Dissertation

submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

Mattia Forneris
born in: Savigliano, Italy

Oral examination: 25 September 2019

Natural sequence variation as a tool to dissect gene expression regulation in *Drosophila melanogaster*

Referees: Prof. Dr. Henrik Kaessmann

Dr. Oliver Stegle

Summary

Genetic variation is a major cause of differences between individuals and it represents a powerful tool to study gene regulation. By interfering with *cis*-Regulatory Modules (CRMs), variants can unravel CRM function. On the other hand, predicting the effect of variants on phenotype by the DNA sequence has proven to be challenging. In this thesis, I use *Drosophila* embryonic development as a model system to study diversity in gene regulation at the transcriptional level.

CRMs can be characterized using multiple genome-wide techniques such as DNase hypersensitivity. However, despite having comprehensive CRM maps, it is still difficult to predict what are the genes regulated by each CRM. Functional methods, such as mutagenesis, are effective but poorly scalable. To address this issue, I developed an eQTL method (called DHS-eQTL) that makes use of naturally occurring genetic variation, to associate CRMs with the genes they regulate. The results reveal 2,967 DHS-eQTLs and indicate a high extent of CRM sharing between genes. We validated the results with *in silico* and *in vitro* approaches and I discuss upcoming *in vivo* experiments. We observed long-range enhancer regulation suggesting that commonly used methods to associate genes and enhancers underestimate their distance. Also, the DHS-eQTLs show that promoter-proximal CRMs have widespread distal activity.

The separation between populations causes an increase in genetic differences by drift and adaptation to different environments. We investigated gene expression differences between *Drosophila* populations from five continents by performing RNA-Seq on 80 inbred fly lines. We performed multiple quality-control tests to ensure that the gene expression dataset is of high quality. Gene expression profiles show detectable diversity among the fly lines from different continents and confirm what has been observed at the genetic level. In particular, the African population is the most separated, while the American, European and Australian ones show less diversity. In addition, we identified 903 gene and 2,021 exon eQTLs.

Genetic variants can interfere with Transcription Factor Binding Sites (TFBS) and this might, in turn, lead to changes in chromatin accessibility. We applied LS-GKM (an SVM method that uses gapped k-mers) to learn sequence features of tissue-specific accessible chromatin and predict the impact of natural sequence variation on

accessibility. We train LS-GKM on six tissue-specific training sets: neuroectodermal, mesodermal and double negative CRMs divided in promoter-proximal and promoter-distal. The method unbiasedly recovers tissue-specific TFBS and shows good performance despite the small training sets. Finally, we score variants from groups of inbred *Drosophila* lines. Interestingly, rare variants have a higher impact on accessibility.

Zusammenfassung

Genetische Variation ist eine der Hauptursachen für die Unterschiede, die zwischen Individuen bestehen, und stellt ein wirksames Mittel zur Untersuchung der Genregulation dar. Variationen, die *cis*-regulatorische Module (CRMs) beeinflussen, können helfen deren Funktion zu entschlüsseln. Allerdings ist die Prognose der Auswirkungen genetischer Variationen auf den Phänotypen anhand der DNA-Sequenz immer noch schwierig. In der vorliegenden Arbeit nutze ich die Embryonalentwicklung der Fruchtfliege *Drosophila* als Modellsystem, um mithilfe der genetischen Diversität die Genregulation auf transkriptioneller Ebene zu untersuchen.

CRMs können durch verschiedene genomweite Methoden, wie beispielsweise der „DNase hypersensitivity“, beschrieben werden. Obwohl so umfassende CRM-Karten erstellt werden konnten, ist die Zuordnung der Gene zu den CRMs, durch die sie reguliert werden, weiterhin schwierig. Funktionale Methoden, wie die Metagenese, sind effektiv, können aber nur unzureichend auf das gesamte Genom angewandt werden. Um auf dieses Problem einzugehen, habe ich eine eQTL Methodik entwickelt (genannt DHS-eQTL), welche sich der natürlich vorkommenden genetischen Variation bedient, um CRMs den von ihnen regulierten Genen zuzuordnen. 2.967 DHS-eQTLs wurden identifiziert und ich konnte zeigen, dass CRMs häufig mehrere Gene regulieren. Die Ergebnisse wurden mit *in silico* und *in vitro* Methoden validiert und ich diskutiere anstehende *in vivo* Experimente. Unsere Beobachtungen zeigen des Weiteren, dass Enhancer ihre Zielgene häufig über größere genomische Distanzen hinweg regulieren, und legen damit nahe, dass gemeinhin verwendete Methoden für die Zuordnung von Genen und Enhancern deren Distanz unterschätzen. Darüberhinaus zeigen die DHS-eQTLs, dass Promoter-proximale CRMs umfassende distale Aktivität aufweisen.

Die räumliche Trennung von Populationen führt zu einer Zunahme der genetischen Unterschiede zwischen diesen, verursacht durch Drift und Adaption an die verschiedenen Umweltfaktoren. Wir haben die Genexpressionsunterschiede zwischen *Drosophila* Populationen von fünf Kontinenten untersucht. Dazu wurde RNA-seq an 80 Inzuchtfliegenlinien durchgeführt. Die hohe Qualität der resultierenden Datensätze wurde durch verschiedene Qualitätskontrollen sichergestellt. Die Genexpressionsprofile zeigen eine nachweisbare Diversität zwischen den Fliegenlinien der verschiedenen Kontinente und bestätigen damit was

bereits auf genetischer Ebene beobachtet wurde: Die afrikanische Population grenzt sich am stärksten ab, während die amerikanische, europäische und australische weniger Diversität aufweisen. Darüberhinaus konnten wir 903 Gen- und 2.021 Exon-eQTLs identifizieren.

Der genetischen Variation liegen Änderungen in der DNA-Sequenz zugrunde und diese Änderungen können Transkriptionsfaktorbindestellen (TFBS) stören. Diese wiederum können zu einer Veränderung des Chromatins führen (offen/geschlossen oder „accessible/inaccessible“). Wir haben LS-GKM angewendet (eine SVM Methode, die „gapped k-mers“ verwendet), um die Sequenzeigenschaften von gewebespezifischer „chromatin accessibility“ zu lernen und den Einfluss von natürlichen Sequenzvariationen auf diese Zugänglichkeit zu Chromatin vorherzusagen. Dafür haben wir LS-GKM mit sechs gewebespezifischen Datensets trainiert: neuroektodermale, mesodermale und doppelt-negative CRMs, jeweils unterteilt in Promoter-proximale und Promoter-distale Sequenzen. Trotz dieses kleinen Trainingssets erbringt die Methode gute Leistungen und findet in unvoreingenommener Weise gewebespezifische TFBS. Abschließend bewerten wir Varianten von verschieden Gruppen inzüchtiger *Drosophila*-Linien. Interessanterweise zeigt sich dabei, dass seltene Varianten einen größeren Einfluss auf die Chromatin Zugänglichkeit haben.

Table of Contents

I - Introduction

1 - Gene expression is regulated by a plethora of *cis* Regulatory Modules

1.1 - Gene expression drives cell diversity

In this section, I will discuss the importance of gene regulation to shape cell diversity during embryonic development. Precise spatio-temporal regulation of gene expression is fundamental for cell differentiation and the definition of anatomical structures. In addition, I will introduce *Drosophila melanogaster* as a model organism to study embryonic development.

1.1.1 - All cells of each organism have the same genome but express different sets of genes

Multicellular organisms are complex systems that can perform many functions. From tissue regeneration and food digestion to memory and movement, each process is carried out by specialized cell types. Cell types are different in both their structure and the purpose they fulfill. For example, neurons are very elongated cells with multiple branches that form connections with other neurons and transmit electrical signals. Hepatocytes, on the other hand, are much smaller and round in shape. Their main functions are linked to metabolism and protein synthesis. Despite huge diversity between cell types, all cells of multicellular organisms contain the same DNA (with a few exceptions, such as erythrocytes or lymphocytes). This means that every cell contains all the information necessary to generate a full organism and, consequently, any other cell type. Differences between cells arise because of the expression of a specific subset of genes in each cell. The messenger RNA (mRNA) is then translated into protein that characterizes cell-specific structure and function. For example,

human cells generally express between 30 and 60% of all genes encoded in the genome[1].

Gene expression is controlled at many stages. The regulation of transcription is the first and arguably the most important step. A gene can be switched “on” or “off” only at the transcriptional level, while the other steps of gene regulation determine the amount, stability and post-translational modifications of the protein product. In the next sections, I will focus on gene expression regulation at the transcriptional level.

1.1.2 - Differential gene expression gives rise to cell type diversity and anatomical structures during development

In order to understand how cell types arise through gene regulation, it is essential to study how development works. In fact, most animals start their life as one single undifferentiated cell. After several replication cycles, cells differentiate into more and more specific cell types. At the same time, the major body plan is defined. Cell modifications and the formation of anatomical structures are strictly associated with gene expression changes.

Drosophila melanogaster has proven to be a powerful model organism to study embryonic development. Besides being cheap to raise and easy to mutagenize[2], *Drosophila* has a compact genome[3] and can produce large amounts of embryos in a short time. In addition, it often harbors a single copy of developmental genes (mammals often have multiple genes with overlapping functions), making it easier to hack developmental processes[2].

As for the majority of animals, the development of *Drosophila* starts from a fertilized cell that undergoes fast replication cycles. At 2 hours post fertilization (hpf), the embryo is made of multipotent cells that contain molecular signals, in form of Transcription Factors (TFs), that will determine their fate (Figure 1a,d). The gradients of molecules in the egg at this stage define the future axes of the embryo - this is the first step to delineate the body plan and it will later lead to the formation of the three main segments: head, thorax and abdomen. Gastrulation forms the three germ layers – ectoderm, mesoderm and endoderm – and is followed by specification of cell types within each germ layer (Figure 1b,e). Finally, after the major body plans have

been established, cells differentiate to their final state, giving rise to the variety of cell types that are found in the larval and adult body (Figure 1c,f). All chapters of this thesis will discuss experiments linked to *Drosophila* embryonic development and, in particular, to the following time intervals: 2-4 hpf (multipotent cells and gastrulation), 6-8 hpf (cell specification) and 10-12 hpf (terminal differentiation).

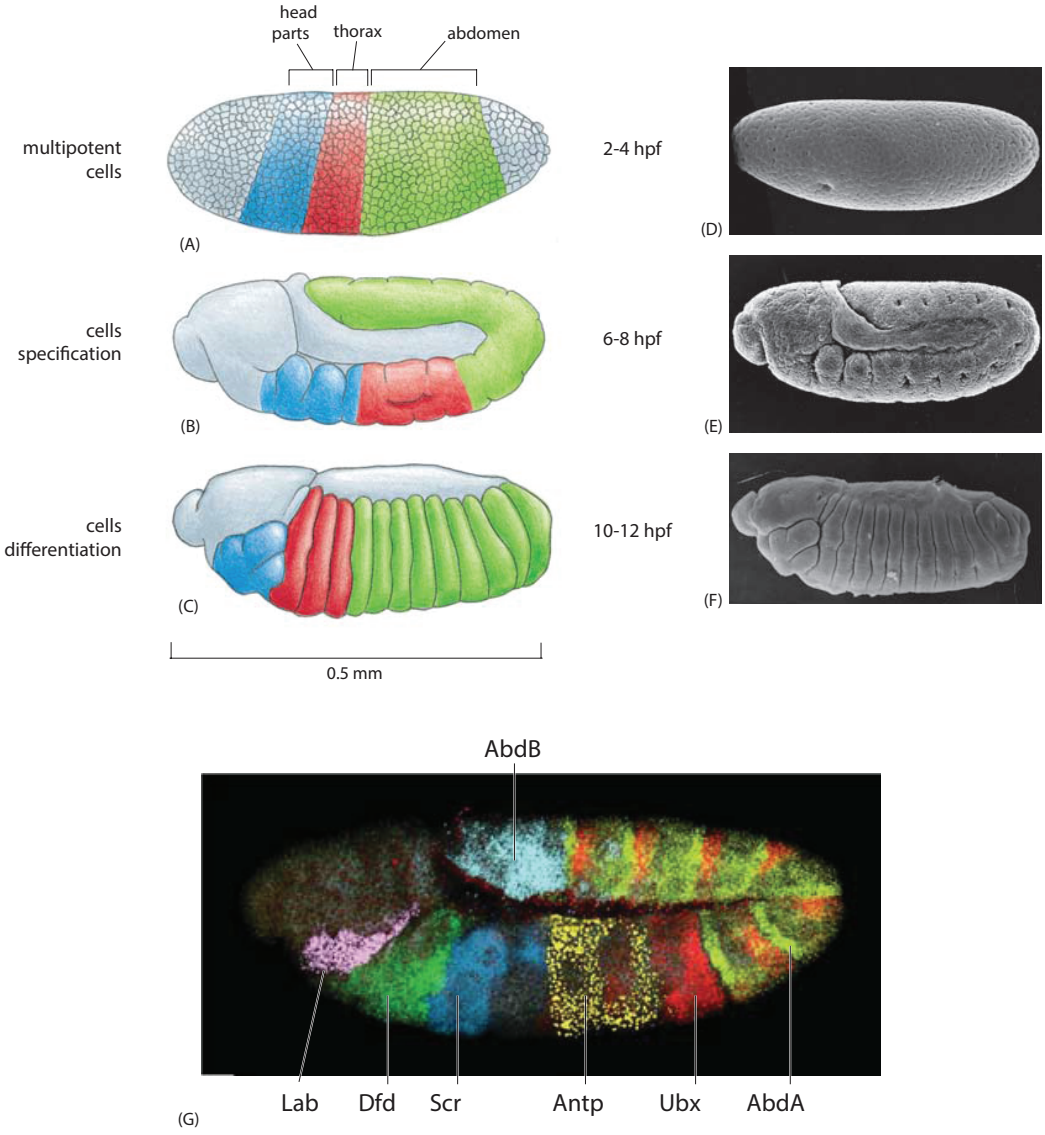


Figure 1 - *Drosophila* embryonic development is guided by precise spatio-temporal gene expression. (A and D) *Drosophila* embryo at 2 hpf. The cells are still in a multipotent state but the major body plans have been already established. (B and E) *Drosophila* embryo at 6 to 8 hpf. Gastrulation has occurred and cells are undergoing specification within germ layers. (C and F) *Drosophila* embryo at 10 to 12 hpf. Terminal differentiation of cell fates is underway. (G) Patterns of *Hox* gene expression in an embryo at 6 to 8 hpf. These precise expression patterns will specify segment identity along the anterior-posterior axis. Adapted from Alberts *et al.*[1], (D) and (E) from Turner *et al.*[4], (F) from Petschek *et al.*[5], (G) from Kosman *et al.*[6]

Specific gene expression is also crucial to define the anatomy of *Drosophila*. Figure 1g shows the overlap between gene expression and future anatomical structures in an embryo at 6-8 hpf. Colors correspond to the expression patterns of different homeotic genes. The precise spatio-temporal expression of these genes is crucial to define the body plan. For example, the expression of *Antennapedia* (*Antp*) and *Ultrabithorax* (*Ubx*) identifies the developing thorax, while *abdominal A* (*abd-A*) and *Abdominal B* (*abd-B*) define the developing abdomen.

1.2 - Gene expression is controlled by the interplay of many *cis* Regulatory Modules

Protein coding sequences only make up 15.9% of the *Drosophila* genome. This proportion goes down to 2% in most mammals, including humans. The remaining part of the genome - referred to as the non-coding genome - is disseminated of *cis* Regulatory Modules (CRMs): discrete genomic regions that regulate gene expression and exert their function by recruiting Transcription Factors to the DNA. There are four major classes of CRMs: promoters[7], enhancers[8], silencers[9] and insulators[10] each with different functions and sequence composition. In the following pages, I will briefly introduce them.

1.2.1 - Promoters

Promoters are short CRMs that regulate the initiation and intensity of gene expression and they integrate the cues from distal elements[7]. Coding and non-coding genes have at least one promoter sequence that ensures robust and preferential transcription in the direction of the gene. The first transcribed base on the DNA sequence is called Transcription Start Site (TSS) and it might vary between transcripts. The surrounding area (about ±50 base pairs) is called core promoter. The core promoter recruits RNA Polymerase II (Pol II) and the General Transcription Factors to assemble the Pre-Initiation Complex (PIC)[11]. Core promoters are not the

only sequences capable of assembling the PIC[12], but the main difference with non-promoter regions is that the latter generally show bi-directional transcription.

Depending on the core promoter type, transcription initiation can be focused or broad[13]. Narrow (or focused) promoters have a specific TSS and produce transcripts that start from the same position. On the other hand, broad promoters have multiple TSS over the core-promoter region: the PIC can be assembled on multiple positions and they generate transcripts with variable starting positions. The sequence and transcriptional properties of core promoters are related to the function of the gene they regulate. Core promoters can be divided into three categories found in all metazoans that depend on the gene that they regulate:

1. Adult tissue-specific genes. These core promoters tend to be associated with TATA-box and Initiator (Inr) motifs[14] and tend to be narrow promoters. They are usually active in terminally differentiated cells.
2. Housekeeping genes. Housekeeping genes are expressed in most cells and across developmental stages. In *Drosophila* they are enriched for Ohler 1, Ohler 6 and DNA replication-related element (DRE) motifs[15].
3. Developmental genes. These genes are mostly expressed during development and they are required to be efficiently activated and inactivated. To achieve fast regulation, they can be "poised" to be transcribed. They contain Inr and Downstream Promoter Element (DPE) motifs and have focused initiation[16].

Promoters can be characterized by measuring their capability to initiate transcription. Methods such as CAGE[17,18] and PRO-Seq[19] precisely identify the very first bases of mRNAs and have been used to map at single base pair resolution the TSS of many genes. In addition, these assays can characterize broad and narrow promoters. Another method to identify promoter activity genome-wide is SuRe[20], a technique used to assess self-transcription capability of DNA sequences.

1.2.2 - Enhancers

Enhancers have been functionally described as sequences capable of increasing transcriptional output regardless of their orientation and distance from the gene[21]. In

general, enhancers are clusters of Transcription Factor Binding Sites (TFBS) and function by attracting multiple TFs to the chromatin[22]. Transcription factors cooperatively bind to enhancers and displace the nearby nucleosomes. The sequence composition of enhancers is more heterogeneous than that of promoters, making it more challenging to categorize them. Another difference between promoters and enhancers is that while the first tend to be broadly active, the second act in a tissue and time-specific manner. The diversity of TFBS composition found in enhancers also explains how they can drive such a variety of transcriptional patterns.

Genes are generally regulated by multiple enhancers and the global expression pattern that we observe is the sum of the action of multiple specific enhancers. In addition, enhancers act redundantly to increase gene expression robustness. A well-characterized example of combinatorial enhancer activity is the regulation of the *Drosophila* gene *even skipped* (*eve*). At 2 hpf *eve* is expressed in seven stripes that, together with the specific expression of other genes, will guide the segmentation of the embryo (Figure 2). This complex expression pattern is driven by multiple enhancers, each responsible for the expression of one or two stripes. Other enhancers are responsible for *eve* expression later during embryogenesis in the brain and muscle precursors.

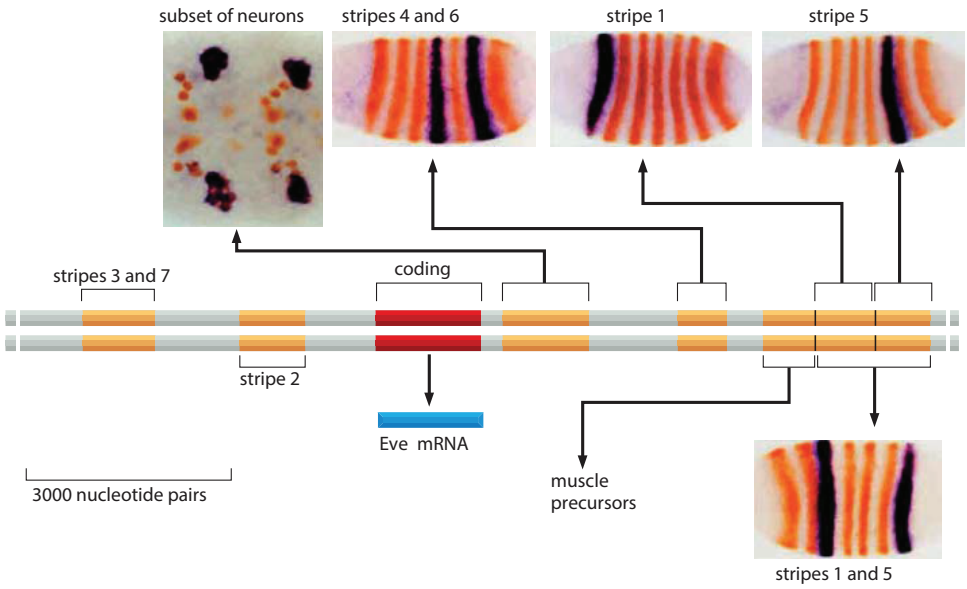


Figure 2 – *eve* expression pattern is guided by the interplay of many CRMs. The figure shows a schematic representation of the locus of the *eve* gene. The coding sequence is represented in red while the surrounding enhancers are shown in orange. Arrows point to the expression patterns driven by each enhancer alone. The *in situ* hybridization images display the total expression of *eve* (orange) and the specific expression driven by each enhancer (dark blue). From Alberts *et al.*[1], adapted from Fujioka *et al.*[23]

By displacing the nucleosomes, the binding of transcription factors on enhancers makes the chromatin more sensitive to cutting by endonucleases and insertion of transposases. This property is assessed respectively by DNase hypersensitivity[24] and ATAC-Seq[25]. In addition, ChIP-Seq[26,27] can be used to characterize transcription factor binding and then identify putative enhancers. Furthermore, STARR-Seq[28] is a method that can assay enhancer activity genome-wide. Finally, both promoter and enhancers are conserved during evolution. A phylogenetic strategy to identify them is to study sequence conservation across species[29]. The expression patterns driven by enhancers can be characterized by enhancer assays such as those shown in Figure 2.

1.2.3 - Silencers

Silencers operational definition is similar to that of enhancers. In fact, silencers can suppress transcription independently of their orientation and distance from the target gene[30]. They are clusters of Transcription Factor Binding Sites and recruit repressor proteins to the chromatin. They can act by contacting the promoter (with the same mechanism as that of enhancers) or can induce epigenetic modifications that suppress transcription. Silencers can be identified with DNase hypersensitivity, ATAC-Seq and ChIP-Seq and display similar properties to enhancers making it difficult to distinguish the two using only molecular assays.

1.2.4 - Insulators

Enhancers can influence gene expression even if they are located hundreds of kilobases from the target gene. Because of this property, enhancers could engage in many unspecific interactions with surrounding genes. Insulators are a category of CRM that can stop the interaction between enhancers and promoters if they are placed in between the two on the linear genome[31]. Insulators do not influence promoter or enhancer intrinsic activity; they act by blocking the physical interaction between the two. In *Drosophila*, there are at least five classes of known insulators[32] each bound by different combinations of insulator proteins, such as CTCF, GAGA binding factor (GAF) and suppressor of hairy wing (Su[Hw]).

Insulators organize the genome in so called Topologically Associating Domains (TADs). TADs are self-interacting regions that usually share the same epigenetic marks and are delineated by insulators at their boundaries. Enhancers are unlikely to interact with genes outside of the TAD they are located in. The disruption of TAD borders can cause non-physiological enhancer-promoter interactions and ectopic gene expression that can lead to pathologies such as cancer[33]. Insulator regions can be characterized by ChIP-Seq and motif analysis since most insulator proteins have clear binding preferences. In addition, TADs can be studied with chromatin conformation capture techniques[34] that reveal interaction frequencies between DNA fragments.

1.2.5 - The signal from multiple *cis* Regulatory Modules is integrated at the promoter level

As we have seen, gene expression is regulated by the interplay of many *cis* Regulatory Modules (Figure 3). CRMs recruit regulatory proteins to DNA and, by looping, they interact with the target gene promoter. This interaction occurs via the mediator complex. The activating and repressing signals from enhancers and silencers are then integrated at the promoter and determine the transcription rate. Each CRM is bound by a combination of TFs that regulate the effect on gene expression and the specificity of interaction with the target promoter. The *Drosophila* genome encodes for more than 1,000 transcription factors allowing for a variety of combinations at the CRM level.

TADs generally include tens of genes and hundreds of CRMs but not all enhancers regulate all genes[35]. In fact, classic models of transcription regulation postulate that each CRM regulates only one gene[1] and only a handful of CRM sharing examples are known[36]. It is still unclear how specificity in CRM-promoter interactions is achieved. Furthermore, developing high throughput assays to discover which CRM regulate which promoter has proven to be challenging. In chapter “II - Genetic variation as a tool to associate *cis* Regulatory Modules with their target genes” I will introduce a method to perform CRM to gene associations. This method identifies almost 3,000 CRM-promoter interactions and reveals a high rate of CRM sharing in *Drosophila*.

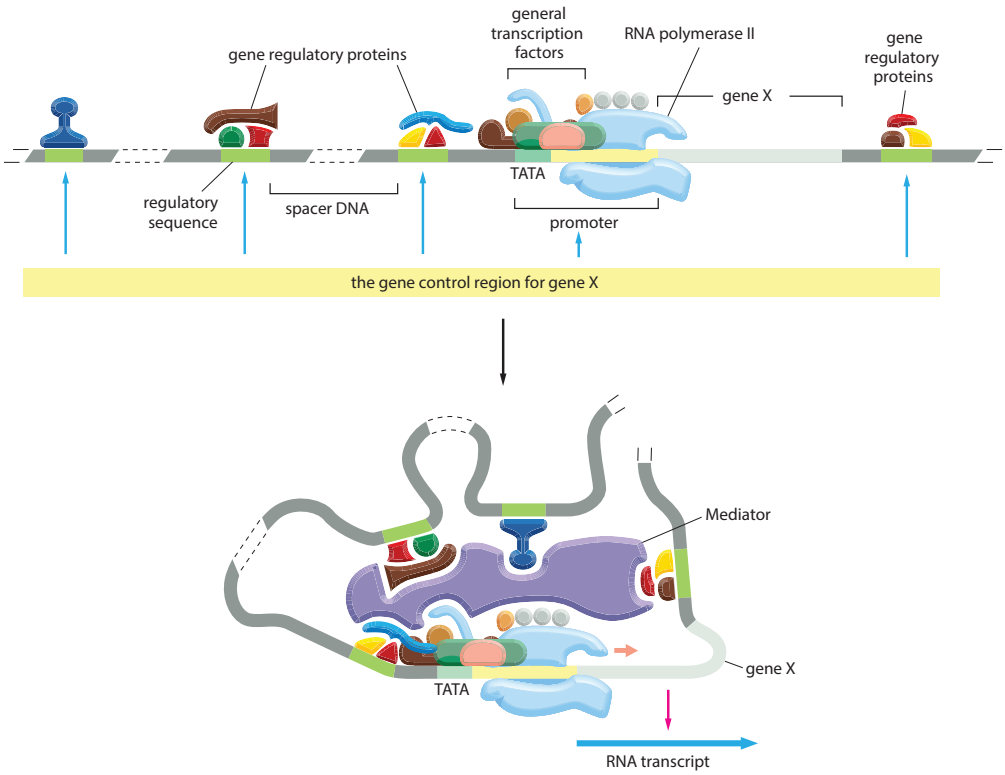


Figure 3 – Model of gene control region for eukaryotic genes. The figure gives a schematic representation of the elements that are involved in gene regulation. The gene X (light grey) is preceded by a promoter that recruits the general transcription factors and the PolII to assemble the PIC. Around the gene, many CRMs (light green) are bound by transcription factors. By looping, CRMs contact the promoter through the mediator complex and activate transcription. From Alberts *et al.*[1]

2 - Genetic variation causes diversity between individuals of the same species

2.1 - Population genetics studies differences between individuals

Despite following the same developmental program, animals of the same species are not identical. Population genetics is the branch of genetics concerned with understanding diversity between individuals. In this section, I will introduce some basic concepts about naturally occurring genetic variation and the consequences on phenotype.

2.1.1 - Differences between individuals of the same species are caused by genetic variation and interactions with the environment

Two main factors shape phenotypic variability: the environment and genetic variation. Diversity between genomes is generated by, for example, DNA polymerase errors during DNA replication and is inherited from parents to the offspring[37]. Genetic variants are locations in the DNA sequence that are polymorphic (different across individuals). Alternative versions of a variant are called alleles and together, they make the genetic pool of a population. The other source of phenotypic variation is the environment. In fact, each individual needs to face different environmental challenges that, especially during development, can modify the phenotype. Depending on what phenotype is considered, the environmental or the genetic component might have a larger influence than the other. A simple way to estimate how much genetics alone explains variability in the phenotype is to measure a phenotype's heritability[38]. For example, in human, height is highly controlled by genetics with an estimated heritability over 80%[39]. On the other hand, body weight is mostly dependent on the environment, especially at a young age[40].

A disadvantage of studying phenotypic variation in natural populations is that it is very challenging to disentangle the genetic from the environmental contributions. In the example seen before, the heritability of human weight is inflated by common eating behaviors within families. This component is independent from genetics but increases the heritability estimates. Model organisms can solve this issue and represent an outstanding resource for population genetics. In fact, it is possible to raise them in controlled conditions that minimize the environmental contribution to phenotypic variation.

2.1.2 - The majority of naturally occurring variants have no effect on fitness

Genetic variants occur at different frequencies across populations. After arising due to replication errors, variants can be inherited by the offspring or disappear from the population. In fact, the allele frequency of variants can change over time because of drift (random changes) or selection (due to impact on fitness). Common alleles are generally favored in the population and have a lower chance of disappearing by drift. In addition, genetic variation is the raw material for evolution. Natural selection favors the reproduction of the fittest individuals and causes an increase in the frequency of variants with positive effects and a decrease in frequency of variants with negative effects.

In 1968, Motoo Kimura estimated an exceedingly higher mutation rate in mammals than expected at the time[41]. His calculation from the mammalian hemoglobin sequences was that each mammalian zygote harbors four novel mutations. We now know that this was an underestimate and that non-coding regions have even higher mutation rates than coding sequences. However, this discovery was sufficient to propose that the majority of new mutations have no effect on fitness. This concept is now known as the neutral theory of molecular evolution[42] and is generally used as the null model when studying selection on genetic variants. Confirmation of the neutral theory of evolution has been accumulating during the years. The most relevant is the observation that biological systems show a high degree of robustness: the majority of variation at the molecular level is compensated at the phenotypic level[43] because of mechanisms such as redundancy[44], non-linear responses and pleiotropy[45].

