarri, M. *et al.* A call for a better understanding of causation in cell biology. *Nature Reviews Molecular Cell Biology* **20,** 261–262 (2019).

201. Collombet, S. *et al.* Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proceedings of the National Academy of Sciences* **114,** 5792–5799 (2017).

RÉSUMÉ

La formation d'un embryon est dictée par la séquence ADN propre à cet organisme. La variabilité génétique donne naissance à une grande diversité morphologique, tout en maintenant une organisation générale robuste. Les mutations présentes dans les régions *cis*-régulatrices impactent la transcription via des mécanismes épigénomiques. La variabilité d'expression génique qui en découle peut être compensée par des mécanismes *trans* de rétrocontrôle au sein du réseau de régulation. L'organisation précise de ces interactions *cis* et *trans* restent encore difficile à déchiffrer.

Afin de mieux saisir l'effet des mutations sur la transcription, j'ai analysé des données génétiques, épigénomiques et transcriptomiques en collaboration avec le laboratoire Furlong (EMBL, Heidelberg). L'utilisation de données allèle-spécifiques de lignées F1 de Drosophile a permis d'inférer les interactions directes en *cis* entre les niveaux de régulation, suggérant une différence d'action des marques épigénétiques H3K27ac et H3K4me3 sur l'expression des gènes.

Pour mieux comprendre l'impact en *trans* de la structure des réseaux de régulation sur l'expression génique, j'ai ensuite construit un modèle logique de la spécification de l'axe dorso-ventral chez l'embryon d'oursin, en collaboration avec le laboratoire Lepage (iBV, Nice). Les analyses multicellulaires et stochastiques ont permis de détecter les composants clés du réseau, notamment la dynamique de répression mutuelle entre Nodal et BMP. En conclusion, l'analyse de données allèle-spécifique et la modélisation logique m'ont permis de d'étudier les mécanismes de la régulation transcriptionnelle sous deux perspectives complémentaires.

MOTS CLÉS

régulation transcriptionnelle ; bioinformatique ; déséquilibre allélique ; modélisation logique ; signalisation TGF-beta ; épigénomique ; transcriptomique ; découverte de motifs ADN

ABSTRACT

The development of an embryo derives from the DNA sequence of this organism. Genetic variability gives rise to great morphological diversity, while maintaining a robust general organisation. Mutations present within *cis*-regulatory regions impact transcription via epigenomic mechanisms. The resulting variability in gene expression can be buffered by *tran* feedback mechanisms within the regulatory network. The precise organisation of these *cis* and *trans* interactions remains difficult to decipher.

In order to better grasp the effect of mutations on transcription, I analysed genetic, epigenomic and transcriptomic data in collaboration with the Furlong laboratory (EMBL, Heidelberg). The use of allele-specific data from Drosophila F1 lines enabled to infer direct *cis*-interactions between the regulatory layers, suggesting a difference in the action of the epigenomic markers H3K27ac and H3K4me3 on gene expression.

To better understand the *trans* impact of the structure of regulatory networks on gene expression, I have built a logical model of the dorsal-ventral axis specification in sea urchin embryo, in collaboration with the Lepage laboratory (iBV, Nice). Multicellular and stochastic analyses permitted to detect key components of the network, including the cross-repression dynamic between Nodal and BMP. To conclude, allele-specific data analysis and logical modelling allowed me to study the mechanisms of transcription regulation from two complementary perspectives.

KEYWORDS

transcriptional regulation ; bioinformatics ; allelic imbalance ; logical modelling ; TGF-beta signalling ; epigenomics ; transcriptomics ; DNA motif discovery