In chapter “IV - Impact of natural sequence variation on *Drosophila melanogaster* chromatin accessibility”, I will discuss prioritization of genetic variants. It is still challenging to predict the effect of genetic variants on phenotype, despite the knowledge accumulated in the last years. To this end, I applied a machine learning technique to predict the impact of variants on chromatin accessibility.

2.1.3 - Variants can be under positive or negative selection

As discussed in the previous paragraph, the vast majority of genetic variation has little or no effect on fitness and it is therefore ignored by natural selection. On the other hand, variants with an effect on phenotype might confer a fitness advantage or a disadvantage. Depending on the effect, the variant will increase in frequency in the population over time (positive selection) or decrease (negative selection). In addition, variant frequencies can change because of drift. This phenomenon hits neutrally evolving variants and has a greater effect on small populations.

Individuals from the same species tend to be separated in populations that are isolated by geographic barriers such as seas or mountains[46]. By occupying new territories, groups of individuals increase their physical distance, which is the major factor causing geographic isolation. Populations behave as independent groups of individuals, with separated genetic pools. In fact, in isolated populations variant frequencies can drift independently and different environments might pose different challenges. If the isolation lasts for long periods, individuals from the same species eventually accumulate genetic incompatibilities that make interbred offspring less and less fit[47].

In chapter “III - Gene expression variation among *Drosophila melanogaster* lines from five continents” I will introduce a novel gene expression dataset that sheds light on gene expression regulation differences between five independent populations of *Drosophila*. Differential expression among continents suggests some extent od adaptation to different environments.

2.2 - Genome Wide Association Studies

Genome Wide Association Studies (GWAS) are a statistical method to identify genomic *loci* associated with a phenotype[48]. If a genetic marker (in present days genetic variants are used as markers) segregates with a phenotype, then the *locus* where the variant is located is in linkage with the phenotype. This is a correlative approach that requires thousands of cases and controls to achieve enough power. GWAS have identified many genes associated with multigenic phenotypes, such as autism[49] and diabetes[50]. The majority of causal variants identified in GWAS are not located in the proximity of genes, making the interpretation of results challenging.

2.2.1 - Quantitative Trait *Loci*

Quantitative Trait Loci (QTLs) are genomic regions that are associated with a quantitative phenotype. Before the genome-sequencing era, QTLs have been identified in a variety of ways that did not require a fine mapping of markers. In present days, the approach to map QTLs is similar to GWAS and it is based on the correlation between the presence of a variant and the phenotype. This process is made easier by inbreeding since inbred fly lines exhibit homozygosity for the majority of *loci*.

QTLs are commonly used to dissect gene regulation and the phenotypes are molecular ones (e.g. gene expression). Expression-QTLs (eQTLs) are mapped by correlating the expression of a given gene with the allelic status of the variants surrounding the gene. A schematic overview of the eQTL statistical process is shown in Figure 4. Other common molecular quantitative phenotypes are chromatin accessibility (caQTLs[51]) and histone binding (hQTLs[52]). By identifying genomic regions linked to molecular phenotypes, QTLs offer a way to understand gene regulation. In addition, they can be used to interpret GWAS results and shed light on how non-coding regions influence complex phenotypes.

eQTL can be functionally mapped when a variant impacts gene regulation at some level. For example, Cannavò *et al.*[53] describe an eQTL whose minor allele disrupts the binding site of Sloppy Paired 1 (Slp1) at the promoter of *CG10396*. Slp1 predominantly functions as a transcriptional repressor, causing *CG10396* expression

to be higher in the minor allele. As in this case, naturally occurring variants can hack gene regulation and reveal the function of CRMs. In chapter "II - Genetic variation as a tool to associate *cis* Regulatory Modules with their target genes" I will describe an eQTL method that makes use of natural variation to understand the function of CRMs.

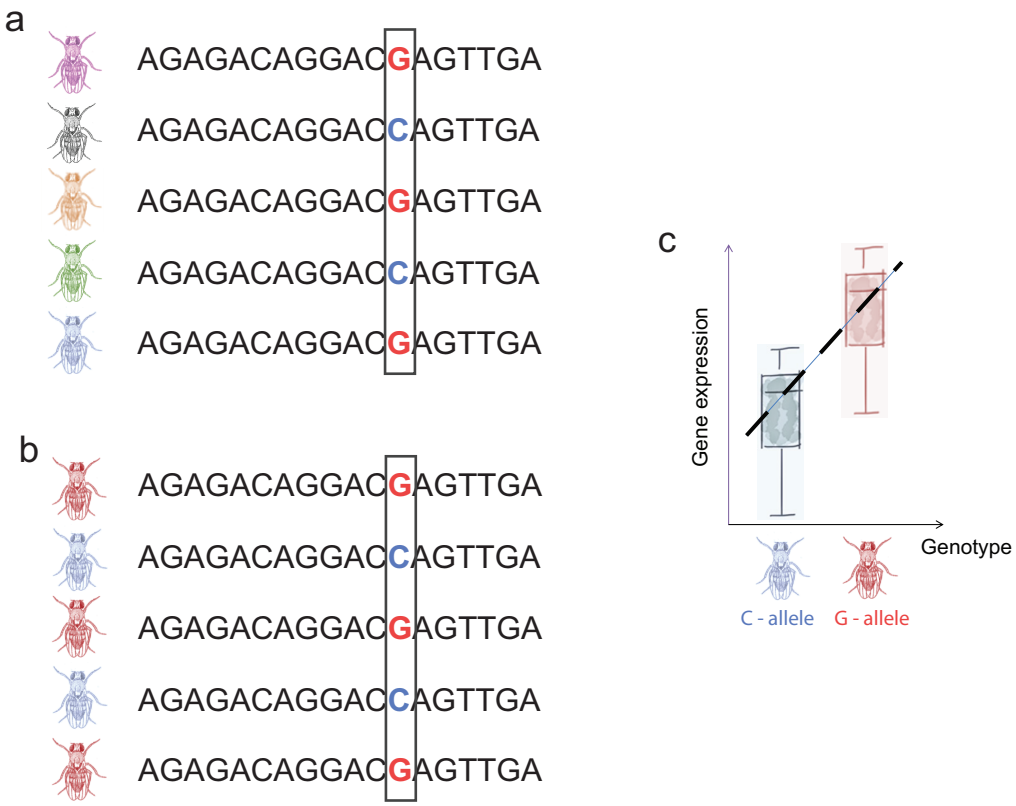


Figure 4 – Schematic representation of Quantitative Trait Loci analysis. The figure gives a simplified overview of the test for association between a genetic variant and the expression of a putative target gene. (a) Different inbred lines are genotyped. The black box indicates a variant with a G (Major) and C (minor) alleles. Here we assume homozygosity for the variant across all fly lines. (b) The individuals are separated into two groups depending on what allele they have. (c) A correlation test is performed between the gene expression values of the two groups. If the correlation is significant, the locus harboring the variant is considered an expression Quantitative Trait Locus of the tested gene. In the case of heterozygosity (C / G alleles), there would be a third group in between the major and minor alleles.

2.2.2 - Inbreeding increases power in QTL studies

Inbred lines can be generated by selecting a few individuals and by crossing them for multiple generations. As described before, genetic drift has a larger impact on small populations. After many generations (from 10 to 50) of replication between siblings,

the genetic pool decreases dramatically, leaving the majority of variant *loci* with only one allele. This, in turn, reduces heterozygosity and makes individuals almost genetically identical to each other. Inbred fly lines are generated by starting from a common natural population and by repeating the inbreeding procedure in parallel to generate multiple lines. If we consider a genetic variant with two alleles in the original outbred population, genetic drift will cause some inbred lines to only have the first allele and some others to only have the second. An example of inbred fly lines is the *Drosophila* Reference Genetic Panel[54] (DGRP). The DGRP is a collection of more than 200 inbred fly lines that have been generated and genotyped for population genetics studies.

Inbreeding gives a statistical advantage when mapping QTLs. Heterozygosity often masks extreme phenotypes that are present only in homozygosis. By studying lines homozygous for one or the other allele it is possible to assess the two extremes of the spectrum. Figure 5 shows how power for the discovery of QTL changes when using large and/or inbred populations. Inbreeding allows for powerful discovery of rare variants even by studying relatively small populations.

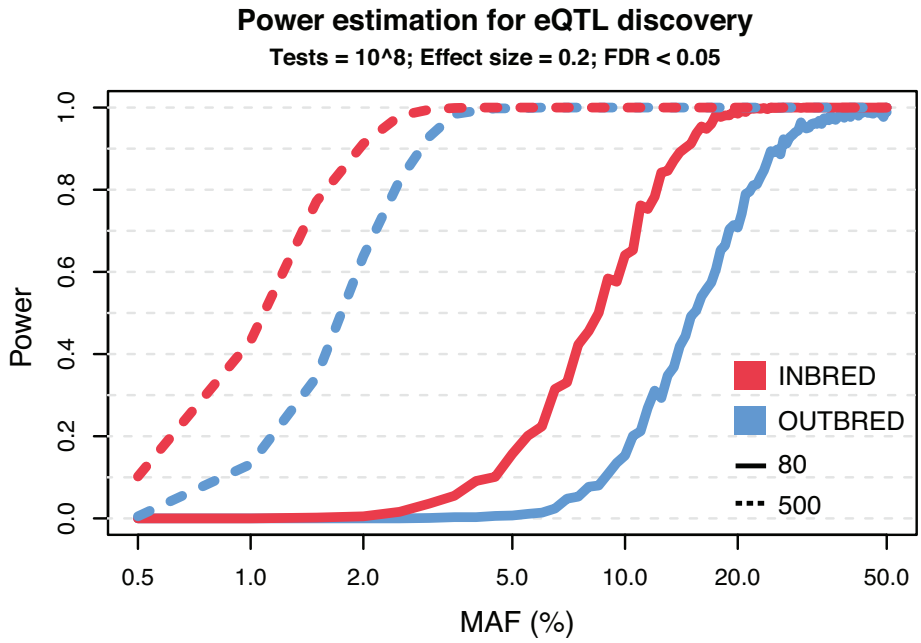


Figure 5 – Estimated power to discover eQTL. The figure shows the estimated power to discover eQTLs for outbred (blue lines) and inbred (red lines) populations, and for large (dotted lines) and small (solid lines) sample sizes. The fixed values in this plot are the number of tests performed (10^8, which is a good estimate for *cis*-eQTL in *Drosophila*), the eQTL effect size (0.2) and the False Discovery Rate (0.05). The power increases as a function of the minor allele frequency (MAF).

2.2.3 - *Drosophila* is a powerful model organism to study population genetics

Drosophila melanogaster has long been used as a model organism to study population genetics. The advantages of *Drosophila* include a compact genome, fast replication time and low cost to maintain. Regarding population genetics, isogenic inbred fly lines are easy to obtain and entire populations can be maintained at relatively low costs and with reasonable human labor[54]. Here is a brief list of the main advantages of using inbred *Drosophila* lines for population genetics studies:

- Isogenic fly lines have homozygous states in the vast majority of variants. This results in larger power for association studies, as discussed in the paragraph above.
- Isogenic fly lines yield the possibility to consider individuals belonging to the same line as clones. This allows to collect multiple individuals and to perform experiments at different times. By pooling individuals from isogenic lines, it is possible to increase the overall sample material. This is especially important when studying development, given that embryos generally offer little amounts of sample.
- The short generation time of *Drosophila* and the rarity of chromatin refractory to recombination reduce the size of linkage disequilibrium blocks. This allows for precise identification of causal variants and increases the information content of variants in the same locus.

3 - Aim of the study

In this thesis, I will describe three projects that I developed during the course of my Ph.D. The projects are all linked to genetic variation and its effects during *Drosophila melanogaster* embryonic development. Each project will be presented in a separate chapter:

- Chapter II - Genetic variation as a tool to associate cis Regulatory Modules *with* their target genes". In this chapter, I will describe a novel application of the eQTL framework to map CRM to gene associations. The method identifies almost 3,000 CRM to gene associations and indicates widespread CRM sharing.

- Chapter III - Gene expression variation among *Drosophila melanogaster* lines from five continents". In this chapter, I will introduce a novel RNA-Seq dataset. We quantified gene and transcript expression of 80 *Drosophila* lines from five continents and identified differentially expressed genes and transcripts. In addition, we mapped gene-eQTLs and exon-eQTLs

- Chapter IV - Impact of natural sequence variation on *Drosophila melanogaster* chromatin accessibility". Here, I applied a machine learning approach to prioritize natural variants by their predicted effect on tissue-specific chromatin accessibility. The method gives insight into CRM sequence composition and predicts tissue-specific variant effects.

II - Genetic variation as a tool to associate *cis* Regulatory Modules with their target genes

1 - Introduction

1.1 - Multiple techniques can be used to identify *cis* Regulatory Modules

Gene expression is a tightly regulated process both in space and time, that begins with the regulation of gene transcription[1]. In order to be transcribed, genes need the activation of proximal and distal regulatory sequences known respectively as promoters and enhancers. In a simplified model: the enhancer is bound by transcription factors, contacts the promoter by chromatin looping, and signals to the Pre-Initiation Complex, bound at the promoter, to initiate transcription.

Gene regulatory sequences, such as promoters and enhancers, are combinatorically bound by transcriptions factors and this property has been used to identify and analyze them. In the past decades, a plethora of techniques has emerged to characterize *cis* Regulatory Modules (CRMs) on a genome-wide scale. For example, CAGE[17] and PRO-seq[19] can shed light on the initiation of transcription, DNase hypersensitivity[24] and ATAC-Seq[25] identify regions of open chromatin and different versions of Chromatin Immunoprecipitation[55] have been used to characterize chromatin states[56] and transcription factors binding sites[27]. The *Drosophila* community has made major efforts in characterizing CRMs across tissues, developmental time points and sexes. These efforts provide a wealth of information about the CRMs used during embryogenesis and other stages of *Drosophila melanogaster* development.

1.2 - Functional and correlative methods to link CRMs and target genes

Despite the amount of information that has been collected on CRMs, knowing which CRMs regulates which genes is far from trivial, but yet crucial to understanding gene regulation and its complexity. Databases such as REDfly[57] collect evidence for enhancer to gene associations, but these annotations come from heterogeneous sources, experimental procedures, biological contexts and genetic backgrounds. These associations mostly rely on the measurement of putative target genes after deleting (or interfering with the function of) the CRM. Genomic alterations are complex to achieve in multicellular organisms. Nevertheless, in the last years, CRISPR-Cas9 technology[58] was introduced as a precise genome editing tool and revolutionized the field of genome manipulation. After its introduction, genome manipulation has become quicker, easier and more affordable, but it is still far away from being applicable on a genome-wide scale for multicellular organisms. For these reasons, the information about the regulation of target genes by CRMs is sparse, derived from heterogeneous sources and is biased towards a few genes of large interest for the community.

Genetic approaches are still not applicable at a large scale. In fact, many studies still rely on assigning CRMs to their closest gene, despite the existence of more sophisticated methods and the fact that several lines of evidence indicate that gene-CRM proximity is not predictive. Scientists have developed correlative methods to link CRMs and their target genes genome-wide. Many of these methods look for an overlap between the tissues where the CRM is active and the neighboring genes are expressed (or the neighboring promoters are active). This approach has been applied to a wide range of data, from enhancer assays (e.g. Kvon *et al.*[59]) to single-cell ATAC-Seq (Cicero[60]). The major shortcomings of correlative approaches are that they are very specific to the cellular context and underestimate CRM to gene distance (always preferring the closest gene that fits the requirements).

1.3 - Quantitative Trait Loci as a functional method to associate CRMs to target genes on a genome-wide scale

Quantitative Trait Loci analysis[61] represents a functional strategy to link genomic regions to target genes. In particular, eQTLs[62] can link genomic variants to variation

in gene expression. eQTL analysis is applied genome-wide and provides functional links between genomic *loci* and target genes. Advantages of the eQTL method are that eQTLs are unbiased toward genomic sequence function and "CRM-to-gene-distance" (though often only elements within a fixed range are tested). Furthermore, they can uncover novel functional elements. On the other hand, eQTL results are often difficult to interpret because the majority of eQTLs fall outside regions with a known function. In addition, eQTLs require an extensive number of tests[53], so that only the strongest associations are discovered.

1.4 - Overview of the project

In this project I build on existing eQTL methods to functionally associate CRMs and target genes, using DNase hypersensitivity regions as a proxy for CRMs. eQTL methods have largely been used in an unbiased way to ask generic questions. Here, I bias the test towards DHS and I compromise on search space and single variant resolution to increase power and reduce the number of tests. The goal is to maximize the number of DHS to gene associations. In this project I make use of the following datasets:

- An extensive map of DNase hypersensitivity during *Drosophila* embryo development generated by James Reddington and David Garfield in the Furlong laboratory (*unpublished*). The data provide time and tissue resolution.
- Full genotype information for 80 inbred *Drosophila* lines belonging to the *Drosophila* Genetic Reference Panel[63] (DGRP).
- Gene expression information for the corresponding 80 DGRP lines for 3 time points during embryo development (2-4 hpf, 6-8 hpf, 10-12 hpf) previously generated in the Furlong laboratory by Cannavò *et al.*[53]

In the following pages, I will compare different methods, show *in silico* and *in vitro* validations of the DHS-eQTLs, and I will describe the results and their implications.

2 - Results

2.1 - An eQTL method to associate *cis* Regulatory Modules to target genes

In this work, I aim to functionally associate DHS to their target gene. To achieve this goal, I developed a new eQTL approach called DHS-eQTL. In this section, I will discuss the rationale behind building on existing eQTL methods with the specific purpose of performing DHS to gene associations. The method described here leverages the information of multiple variants overlapping the same DHS and reduces the number of tests to maximize the number of DHS to gene associations. I will examine the sources of gene expression variation in our dataset and describe how to control for non-*cis* components. I will finally compare three statistical methods to identify DHS to gene associations, their advantages and disadvantages.

2.1.1 - DNase hypersensitivity as a proxy for *cis* Regulatory Modules

In this project, I used a comprehensive DHS atlas with tissue and time resolution to identify CRMs during *Drosophila* embryogenesis. The dataset included 19 samples, all in duplicates: it spanned 5 tiling time points during *Drosophila* embryo development (2-4 hpf, 4-6 hpf, 6-8 hpf, 8-10 hpf, 10-12 hpf) and 3 FACS sorted tissues: neuroectoderm, mesoderm and non-neuroectoderm/non-mesoderm tissues (Figure 6). The full dataset was analyzed *de novo* and mapped to the latest *Drosophila melanogaster* genome assembly (BDGP6). A total of 63,157 DHS was identified.

DHS were separated into two groups depending on their vicinity to a known Transcription Start Site (TSS):

- Promoter-proximal DHS: are located within 500 base pairs of a known TSS. Since *Drosophila* TSS are highly enriched at Topologically Associating Domains (TAD) borders[64,65], the promoter-proximal DHS represent a heterogeneous group. Promoter-proximal DHS include core promoters,

promoter-proximal insulators and enhancers. There are 23,268 promoter-proximal DHS.

- Enhancers: are more distal than 500 bases from any annotated TSS. TSS distal DHS are largely enhancers. There are 39,889 enhancers. Only a small proportion of enhancers is bound by insulator proteins.

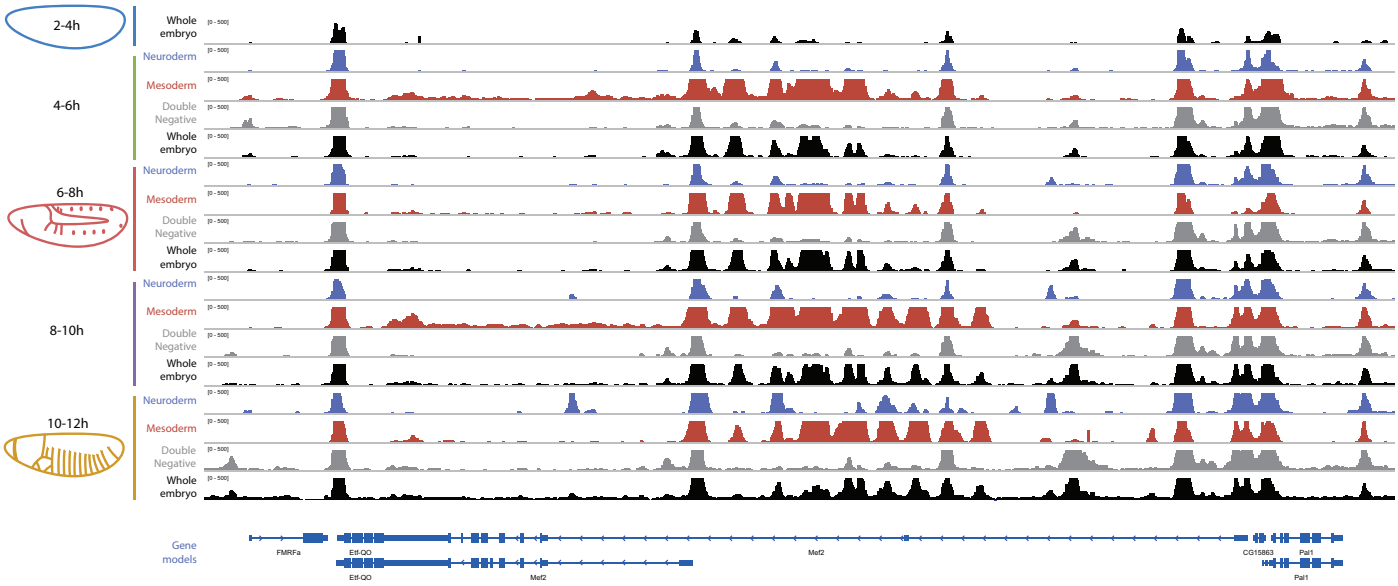


Figure 6 - DNase hypersensitivity assay. DNase profiles around the mesodermal gene *Mef2*. The tracks show DNA accessibility in different tissues and time points during *Drosophila* embryonic development. Blue: neuroectoderm, Red: mesoderm, Grey: double negative, black: whole embryo.

2.1.2 - High variant density and small linkage disequilibrium blocks in *Drosophila melanogaster* support the use of an eQTL multivariate model

The *Drosophila* Genetic Reference Panel (DGRP) lines are a panel of *Drosophila melanogaster* inbred lines that come from a uniform geographic location. They harbor more than 6,1 million variants that correspond to a density of one variant every 29 genomic bases. Furthermore, variants are slightly enriched in open chromatin by a factor of 1.18. Figure 7a shows the density of variants overlapping the DHS. Although 16.1% of DHS do not overlap any variant (and are therefore ignored in our following analyses), the majority of DHS (>80%) contained one or more variants. Genetic variants located closely in the genome are partially redundant due to linkage disequilibrium (LD). While LD blocks generally span tens of kilobases in mammals, in the *Drosophila* genome they are on average smaller than 100 bp[63]. To

test for independence of variants overlapping the same DHS, I performed a Principal Component Analysis (PCA) and calculated how many Principal Components (PCs) were necessary to explain more than 99% of the variance for each DHS. In the extreme case of perfect LD between all variants overlapping the same DHS, one PC would explain all of the variance. Figure 7b shows that on average 3 or more PCs were necessary to explain the variance on each DHS, indicating high independence between variants. The high independence between variants is caused by the short reproductive cycles (15 days) of *Drosophila melanogaster* and the low proportion of compacted heterochromatin that is refractory to recombination. Independence of variants not only allows for single base resolution eQTL identification[53] but it also represents a wealth of information that can be leveraged on. In this study, we included the variants overlapping the same DHS in one association test to reduce the total number of tests and to increase power (Figure 8). The rationale is that variants that impair or enhance the function of the same Regulatory Element will affect the same phenotype.

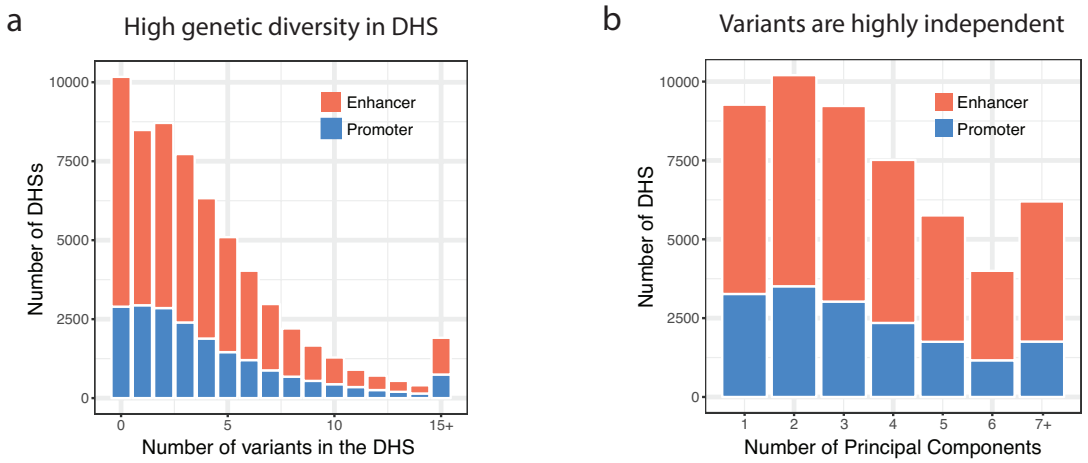


Figure 7 - Genetic variants are dense over DHS and are highly independent. (a) Distribution of the number of variants overlapping each DHS. About 10,000 DHS do not overlap any variant and cannot be analyzed with any eQTL method. The majority of DHS overlaps multiple variants. (b) Number of Principal Components (PCs) necessary to explain >99% of the variants overlapping each DHS. If variants are in perfect linkage disequilibrium, one PC will explain them all. In the majority of cases >1 PC is necessary to describe the variation overlapping each DHS showing high independence between variants.

In order to further reduce the number of statistical tests, increase power in DHS to gene associations and uncover complex DHS to gene associations, I took the following steps (Figure 8):

- Variants that did not overlap any DHS were excluded from testing
- If multiple variants overlapped the same DHS they were included in one multivariate test
- All DHS-eQTL statistical tests were corrected for multiple testing in a joint FDR approach. This allowed the DHS-eQTL method to discover multiple DHS associated with the same gene and vice-versa.

In our case, these adjustments led to a decrease from 17,234,822 to 483,064 tests.

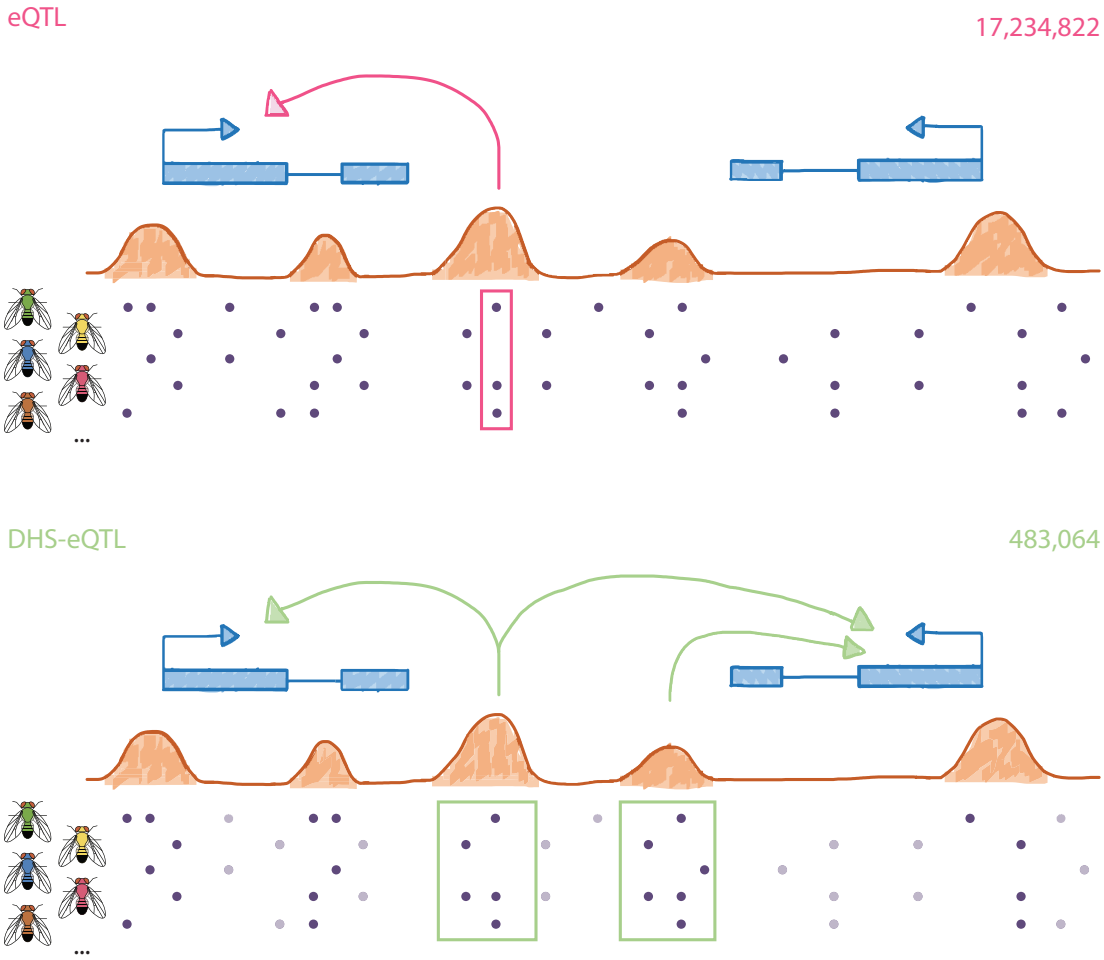


Figure 8 - Schematic representation and comparison of classic eQTL and DHS-eQTL methods. The top panel shows a schematic representation of the classic eQTL approach: we perform one test for each variant-to-gene association; all variants are tested for association independently of their genomic location; usually one association is identified for each gene (arrows). The bottom panel shows a schematic of the DHS-eQTL approach: I ignored variants that do not overlap any DHS; I performed one test for each DHS to gene association; if multiple variants overlap the same DHS they are pooled in a multivariate test; I could find multiple eQTLs for the same gene and for the same DHS. On the right, the total number of tests performed by the two methods. Blue: gene models; brown: DNase Hypersensitivity; purple dots: genetic variants; magenta and green boxes and arrows: eQTLs.

2.1.3 - Variance decomposition analysis shows that population structure is a major driver of gene expression variation

Gene expression is a tightly regulated process but it is not immune to variation. Environment, genetic variation and noise are among the causes of gene expression variation[61]. To further investigate what are the main sources of variation in our gene expression dataset I performed a variance decomposition analysis[66]. This method quantifies the amount of variance of a dependent matrix that can be explained by many independent matrices. For each gene, expression variation was split into four components:

1. Environment: represents expression changes during developmental times. This value captures differences among time points in the gene expression matrix and it includes batch effects.
2. Population structure: identifies genetic similarities between individuals. The DGRP fly lines come from a uniform geographic location and share the same population history. For this reason, population structure and *trans* variation (whole genome variation) could not be distinguished. This feature encompassed both population structure and *trans* variation.
3. *cis* variation: corresponds to genetic variation around the gene of interest (in this case ±50 kb from the gene).
4. Noise: is the remaining component and it corresponds to unexplained variation.

For each gene, the four components sum up to one.

eQTL studies are focused on dissecting the role of genetic variants in changing gene expression. Removing confounding factors from the association tests is crucial to avoid false positive associations. The variance decomposition analysis is useful to estimate the amount of gene expression variation explained by components other than *cis* genetic variation. In this project, I focused on eQTLs in the vicinity of the target gene, making developmental stage, batch effects, population structure and *trans* variation potential sources of false positive associations.

Figure 9 shows the amount of gene expression explained by the four components mentioned above. Environment (i.e. in our case developmental time) was the main driver of variation with a mean explanatory power of 36%. The *Drosophila* embryos are undergoing great anatomical modifications during the stages in this study making developmental time the largest predictor of gene expression changes. To remove

environment, together with batch effects, as a confounding factor in the DHS-eQTL test, I used PEER (see Methods). To confirm that PEER is actually removing the environment component I performed a variance decomposition of PEER residuals: the median amount of variance explained by the environment component was 0, proving that PEER effectively removes environmental sources of variation. Population structure and *trans* variation were also major drivers of gene expression variation with an average explanatory power of 33%. This corresponds to line-to-line similarity and it is removed by LIMIX and mtSet as I will discuss in the next sections. *cis* variation explained on average 13% of gene expression variation, with a very wide distribution: for 326 genes, this component explained more than 50% of gene expression variation. Finally, noise is completely orthogonal to the other three components and it does not represent a source of false positive eQTLs. It had an average explanatory power of 18%. These results show how *cis* variation could be confounded by other sources of gene expression variation (in particular population structure, environment and batch effects), calling for methods that can control them.

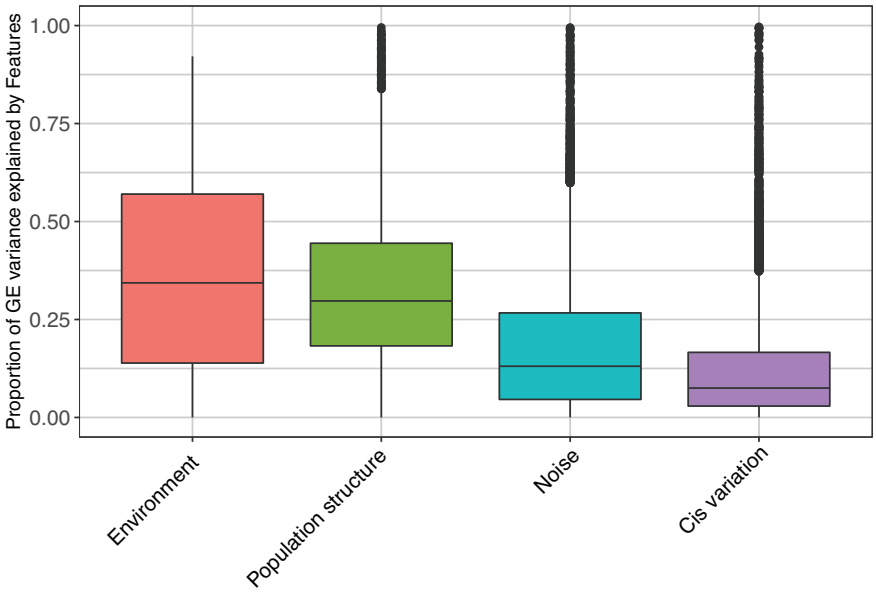


Figure 9 – Variance decomposition of gene expression. The plot shows the amount of gene expression variation explained by three features: environment, population structure and cis variation. For each gene, gene expression is decomposed by these three components and residual noise (up to 1). Each dot corresponds to an expressed gene.

2.1.4 - Testing three eQTL methods within the DHS-eQTL pipeline

We assayed three QTL methods to identify DHS to gene associations: Linear Mixed Model (LIMIX[67]), Principal Component multiple Regression (PC-regression) and multiple Set test (mtSet[68]). The three QTL methods are different in many aspects. In particular, LIMIX performs a univariate test: it fits one test for every variant-to-gene association. On the other hand, PC-regression and mtSet are multivariate approaches that can integrate the information from multiple variants. The two multivariate approaches are different in some regards. In particular, mtSet can model multiple phenotypes (in our case we have gene expression values for three time points during development) as the sum of the variants in the genetic region, population structure and residual noise. mtSet performs a single test for each DHS to gene association and it can leverage the three gene expression measurements for each gene. Finally, PC-regression can estimate the effect size of genetic variation on gene expression but it cannot model multiple phenotypes requiring 3 tests for each DHS to gene association. In order to compare the performance of LIMIX, mtSet and PC-regression, I performed an eQTL call using the same inputs for the three methods. Figure 10 outlines the pipeline used to call DHS-eQTL (see also *Methods*). Briefly, I quantified gene expression from 3'-Tagged Sequencing reads published by Cannavò *et al.*[53]. This dataset reports gene expression for 80 inbred DGRP lines and spans three time points during development. I corrected for hidden batch effects with PEER[69] and for mapping biases using WASP[70]. Finally, association tests were performed between corrected gene expression and sets of variants overlapping each DHS. I performed (i) one test for each variant-to-gene association using LIMIX, (ii) one test for each DHS-to-gene association using mtSet and (iii) three tests for each DHS-to-gene association using PC-regression. The three association methods were well calibrated as shown by their qqplot (Supplementary Figure 2).

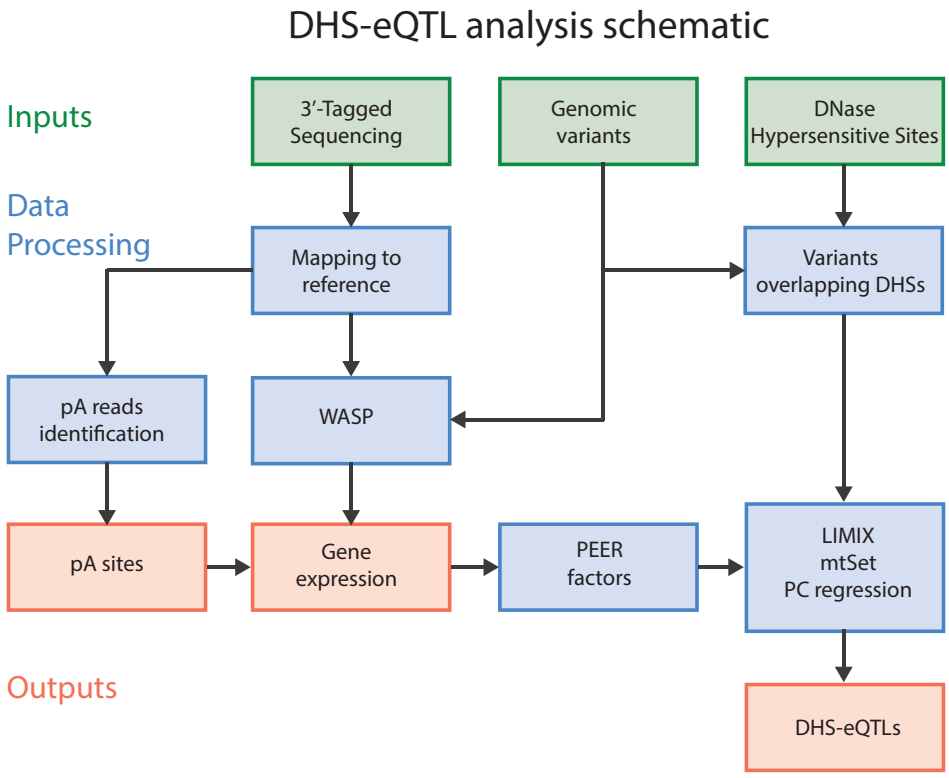


Figure 10 - Schematic of the pipeline used to call DHS-eQTLs. The three eQTL methods (LIMIX, PC-regression and LIMIX) were applied to the same inputs. 3'-Tagged sequencing reads were processed to identify polyadenylation (pA) sites. pA sites were quantified using mappability filtered reads (WASP) and the sum of pA usage was used as a measure of gene expression. Hidden factors and batch effect were removed with PEER. We include in the test only variants that overlap DHS. The three eQTL methods use the same phenotypes. Regarding the genotypes, LIMIX uses single variants while PC-regression and mtSet all the variants that lay on the same DHS.

2.1.4 - mtSet is the only multivariate eQTL method that controls for population structure

After correcting for residual linkage disequilibrium, PC-regression identified 5,478 DHS-eQTLs while mtSet finds 2,967. Both multivariate approaches discovered more DHS-eQTLs than the univariate approach LIMIX: 1,449 (Figure 11a). The larger power of multivariate models and the reduced number of tests explains this difference.

PC-regression might identify more DHS-eQTLs than mtSet because it does not correct for variation in gene expression caused by population structure. On average, population structure explained 32% of gene expression variation (Figure 9). To assess if population structure could drive false DHS-eQTLs, we tested if DHS-eQTLs identified only by one statistical method have a larger proportion of variance

explained by population structure. Figure 11b shows that DHS-eQTLs discovered by both mtSet and PC-regression have a lower proportion of variance explained by population structure than DHS-eQTLs found exclusively by mtSet ($p < 5.57e-06$) and PC-regression ($p < 0.036$). This indicates that DHS-eQTLs discovered by PC-regression and not by mtSet have higher chances to be driven by population structure. On the other hand, even if DHS-eQTLs found only by mtSet have a higher proportion of variance explained by population structure, this component is removed from the association test. These observations justify the use of a multivariate approach that corrects for population structure. We then chose to use mtSet results as our set of DHS-eQTLs.

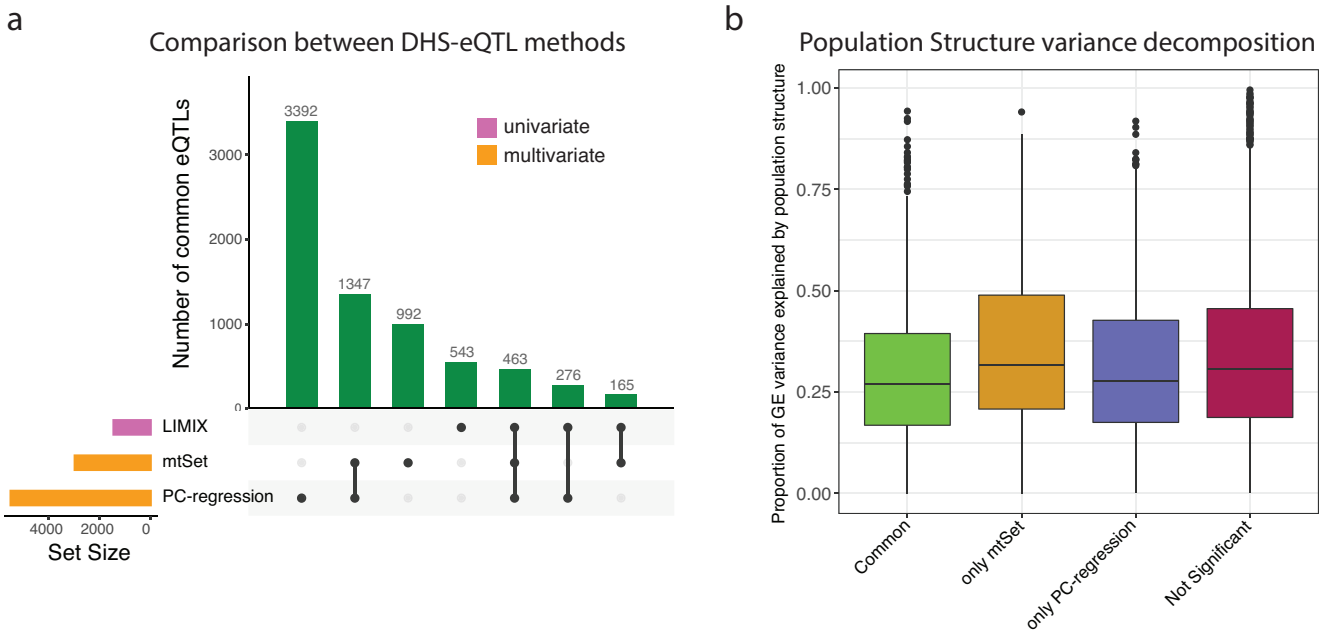


Figure 11 - Comparison of results between eQTL methods. (a) The UpSet plot shows the intersection between the DHS-eQTLs identified by the 3 methods. PC-regression discovers more DHS-eQTL than any other methods, most of which are unique. LIMIX, the univariate method tested, is the least powerful. (b) Proportion of variance explained by population structure for genes involved in DHS-eQTLs discovered by both mtSet and PC-regression (Common), only mtSet, only PC-regression and genes not involved in any DHS-eQTL (Not Significant).

2.1.5 - DHS-eQTLs discovered by the univariate model are a subset of those discovered by the multivariate model

Figure 11 shows that LIMIX identified 543 unique eQTLs. This is unexpected given the higher power and similarity to mtSet: both implement a population structure correction, but mtSet leverages on multiple variants and requires fewer tests. To investigate this, I plotted the uncorrected p-values for the significant associations from mtSet (Figure 12a) and LIMIX (Figure 12b). LIMIX did not reproduce well the results from mtSet. The distribution of LIMIX p-values from DHS-eQTLs discovered by mtSet (Figure 12c) showed that sub-threshold p-values from LIMIX are semi-randomly distributed. In contrast, mtSet reproduced to a higher extent the significant results from LIMIX (Pearson correlation: 0.662). The DHS-eQTLs unique to LIMIX had p-values from mtSet close to the FDR cutoff (Figure 12d).

Changes in gene expression can be driven by a single variant or by multiple variants acting in a cooperative or antagonistic manner. mtSet, by testing the effect of multiple variants at the same time, can capture complex scenarios while LIMIX tests are confined to the effect of single variants. We can then expect simple scenarios to be captured by both methods (LIMIX is more powerful here with some tests just below the FDR cutoff for mtSet) but the complex scenarios are captured by mtSet alone (shown by low correlation between the two methods when looking at mtSet significant DHS-eQTLs). It is worth noting that this reasoning is confined to causal variants only. In fact, there were on average more variants on DHS-eQTLs found only by LIMIX than on those found only by mtSet (Supplementary Figure 3). This counterintuitive observation may be explained by the fact that a high number of neighboring variants might dilute the effect of a single causal variant and decrease the power of mtSet. In conclusion, these observations confirm that LIMIX results are always captured by mtSet (even if under the FDR cutoff), while the reverse is not always true.

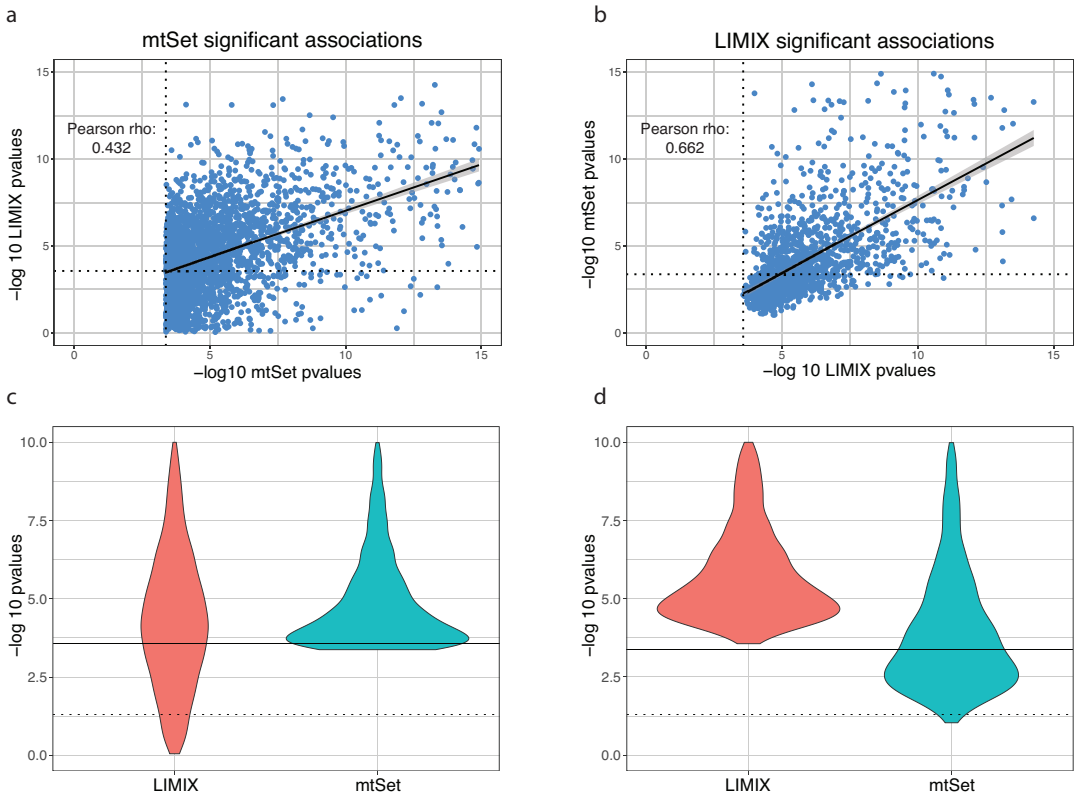


Figure 12 – LIMIX unique eQTLs are subthreshold eQTLs in mtSet. The scatterplots show the p-values for significant eQTLs for either mtSet or LIMIX. (a) The scatterplot shows the uncorrected -log10 p-values from mtSet (x-axis) and LIMIX (y-axis) for DHS-eQTLs discovered by mtSet. The vertical dashed line indicates the FDR cutoff for mtSet, while the horizontal dashed line indicates the FDR cutoff for LIMIX. The p-values have a Person correlation of 0.432 (b) The scatterplot shows the uncorrected -log10 p-values from mtSet (x-axis) and LIMIX (y-axis) for DHS-eQTLs discovered by LIMIX running on DHS only. The vertical dashed line indicates the FDR cutoff for LIMIX while the horizontal dashed line indicates the FDR cutoff for mtSet. The p-values have a Person correlation of 0.662. (c) The violin plot represents the -log10 p-value distribution for mtSet and LIMIX of DHS-eQTL discovered by mtSet. The horizontal solid line indicates the FDR cutoff for LIMIX. The dotted line corresponds to a p-value of 0.05. (d) The violin plot represents the -log10 p-value distribution for mtSet and LIMIX of DHS-eQTL discovered by LIMIX. The horizontal solid line indicates the FDR cutoff for mtSet. The dotted line corresponds to a p-value of 0.05.

2.2 - Overview of the DHS-eQTL results

mtSet discovered a total of 2,967 DHS-eQTL. The nature of the method allowed for the discovery of multiple DHS associated to the same gene and multiple genes linked to the same DHS, in particular, the 2,967 DHS-eQTL involve 1,673 genes and 2,693 DHS. Depending on the type of DHS involved in the eQTL, we can split the results in promoter-proximal-eQTLs and distal, putative enhancer-eQTLs. 1,805

promoter-proximal-DHS (7.8% of all promoter-proximal DHS) were involved in 2,005 promoter-proximal-eQTLs. 888 distal-DHS, which I refer to as enhancers from now on (representing 2.4% of all enhancers) were involved in 962 enhancer-eQTLs.

Figure 13a shows the number of DHS associated to each gene. While the majority of genes were involved in one DHS-eQTL, 635 (38%) of them were linked to two or more DHS, up to 19 DHS. The majority of eQTL studies have focused on the best association for each gene. By including all DHS to gene associations in the multiple testing correction, the DHS-eQTL approach represents a step forward in capturing the complexity of gene regulation.

Figure 13b displays the number of genes associated to each DHS. The ground assumption is that every CRM regulates only one target gene, because only a handful of examples of enhancers that regulate multiple genes are reported in the literature[71]. Our results showed that 231 DHS (8.6% of DHS involved in DHS-eQTLs) were associated to 2 genes and 21 (0.8%) are associated to 3 or 4 genes. Of those, 173 were promoter-proximal-eQTLs and 58 were enhancer-eQTLs. This result gives functional evidence that CRM sharing is more widespread than previously anticipated and gives new functional insights into the complexity of a gene's regulatory landscape. One could argue that DHS that are involved in multiple DHS-eQTLs represent *trans* interactions. Unfortunately, it is impossible to separate *cis* from *trans* interactions within populations of inbred individuals so we cannot directly estimate the amount of *trans*-eQTL within these results. On the other hand, testing *cis* window of ±50 kb around each gene makes it unlikely to discover *trans* associations. I will expand more on the estimation of *trans*-eQTLs in "2.6.1 - Different types of activity from promoter-proximal DHS".

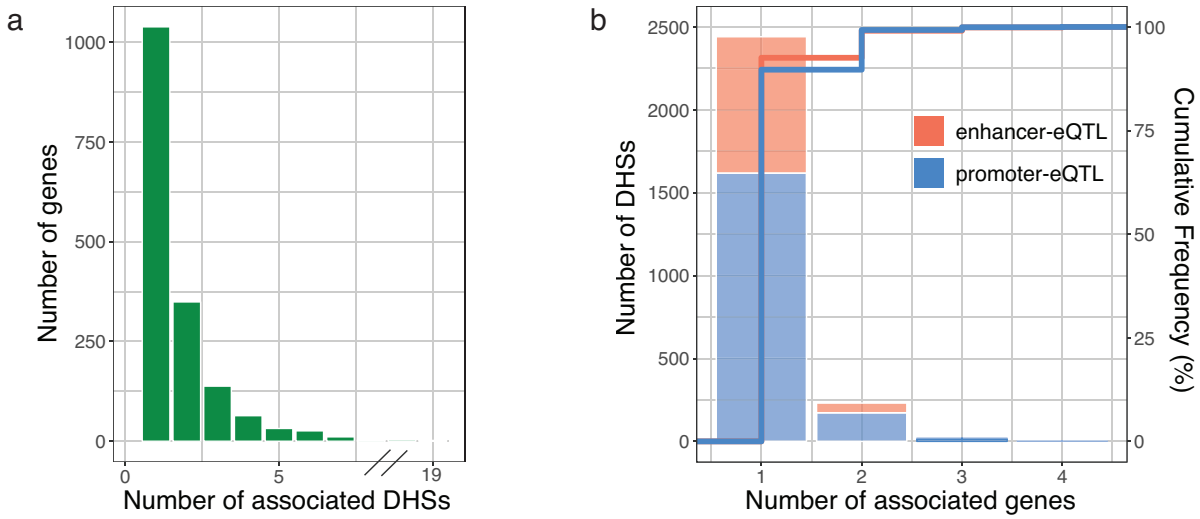


Figure 13 – Number of genes associated to each DHS and vice versa. (a) Number of DHS associated to each gene. (b) The plot shows the number of genes associated to the same DHS. DHS are divided in promoter-proximal DHS and enhancers. Enhancer-eQTL are shown in red, promoter-proximal-eQTL are shown in blue.

2.2.1 - Distal enhancer-eQTLs

A current key challenge in our understanding of genome regulation is how to link regulatory elements (CRMs) to their target genes. This is particularly difficult for distal enhancers, which can be separated by tens or hundreds of kb from their target gene, with many non-target genes in between. Given the increased statistical power of our approach, mtSet identified 962 enhancer-eQTL, one enhancer of which is shown in Figure 14. An enhancer, open only in the neural tissue, is associated with the gene *aret* (also called *bruno 1*), an RNA binding protein involved in splicing. The DHS is overlapping an enhancer genomic fragment tested by Kvon *et al.* The fragment has an enhancer activity specific to the ventral nerve cord (Figure 14). In addition, the expression patterns of the gene *aret* have been characterized and available on BDGP[72]. The gene is expressed in the ventral nerve cord and other tissues. The DHS-eQTL approach is completely unbiased towards this orthogonal information that in turn validates this association. We can conclude that the DHS-eQTL method identifies a novel distal enhancer for *aret* located 31 kb away, with 7 genes in between.

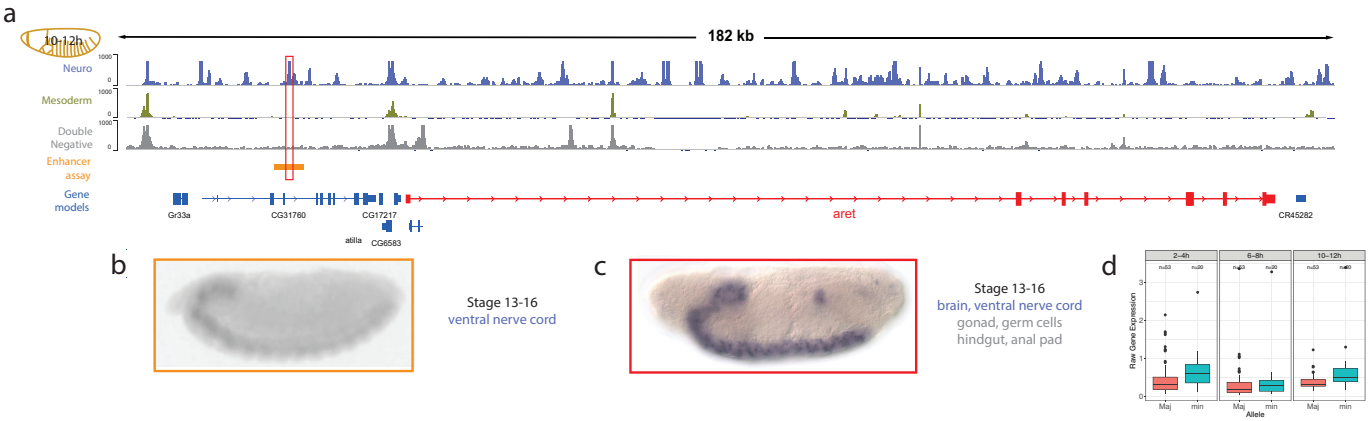


Figure 14 – Example of DHS-eQTL. (a) Screenshot displaying an example of a DHS-eQTL. A neuroectoderm specific enhancer is associated to *aret*. The tracks on top represent the DNase assay in neuroectoderm, mesoderm and double negative tissues at 10-12 hpf. The DHS overlaps a known enhancer (the region represented by the orange track was tested in an enhancer assay[59]). (b) Expression pattern induced by the region in orange when tested in an enhancer assay. (c) Gene expression pattern of gene *aret*. The gene expression induced by the enhancer region overlapped the expression pattern of *aret*. (d) Gene expression difference between major and minor alleles of the causal variant. The minor allele has consistently higher expression in all three time points.

2.3 - Validations of the results

The DHS-eQTL approach is intentionally focused on regulatory regions, to achieve more power in DHS to gene associations. In this section, I will compare the results from the DHS-eQTL with a Tiling Window eQTL method that takes the entire *cis* window into account. In addition, I will present two *in silico* validations of the results (overlap of DHS and gene tissue-specificity and Hi-C signal enrichment between DHS-eQTL), one *in vitro* validation (qPCR confirmation of major/minor allele expression) and one *in vivo* (validation by CRISPR mutagenesis).

2.3.1 - DHS are enriched for eQTL signal

By focusing on DHS, we exclude a substantial portion of the genome. The advantages of this approach are a 6-fold reduction in the number of tests and a definition of genomic windows that correspond to biologically functional units. To assess the impact of these two features, I tiled the genome in 300 bp large windows

(equal to the median DHS size) and tested all Tiling Windows within 50 kb of each gene for association, using mtSet (Figure 15). This approach considered the entire *cis* genomic region and is completely agnostic to DHS information, making it a traditional multivariate eQTL approach. The pipeline used to test for association is identical to the DHS-eQTL one. Tiling Windows have a similar variant frequency to DHS, but variants were less independent, indicating that DHS had smaller linkage disequilibrium blocks than the rest of the genome (Supplementary Figure 4). In total, I identified 9,855 Tiling Windows eQTLs. 1,237 TW-eQTL overlapped a DHS and reassuringly, 1,005 (81.2%) of them were associated to the same gene as in the DHS-eQTL (Figure 16).

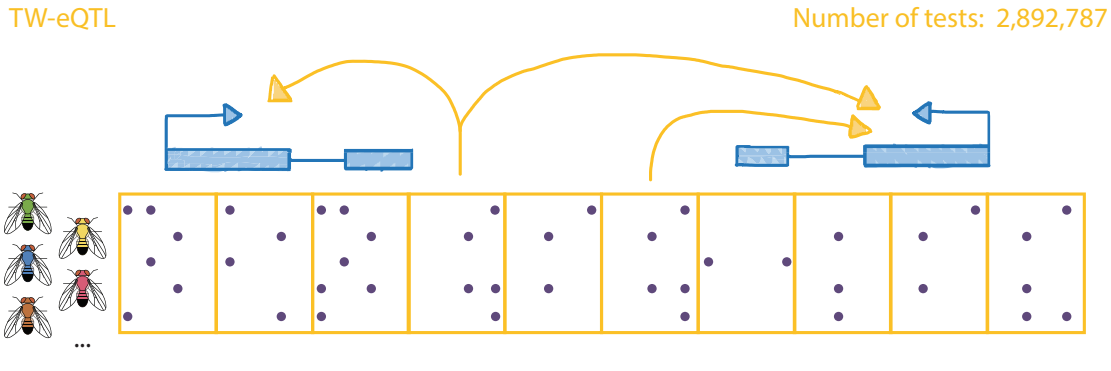


Figure 15 – Schematic representation of the Tiling Window-eQTL (TW-eQTL) approach. The entire genome is divided into Tiling Windows of 300 bp each. I perform one association test for each window. This method is similar to the DHS-eQTL method; the only difference is on the definition of the tested regions.

Considering the entirety of the genome identified many more eQTLs (9,855 TW-eQTLs vs 2,957 DHS-eQTLs). On the other hand, the DHS-eQTL discovered a higher number of DHS to gene associations (1,237 TW-eQTLs overlap a DHS vs 2,957 DHS-eQTLs): this comparison was the most relevant given the goal of this project. In addition, despite increasing the number of tests by six times (2,892,787 vs 483,064), the TW-eQTL only provided 3.3 times more associations than the DHS-eQTL. By comparing DHS-eQTL and TW-QTL performance directly with a joint test, I observed a 1.46-fold enrichment of DHS-eQTL, demonstrating that p-values from the DHS-eQTL tests are globally lower.

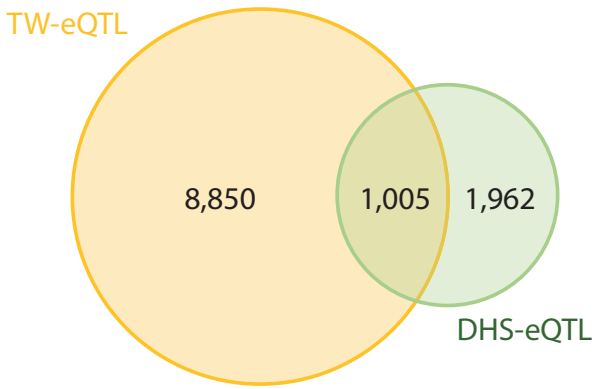


Figure 16 – Venn diagram comparing the results from TW-eQTL and DHS-eQTL. The TW-eQTL approach identifies more than three times more associations than the DHS-eQTL. However, by focusing on DHS only, the DHS-eQTL method is the best to associate CRMs to target genes.

TW-eQTL are unbiased toward what genomic region is tested. This allowed us to identify genomic features that were enriched or depleted for eQTLs. Figure 17 shows how frequently TW-eQTLs overlap different genomic features, compared to the background. TW-eQTLs were enriched at promoter-proximal DHS and exons. Genetic variants that overlap promoters are known to have larger effects on gene expression compared to enhancers[53] thus making promoter-proximal DHS more likely to harbor an eQTL. Exons are also known to be enriched in eQTLs[53]: 42% of the TW-eQTLs overlapped exons. On the other hand, exons are depleted in DHS and had little relevance within the scope of this project. TW-eQTLs were depleted in introns and intergenic regions; they were also slightly depleted in enhancers. While introns and intergenic regions are known to be depleted for eQTLs (because they are depleted for regulatory elements), we expected enhancers to be enriched for eQTLs. A possible explanation for depletion of eQTLs on enhancers is that genes are generally regulated by multiple enhancers with overlapping activities[44]. Mutations on enhancers can be buffered more effectively than on promoters, making it less likely to cause measurable changes on gene expression.

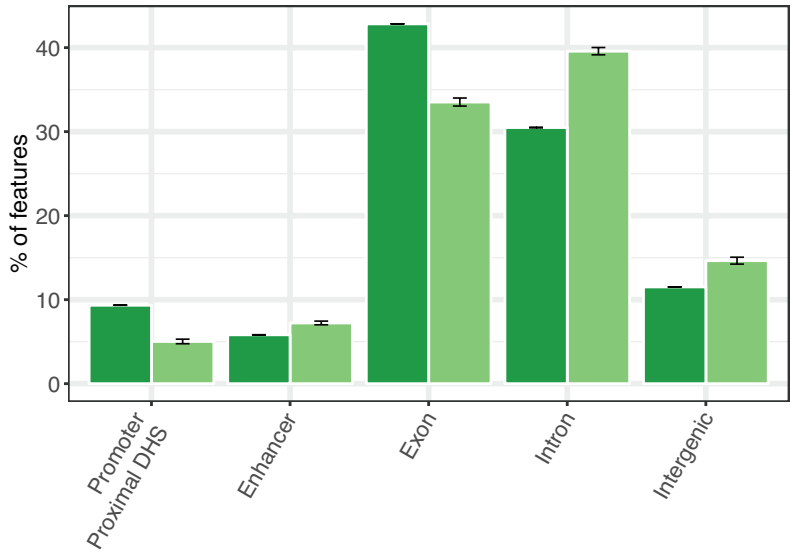


Figure 17 – Enrichment of TW-eQTL on genomic features. The dark green bars show the observed frequency of TW-eQTLs on each genomic feature. The light green bars show a randomized background of frequencies. All frequencies sum up to 100%.

2.3.2 - DHS-eQTL method is sensitive to precise DHS identification

Tiling Windows are defined as consecutive windows on the genome and create breaks that do not take into account genomic function. On the other hand, DHS are biologically meaningful and correspond to separate CRMs. To test if pooling variants in a biologically meaningful way leads to a difference in test significance, I performed a pairwise comparison between p-values from the DHS-eQTLs and the TW that overlaps most bases. The DHS-eQTL p-values were globally significantly lower (one tailed Wilcoxon test: $p<0.0034$) than those from the TW-eQTL that overlaps most bases. This result indicates that a precise definition of window borders leads to stronger associations.

In conclusion, focusing on biologically defined genomic regions and decreasing the number of tests enabled the DHS-eQTL approach to identify more than twice DHS to gene associations than the TW-eQTL approach. The TW-eQTL approach is perfectly valid and it gave insights into how multivariate approaches work, but it identifies less DHS to gene associations compared to the DHS-eQTL method.

2.3.3 - Enhancer and gene tissues of activity overlap more than at random

To validate our DHS to gene associations we tested for a correspondence between the tissues where the DHS is active and the target gene is expressed. Since the DHS-eQTL method does not incorporate information about tissue-specific activity of DHS and gene expression, we can use this important feature to test if our DHS-eQTLs are enriched for concordant DHS and gene tissues of activity. As many genes are annotated as having ubiquitous expression at early developmental stages, we focused on the latest time point of the DNase hypersensitivity data. We complemented our DHS data with gene expression information from the Berkeley *Drosophila* Genome Project[72] (BDGP). This resource provides annotated expression patterns for more than 8,000 genes. Figure 18a shows an example of a coherent tissue match between a DHS and gene pair in a DHS-eQTL. The gene *Mp20* is expressed in the visceral muscle at stage 16 (Figure 18b) and was associated to a mesoderm specific DHS about a kilobase away from its Transcription Start Site. On the other hand, Figure 18d shows an example of an incoherent association: *CG10602* is a gene expressed in the crystal cells at stage 16 but was associated to a neuroectoderm specific enhancer by mtSet (Figure 18e).

There are multiple reasons why we might find incoherent associations. First, *in situ* hybridization may not target all gene isoforms and it might not capture some tissues where the gene is expressed. Second, BDGP annotation is based on manual annotation of *in situ* hybridization images and it is prone to human errors. The BDGP dataset provides invaluable insight and it is powerful enough for a global validation, but the very nature of the assay does not enable assumptions about every single case. Third, DHS data, despite being tissue-specific, might miss rare cell types (for example, crystal cells represent only a small proportion of the Double Negative tissue). Fourth, the TSS distal DHS might be a silencer and therefore it could be active in tissues where the gene is not expressed and should be inversely associated with the gene's expression.

I globally tested if DHS-eQTLs are enriched for coherent associations. I defined tissue-specific enhancers at 10-12 hpf as those having a summit only in one time point and having a significantly different coverage using DESeq2[73]. Gene expression annotation was downloaded from BDGP and we mapped the anatomical tissue terms at stages from 13 to 16 (that correspond to 10-12 hpf) to the 3 tissues in the DHS study. I focused on DHS-eQTL that link tissue-specific enhancers to tissue

specifically expressed genes. Among these DHS-eQTLs there were 38 coherent and 24 incoherent associations. To test for significance, I performed 10,000 permutations of random DHS to gene associations and obtained a background distribution of coherent and incoherent associations (Figure 18c,f). Coherent associations were enriched in our results while incoherent ones were depleted. Incoherent associations might point to false discoveries but they do not display lower quality signatures. In fact, they have the same DHS to gene distance and p-value distributions as coherent associations. These results show an enrichment of coherent tissue of activity between DHS and target gene among the DHS-eQTLs.

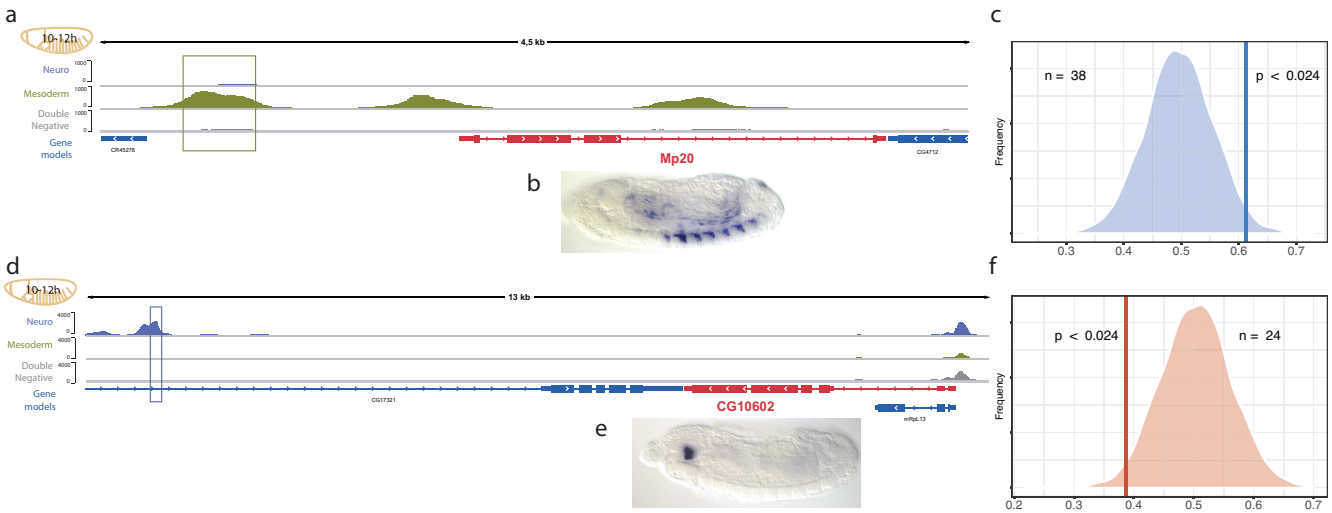


Figure 18 – Enhancer and gene tissues of activity overlap more than at random. (a) The gene Mp20 is associated to a mesoderm specific enhancer. (b) Expression pattern of *Mp20* at stage 16 (corresponding to 10-12 hpf). (c) The vertical bar shows the proportion of observed coherent associations. The transparent distribution represents the permutation background. (d) The gene *CG10602* is associated to a neuro specific enhancer. (e) Expression pattern of *CG10602* at stage 16 (corresponding to 10-12 hpf). (f) The vertical bar shows the proportion of observed incoherent associations. The transparent distribution represents the permutation background.

2.3.4 - DHS involved in DHS-eQTL are enriched for Hi-C contacts

Distal regulatory elements – such as enhancers – can be located hundreds of kilobases away from their target promoter, but they are known to function by coming in physical proximity with the target gene's promoter. In physiological contexts, enhancers known to regulate a gene are closer in the 3D space to the target gene promoter than those that do not regulate it, independently of the linear distance[35]. In the past years, many chromosome conformation capture techniques have been

developed to investigate the genome spatial arrangement. Among them, Hi-C[34] quantifies all pairwise contact frequencies between genomic regions, providing a measure that can be interpreted as spatial distance. With the help from Aleksander Jankowski, a post-doc in the Furlong laboratory, we analyzed Hi-C in wild type embryos staged at 6-8 hpf to quantify chromosome contact frequency. DHS to gene associations are expected to be closer in space than random associations. We globally observed that DHS involved in a DHS-eQTL contact their target promoter more frequently than random DHS at the same linear distance from the promoter (Wilcoxon p<10^-188) (Figure 19). This indicates that the DHS are regulating the gene there are associated with, by the DHS-eQTL method.

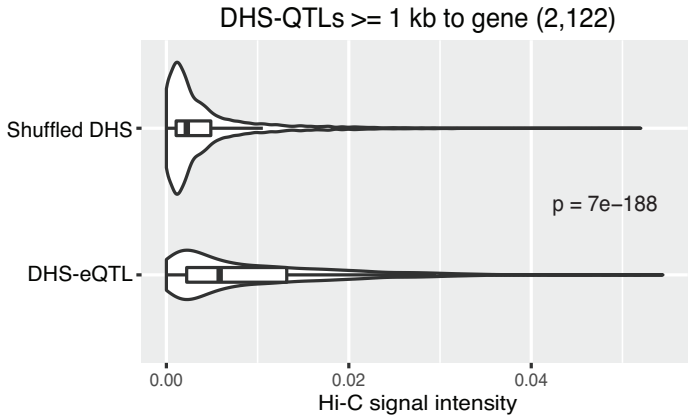


Figure 19 – DHS-eQTL have stronger Hi-C signal than random DHS to gene associations. The plot shows the Hi-C signal intensity distribution between DHS and target gene TSS. The top violin plot represents the distribution of shuffled DHS-eQTL (DHS are shuffled while genes are kept constant, DHS to gene distance is matched). The bottom violin plot represents DHS-eQTLs intensity signal distribution between DHS and target gene TSS. Hi-C signal correlates with proximity.

2.3.5 - qPCR validates gene expression differences between major and minor alleles

Gene expression differences between genotypes are essential to identify eQTLs. But since the test is correlation based, differences might be subtle and challenging to reproduce. In this study, we used mtSet, a statistical method that can take multiple variants into account. By performing a multivariate test, mtSet cannot identify a putative causal variant among the ones that are tested. To achieve this, I defined the variant with the lowest p-value in the univariate (LIMIX) test, as the putative causal variant. This allowed for splitting gene expression between alleles and estimating the

effect size of the variant on gene expression (Supplementary Figure 5). To confirm both the direction of the causal variant effect on gene expression and the fold change in gene expression, we used Real Time quantitative PCR (RT-qPCR). The experiments were performed by Rebecca Rodriguez Viales, a technician in the Furlong laboratory. We selected 5 promoter-proximal-eQTLs and 5 enhancer-eQTLs in which the target gene TSS is at least 10 kb away from the DHS (Supplementary Figure 5a-b). For each DHS-eQTL, we chose two DGRP lines harboring different alleles of the causal variant. The median fold change in gene expression between the two genotypes is lower than 2, indicating that the causal variants had minor effects on gene expression. We could reliably confirm mtSets associated expression differences for 9 out of 10 genes (*eIF3f1* gave discordant results in different replicates and with different primers) and we confirmed our fold change estimates by quantifying expression in the two genotypes with RT-qPCR. All genes were differentially expressed between the two genotypes and in 8 out of 9 cases the direction of expression was the same as measured with 3'-Tagged Sequencing (Supplementary Figure 5c,d). RT-qPCR and 3'-Tagged Sequencing fold changes in gene expression between major and minor alleles are highly correlated (Pearson correlation: 0.85) (Figure 20). RT-qPCR validates the gene expression measurements from 3'-Tagged Sequencing, the direction of the eQTL effect and the fold-change in gene expression between lines harboring the Major and minor alleles.

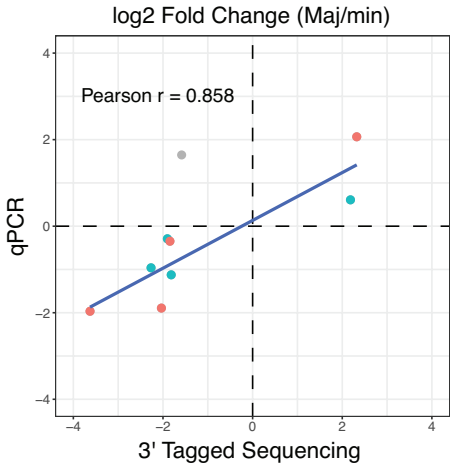


Figure 20 – qPCR maj/min fold change correlates with 3'-Tagged Sequencing measurement. The plot shows the correlation between log2 fold changes in gene expression between major and minor allele as measured by 3'-Tagged sequencing and quantitative Real Time PCR (qPCR). The orange dots represent enhancer-eQTLs while the blue dots represent promoter-eQTLs. The correlation between qPCR and 3' Tagged-Sequencing log2 fold changes is: 0.858. The grey dot represents the fold change in expression of RpS5b, for this sample the fold change in gene expression does not agree and it is excluded from the correlation.

2.4 - The DHS-eQTL approach indicates that correlative methods underestimate the distance between enhancers and target genes

Direct methods to associate enhancers to target genes (such as enhancer deletion) tend to be poorly scalable and time consuming: it is currently impossible to apply them at a genome-wide scale to a developing organism. For these reasons, molecular biologists have relied on correlative methods or on cell-culture based *in vitro* proxies of development to associate enhancers and target genes genome-wide. The most trivial approach is to associate enhancers to their closest gene. This simplistic method represents the best guess when nothing is known about the biology of the enhancer or of the gene, but it is poorly accurate. Another approach is to link enhancers and target genes based on the patterns of enhancer activity and gene expression. In a seminal study, Kvon *et al*.[59] tested more than 10% of the *Drosophila* genome in an enhancer assay that reveals what tissues the enhancer is capable of driving expression in. They assigned enhancers to the closest gene with a compatible expression pattern. The major drawback of the method is that it relies on sparse data: tissues of expression is annotated for only a fraction of genes. This method is more accurate than linking enhancers to their closest gene, but it is still biased towards short distance.

On the other hand, the DHS-eQTL approach is completely agnostic about the distance between the enhancer and the target gene (within our tested +/- 50 kb *cis*-window). Figure 21 shows an example of long-distance enhancer-eQTL. A mesoderm specific enhancer was associated to *PH4alphaSG1*, about 30 kb away. The enhancer was associated to its 11th closest gene. In this section, I will compare the distribution of enhancer-to-gene distances obtained from different association methods.

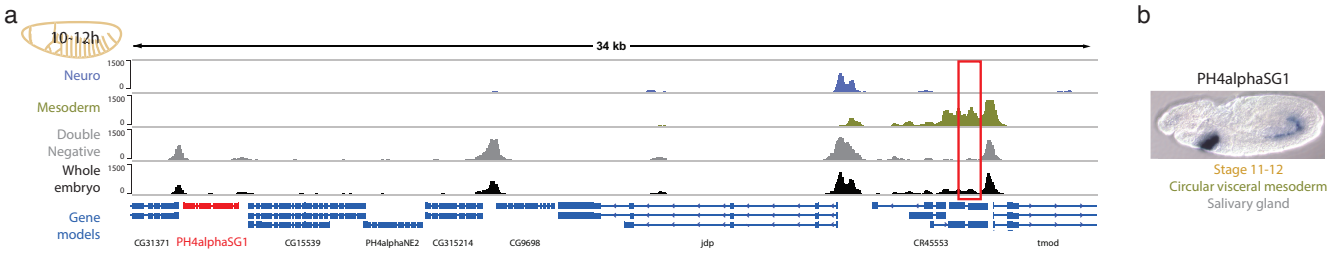


Figure 21 - Mesoderm specific enhancer skips 6 genes to reach its target. (a) Browser shot of a mesoderm specific enhancer associated to the gene PH4alphaSG1. The enhancer (red box) is located 30 kb away from the gene promoter and skips 6 genes to reach it. (b) Gene expression pattern of PH4alphaSG1: the gene is expressed in the circular visceral mesoderm and in the salivary gland at 10-12 hpf.

2.4.1 - Enhancers act over longer distances than promoter-proximal DHS

The eukaryotic genome is highly compacted in the nucleus making linear distance a poor estimate of 3D-space distance. Chromatin Conformation Capture technologies show that long distance interactions are frequent[35]. These observations suggest that the idea of enhancers regulating the closest gene in linear distance might not represent a general rule. When looking at the distribution of distances between the DHS and the associated target gene TSS among the DHS-eQTL, we see that they are skewed toward short distances (Figure 22a). On the other hand, the same figure shows that the majority of DHS to gene interactions span more than 10 kb indicating that long distance (within the compact *Drosophila* genome) DHS to gene interactions are very common. The same result can be observed in Figure 22b: the figure shows the number of genes that a DHS skips to reach its target. 74.2% of DHS were not associated to their closest gene, with 14.4% of the DHS skipping more than 10 genes. Virtually all eQTL studies showed the same distributions in Figure 22[53] and this is consistent with the idea that DHS are still more likely to regulate genes in their proximity. It is important to stress again that DHS to gene distance is not a parameter in the eQTL model, making the distributions in Figure 22a a completely unbiased result. If we separate enhancers from promoter-proximal DHS we observe distinct behavior for the two classes. In particular, enhancers act more distally than promoters. To summarize, Figure 22 indicates that: (1) Both enhancers and promoter-proximal DHS are more likely to regulate a gene in their proximity than a distal one. (2) Enhancers often span long distances to reach their target genes. (3) Enhancers act over longer distances than promoter-proximal DHS. (4) There is widespread distal activity from promoter-proximal DHS.

One caveat of the results presented here is that the eQTL approach does not make it possible to distinguish *cis* from *trans* associations. For example, a variant might marginally affect the expression of a transcription factor in its vicinity, which in turn might have larger effects on the regulation of a gene many kilobases away. The proximal effect would not be detected while the distal would: this association would show as a long distance eQTL. A second caveat is that eQTLs are just capturing a small proportion of DHS to gene interactions. In particular, eQTLs can be found when a variant (or a combination of variants) changes the property of a regulatory region that in turn influences gene expression. Variant density is not uniform across the genome making some regions more likely to harbor an eQTL. The variant also has to affect expression in a consistent manner in order to detect a significant association. eQTLs in general (including DHS-eQTL) are enriched for metabolic genes whose changes in expression have a smaller effect on fitness. Considering these limitations, the results of the DHS-eQTL approach are not necessarily representative of the entire regulatory landscape.

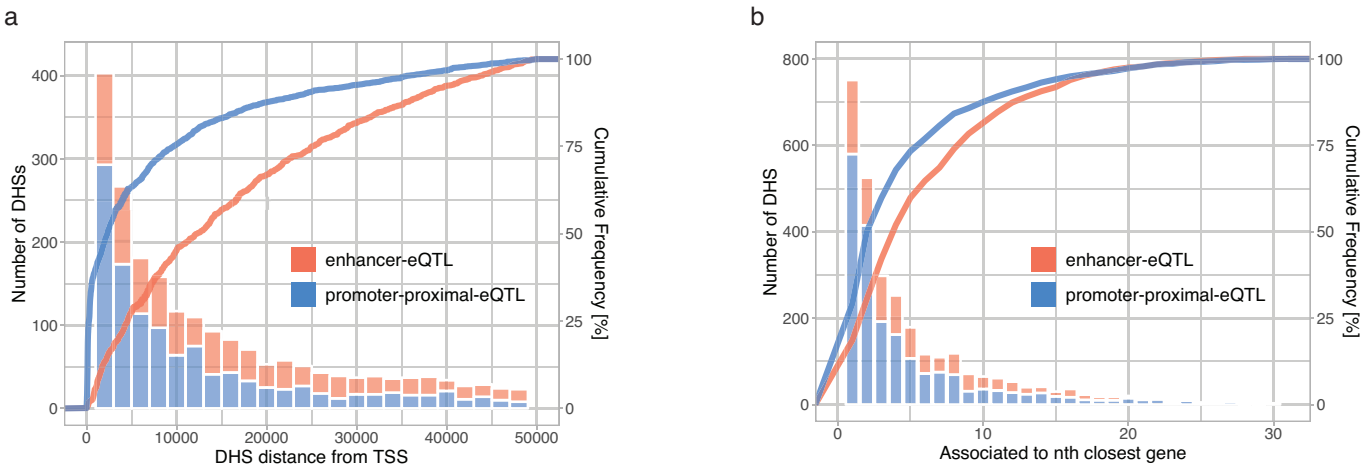


Figure 22 – Distribution of distances between DHS and target genes for enhancer-eQTLs and promoter-proximal-eQTLs. (a) Distribution of DHS to target gene TSS distance split by enhancer-eQTLs and promoter-proximal-eQTLs. (b) Genes around each DHS-eQTL are ranked from the closest to the farthest (based on their TSS position). The plot shows the rank of the target gene for each DHS-eQTL. 751 DHS are associated to their closest gene, 525 to the second closest and so on.

2.4.2 - Annotation methods based on expression patterns underestimate enhancer to gene distance

Enhancer activity can be tested in an ectopic context via an enhancer-reporter assay. This assay works by cloning a genomic region with unknown enhancer activity in

front (or downstream) of a reporter gene with a minimal promoter. Without enhancer activity from the assayed region, the reporter gene will not be expressed. When integrated into the genome, an enhancer assay also reveals the tissues and times where the enhancer is active. This is achieved by visualizing the reporter gene expression. Kvon *et al.*[59] have used such transgenic enhancer assays to study enhancer activity for thousands of elements during embryogenesis. They tested more than 10% of the *Drosophila* non-coding genome by dividing it in more than 7,000 Tiles (so called Vienna Tiles) each one tested in an enhancer assay. In their work, the researchers associated enhancers to the closest gene with a known expression pattern (from BDGP[72]) that overlapped the enhancer assay expression. This gene assignment approach has a few shortcomings: (1) The Tiles span a few kilobases, while known enhancers are generally smaller, in the range of 300-500 bp. Tiles might include multiple enhancers with different activity patterns and each of them might regulate different genes. (2) The method relies on sparse gene expression data from BDGP. Less than half of *Drosophila melanogaster* genes had entries in the database at the time of the study. (3) Genes are not randomly dispersed in the genome; this is especially relevant for the compact *Drosophila* genome. In fact, neighboring genes tend to have similar expression patterns, meaning that around each enhancer there can be many genes with expression patterns that overlap the enhancer activity.

Figure 23 shows the distance distribution between enhancers and target genes assigned by 3 methods: closest gene, the Vienna Tiles approach from Kvon *et al.* and the enhancer-eQTLs identified here. Both Kvon *et al.* approach and enhancer-eQTLs show that associating enhancers to their closest genes largely underestimates enhancer to gene distance (Figure 23a). The main difference between the Vienna Tiles and the enhancer-eQTL results is in the number of genes skipped by each enhancer (Figure 23b). Kvon *et al.* associate 47.8% of enhancers to their closest gene while this correlation could only be found for 19.2% of enhancer-eQTLs. This difference might be explained by the fact that the search stops at the first gene whose expression overlaps the enhancer activity. The DHS-eQTL approach shows that enhancers act over longer distances than suggested by correlative methods.

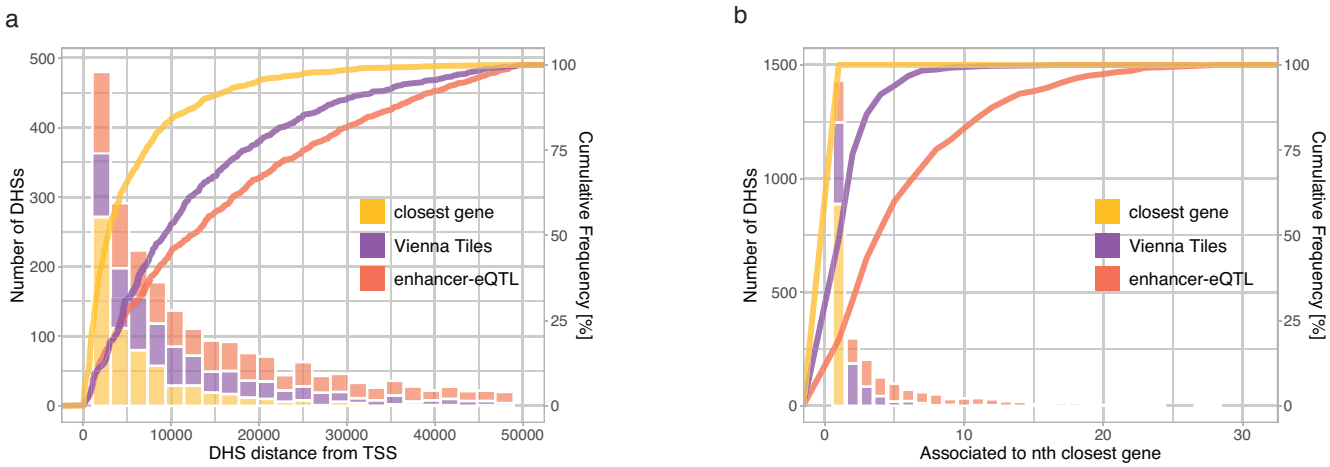


Figure 23 - Distribution of distances between enhancers and target genes from three association methods. (a) Distribution of enhancer to target gene TSS distance split by the method of enhancer to gene association: closest gene, Vienna Tiles and enhancer-eQTLs. Enhancers-eQTLs span longer distances than the other methods. (b) Genes around each enhancer are ranked from the closest to the farthest (based on their TSS position). The plot shows the rank of the target gene for each enhancer to gene association method. Enhancers associated to their target the "closest gene" method fall in the first bar by definition.

2.5 - Enhancers and promoter-proximal DHS can regulate multiple genes

In the textbook view of gene regulation, CRMs control only one target gene[1]. In fact, biological systems need to regulate gene expression in a very specific way, especially during embryonic development[74]. The assumption is that to achieve this accuracy, each gene has a set of unique regulators. This view is challenged by many observations coming from topological studies and co-expression. In fact, breaking insulator elements brings enhancers in contact with new target genes[75], suggesting that some enhancers are rather promiscuous if they are given the chance to interact with new genes. In addition, genes that are close in the linear genome tend to be co-expressed[76], suggesting regulatory elements sharing among genes[77]. Despite these global evidences, there are only a few known examples of enhancers that regulate two genes[71]. The DHS-eQTL method tests, on a genome-wide scale, associations between DHS and target genes. The results provide hundreds of examples of CRM sharing and estimate that at least 8% of CRMs are shared by genes during *Drosophila* embryonic development.

2.5.1 - Example of enhancer sharing

Figure 24 shows an example of enhancer sharing. An enhancer, open only in the Double Negative (non-neuro non-meso) tissue, is associated to both *Caf1* and *mRpS10* genes. BDGP reports the expression patterns of the two target genes and they are both expressed in the midgut, an endoderm derived tissue that is included in the Double Negative tissue from the FACS sorting. The DHS-eQTL model is unbiased towards the tissues of expression of the genes, making this an independent validation of the results.

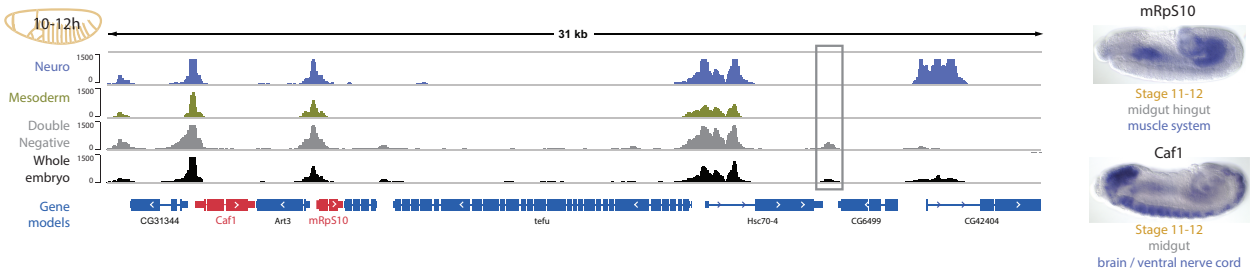


Figure 24 – A Double Negative specific enhancer is associated with two distal genes both expressed in the midgut. The figure shows an enhancer open only in the Double Negative tissue at 10-12 hpf, associated to two distal genes: Caf1 and mRpS10. The expression patterns at stage 11-12 (corresponding to 8-10 hpf) are displayed on the right. Both genes are expressed in the midgut.

2.5.2 - The expression of genes linked to the same DHS is highly correlated

The DHS-eQTL approach discovered hundreds of promoters and enhancers associated to two genes i.e. *cis* Regulatory Modules sharing. Variants that modify the function of a DHS that regulates two genes should have an impact on both genes, leading to co-expression across DGRP lines. To test this hypothesis, I performed a Pearson correlation between expression (here I used gaussianized PEER residuals) of genes associated to the same DHS. The qqplot in Figure 25a shows the correlation p-values in red. 100 (43.5%) genes associated to the same DHS were indeed correlated in gene expression and the p-values were lower than expected by chance. Genes regulated by the same DHS are not randomly located in the genome but tend to be in proximity of each other (Figure 25b). This result was unbiased since the model does not take into account the distance between genes and indicated that genes in the same neighborhood are more likely to be coregulated. To assess the significance of this result I added two controls. Firstly, the light grey dots (Figure 25a) show correlation p-values of random pairs of genes across the whole genome. The

p-values from this correlation follow closely the distribution expected for random correlations. Secondly, since genes located in the same neighborhood tend to be co-expressed[76], I tested if the correlation between genes that share a CRM could be explained by vicinity alone. I selected random pairs of genes whose distance matches the distribution in Figure 25b and plotted the p-values in dark gray in Figure 25a. Genes located closely in the genome were correlated to a higher extent than random pairs of genes, but to a lower extent than genes that share a DHS. This shows that genes that share a DHS had a high degree of co-expression that cannot be explained by vicinity alone.

The correlation of expression between genes associated to the same DHS can lead to a circular argument. In fact, the DHS-eQTL is a correlation-based test and the three elements (two genes and one DHS) all correlate with each other. On one hand, one could argue that since the eQTL tests are based on correlation, if the dependent variables are correlated, then they will be associated to the same independent variables. This would mean that only one of the genes is truly regulated by the DHS, while the other gene's expression is correlated to the first. But on the other hand, gene expression does not correlate across the whole genome (Figure 25a) indicating that co-expression is caused by *cis* regulation. In addition, this argument is of concern only if the correlation of expression is caused by a factor other than *cis* regulation, such as batch effects, population structure or *trans* effects. The pipeline adopted here removed these confounding effects to the best of our knowledge, leaving only *cis* regulation as an explanation for co-expression. Other than logical arguments, the best way to understand causality in such an interconnected system is to interfere with one element and observe what happens to the others. Therefore, we are performing CRISPR deletion of 11 DHS associated to two genes and we will quantify the changes in gene expression for both targets (see "3.1 - Validation of complex DHS-eQTLs by *in vivo* CRISPR-Cas9 mutagenesis").

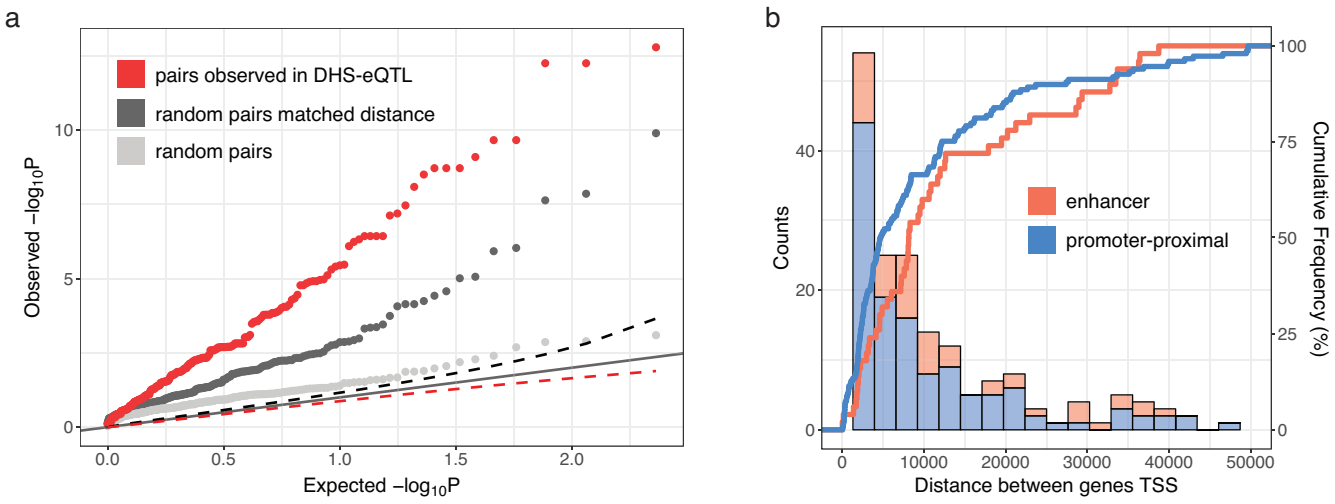


Figure 25 – Genes associated to the same DHS are close and have highly correlated gene expression. (a) qqplot of Pearson correlation p-values of expression from gene pairs. The correlation is measured among DGRP lines. The p-value reported here is the lowest among the three time points. The red dots represent correlation between genes associated to the same DHS in the DHS-eQTL results. The light grey dots represent the correlation between random pairs of genes. The dark grey dots represent the correlation between random pairs of genes with a distance distribution matched to the genes pairs in the DHS-eQTL results. The solid line shows the expected p-value distribution for non-significant tests, the dashed lines show 95% confidence interval. (b) Distance distribution between pairs of genes associated to the same DHS. The gene pairs are divided into two categories depending on if they are associated to an enhancer (red) or a promoter-proximal DHS (blue). Cumulative curves are shown on top.

2.5.3 - Promoters and enhancers that regulate multiple genes show different relationships with target genes

Promoter-proximal-DHS and enhancers that are associated to two genes, can have different physical relationships with the target genes. To better understand their behaviour, I separated the DHS associated to two genes in four categories depending on where the DHS and the target genes are located:

- Overlapping both genes: the DHS is located in proximity (< 500 bp distance or overlaps) of both genes.
- Overlapping one gene: the DHS is located in proximity (< 500 bp distance or overlaps) of one gene but it is distal to the other.
- Between genes: the DHS is distal to both genes and is located in between them.
- External: the DHS is distal to both genes and is located externally to them.

Figure 26a shows a schematic of these four categories and Figure 26b displays the counts for promoter-proximal DHS and enhancers. While promoter-proximal DHS

were frequently located close to at least one of the two targets, enhancers generally did not overlap any of the target genes and are more often external. The promoter-proximal DHS that overlap one gene might point to *trans*-eQTLs. In fact, the DHS could have a *cis* effect on the expression of the proximal gene that in turn has a *trans* effect on the second distal gene. I will further discuss this observation in the next section.

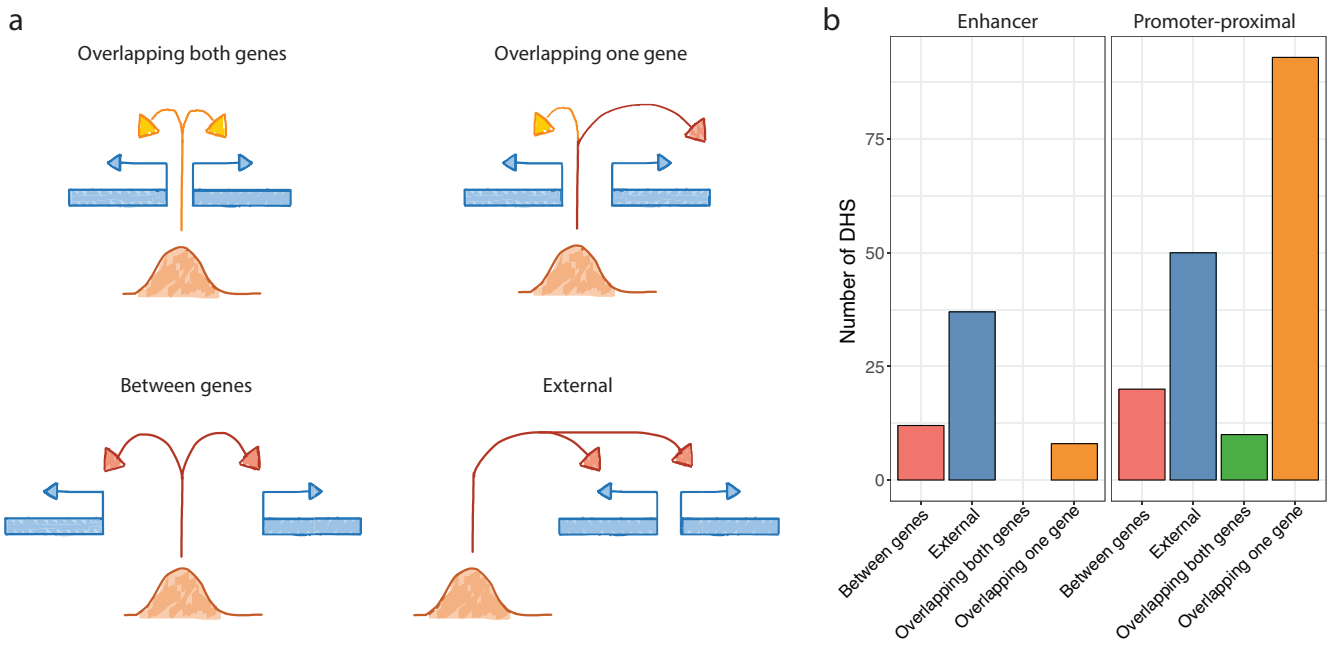


Figure 26 – Position of DHS associated to two genes relative to the targets. (a) A schematic representation of four possible relationships between the DHS and the two target genes. Yellow arrows represent short (< 500 bp) distance between the DHS and the target gene and red arrows represent long distance. I classified the relationships in four categories: the DHS can overlap both genes, overlap one gene and be distal to the second, be distal to both genes and be located either in between the two genes or externally. (b) The bar plot shows the counts of DHS belonging to the four categories. DHS are divided in enhancers and promoter-proximal DHS.

2.6 - Promoter-proximal DHS have widespread distal activity

In this project, I divided DHS in two groups depending on their proximity to known TSS: enhancers are TSS distal and promoter-proximal DHS are TSS proximal. The promoter-proximal group included a broad range of genomic elements other than core promoters that are challenging to separate from each other. For example, promoter-proximal DHS were highly enriched for insulators (59.7% of them have an insulator binding site), compared to enhancers (15.2%). Within the DHS-eQTL results, we unexpectedly observed a high proportion of distal activity from promoter-proximal DHS. Distal activity could come from non-core promoter elements such as insulators or promoter-proximal enhancers. In this section, I will expand more on the observation that promoter-proximal DHS are often associated with distal genes.

2.6.1 - Different types of activity from promoter-proximal DHS

Promoter-proximal DHS can have different modalities of activity depending on the distance to the target gene. I divided promoter-proximal DHS activity into four categories: local, distal, local and distal, and multiple distal. Figure 27 shows a schematic representation of these types of activity and an example for each.

Assuming that promoter-proximal DHS behave as core promoters, we expect them to regulate genes in their vicinity. Figure 27 shows an example of promoter-proximal DHS with local activity: the gene *Tfb4* was linked to a DHS that overlaps its TSS. Promoter-proximal DHS have a 3,68 odds-ratio enrichment for local activity compared to any other distal activity. In fact, 615 DHS-eQTLs described an association between a promoter-proximal DHS and a gene with a TSS within 500 bp from the DHS. This result is not novel, but it is very reassuring.

1,004 DHS-eQTL describe distal activity from promoter-proximal DHS. Figure 27 shows an example of promoter-proximal DHS with distal activity: the promoter of gene *CG8378* was linked to the distal gene *Hen1*. This was an unexpected behavior for a core promoter, but promoter-proximal DHS also include enhancers and insulators that can have distal activity. One could argue that these DHS-eQTLs represent *trans*-eQTLs. In fact, a variant on a promoter might have a mild effect on the expression of the gene regulated by the promoter itself, that in turn leads to

larger effects on the gene's targets. For example, returning to the example in Figure 27, *CG8378* is a transcription factor belonging to the Smyd gene family. If we suppose that *CG8378* regulates *Hen1*, the variant causing the DHS-eQTL between the promoter of *CG8378* and *Hen1* might cause an undetectable expression change to *CG8378* itself (mtSet uncorrected p-value > 0.1); that in turn would cause larger differences in the expression of *Hen1*. Then we will discover only the association between the promoter of *CG8378* and *Hen1*. To assess if *trans* effects could globally explain distal regulation from promoter-proximal DHS, I tested if this category of promoter-proximal DHS is enriched to be in the vicinity of a transcription factor's TSS. Compared to all the expressed genes, promoter-proximal DHS with distal activity were not enriched for being located at the 5' of Transcription Factors (3.60% were at a transcription factor TSS, within 500 bp, compared to 3.62% of all expressed genes that are annotated as transcription factors). On the other hand, promoter-proximal DHS with local activity were depleted for being close to a Transcription Factors TSS (1.26%). The proportion of promoter-proximal DHS with distal activity in the vicinity of a transcription factor is marginal, suggesting that this mechanism cannot be explained by *trans* effects alone.

Out of the 173 promoter-proximal DHS associated to two genes, 96 were linked to a gene in their vicinity together with a distal gene, while 77 were linked to two distal genes. Figure 27 shows an example of a promoter-proximal DHS associated to a local and a distal gene: the promoter of *e(y)2b* was associated both to itself and the distal gene *CG1091*. It also shows an example of promoter-proximal DHS associated to two distal genes: the promoter of *gammaTub23C* was associated to *CG9641* and *CG3165*. Again, there was no evidence for enrichment of transcription factors TSS close to these DHS (4.40% of promoter-proximal DHS with local and distal activity were promoters of transcription factors, as were 2.74% of promoter-proximal DHS linked to two distal genes).

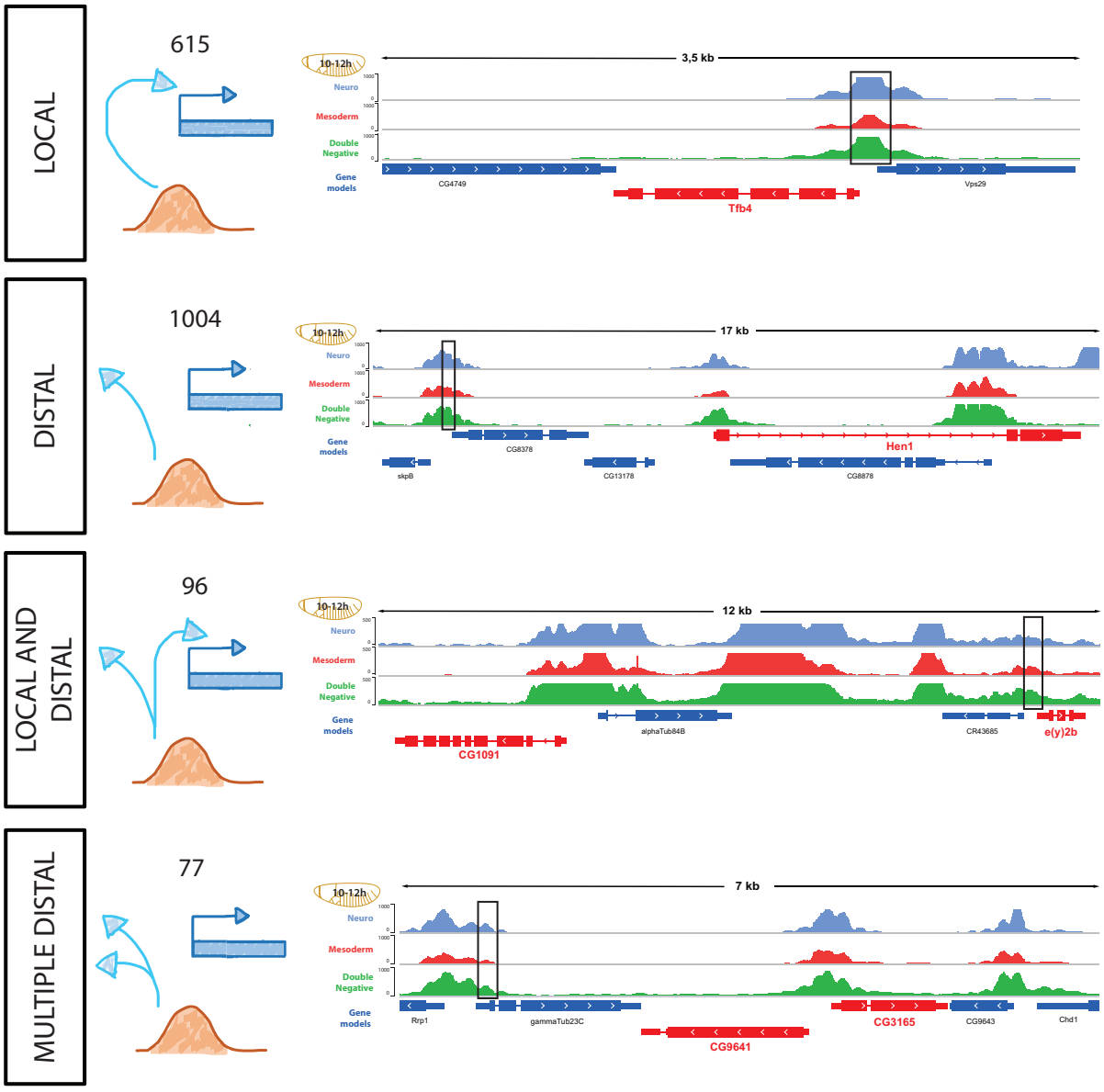


Figure 27 – Four types of activity from promoter-proximal DHS. Promoter-proximal-eQTLs can be divided into four categories. PROXIMAL: 615 promoter-proximal DHS are associated to a gene that has a TSS within 500 bp. On the right, a DHS-eQTL between the promoter of *Tfb4* and the gene *Tfb4* itself. DISTAL: 1004 promoter-proximal DHS are associated to a gene that does not have a TSS within 500 bp. On the right, a DHS-eQTL between the promoter of *CG8378* and the distal gene Hen1. PROXIMAL AND/OR DISTAL: 96 promoter-proximal DHS are associated to two genes, at least one of which has a TSS within 500 bp. On the right, the promoter of *e(y)2b* is associated to *e(y)2b* itself and the distal gene CG1091. TWO DISTAL: 77 promoter-proximal DHS are associated to two genes whose TSS are further than 500 bp. On the right, the promoter of *gammaTub23C* is linked to the two distal genes *CG9641* and *CG3165*. The two genes are divergently transcribed from the same promoter and their expression is highly correlated.

2.6.2 - Promoters of convergently transcribed genes are associated to each other

Within our DHS-eQTL dataset, we observed 3 cases of convergently transcribed genes whose promoters are associated to both genes. These may be special cases of promoter-proximal DHS that have both proximal and distal activity (Figure 27). Figure 28 shows an example: *Sec61gamma* and *Arp10* were short genes transcribed in a convergent direction. Their 3'-UTR did not overlap because they use independent polyadenylation sites. Expression of the two genes was positively correlated (Pearson correlation = 0.49), indicating strong coregulation. The DHS-eQTL results imply that both genes were linked to each other's promoters (the promoter of *Sec61gamma* was associated to itself and *Arp10* and the promoter of *Arp10* was associated to itself and *Sec61gamma*). In addition, both promoters were bound by BEAF and CP190, two insulator proteins. These independent lines of evidence suggest that the promoters might loop together and regulate each other's expression. These features are common to the other two promoter couples (*dnk* and *snRNP-U1-C*, *Cdc2rk* and *mRpL42*) indicating that this might be a rather common mechanism of gene regulation. We will test these three cases performing six independent deletions of the promoters (as discussed in "3.1 - Validation of complex DHS-eQTLs by *in vivo* CRISPR-Cas9 mutagenesis").

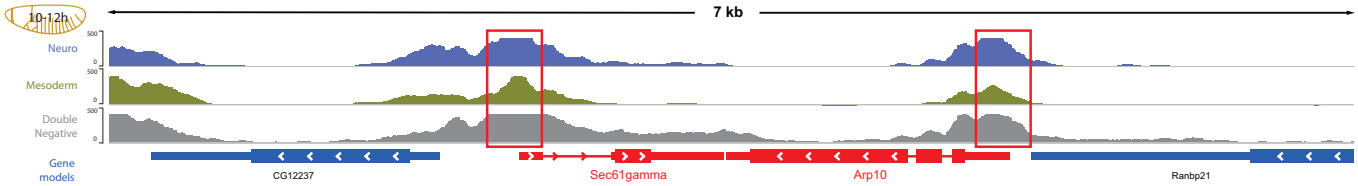


Figure 28 – Sec61gamma and Arp10 promoters are associated to both genes. The genes Sec61gamma and Arp10 are transcribed in convergent direction, but their polyadenylation sites do not overlap. Both promoter-proximal DHS (in red boxes) are associated to both genes, suggesting coregulation from promoter-proximal elements. The expression of the two genes is highly and positively correlated across DGRP lines.

2.6.3 - Promoter-proximal DHS with distal activity show weak evidence of enhancer behavior

Promoter-proximal DHS are a heterogeneous group of CRMs that encompasses core promoters, enhancers and insulators. We observe widespread distal activity from promoter-proximal DHS that cannot be explained by *trans* effects alone.

Another possible explanation is that promoter-proximal DHS with distal activity behave as enhancers. To assess this, I performed three tests to measure if promoter-proximal DHS with distal activity are in some ways different from those with local activity.

Firstly, recent work from the Furlong laboratory shows that the textbook definition of promoters and enhancers just represents two extremes of a continuum where most CRMs are placed. Mikhaylichenko *et al*.[78] showed that transcriptional Orientation Index (OI) is an indication of enhancer/promoter activity. Classically defined enhancers have bidirectional transcription while promoters are transcribed in one direction only. In their study, Mikhaylichenko *et al.* observe a continuum between these two extremes reflected both at the OI and the activity level. For this analysis I used the OI data from Mikhaylichenko *et al*., obtained genome-wide at 6-8 hpf, and I assigned an OI to each DHS in our study. Promoter-proximal DHS with strong distal activity (skipping at least 10 genes) showed more bidirectional transcription than promoter-proximal DHS with only local activity (Figure 29). The difference, despite being significant, is minor and it is lost when considering all promoter-proximal DHS with distal activity.

Secondly, STARR-Seq[79] is an *in vitro* technique that assesses in a direct and systematic way, the whole genome for enhancer function in cell culture based assays. STARR-Seq tests genomic fragments for enhancer function by placing them in a construct downstream of a core promoter. If the fragment has enhancer activity, it will activate the core promoter and self-transcribe. I downloaded the STARR-Seq data from Arnold *et al.*[28] which was conducted in S2 cells and compared differences in STARR-Seq signal between promoter-proximal DHS with local and distal activity. I could not find any difference between the two groups, both by using the continuous signal from the assay or counts of STARR-Seq peaks. Enhancers act in a tissue-specific way and the cellular context in which they are tested can bias the results. In particular, the STARR-Seq experiments were performed in S2 cells making it difficult to interpret the results in an embryonal context.

Thirdly, I tested if the promoter-proximal DHS with proximal or distal activity show differences in their Transcription Factors Binding Sites (TFBS) motifs. I used a *de novo* motif discovery approach (DREME[80]: using the promoter-proximal DHS with distal activity as tested sequences and the ones with local activity as background) as

well as a motif enrichment approach (AME[80]: with the same setup). Both analyses do not show any noticeable difference in motif composition between the two groups.

These results show that promoter-proximal DHS with distal activity have only weak enhancer signatures. Distal activity from promoter-proximal DHS does not globally come from elements that have enhancer-like behavior, although a portion of them might still be enhancers. One caveat is that DHS-eQTL described only a small proportion of all CRM interactions, meaning that some promoter-proximal DHS with local activity might have undetected distal activity as well, this would make our comparisons less powerful. Another explanation is that most promoter-proximal DHS are bound by insulator proteins, suggesting that distal activity might come from alterations of chromatin looping (despite we do not observe insulator binding enrichment in promoter-proximal DHS with distal activity).

In conclusion, distal activity from promoter-proximal DHS probably arises from a combination of factors: modifications of chromatin 3D structure, enhancer activity from promoter-proximal DHS and *trans* regulation from proximal genes.

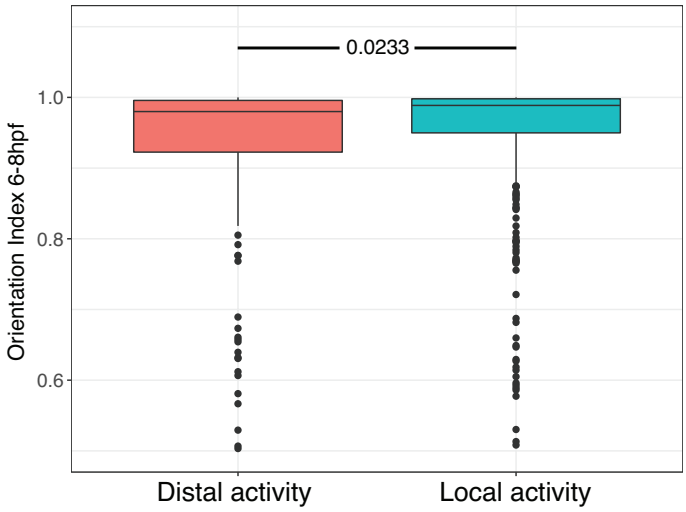


Figure 29 – Promoter-proximal DHS with strong distal activity have significantly higher bidirectional transcription than promoters with proximal only activity. Promoter-proximal DHS are divided in two groups: distal activity, if they are associated to a gene that is the tenth closest or more distal to the DHS; proximal only if they are only associated to one of the two closest genes. The y-axis reports the distribution of transcription Orientation Index from PRO-Cap of the promoter-proximal DHS at 6-8h.

3 - Perspectives

We are currently performing *in vivo* experiments to further validate the DHS-eQTL results. In particular, 12 DHS involved in complex interactions are being deleted with CRISPR-Cas9 mutagenesis and expression changes will be measured in the target genes. Furthermore, we will perform double *in situ* hybridization to observe the expression patterns of genes regulated by the same DHS.

3.1 - Validation of complex DHS-eQTLs by *in vivo* CRISPR-Cas9 mutagenesis

In order to validate the DHS-to-gene associations from the DHS-eQTL results, we are currently performing 12 DHS deletions using CRISPR technology in collaboration with Katharina Bender from the Furlong laboratory. The elements that will be deleted represent a collection of interesting results from the DHS-eQTLs and focus especially on CRM sharing. We plan to remove the DHS and measure gene expression of the genes linked by the DHS-eQTL, to further validate our discoveries. Among the 12 CRISPR deletions there are:

- One enhancer associated to a distal gene as discussed in “2.2.1 - Distal enhancer-eQTL”. We will perform one CRISPR deletion of the enhancer.
- Two cases of enhancers associated to two distal genes (further discussed in “2.5.1 - Example of enhancer sharing”). We will perform 2 CRISPR deletion of the enhancers.
- One case of genes whose promoters-proximal DHS are associated with each other but the genes are distal (further discussed in “2.6.1 - Different types of activity from promoter-proximal DHS”). We will perform a total of 2 independent CRISPR deletions of the promoters.
- Two promoter-proximal DHS associated to two distal genes (further discussed in “2.6.1 - Different types of activity from promoter-proximal DHS”). We will perform a CRISPR deletion of the promoter-proximal DHS.
- Three cases of convergently transcribed genes, in which both promoter-proximal DHS are associated to both genes (further discussed in “2.6.2 - Promoters of convergently transcribed genes are associated to each other”). A total of 6 independent CRISPR deletions of the promoters will be performed.

3.2 - Double in situ hybridization to validate CRM sharing

To further validate CRM sharing, Katharina Bender, a technician from the Furlong laboratory, is performing double *in situ* hybridizations for 6 pairs of genes associated to the same promoter-proximal DHS and 6 pairs of genes associated to the same enhancer. The rationale of this experiment is that genes that are regulated by the same elements should be co-expressed in similar tissues and time points. We also know the activity patterns of the DHS, so we can expect the two targets to be co-expressed in the same tissues where the DHS is active. For each of the 12 experiments, the setup is similar to Figure 24 but the expression patterns will be observed from the same embryos.

4 - Discussion

DHS-eQTLs is an effective method to associate CRMs to their target genes. By making use of a multivariate eQTL framework and focusing on specific genomic regions, it increases power and reduces the number of tests necessary in a traditional eQTL approach. I identified 2,973 DHS-eQTLs that describe functional CRM to gene associations: 2,005 are promoter-proximal DHS to gene associations and 962 are enhancer to gene associations. Both enhancers and promoters appeared to act over longer distances than suggested by other methods and I discussed how the field tends to underestimate long range interactions. In addition, the results showed hundreds of examples of promoter and enhancer sharing among genes, indicating that co-expression might come from sharing the same regulatory elements. The results also indicated that promoter-proximal DHS have widespread distal activity, though it is not clear through what mechanism. I described different modalities of regulation from enhancers and promoter-proximal DHS associated to multiple genes. In addition, I *in silico* validated the results by (1) overlap of tissue activity for enhancers and target genes, (2) enrichment of eQTL signal on DHS, and (3) Hi-C contacts enrichment, and experimentally by (4) qPCR and (5) *in situ* hybridization. Finally, I discussed 12 upcoming CRISPR deletions of DHS that will serve as further validations of complex DHS-eQTLs.

The DHS-eQTL method compromised on search space and unbiased testing to maximize the number of DHS to gene associations. It is not an alternative to traditional eQTL testing but it represents a functional strategy to address a specific issue. This method could be successfully applied to other model organisms, including mammals. The genome of mammals has different properties than the *Drosophila melanogaster* one and it will present different challenges including larger linkage disequilibrium blocks, heterozygosity and increased genome size. Larger linkage disequilibrium blocks might reduce power and single CRM resolution, while the total number of tests will be much higher than in *Drosophila*. On the other hand, reducing the search space to DHS will have a greater impact on mammalian systems (since a smaller proportion of mammalian genomes is functional compared to *Drosophila*). In conclusion, the DHS-eQTL method can associate regulatory elements to target genes and shed light on gene expression regulation.

5 - Methods

5.1 - Identification and quantification of polyadenylation sites and gene expression

Gene expression is measured through Tagged Sequencing of polyadenylated RNA. The technique was used to redefine polyadenylation (pA) sites in Cannavò *et al.*[53]. It can also be used to quantify gene expression by summing the expression of all pA of each gene with a correlation of 0.90 to RNA-Seq[53]. For the best performance, it is necessary to refine the existing annotation of pA sites. The work from Cannavò *et al.*[53] was performed by mapping reads to the *Drosophila* assembly BDGP5, so I repeated the analysis refining the methods used by Nils Koelling in Ewan Birney laboratory, using the more recent Dm6 (BDGP6) assembly.

5.1.1 Polyadenylation sites definition

In this work I used the Tagged sequencing data published in Cannavò *et al.* This is a collection of 3' Tagged sequencing across 80 DGRP (*Drosophila* Genetic Reference Panel) lines and 3 developmental time points. To identify the pA sites, I pulled all reads across lines and time points. Since 14 samples were re-sequenced there are a total of 254 samples.

3'-Tagged-Sequencing reads were trimmed to remove adapters to a uniform length of 44 bp using Trimmomatic[81] (version 0.33). pA sites were identified with two rounds of mapping. In the first round, I mapped the reads to the *Drosophila* reference genome (version BDGP6) using bwa mem[82] (version 0.6.1) and options "-n 5 -e 10 -q 20". I then excluded the reads that were mapping at this stage, as they were likely to be 5' of the termination site. I selected the unmapped reads that contained at least 5 terminal A nucleotides. In the second round, I remapped these unmapped reads in the same way as before after removing the terminal As. Reads that were mapped at this stage were defined as polyadenylation reads (pA reads). We produced a strand specific coverage of pA reads and defined polyadenylation sites (pA sites) as regions

with coverage ≥ 15.

We then applied a combination of filters to exclude low quality pA sites. Due to a partial failure of strand specificity in generating the sequencing libraries, highly expressed pA sites showed a corresponding antisense site. To remove these artifacts, I excluded pA sites that were perfectly included in an antisense site. In addition, to remove noisy and genotype specific pA sites, they are required to have non-zero expression (check "Quantification of pA sites and gene expression") in at least 50% of the DGRP lines in at least one of the time points.

Finally, the pA sites were expanded 200 bases upstream or until the next pA site, whichever is the shortest.

5.1.2 - Association of pA sites to genes

I downloaded the genome annotation from Flybase[83], version 6.13. I then selected all annotated polyadenylated transcripts (mRNAs and ncRNAs) and produced a series of strand-specific genomic intervals from this subset of Flybase annotation. In particular, I defined:

- mRNA pA sites: as regions 500 bases upstream and downstream from the last base of annotated mRNA.
- ncRNA pA sites: as regions 500 bases upstream and downstream from the last base of annotated ncRNA.
- exon: as annotated exons plus 500 bases downstream.
- introns: as annotated introns plus 500 bases downstream.
- distal mRNA pA sites: as regions 500 bases upstream and 10,000 bases downstream from the last base of annotated mRNA.

Each of these intervals was associated to their corresponding gene according to Flybase 6.13 annotation from which they were derived. The pA sites identified as in the previous sections were intersected with these intervals. If a single pA site multiple features it was assigned to the highest priority feature. The priority order corresponds to the list above, with mRNA pA sites having the highest priority and annotated distal mRNA pA sites having the lowest. In the rare cases when a single pA site was

assigned to multiple features of the same priority but belonging to different genes, the pA site was left without association.

Through this approach, a total of 24,169 (93%) pA sites were uniquely associated to a gene. 1,884 (7%) could not be associated to a single gene.

5.1.3 - Quantification of pA site usage and gene expression

To quantify pA site usage, I utilized the reads mapped in the first round of alignments described in the section in "5.1.1 Polyadenylation sites definition". These are reads that map to the genome without requiring to be trimmed of terminal A nucleotides.

Reads are then filtered for mappability issues. This step ensures comparability of gene expression across DGRP lines. Since all reads are mapped to the same reference, variants across lines might favor mapping of reads harboring the reference allele. These bias in quantification caused by mapping to a common reference are globally called mappability issues. To address this issue I used WASP[70] (downloaded from GitHub on 2 November 2015) pipeline for mappability bias filtering. All reads are tested for mapping as if they were harboring any combination of variants across all 80 DGRP lines. Firstly, I use find_intersecting_snps.py to identify all variants across the DGRP lines that intersect any mapped read. Reads that do not overlap variants are kept. Reads that overlap variants are changed to all haplotype combinations across the population and remapped as in "5.1.1 Polyadenylation sites definition". I then run filter_remapped_reads.py. The script checks that all version of each read mapped with a MQ>20, if so, the read is kept. I then merge all reads that passed WASP filtering and obtain WASP filtered bam.

I assigned the reads to the corresponding expanded pA sites using htseq-count[84] (version 0.7.2) with options "-m intersection-nonempty -s yes –q". I proceeded in a sample specific manner: I quantified pA sites usage and gene expression for our 254 samples separately. I reasoned that gene expression could be computed by summing all isoforms expression following Cannavò *et al.* In fact, each mRNA molecule can generate a single unique read from its pA tail. I then summed the counts of all pA sites assigned to each gene to compute gene expression. Finally, I

library-size normalized the counts by scaling the counts by the ninetieth percentile (equivalent to DESeq2 library size normalization, Pearson r = 0.98). I consider expressed those genes that have non-zero expression for half of the DGRP lines in at least one time point. There is a total of 9,054 expressed genes.

5.1.4 - Removal of hidden factors within the expression matrix

I used PEER[69] to correct for batch effects and hidden factors within the gene expression data. PEER discovers hidden factors within a matrix, fits them and can subtract them from the original matrix by outputting residuals. It is a useful tool to remove batch effects and increase power in eQTL discovery. First, I gaussianized the gene expression data by gene and time point (I substituted gene expression with the rank and fit the ranks on a gaussian distribution). I then ran PEER on gaussianized expression full matrix (75 lines, by 9,054 expressed genes and 3 time points). I used PEER with 10 hidden factors to obtain the residuals. Finally, I gaussianized the residuals.

5.2 - DNase Hypersensitivity Sites analysis

In this work I used the DNase Hypersensitivity dataset from Reddington *et al*. The data was obtained from staged *Drosophila melanogaster* embryos at tiling intervals from 2-4 hpf to 10-12 hpf. Embryos were fixed and FACS sorted for different markers: Mef2 was used as mesodermal marker; Wormiu at 4-6 hpf and elav from 6-8 hpf to 10-12 hpf were used as neuroectoderm markers; cells that were not sorted for either mesodermal or neuroectodermal marker belong to the Double Negative tissue. All time points were also complemented by whole embryo (not FACS sorted) data. 2-4 hpf time only has whole embryo data. 6-8 hpf time mesodermal tissue was further sorted for bin positive and bin negative cells. This leads to a total of 19 samples all in duplicates.

5.2.1 - DHS identification

In a DNase Hypersensitivity assay, an endonuclease (DNase) digests nucleosome free DNA, cleaving DNA regions that are not protected by the binding of TFs. By analyzing sequence reads, cleaved and uncleaved sites can be identified, thus revealing which sites bind a lot of transcription factors (e.g. an enhancer) and which do not. Sites that are protected from cleavage are called hypersensitive sites. Although typically referred to as open chromatin regions, DNase Hypersensitive Sites (DHS) are by definition TF-bound regions throughout the genome.

I reanalyzed the DHS data to move the annotation from the old genome assembly Dm3 (BDGP5) to the newer Dm6 (BDGP6) assembly that is used throughout this project. I followed the same analysis pipeline in Reddington *et al*. with the help of Charles Girardot from the Genome Biology Computational Support and Sascha Meiers from the Korbel laboratory at EMBL. We mapped the reads to Dm6 assembly using bwa mem keeping duplicates separate, sorted and removed duplicates and unmapped reads.

5.2.2 - Peak calling, IDR, summit merge across samples.

DHS peaks and summits were identified for each of the 19 biological conditions separately using the Irreproducible Discovery Rate (IDR) workflow described in Landt *et al*.[85] and implemented with the following details. Reads in the form of BAMPE files devoid of duplicates were used (two biological replicates were available for each condition) as workflow input. MACS2[86] version 2.1.1.20160309 was used as the peak caller with parameters “-g 1.2e8 -p 0.5 --keep-dup all --call-summits”; and a maximum of 100,000 peaks were passed to subsequent IDR analysis. The IDR analysis was executed using summits reported by MACS2 “slopped” by 30 bp (resulting in 60 pb regions centered on MACS2 reported summits). Merging and read shuffling operations were performed with SAMTools[87] 1.3.1 (merge & bamshuf). This procedure resulted in 19 DHS peak sets defined as the peaks passing an IDR threshold of 0.05 from the IDR analysis executed on the pooled pseudo-replicates. This peak set is often referred to as the “optimal” peak set.

We then merged these 19 DHS optimal sets into the final set of DHS peaks using the following custom procedure. The 1 bp summits from the optimal DHS sets were pooled together in a coordinate-sorted BAM file using BEDTools's bedtobam[88] 2.24.0. A smoothed coverage bedgraph (representing the summit density) was then generated using bamCoverage from deepTools[89] version 2.5.1 with parameters "--outFileFormat bedgraph --fragmentLength 1 --binSize 1 –smoothLength 80 --missingDataAsZero yes". Continuous stretches of bases with non-zero scores were extracted as the final DHS peaks. In each final DHS peak, the summit is defined as the base with the highest coverage. Visual inspection of the resulting DHS peaks revealed that larger DHS peaks were sometimes made of two or more sub-regions. DHS peaks larger than 300 bp were therefore post-processed and split into different DHS peaks provided that each resulting peak contains at least 2 DHS summits (from the 19 DHS optimal sets) and is located at least 80 bp apart from another DHS peak.

After identifying summits across conditions, I expanded them ±150 bases to define DHS peaks. If two summits were closer than 300 bases, the boundary between them was set at the midpoint, so that the DHS peaks would not overlap. In the text, I always refer to DHS as the here defined DHS peaks. The final set contains 63,157 DHS peaks.

5.2.3 - Tissue-specific DHS

I defined tissue-specific DHS in a time point specific manner. For each time point, I tested if a DHS is open only in one of the three tissues (Neuroectoderm, Mesoderm, Double Negative). In order to be called tissue-specific, a DHS is required to have a summit only in one tissue and to be differentially accessible in the same tissue. Summits are defined for each condition as described in the previous section. I calculated differential accessibility with DeSeq2[73]. For each peak and condition (time point and tissue), I obtained a coverage track from the mapped bam files. I then quantified the coverage for each condition with Rsubread[90] package and performed all pairwise comparisons between tissues at the same time point with DeSeq2. A DHS was considered to be differentially accessible if it had significantly higher accessibility in both comparisons.

5.2.4 - Intersection with variants

For the DHS-eQTL tests, I tested for association only variants that were overlapping DHS. Firstly, I moved the DGRP vcf (Freeze 2) coordinates to the Dm6 assembly using GATK[91]. During this process, 99.7% of the variants were successfully moved to Dm6. Then I subset the vcf to the 75 DGRP lines included in this work. Finally, I filtered the vcf using vcftools[92] with the following options: "--maf 0.05 --max-maf 0.95 --min-alleles 2 --max-alleles 2 --max-missing 0.2" to include only biallelic variants with a minor allele frequency greater than 5% and a maximum of 20% unknown genotypes. The filtered vcf was then intersected with the DHS using bedtools.

5.3 - Comparison of alternative methods to perform DHS-eQTL

In "2.1.4 - Testing three eQTL methods within the DHS-eQTL pipeline" I tested three eQTL association methods to get the best performance and reliability in discovering DHS to gene associations. The three methods used the same input gene expression matrix (gaussianized PEER residuals as discussed in the section above) and genotype matrix. LIMIX tests for association all variants within 50 kb of the gene body, while mtSet and PC-regression tested for association all DHS within 50 kb of the gene body. In the following tests, I only included the 75 DGRP lines with complete gene expression data across the three time points to avoid the imputation of phenotypes.

5.3.1 - Linear mixed model (LIMIX)

LIMIX[67] was the only univariate method tested here. It is based on a linear mixed model that explains a matrix (samples by conditions) of gene expression as a sum of a fixed effect (F) and random effects (U).

$$Y = \sum_{j=1}^{J} F_j + \sum_{i=1}^{I} U_i$$

The phenotype matrix is a 75 (DGRP lines) x 3 (time points) matrix. I used a simple genotype vector as fixed effect and a Kinship relatedness matrix as random effect. The relatedness matrix is obtained from neutrally evolving variants only. I defined neutrally evolving variants as those that only overlap short introns (>65 bp) following Grenier *et al*.[93] and I filtered them with the same criteria as in "5.2.4 - Intersection with variants". The neutrally evolving variants are then used to obtain a Kinship 75 X 75 (sample by sample) matrix that describes the similarity between each DGRP line couple.

LIMIX is used with the "*any effect test*" to test for any association (time-point specific or not) between the variant and the gene expression. I performed a total of 17,234,822 variant-to-gene association tests. For each gene, I selected the lowest p-value variant and corrected the p-value with Bonferroni correction (I multiplied the p-value by the number of variants tested for association with that gene). Then, I corrected the best p-values for each gene with Benjamini-Hochberg FDR approach[94]. I considered FDR corrected p-values lower than 0.05 as significant. Finally, variants significantly associated with a gene were mapped to the DHS they overlap using bedtools[88] to identify DHS-to-gene associations.

5.3.2 - PC-regression

I used Principal Component Regression (PC-regression) as a simple implementation of multiple regression. This method directly tested for association between DHS and the target gene. Since the number of variants overlapping the same DHS is variable and goes up to 90, I used Principal Components to reduce the number of independent variables in the regression and avoid overfitting. In particular, for each DHS I obtained a N x V (where N is the number DGRP lines in this study and V the number of variants overlapping the DHS) matrix of variants and perform principal components on this matrix. I then sorted the PCs by the amount of variance explained and defined an N x P matrix, where p is the number of principal components. I considered as many PC as necessary to explain 99% of the variance up to a maximum of 7, to avoid overfitting. The PCs are then used as independent variables in a multiple regression with a number of regressors from 1 to 7, depending on the DHS.

$$y = \beta_0 + \sum_{i=1}^{7} \beta_i PC_i + \epsilon$$

The phenotype matrix was the same used for LIMIX. In this case, I did not use any fixed effect. The genotype matrix corresponded to a N x V matrix where N is the number of samples and V the number of variants overlapping a DHS. mtSet is based on set test that does not incur into overfitting. For this reason, I could enter the full genotype matrix. The relatedness matrix was the same used for LIMIX. mtSet provided a p-value for each DHS-to-gene association and performed a total of 483,064 tests. All p-values were then corrected with Benjamini-Hochberg FDR correction and I considered FDR corrected p-values lower than 0.05 as significant.

5.3.3 - mtSet

mtSet[68] is a mixed model approach that allows for the association between multiple variants and multiple phenotypes while accounting for population structure. mtSet models a matrix of gene expression (in our case N x 3) as the sum of fixed effects (FB), the genotype matrix (U_r), the relatedness matrix (U_g) and residual noise (ψ).

$$Y = FB + U_r + U_g + \psi$$

The phenotype matrix is the same used for LIMIX. In this case, I did not use any fixed effect. The genotype matrix corresponds to a N x V matrix where N is the number of samples and V the number of variants overlapping a DHS. mtSet is based on set test that does not incur into overfitting. For this reason, I can input the full genotype matrix. The relatedness matrix is the same used for LIMIX. mtSet provides a p-value for each DHS-to-gene association and performs a total of 483,064 tests. All p-values are then corrected with Benjamini-Hochberg FDR correction and I consider as significant associations those with FDR corrected p-values lower than 0.05.

5.3.4 - Removal of associations in linkage disequilibrium

Despite the small LD blocks in *Drosophila*, neighboring DHS can still be in linkage. In order to avoid spurious associations caused by LD, I removed any DHS-eQTL whose DHS was in LD with another DHS-eQTL with strong association. In fact, both mtSet and PC-regression reported multiple associations for the same gene. To remove associations in LD I applied a stringent approach. I performed all variant to variant correlation and considered as in LD all correlation with a Pearson coefficient greater than 0.8. Any couple of DHS was considered to be in LD, if at least one variant on the first DHS was in LD with a variant on the second DHS. I then ranked all DHS-eQTL from the lowest to the highest p-value and I discarded any DHS-eQTL in LD with another having a lower p-value. Following this procedure, 28% of DHS-eQTL from mtSet and 31% from PC-regression were discarded.

5.4 - Validations of results

5.4.1 - Tiling-windows-eQTL

Tiling Windows were defined as consecutive windows of 300 bp. Each chromosome was split into 300 bp tiling windows and all windows within 50 kb of a gene's body were tested for association with that gene. The association pipeline was identical to "5.3.3 - mtSet".

The enrichment of eQTL signal on DHS was obtained by merging the uncorrected p-values from the 483,064 DHS-eQTL tests and the 2,892,787 TW-eQTLs tests. The raw p-values were corrected for multiple testing using Benjamini-Hochberg FDR and the enrichment was obtained by odds ratio.

Enrichment of TW-eQTLs on genomic features was obtained by overlapping TW-eQTLs with the BDGP6 genome annotation from FlyBase 6.13. The number of bases between the TW-eQTL and the feature was counted for every TW-eQTL and divided by the total number of TW-eQTL bases. Enrichments were obtained by using 10 random sets of TWs.

5.4.2 - DHS and genes tissue overlap

I downloaded the BDGP annotation[72] on the 24 of July 2017. The annotation reported a fixed term tissues of expression annotation by stage of development. I focused on stages from 13 to 16 that overlap the 10-12 hpf time point. All BDGP specific terms were mapped to more general terms, that were in turn assigned to the 3 tissues from the DHS FACS sorting to have matching terms between datasets. In particular, Table 4 showed the correspondence between general BDGP terms and the 3 FACS sorted tissues. I ignored terms mapped to "none" annotation. For each gene, I took the union of all annotations it was mapped to. I excluded genes that are expressed in "ambiguous" tissues (these tissues cannot clearly be assigned to only one of the three FACS sorted tissues). Finally, I defined as tissue-specific, those genes that are expressed in only one of the three FACS sorted tissues.

I identified tissue-specific enhancers as described in "5.2.3 - Tissue-specific DHS" and considered only enhancer eQTLs where both the gene and the DHS are tissue-specific. I excluded promoters from this analysis because the majority of them are ubiquitously open and in close proximity to the target gene. I then divided the tissue-specific enhancer-eQTL into two categories:

- coherent enhancer-eQTLs were associations between a tissue-specific enhancer and a tissue-specific gene active in the same tissue.
- incoherent enhancer-eQTLs were associations between a tissue-specific enhancer and a tissue-specific gene active in different tissues.

The enrichment of coherent enhancer-eQTLs was assessed by performing 10,000 random enhancers to gene associations. In particular, I selected 2,973 random DHS to gene associations among all those tested in the DHS-eQTL. I then proceeded as for the enhancer-eQTLs and obtained the proportion of coherent versus incoherent enhancer-eQTLs. Finally, I obtained an empirical p-value by comparing the coherent versus incoherent ratios from the permutations to that observed from the enhancer-eQTLs.

5.4.3 - CRISPR deletions

The entire CRISPR deletion protocol was performed together with Katharina Bender and Songjie Feng. The deletions were performed in the vasa-Cas9 *Drosophila*

melanogaster line[95] (Bloomington ID: 51324). The guide RNAs were designed using Target Finder website (http://targetfinder.flycrispr.neuro.brown.edu) using the fully sequenced vasaCas9 personalized genome to identify potential off-target sites. We always deleted the entire DHS involved in the DHS-eQTL and designed the guides so that the entire peak signal in all tissues would be deleted. The deletions sizes range from 450 to 2,000 bases. The guide RNAs are 20 nucleotides long. In order to minimize the chance of off-target cuts, we always preferred guide RNAs identified with “maximum” stringency setting. In case this was not possible, we used guide RNAs with “high” stringency and 0 predicted off-targets. The guide RNAs were cloned in the bacterial plasmid #1823 pBs-U6-gRNA-BbsI.

In order to have an efficient deletion of the target sites after CRISPR cuts the genome, we designed homology arms flanking the deleted region. The homology arms were PCR amplified from the vasa-Cas9 genomic DNA using oligos with a melting temperature of approximately 60°C and a GC content ranging from 40 to 60%. The homology arms are approximately 1 kb long. The homology arms were cloned into the bacterial vector pHD-dsRed-attP #1473. To insert the homology arms into the vector, we used AarI restriction enzyme for the left and SapI for the right homology arms. If the restriction site was not present, restriction cloning was used. Finally, if there were internal cut sites, we either switched to InFusion cloning or changed the order of homology arms insertion.

After bacterial amplification of the homology arms, they were injected, together with the guide RNAs, into approximately 100 embryos of the vasa-Cas9 line. The injection mixes contained 150ng/µL donor plasmid for recombination and 75ng/µL of each gRNA in 20µL injection buffer.

The vasa-Cas9 line expresses the Cas9 protein in the ovary giving rise to chimeric progeny. The first cross of F0 will be crossed with yellow-white flies to amplify the transgenic flies. The offspring will be crossed with the respective balancer lines, depending on what chromosome harbors the deletion. Finally, the third cross will be performed between siblings to obtain a homozygous stock if it is viable.

5.5 - Comparison with external databases

The enhancer to gene associations from Kvon *et al*.[59] were obtained from Supplementary Table 4 at: https://doi.org/10.1038/nature13395. The coordinates were moved to BDGP6 genome assembly. The Tile to gene distance was measured from the center of the Tile to the major gene TSS. The “closest gene” associations were obtained by assigning all enhancers involved in an enhancer-eQTL to their closest gene. Enhancer to gene distances were measured from the center of the enhancer to the major TSS of the gene both for the DHS-eQTL associations and the “closest gene” associations.

III - Gene expression variation among *Drosophila melanogaster* lines from five continents

1 - Introduction

1.1 - Geographic isolation causes population structure

Individuals from the same species often cannot intermix as they are restricted by geographic limitations. Geographic isolation leads to the formation of populations, that behave as semi-independent groups of individuals. In fact, geographic isolation of populations causes an accumulation of differences that lead to population structure. If individuals from two populations remain isolated and therefore cannot mate, their variant pool will accumulate differences through two mechanisms. Firstly, variants that appear after the separation of the two populations will not be shared among them. Secondly, the allele frequency of the variants they shared before the separation will change due to drift or selection. Therefore, geographic isolation causes uneven allele frequency among populations, which can be assessed by statistical tests such as the fixation index (F_{ST}). The accumulation of differences among populations may be under selective pressure, which could ultimately lead to phenotypic differences between populations. Measuring differences between populations is crucial to retrieve the migration history of a species and understand the mechanism through which species adapt to new environments.

1.2 - The Global Diversity Lines are a panel of *Drosophila* lines from five continents

Drosophila melanogaster has long been used as a model organism to study population genetics. The majority of population genetics studies on *D. melanogaster* have focused on inbred lines from a defined geographic location[63]. This setup increases power in association studies by removing population structure and allows

for the application of models that assume random mating. However, a few studies have focused on capturing differences between *D. melanogaster* populations. In particular, the laboratory of Andrew Clark has generated a panel of *D. melanogaster* inbred lines from five continents called the Global Diversity Lines[93] (GDL) to further investigate population structure and infer *D. melanogaster* migration history. The GDL include a panel of 85 populations from 5 defined geographic locations: Netherlands, Tasmania, Ithaca, Beijing and Zimbabwe. The GDL were obtained by collecting multiple individuals from the same locations, which were allowed to inter-breed to create a stable genetic pool. Subsequently, a few individuals from the pool were selected and went through ten cycles of inbreeding, giving rise to each line.

To infer population structure among the GDL, Grenier *et al.*[93] performed a Principal Component analysis on neutrally evolving variants. Variants evolving neutrally arise and drift at a constant speed because they are not under positive or negative selective pressure. Figure 30 shows the results of the PCA. The first PC separates the Zimbabwe lines from the non-African populations, indicating that it is the most different from the others. The second PC separates the Beijing population. European, Australian and North American populations cluster together, reflecting their more recent separation from a common population.

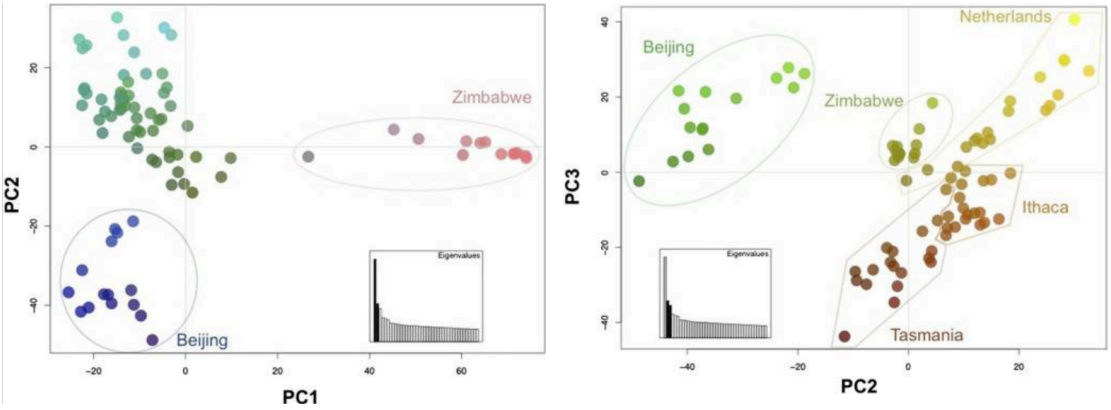


Figure 30 – Population structure among the Global Diversity Lines (from Grenier *et al.*[93]). The plots show the first three PCs from the neutrally evolving variants of the GDL.

1.3 - Migration history of *Drosophila melanogaster*

Drosophila melanogaster is a human commensal and its migrations history is closely related to that of our species[96]. This fly species originated in the sub-Saharan

region[97] and colonized all continents in recent times. It is estimated that *D. melanogaster* left Africa for the first time after the end of the last ice age and quickly spread in Europe and Asia[98]. Following human migration, *D. melanogaster* recently colonized the Americas and Oceania[99]. Similarly to human, the African populations represent the most ancient and diverse population from the genetic variation standpoint. Moreover, they are the most subdivided populations from the other continents[93]. Different populations are not completely isolated from each other. Arguello *et al*.[99] estimated the extent of admixture between *D. melanogaster* populations finding a high correlation with commercial routes. In particular, the European, American and Australian populations have inter-mixed to a high extent, while the African and Asian populations have remained more isolated.

The population genetics studies presented so far are based on genetic variation. As discussed in chapter "IV - Impact of natural sequence variation on *Drosophila melanogaster* chromatin accessibility", the majority of genetic variants do not have any obvious impact on phenotype. In this chapter, I will present a newly generated gene expression database of the Global Diversity Lines. The goal is to assess the impact of genetic variation on gene expression during embryonic development within this diverse genetic populations.

1.4 - Overview of the project

In this project, we aim to measure and analyze gene expression patterns across the Global Diversity Lines during embryonic development. We performed RNA-Seq on 83 GDL whole embryos staged at 10-12 hpf. The goal of the project is to gain insight into the transcriptome differences between *D. melanogaster* inbred lines from five continents. In addition, we used the newly generated gene expression dataset to call expression QTLs.

The project has been developed in collaboration with the laboratory of Andrew Clark at Cornell University, USA. They performed the staged embryo collections and shipped the samples. The RNA extraction and library preparation were performed by Lucia Ciglar, a technician from the Furlong laboratory. I have performed the analysis of the RNA-Seq data with the collaboration of Federica Mantica, a visiting scientist in the Furlong laboratory.

2 - Results

We performed RNA-Seq on whole embryos from the GDL lines staged at 10-12 hours post fertilization and performed multiple quality control steps to ensure high data quality. We identified genes that were differentially expressed among populations and characterized them. In addition, gene expression-QTL and exon expression-QTL were called to dissect the regulatory landscape of *D. melanogaster*.

2.1 - A panel of *Drosophila* gene expression from 5 continents

The Global Diversity Lines constitute a panel of fully genotyped *D. melanogaster* lines from five continents[93]. To complement the genotype information and gain insight into developmental patterns across continents, we performed RNA-Seq on whole embryos staged at 10-12 hpf. The RNA-Seq dataset includes (Figure 31):

- 18 lines from Tasmania (Oceania)
- 14 lines from Ithaca (North America)
- 18 lines from the Netherlands (Europe)
- 16 lines from Beijing (East Asia)
- 17 lines from Zimbabwe (Africa)

Figure 31 – Schematic representation of the Global Diversity Lines RNA-Seq dataset. We performed RNA-Seq on 83 Global Diversity Lines coming from five continents. The figure indicates the geographic location of origin of the lines and the number of sequenced lines belonging to that region.

In addition, to gain insight into the dynamical expression changes during development, we performed RNA-Seq on 6 lines (two lines from Ithaca and one line of the other populations) staged at 2-4 hpf and 5 lines (one per population) staged at 6-8 hpf. Finally, two lines were sequenced in duplicates for a total of 96 RNA-Seq samples. In the following pages, I will discuss the quality control pipeline and the filtering steps that were adopted to remove problematic samples. In particular, we controlled for:

- RNA-Seq mapping quality by analyzing multiple metrics from Samtools[87] and Picard[100];
- RNA-Seq protocol reproducibility by correlation of technical duplicates;
- Sample staging by comparing each sample with the corresponding modENCODE RNA-Seq data[101];
- Potential batch effects caused by sample transportation, RNA extraction, mRNA isolation and sequencing.

2.1.1 - RNA-Seq mapping and quality control

The raw reads files were mapped to the BDGP6 assembly using STAR[102]. Gene expression was quantified with RSEM[103] (See "5 - Methods" for a complete discussion of the mapping pipeline). We assessed the mapping quality using multiple metrics from Samtools[87] and Picard[100] and we summarized and compared them using MultiQC[104]. All samples showed high quality in all metrics except for one sample from Tasmania (T36B) that showed evidence of RNA degradation. The sample was removed from the subsequent analysis.

2.1.2 - Batch effect quality control and correction

Samples from different populations were processed and sequenced in a randomized manner to better dissect potential sources of batch effects. In particular, we controlled for four batch effects by Principal Component Analysis:

- Transportation box (the embryos staged at 10-12 hpf were shipped in 10 separate boxes)
- RNA extraction batch (the RNA extraction was performed in 10 separate days)
- mRNA isolation batch (the mRNA isolation was performed in 10 separate days, randomizing the batches for the RNA extraction)
- Sequencing set (the samples were sequenced on 4 independent Illumina lanes)

Supplementary Figure 8 displays the RNA-Seq data separated by the first three PCs and colored by the four potential batch effects. The samples showed no structure when divided by mRNA isolation date and sequencing sets. On the other hand, the first principal component clearly separated the samples shipped in box 9 and whose RNA was extracted on the 18/01/2017 (Box 9 was processed on day 18/01/2017, so it was impossible to disentangle the two effects). We concluded that the RNA samples shipped in box 9 were corrupted and removed them from the following analysis. Further analysis showed that these same samples have high rRNA percentage despite mRNA isolation by polyA enrichment. Following this observation, we excluded all samples that had a rRNA content higher than 35% of total RNA. This step removed all visible batch effects from our data.

2.1.3 - RNA-Seq provides a reliable measure of gene expression

Two RNA-Seq samples were sequenced in duplicates to assess technical variability. The RNA extractions and sample preparations were performed independently from the same embryo collections ensuring a control of the entire RNA-Seq protocol. Figure 32 shows the correlation between ZS29 and ZW190 technical duplicates. The correlations exceed 0.98 showing high reproducibility of the RNA-Seq measurements.

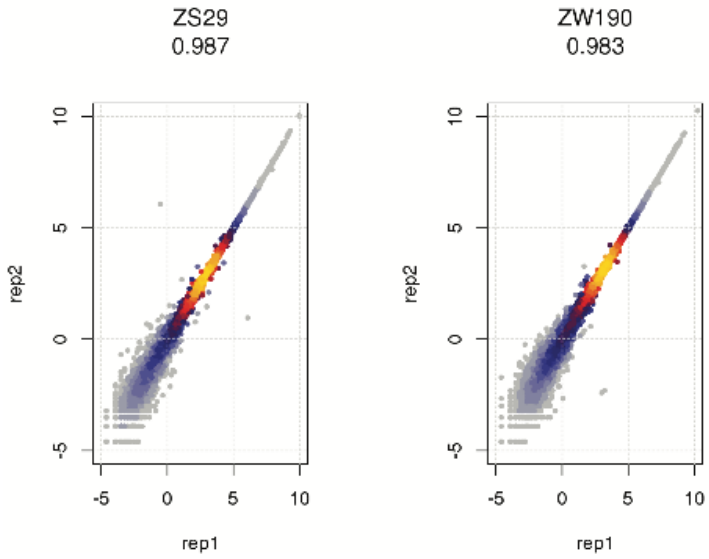


Figure 32 – Biological replicates show high correlation. The scatterplot shows the correlation between two biological replicates for the Zimbabwe lines ZS29 and ZW190. The Spearman correlation coefficient is shown under the line's names.

2.1.4 - Staging of samples by comparison with modENCODE gene expression

In order to confirm that the staging of the embryo collections was at the expected time point during development, we performed an *in silico* staging analysis, as we've performed previously[53]. In particular, we correlated the gene expression measurements from our GDL samples with a two-hour time-course done in reference strain (modENCODE[105]). The modENCODE gene expression data were generated at time points ranging from 2-4 hpf to 22-24 hpf. We downloaded and processed the modENCODE RNA-Seq samples in the same way as the GDL samples (see "5 - Methods") and performed a Spearman correlation between gene expression values. Figure 33 shows the correlation of GDL samples staged at 10-12 hpf with modENCODE samples from 2-4 hpf to 12-14 hpf. The GDL samples were highly correlated with the corresponding modENCODE samples and showed a slight shift towards the later time window (12-14 hpf). This is consistent with the results from Cannavò *et al.*[53] (Supplementary Figure 1) confirming that the samples are tightly staged. All GDL samples displayed the same trend except for one sample that was removed from subsequent analysis (the unstaged sample is also visible in Figure 34 as clustering with samples at 6-8 hpf). We repeated the same analysis for the GDL samples staged at 2-4 hpf and 6-8 hpf obtaining the highest correlations at the

expected time point for all samples. These results confirm that the collections have been performed ensuring precise staging of the embryos.

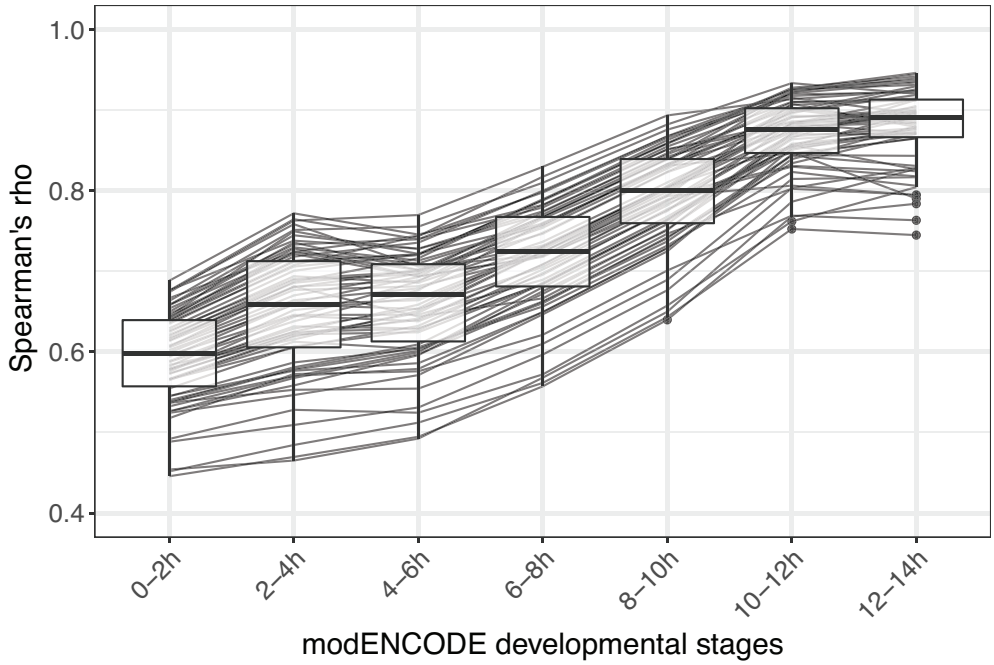


Figure 33 – Staging of samples at 10-12 hpf. The plot shows the correlation between GDL whole embryo samples staged at 10-12 hpf and modENCODE whole embryo samples staged from 0-2 hpf to 12-14 hpf.

2.1.5 - Samples separate by stage

As a final quality control, we assessed if all GDL samples separated by stage. We included all sample from 2-4 hpf to 10-12 hpf and performed a PCA. Figure 34 shows the results of the PCA. The samples are clearly separated by developmental stage by the first PC. The first PC explains a high proportion (38.2%) of gene expression variation. This is a comforting result that confirms the high quality of the final dataset.

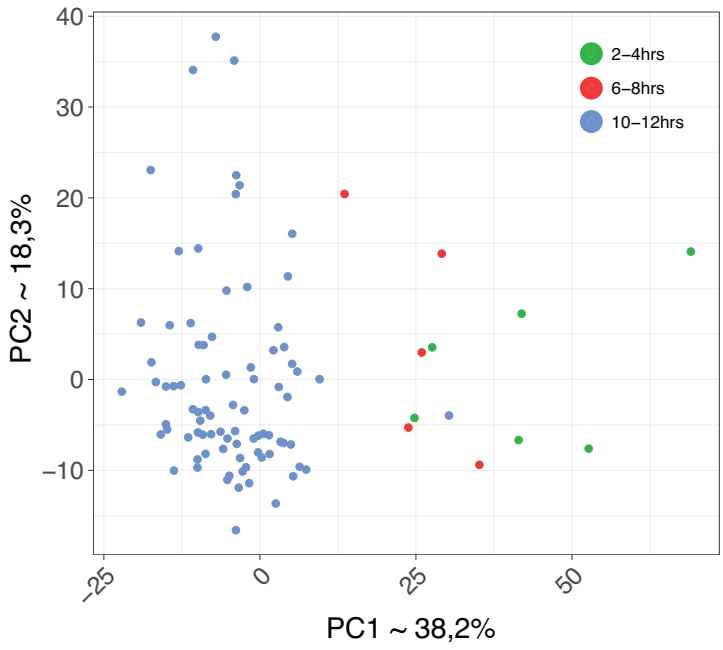


Figure 34 – PCA separates samples by stage. Principal Component Analysis of all sequenced samples. The samples are colored by stage. The first PC explains 38.2% of gene expression variation and clearly separates samples by stage.

2.2 - Transcriptome differences among continents

We performed a differential expression analysis across populations to identify patterns of expression that are population specific. Through differentially expressed genes, we quantify the differences across transcriptomes. Below, I discuss the strategy to identify and classify differentially expressed genes.

2.2.1 - Differential expression is more accentuated at the transcript level

The GDL populations have accumulated genetic differences during thousands of reproductive cycles. The larger dissimilarities at the genetic level can be seen between the African lines and the other populations. In addition, the African and Asian lines have experienced little admixture, meaning that their genetic pool is more isolated[99]. On the other hand, the European, American and Australian lines were separated recently and are less differentiated. We investigated if the population to

population distances visible at the genetic level is also present at the transcriptomic level.

Firstly, we performed a PCA on gene expression levels and were surprised to obtain no clear separation between populations. This indicates that there is no global structure at the transcriptome level. Performing the PCA on specific gene groups that are more likely to be under selective pressure (such as metabolic or olfactory genes) also gave no clear population structure.

Secondly, we implemented a differential expression analysis both at the gene and transcript level. Gene and transcript expression was quantified using RSEM[103] and differential expression was computed with DESeq2[73]. We compared each population against all the others taken together to identify population specific patterns of expression. Figure 35 a and c respectively show the number of differentially expressed genes and transcripts per population. In both cases, the Zimbabwe population has the most differentially expressed elements compared to the others, consistent with the findings at the genetic level. Surprisingly, the Beijing line, despite being more recent, has a similar number of differentially expressed genes. This could be explained by the relative isolation that this population has experienced[99]. On the other hand, the most recently separated populations (Ithaca and Tasmania) only have a handful of differentially expressed genes and transcripts. Another interesting observation is that differences are more accentuated at the transcript level compared to the whole gene level.

A technical caveat of the “1 vs all” setup is that it might underestimate differential expression for the populations of European descendance (Netherlands, Ithaca and Tasmania). In fact, these three populations are similar to each other and by having two lines with similar expression patterns in the background group, DESeq2 is less likely to find differentially expressed elements. To address this issue, we performed a “1 vs 1” differential expression analysis between all population pairs (Figure 35b,d). When comparing populations of European descendance with the African or Asian populations directly, we identify a comparable number of differentially expressed gene except for the Tasmania lines, that show more differentially expressed genes than in the global analysis. The strongest effect is seen in the “Zimbabwe vs Tasmania” comparison. In addition, the pairwise comparisons between populations of European descendance show that the “1 vs all” test underestimates the number of

differentially expressed genes in these populations. In conclusion, the “1 vs 1” setup shows a higher extent of differential expression for the Tasmania lines than estimated “1 vs all” setup, but it confirms that Ithaca and Netherlands populations have less transcriptional differences from the other lines.

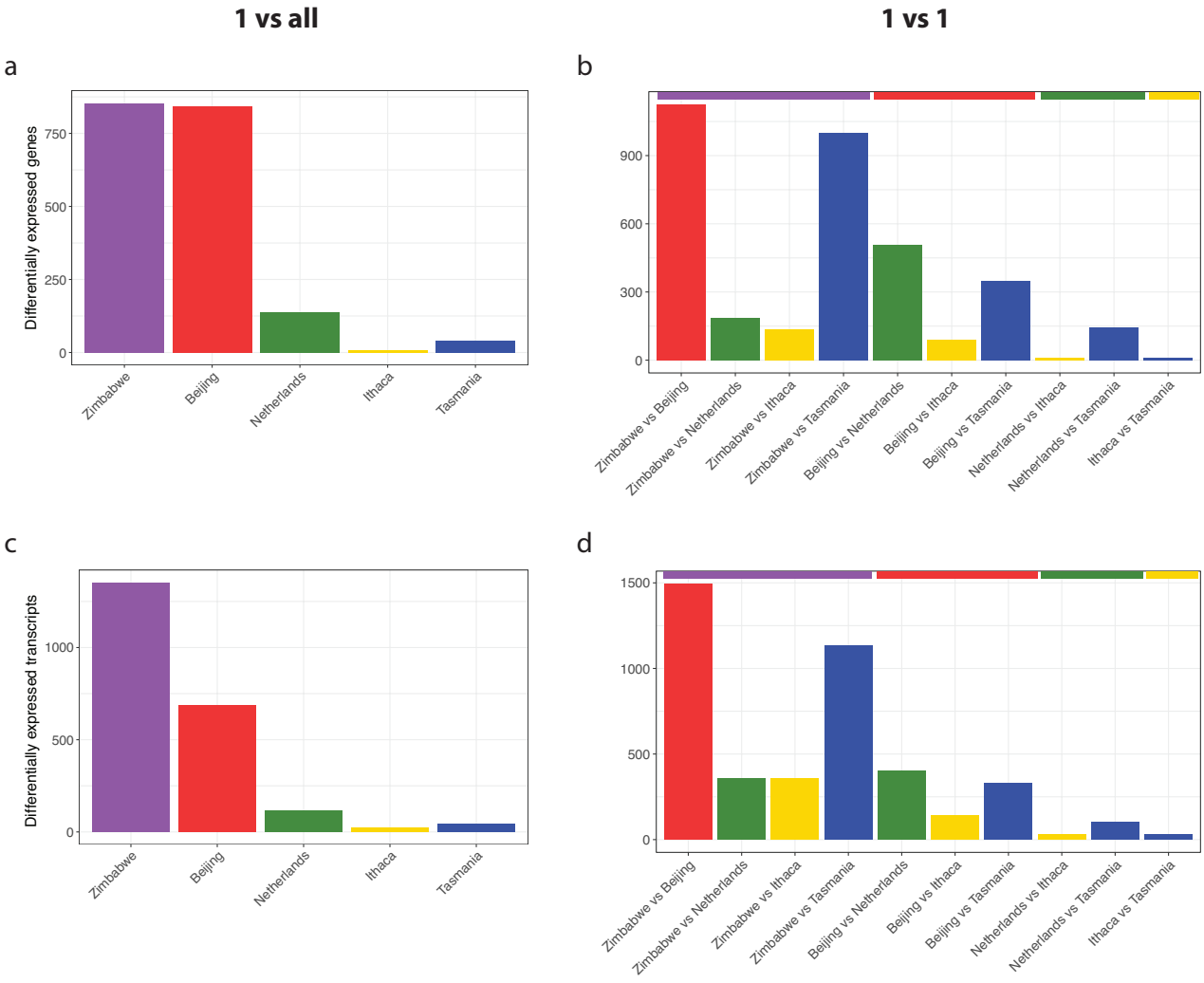


Figure 35 – Number of differentially expressed genes and transcripts. (a) The bar plot shows the number of differentially expressed genes for each population within the GDL. (b) The bar plot shows the number of differentially expressed genes for each one to one contrast between populations. (c) The bar plot shows the number of differentially expressed transcripts for each population within the GDL. (d) The bar plot shows the number of differentially expressed transcripts for each one to one contrast between populations. Differential gene and transcript expression were obtained by comparing one line with all the others.

2.2.2 - Genes enrichment among differentially expressed genes

To characterize the differentially expressed genes among populations we performed a Gene Ontology enrichment analysis. In all the population except for the Netherland and the Tasmania, we did not identify significantly enriched groups of genes among the differentially expressed ones. This suggests that a large portion of differential expression is caused by drift and is not necessarily functional, though further analyses are required to corroborate this statement.

On the other hand, the differentially expressed genes in the Netherland lines are more than 20-fold enriched in genes linked to cuticle development. This is observed both in the global and pairwise tests. The cuticle is secreted from embryonic stage 16[106] (corresponding to 13 hpf to 16 hpf) until later larval stages and it confers the embryo a protection from water and external stresses. The cuticle production genes are overexpressed in the Netherland population at 10-12 hpf indicating an anticipated or more abundant secretion of cuticular proteins in the European lines. The overexpression of cuticle genes might confer resistance to environmental conditions specific to the Netherlands, such as the colder weather.

In addition, a Gene Ontology analysis of the differentially expressed genes in the Tasmania lines shows that they are enriched for translation, mitochondrial functions and energy metabolism in general. These results are seen only when contrasting the Tasmania line against the African and Asian populations (Figure 35b,d) and indicate metabolic differences between the Australian lines and the others.

Taken together these results support the hypothesis that a large proportion of differential gene expression might be caused by drift. Nevertheless, the Netherlands and Tasmania lines show evidence of adaptation by having differentially expressed genes associated respectively to cuticle secretion and energy metabolism.

2.3 - Identification of gene and exon eQTLs

The GDL expression dataset represents a wealth of information that can be used for association studies. In contrast to our previous eQTL study on the DGRP lines, which

used 3'-Tagged-Sequencing as a measure for gene expression[53], here we can use the full length RNA-seq data described above. This has the advantage of reducing the impact of mapping biases, due to the extensive RNA-seq coverage over the body of the gene. However, the major challenge of this dataset is its inherent high degree of population structure. To address this, we used LIMIX, a linear mixed model that accounts for population structure in the association test to identify eQTLs. As input we used the filtered set of RNA-seq data for 65 lines that were fully genotyped (Grenier *et al.*[93]) and passed all quality control steps, and then corrected for hidden batch effects with PEER[69] and for mapping biases using WASP[70] (see "5 - Methods"). In total, we quantified the expression for 11,382 genes and 63,607 exons.

We separately tested for eQTLs that are linked to gene expression (gene-eQTLs) and exon coverage (exon-eQTLs). In total, we discovered 903 gene-eQTLs and 2,021 exon-eQTLs. The gene-eQTLs and exon-eQTLs show a similar distribution of distances from the target gene TSS (Figure 36a). The distribution is consistent with previous eQTL studies in *D. melanogaster*[53] and in vertebrates[107] (see also "II - Genetic variation as a tool to associate *cis* Regulatory Modules with their target genes"). Exon-eQTLs are more concentrated on the target gene body than gene-eQTLs (Figure 36b). Although expected, this is very reassuring and suggests a direct regulation of exon usage at splice junction sites. On the other hand, gene-eQTLs are found more often at the TSS and in the target gene surroundings.

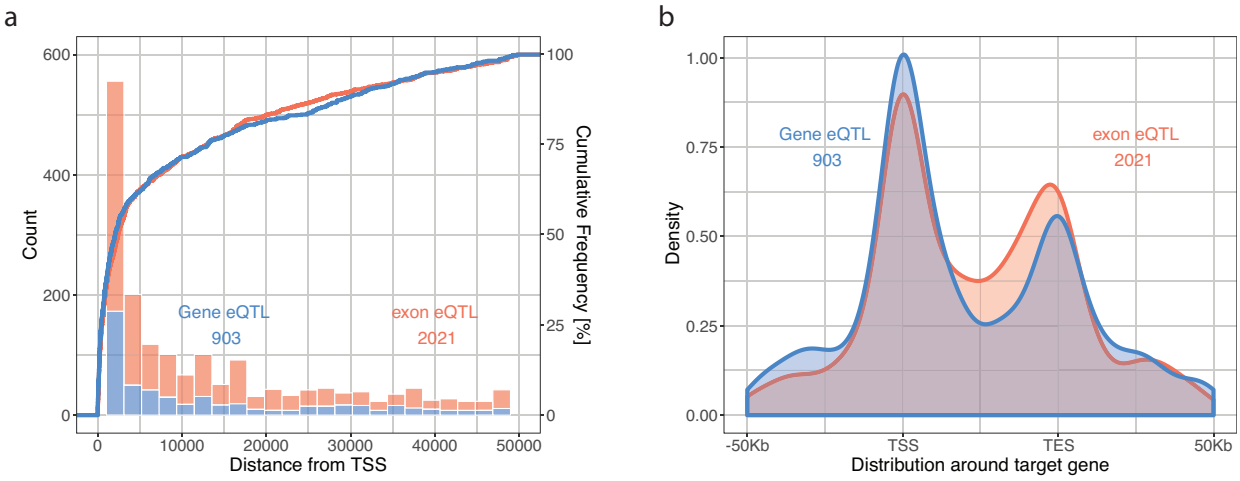


Figure 36 – Distribution of gene-eQTL and exon-QTL distance from TSS and around target gene. (a) Distribution of eQTL distance from target gene TSS. gene-eQTL are shown in blue, exon-eQTL in red. (b) Distribution of eQTLs around target gene. TSS: Transcription Start Site; TES: Transcription End Site

2.3.1 - gene-eQTL

64.7% of genes with a gene-eQTL also have an exon-eQTL. If the total level of gene expression changes, this should be reflected at the exon level as well. On the other hand, the exon-eQTL require a 6-fold higher number of statistical tests and exons are covered by fewer reads than genes, making the exon-eQTL analysis less powerful. As seen in other studies[53], eQTL genes are weakly enriched for metabolic processes. Figure 37 shows an example of gene-eQTL where the coverage is uniformly higher for all exons in the minor allele flies.

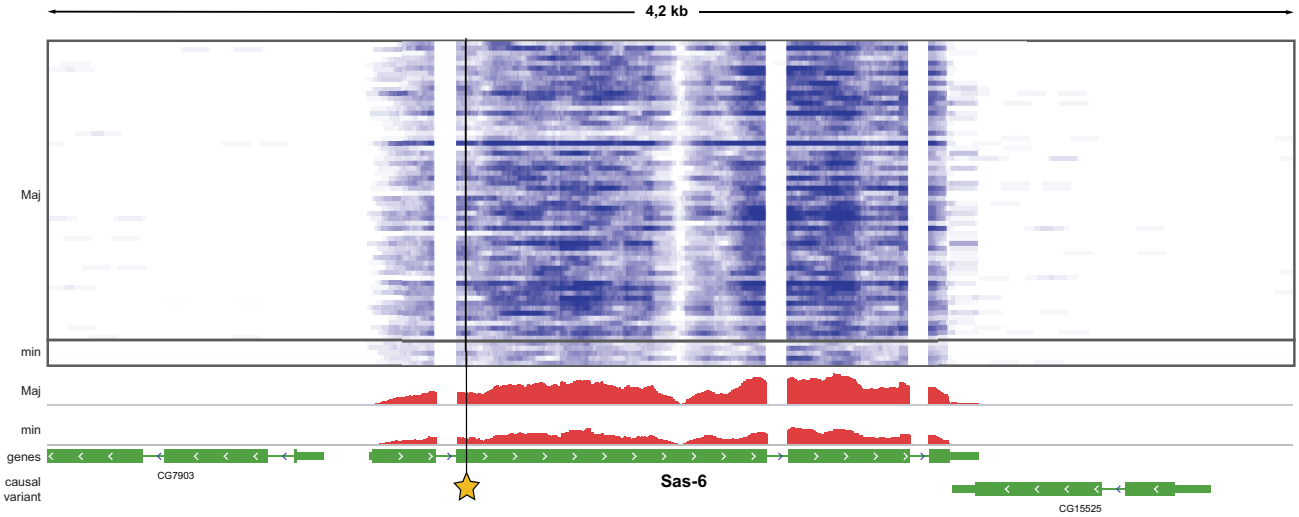


Figure 37 – A gene-eQTL in the second intron of Sas-6 increases the gene expression. The heatmap shows in blue the forward RNA-Seq coverage for all GDL lines divided by Major and minor allele. The average coverage for the Major and minor alleles is shown in red. The star indicates the position of the causal variant.

2.3.2 - exon-eQTL

45.6% of genes with an exon-eQTL also have a gene-eQTL. The tests are performed independently from each other making it likely for exon-eQTLs to be discovered for genes with a gene-eQTL. The remaining 54.4% of exon-eQTLs point to specific exon usage changes that do not globally affect gene expression. This fraction of exon-eQTLs has an effect on alternative splicing alone because it is not associated with changes in total gene expression large enough to lead to a gene-eQTL. Genes involved in exon-eQTLs are weakly enriched for RNA binding and catalytic activities.

An example of exon-eQTL is shown in Figure 38, where the 10th exon of the gene FANCI has 2.3 times higher coverage in the minor allele compared to the major. Coverage differences supporting differential inclusion are visible exclusively at the level of this exon. The exon coverage change is associated with a variant in a neighboring exon that might cause differences in splicing preferences between the two alleles.

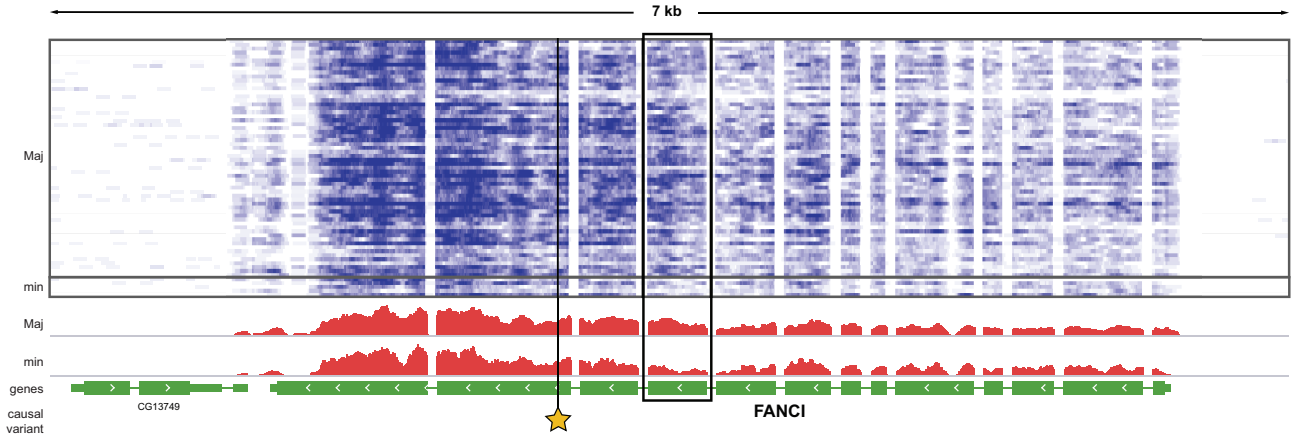


Figure 38 – An exon-eQTL in the 12th exon of the gene *FANCI* decreases the usage of its 10th exon. The heatmap shows in blue the reverse RNA-Seq coverage for all GDL lines divided by Major and minor allele. The average coverage for the Major and minor alleles is shown in red. The star indicates the position of the causal variant. The black box indicates the exon with differential coverage between Major and minor alleles.

3 - Perspectives

3.1 - RNA sequencing of whole embryo samples at 2-4 hpf and 6-8 hpf

The gene expression dataset presented in this section is staged during cell differentiation (10-12 hpf). To complement the dataset with additional time points during development, we plan to obtain samples stage at 2-4 hpf (during multipotent stage) and at 6-8 hpf (during cell specification) for the same GDL lines. This will increase our time resolution and allow us to make gene expression comparisons during development. In particular, since development is a canalized process, we aim to assess if gene expression diversity increases during development. To better achieve this goal and control for batch effects, we already sequenced one line per population at 2-4 hpf and 6-8 hpf. This will allow us to disentangle clustering by hours post fertilization from sequencing batch.

3.2 - Identification of selective pressure on gene expression

In addition, in collaboration with other groups, we plan to expand on the evolutionary insights that can be acquired by analyzing this dataset. The differential expression analysis confirms the population separation observed at the genetic level. On the other hand, except for the cuticle gene overexpression in the Netherlands population and the energy metabolism genes in the Tasmania population, the genetic signatures of differentially expressed genes are difficult to interpret. We aim to identify genes under selective pressure to find patterns of adaptation among the five populations.

4 - Discussion

In this chapter, I have discussed the generation of an RNA-Seq dataset of *D. melanogaster* staged whole embryos from the Global Diversity Lines. The dataset included a few problematic samples that were carefully identified and excluded from the analysis. The other samples displayed high quality scores on multiple metrics and are at the expected stage of embryogenesis (comparison to modENCODE data).

In contrast to the DNA variants, Principal Component Analysis of gene expression does not show any visible population structure. The same is true when sub-setting for genes in functional categories, suggesting that the majority of gene expression variation is not under selection. Differential expression analysis confirmed that the African population is the most separated from the other populations, with the difference being stronger at the transcript compared to the gene level. In addition, the differential expression analysis revealed an overexpression of cuticle related genes in the Netherlands population at 10-12 hpf and differential expression of genes linked to energy metabolism in the Tasmania population. Finally, we identified 903 gene-eQTLs and 2,021 exon-eQTLs. By using a linear mixed model, the high degree of population structure was controlled for. Variants with an effect on exon usage tended to be spread along the gene body.

The gene expression dataset presented here represents a wealth of information for the *Drosophila* population genetics community interested in migration and evolution of the *melanogaster* species.

5 - Methods

5.1 - RNA-Sequencing and mapping

5.1.1 - Generation of staged, high quality RNA-Seq libraries

The samples were collected by the laboratory of Andrew Clark at Cornell University. The embryos were fixed in formaldehyde, frozen and shipped with dry ice to the Furlong laboratory. RNA was extracted and prepared for sequencing using NEBNext ultra directional RNA library prep kit for Illumina sequencing (NEB). The NEBNext protocol performs mRNA isolation by poly-A enrichment. The RNA was directionally reverse transcribed to cDNA. The cDNA was amplified with 11 PCR cycles. The samples were multiplexed and sequenced in 4 batches on Illumina NextSeq 500 HI. Reads were 75 bp long and paired-end. We achieved a median of 10 million unique mapping reads per sample.

5.1.2 - Mapping and gene expression quantification

We built the gene expression quantification pipeline around RSEM[103]. RSEM uses STAR[102] to map directly on the transcriptome. It can also model the likelihood of a read coming from different isoforms if the read does not clearly belong to any. We merged the fastq of technical duplicates to increase coverage. Reads were mapped using STAR (version 2.5.2a) with default parameters and the quantification of both genes and isoforms was performed using RSEM (rsem-calculate-expression). The transcript assembly was based on BDGP6 genome assembly and Flybase 6.13 genome annotation. RSEM provides Transcript Per Million (TPM) as a measurement of transcript and gene expression.

5.2 - RNA-Seq quality control

5.2.1 - FastQC and Picard quality control statistics

We collected a wealth of quality control statistics from unmapped (fastq files) and mapped (bam files) reads. In particular, we used fastqc to collect metrics about the sequencing. In addition, we run Picard[100] MarkDuplicates, CollectRnaSeqMetrics, CollectAlignmentSummaryMetrics and CollectMultipleMetrics on the mapped files from STAR. The statistics were collected with MultiQC[104] for easier visualization.

5.2.2 - Principal component analysis to control for batch effects

We performed principal component analysis on TPM gene expression measurement from RSEM. We included samples staged at 10-12 hpf and colored them by: transportation box, RNA extraction day, mRNA isolation day and sequence set. After removing all samples with high rRNA content, mRNA degradation and incorrect staging, we were left with 72 high quality RNA-Seq samples.

5.2.3 - Staging comparison with modENCODE data

We downloaded the RNA-Seq data from modENCODE (Celniker *et al.*[101]) as a reference for staging. Gene expression was quantified using RSEM as in "5.1.2 - Mapping and gene expression quantification". We performed Spearman correlation between gene expression for each GDL sample with each modENCODE sample to check if samples had the expected gene expression signature associated with their time point. To this end, we ascertained that the GDL samples had the best correlation scores with the corresponding modENCODE sample.

5.3 - Differential gene expression between continents and gene enrichments

5.3.1 - Retrieving population structure from gene expression

We performed a Principal Component analysis on gene expression values to separate the lines by population. No visible structure was present. We then performed the same analysis on gene groups that could potentially be under selective pressure. We performed a PCA on genes annotated as belonging to the following Gene Ontologies: “Metabolic Process”, “Heat-Related”, “Olfact-related”,

5.3.2 - Differential gene and transcript expression

We quantified gene expression using RSEM (as described above) for the 72 lines staged at 10-12 hpf that passed all quality control steps. The “expected counts” from RSEM were rounded to the next unit and used as a measure for gene expression. Differentially expressed genes and transcripts were identified with DESeq2[73]. We compared each population against all others (e.g. Ithaca vs Netherlands, Tasmania, Zimbabwe and Beijing) performing 5 tests in total. In this way population specific patterns of gene expression could be identified. On the other hand, Ithaca, Netherland and Tasmania populations are very similar to each other: to assess the impact of this setup, we performed all 10 pairwise comparisons.

5.3.3 - GO enrichment of differentially expressed genes

We performed Gene Ontology enrichments using R package GOstats[108]. The enrichments were performed separately for the 3 major GO categories: “Molecular Function”, “Cellular Component” and “Biological Process”. The enrichments were calculated using Fisher exact tests and the p-values were corrected using the Benjamini-Yekutieli procedure.

5.4 - QTL call

We built an eQTL pipeline to discover gene expression-QTL (gene-eQTL) and exon usage-QTL (exon-eQTL). We mapped reads to the reference genome and removed controlled mappability biases using WASP[70]. The major challenge for association studies represented by the Global Diversity Lines is the high degree of population structure. Relationships between samples create spurious associations between both variants and gene expression. To this end, we used a linear mixed model approach that includes the population structure and excludes associations that arise from the population structure alone. We used LIMIX[67], a python implementation of linear mixed models.

5.4.1 - Mapping and WASP filtering

We mapped the reads using STAR[102] (version 2.5.2a) with options: "--alignIntronMax 100000 --outFilterIntronMotifs RemoveNoncanonical --outFilterType BySJout --outSAMunmapped Within --readFilesCommand zcat". We filtered reads with mapping biases using WASP[70]. We then removed duplicates reads with Picard tool. This pipeline didn't allow us to use RSEM for quantification since WASP requires two rounds of mapping to the genome.

5.4.2 - Gene and exon expression

To quantify gene and transcript expression we used htseq-count[84] both on genes and exons from Flybase 6.13 genome annotation. Gene and exon expression were library size normalized using DESeq2[73]. The following pipeline was the same for gene and exon expression measurement. We filtered expressed genes and exons as those with expression greater than 0 in at least half of the GDL. Gene expression values were gaussianized to increase power in QTL discovery. To remove batch effects and hidden factors in the data and increase power we used PEER[69] with 10 hidden factors. PEER is a Bayesian method that discovers hidden factors in the gene expression data matrix and outputs residuals that are free of those effects. Finally, we gaussianized the residuals and used them as inputs in the eQTL call.

5.4.3 - Variants filtering and population structure

We used the variant calls from Grenier *et al*.[93] and subsetted it to the lines we had gene expression data for. We used GATK[109] LiftOverVcf to move the vcf coordinates from BDGP5 to BDGP6 assembly. We removed variants with minor allele frequency smaller than 5% and more than 20% missing values. Multiallelic loci were excluded. For each gene we defined a *cis* window as ±50 kb from both gene's ends and tested the variants within this window for association. To better identify population structure we replicated the work in Grenier *et al*.[93]. Starting from the filtered vcf, as described above, we annotated variants based on the genomic features they overlap using SnpEff[110]. We defined neutrally evolving variants as those overlapping introns shorter than 65 bp. The data recapitulated the population structure seen in Grenier *et al*.[93]. The neutrally evolving variants were used to obtain the Kinship matrix to control for population structure in LIMIX.

5.4.4 - gene-eQTL and exon-eQTL call with LIMIX

To perform the eQTL call we used LIMIX[67]. LIMIX implements the following linear mixed model:

$$Y = \sum_{j=1}^{J} F_j + \sum_{i=1}^{I} U_i$$

Where Y is the gene expression vector, F is a NxN kinship (Fixed Effect) matrix representing sample by sample relationship and U is the genotype vector (Unknown Effect). The kinship matrix is obtained by calculating the sample by sample similarity from neutrally evolving variants. We computed empirical p-values by permuting 10,000 times the Y vector and obtaining a background p-value distribution. The test p-value was then ranked among the background p-value. The empirical p-value was obtained with the formula:

$$p_{empirical} = \frac{R_{p-value}}{N_{permutations} + 1}$$

Where $R_{p-value}$ is the rank of the test p-value when compared the permuted p-values.

We considered only the lowest raw p-value association for each gene (one gene-eQTL and one exon-eQTL per gene) and we correct the empirical p-values with Benjamini-Hochberg FDR using a 0.1 cutoff. If multiple variants were in linkage disequilibrium by showing an uncorrected p-value within one order of magnitude from the lowest, we consider them as belonging to an eQTL cloud.

IV - Impact of natural sequence variation on *Drosophila melanogaster* chromatin accessibility

1 - Introduction

1.1 - Non-coding variants are a major source of phenotypic variation

Individuals from the same species share genes and follow the same developmental plan, but have different phenotypes. Phenotypic diversity is caused by the interaction of the environment (especially during development) and genetic variation. While the environmental cues can be measured and controlled to a certain extent, the genetic variants constitute an innate and hidden pool of variation.

Although genetic variants that impact coding sequences can be interpreted based on the changes that they cause on protein sequence, non-coding variants provide a greater challenge. GWAS[111,112] and QTL[53,113] studies consistently reveal that the majority of functional variants are located in the non-coding genome. This phenomenon is conserved across organisms, from humans to *Drosophila*.

The mechanisms through which non-coding variants lead to variation in the phenotype are still poorly understood. The main hypothesis is that non-coding variants influence the phenotype by modifying gene regulation. An example is the human rs11708067 variant, located in an enhancer of the gene *ADCY5*. The variant has a common G allele and a rare A allele: the A allele disrupts the enhancer function and causes lower expression of the gene *ADCY5*, that in turn leads to higher diabetes risk[114]. In addition to altering enhancer function, variants can influence gene expression by causing changes in chromatin topology[115] and epigenetic marks[52]. In general, functional non-coding variants modify gene regulation by interfering with *cis* Regulatory Module (CRM) function.

1.2 - Specific activation of CRMs drives tissue development

The non-coding genome harbors a plethora of CRMs that perform a variety of functions related to gene expression regulation (see “I - Introduction”). CRMs function by recruiting Transcription Factors to the DNA, this leads to higher chromatin accessibility that can be measured with many techniques, including DNAse hypersensitivity assay[24]. Recent single-cell studies[116] show that the majority of CRMs are active only in a subset of tissues. In fact, tissue-specific features emerge because of precise gene regulation[117] and, as a result, every cell type expresses a unique set of genes. Spatial gene expression specificity is achieved by the precise activation of CRMs. Consistent with the observation that a high proportion of CRMs are activated in a tissue-specific manner, the effect of many functional variants differs among tissues[118], and is, therefore, context dependent.

1.3 - Methods to predict the effect of variants on regulatory regions

Each individual harbors millions of variants, but the vast majority of them has little to no effect on gene expression[42]. Since the regulatory code is still poorly understood, it is difficult to predict which variants have an effect on gene expression based solely on DNA sequence. In recent years, many machine learning methods have been developed to summarize the features of regulatory modules and have been successfully applied to estimate the impact of variants on chromatin accessibility. These methods scan the DNA sequences of CRMs and look for enrichment of features that distinguish them from the rest of the genome. This knowledge can then be used to prioritize variants for their predicted effect on CRM function.

Machine learning approaches look either for k-mer or Position Weight Matrix (PWM) enrichment within regulatory sequences. For example, gkm-SVM[119] is a support vector machine method based on k-mer enrichment while PRIME[120] is a random forest method based on PWMs enrichment. Methods based on k-mers outperform those based on PWMs because they describe Transcription Factors Binding Sites (TFBS) in a more flexible way[119]. Recently, Neural network approaches have gained popularity because they achieve better performances than other methods on very large datasets. Basset[121] is a neural network approach based on PWMs enrichments, which was successfully applied to ENCODE human data showing better performance

than k-mer based approaches. The main disadvantage of neural networks is that they require large training sets and their results are challenging to interpret.

1.4 - Overview of the project

In this project, I apply LS-GKM[122], an enhanced version of gkm-SVM, to prioritize genetic variants for their effect on *Drosophila melanogaster* chromatin accessibility during embryogenesis. I applied a gapped k-mer support vector machine method (LS-GKM) developed in Michael Beer's laboratory on a set of tissue-specific DNase Hypersensitive Sites (DHS). The small genome size of *Drosophila* proves to be a challenge for machine learning methods, but LS-GKM shows good performance on small training sets. The DHS dataset has been introduced in "II - Genetic variation as a tool to associate *cis* Regulatory Modules with their target genes". The SVM model is then used to score variants for their tissue-specific effect on chromatin accessibility. The genetic variants analyzed here come from two different populations:

- The *Drosophila* Genetic Reference Panel (DGRP[63]) is a panel of more than 200 *Drosophila melanogaster* lines coming from a unique geographic location. The population has 6,131,648 mapped variants.
- The Global Diversity Lines (GDL[93]) are a group of 80 *Drosophila melanogaster* lines from 5 continents (see also chapter "III - Gene expression variation among *Drosophila melanogaster* lines from five continents"). They capture genetic variation caused by geographic isolation and adaptation to different environments. The population has 6,752,029 mapped variants.

The variant sets for the two populations have 2,828,011 (28%) variants in common.

The goal of this project is to gain insights into the tissue-specific effects of variants and to provide a resource to the *Drosophila* population genetics community for prioritizing causal variants.

2 - Results

2.1 - A machine learning approach to uncover tissue-specific features of chromatin accessibility

In this work we aim to prioritize genetic variants for their predicted effect on chromatin accessibility to produce a resource database for the *Drosophila* community. Genetic variation can affect phenotype by altering gene regulation. In particular, genetic variation can alter the function of CRMs by modifying Transcription Factor Binding Sites (TFBS). Here we apply a gapped k-mer support vector machine method developed by Beer, Gandhi, Lee *et al*.[119,123], to discover tissue-specific open chromatin features. These features are then used to score genetic variants for their predicted effect on chromatin accessibility. The entire project was developed in collaboration with Federica Mantica, a visiting scientist in the Furlong laboratory.

2.1.1 - Identification of tissue-specific DHS

The LS-GKM model learns features that distinguish a positive from a negative set of sequences. In this project, we aimed at distinguishing open from closed chromatin in order to score variants for their ability to increase or decrease accessibility. To dissect the regulatory landscape of the developing *Drosophila* embryo, we used the DNase hypersensitivity dataset generated in the Furlong laboratory by James Reddington and David Garfield (described in “II - Genetic variation as a tool to associate *cis* Regulatory Modules with their target genes”). The dataset offers both time and tissue resolution for a total of 19 samples, all in duplicates. Since the majority of non-coding variants has an effect only on a subset of tissues, we trained LS-GKM on tissue-specific DHS. This allowed us to score variants for their tissue-specific effect on accessibility. We obtained a collection of DHS that were exclusively open in one of the three FACS sorted tissues in at least one of the 5 time points. DHS were further separated into promoter-proximal DHS (closer than 500 bp to a

known TSS) and promoter-distal, putative enhancers (distant more than 500 bp to a known TSS). In total we identify:

- 1,466 promoter-proximal DHS and 9,658 (putative) enhancers that are accessible exclusively in the neuroectoderm tissue.
- 436 promoter-proximal DHS and 2,937 (putative) enhancers that are accessible exclusively in the mesoderm tissue.
- 1,105 promoter-proximal DHS and 4,811 (putative) enhancers that are accessible exclusively in the Double Negative (non-neuro, non-meso) tissue.

Figure 39 shows an example of DNase hypersensitivity coverage tracks and depicts tissue-specific DHS.

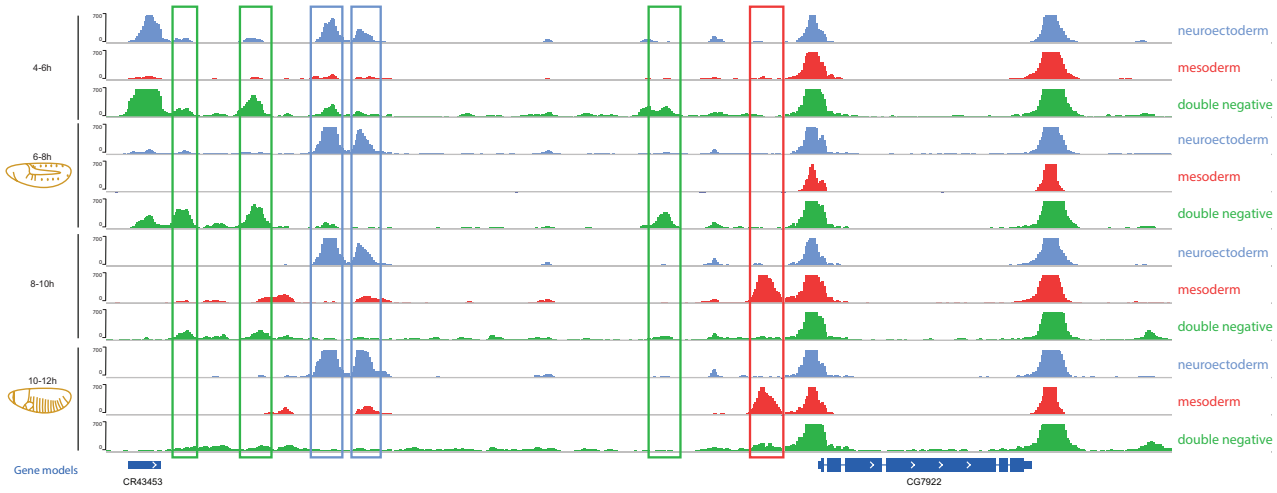


Figure 39 – Identification of tissue-specific DHS. The figure shows DNA Hypersensitivity tracks in 3 tissues across 4 time points (from 4-6 hpf to 10-12 hpf). Colors correspond to different FACS sorted tissues. Blue: neuroectoderm; Red: mesoderm; Green: Double Negative. The boxes identify tissue-specific DHS (DHS open exclusively in one tissue at least at one time points). The box color indicates in what tissue the DHS is exclusively accessible. Blue: neuroectoderm, red: mesoderm, green: double negative tissue.

2.1.2 - A machine learning approach to distinguish open from closed chromatin

In this project, we used large-scale gkm-SVM (LS-GKM[122]) following the pipeline shown in Figure 40. LS-GKM was trained using our set of tissue-specific DHS and a mono- and di-nucleotide matched background. LS-GKM fits an SVM model to separate the positive and negative sets, which can then be used to score all possible 10mers and learn the vocabulary of the regulatory sequences, in our case, tissue-specific motifs (as discussed in “2.3 - Enrichment of tissue-specific Transcription

Factors motifs"). Finally, the regulatory vocabulary can be used to score DNA sequences for their likelihood to belonging to DHS or inaccessible chromatin. By scoring sequences that incorporate the different alleles of a variant, and computing the difference between the two, LS-GKM outputs a deltaSVM: a score that summarizes the impact of the non-reference allele on chromatin accessibility.

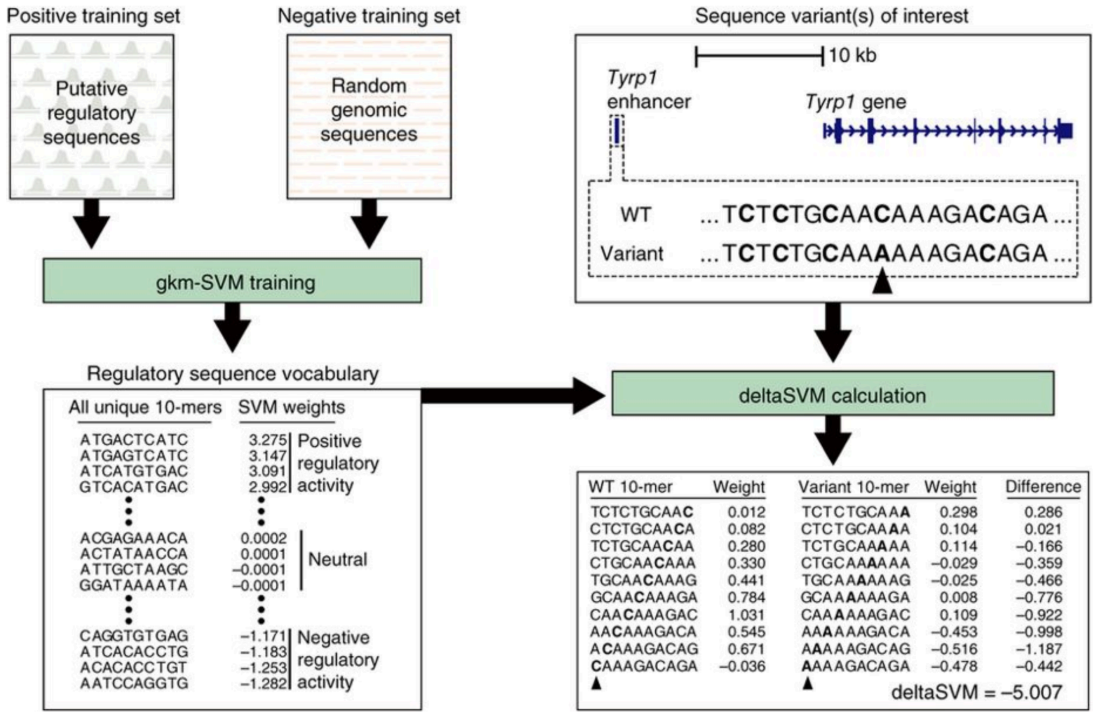


Figure 40 – gkm-SVM pipeline to train the gapped k-mer SVM model and score variants for their predicted effect on open chromatin (from Lee *et al*., *Nature Genetics*, 2015). The figure shows the pipeline implemented in gkm-SVM. Firstly, the model is trained on a positive and a negative training set of sequences. The gapped k-mer weights are used to score all possible 10-mers. These scores are then used to assess the impact of variants on the local genomic sequence. Finally, the model provides a deltaSVM for each variant. If a variant has positive deltaSVM, it means that the sequence with the variant is more similar to the positive training set, while a negative deltaSVM means that the sequence with the variant is more similar to the negative training set.

LS-GKM was trained on our six sets of tissue-specific DHS. For each positive set, we selected a background that matched the same nucleotide and di-nucleotide composition. The training was repeated five times with different backgrounds to increase the stability of the results. The background is composed of intergenic sequences selected to match the nucleotide and di-nucleotide composition of the positive set. Figure 41 shows the ROC curves from ten cross-validations of the six trainings. Only the ROC from the first training are shown but the results are very

similar across the five trainings. The Area Under the Curve (AUC) are higher for enhancers than for promoter-proximal DHS, probably because the enhancer sets include more sequences. After training the models on tissue-specific DHS, all variants can be scored for their tissue-specific effect on chromatin accessibility. The 10-mers scores were averaged across the five trainings and the deltaSVM were computed from the averaged 10-mers scores.

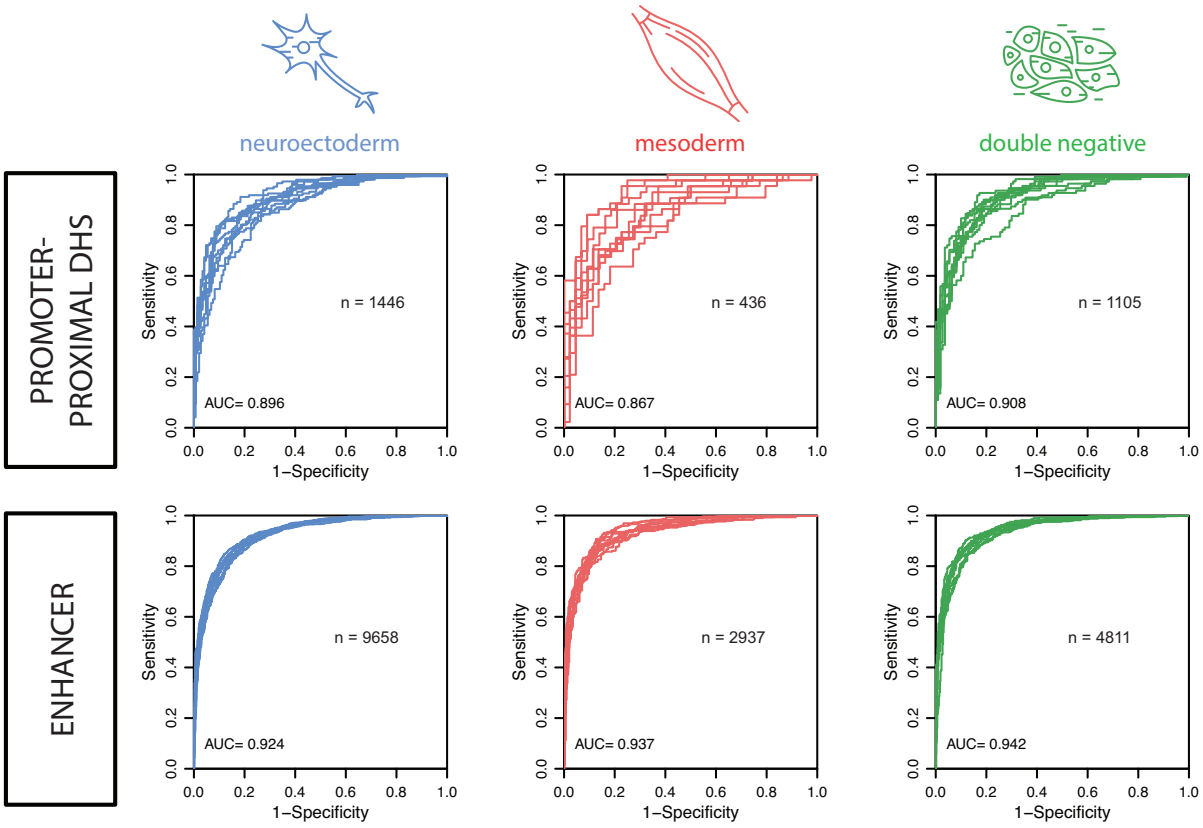


Figure 41 – ROC curves from 10 cross-validations of LS-GKM training. The figure shows the ROC curves from LS-GKM training on tissue-specific DHS divided in promoter-proximal DHS and enhancer. Only the results from the first training out of five performed are shown here. n: number of positive (and negative) sequences in the training set. AUC: Area Under the Curve.

2.2 - Prediction of chromatin accessibility QTLs

The genome of *Drosophila melanogaster* is over 25 times smaller than the one of human and includes far fewer CRMs. This results in a smaller amount of sequences that can be used for training compared to mammalian systems. To corroborate the

good training performance shown by the AUC (Figure 41), we tested the predictive power of the deltaSVM scores. Therefore, we assessed how well the deltaSVM could predict the direction of the effect of chromatin accessibility QTLs (caQTLs). caQTLs describe the associations between a genetic variant and changes in chromatin accessibility. Specifically, we used LS-GKM to predict the caQTLs in the eye-antennal imaginal discs of adult *Drosophila* published by Jacobs *et al*.[124]. We identified eye-antennal imaginal discs specific DHS and trained LS-GKM on three sets of tissue-specific DHS: promoter-proximal DHS (1,863 sequences), enhancers (12,819) and all DHS (14,682). The deltaSVM scores could predict the direction of the caQTLs in the enhancer and all DHS, but not in the promoter-proximal DHS (Figure 42). The predictive power of LS-GKM is lower than for human caQTLs[123]. In addition, by analyzing the results of Lee *et al*.[123] we observed that LS-GKM predicts caQTL effect on promoter-proximal DHS as well as on enhancers in human. These observations could be caused by the smaller amount of sequences used in training because of the smaller *Drosophila* genome size.

We repeated the analysis by training LS-GKM on all (tissue and non-tissue-specific) ATAC-Seq peaks (9,049 promoter-proximal, 21,725 enhancer and 30,774 all DHS) obtaining very comparable results to the eye-antennal imaginal disc specific set. This confirmed that it is possible to train LS-GKM on a subset of sequences and use the model to score variants not included in the training set.

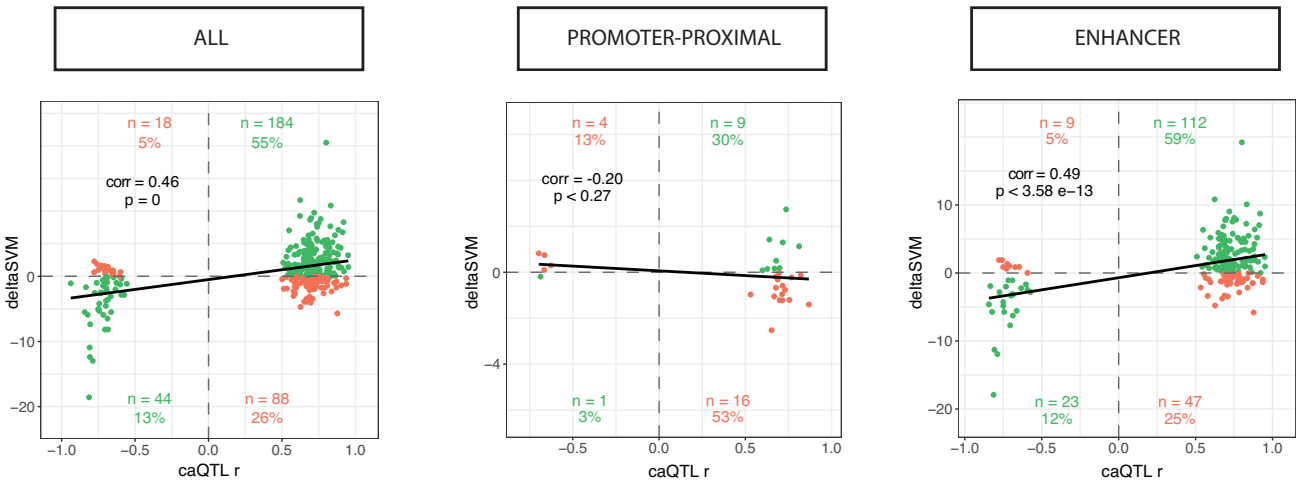


Figure 42 – Prediction of caQTL direction of effect. The plot shows the correlation between caQTL Pearson r and deltaSVM for the causal variant. The ATAC-Seq peaks from Jacobs *et al*.[124] were divided in promoter-proximal and distal. Only the ATAC-Seq peaks that did not overlap any embryonic DHS or coding region were used in the training set. The "ALL" category includes both promoter-proximal and enhancer peaks.

2.3 - Enrichment of tissue-specific Transcription Factors motifs

LS-GKM learns the regulatory motifs enriched in CRMs by scoring gapped k-mers for their likelihood of belonging to CRMs or not. After training the model, we can retrieve SVM weights for all ungapped 10-mers. This set is composed of short sequences that distinguish CRMs from background DNA. LS-GKM was trained on 6 sets of sequences belonging to: neuroectoderm, mesoderm and Double Negative tissue divided in promoter-proximal and enhancer DHS. To retrieve the regulatory features of each set of sequence, we looked for PWMs enrichments within the scored 10-mers.

Figure 43 shows enrichment scores for many regulatory TFs across the training sets. Erect wings (Ewg) is a transcription factor involved in synaptic growth[125] whose motif was strongly enriched within the top k-mers of the neural tissue. The mesoderm shows strong enrichment of many known regulators of muscle development. Mef2 is an essential regulator of muscle differentiation, being required for all muscle types[126] and its motif is strongly enriched in both promoter-proximal DHS and enhancers exclusive to the mesodermal tissue. In the same way, motifs for the transcription factors Biniou (necessary for visceral mesoderm development) and Tinman (required for dorsal somatic muscles and heart formation) are both strongly enriched in the mesoderm. Finally, the Double Negative tissue represents a pool of tissues dominated by ectoderm and endoderm. The motif for GATAe (a TF required for endoderm development[127]) is enriched in the Double Negative tissue together with grainy head, a regulator of epithelial development[128].

The majority of PWMs had similar enrichment patterns in promoter-proximal DHS and enhancers, with some exceptions. For example, Ewg enrichment was stronger in promoter-proximal DHS, suggesting a direct regulation of transcription at the promoter level. On the other hand, Hairy and Tinman motifs showed a stronger enrichment in enhancers, suggesting that they mainly function by binding to promoter distal elements. In fact, Tinman regulates heart formation by specifically binding to enhancer regions[129]. In addition, almost all TFs are specifically enriched in one tissue. One exception is Sloppy paired 1 (Slp1) that is weakly enriched in all promoter-proximal tissues. Slp1 is expressed across tissues during development and it mainly represses gene expression by binding at the promoter of target genes as a

cofactor of Hox proteins[130]. Taken together these results speak for the high quality of our DHS dataset and prove that LS-GKM can unbiasedly learn tissue-specific motif features.

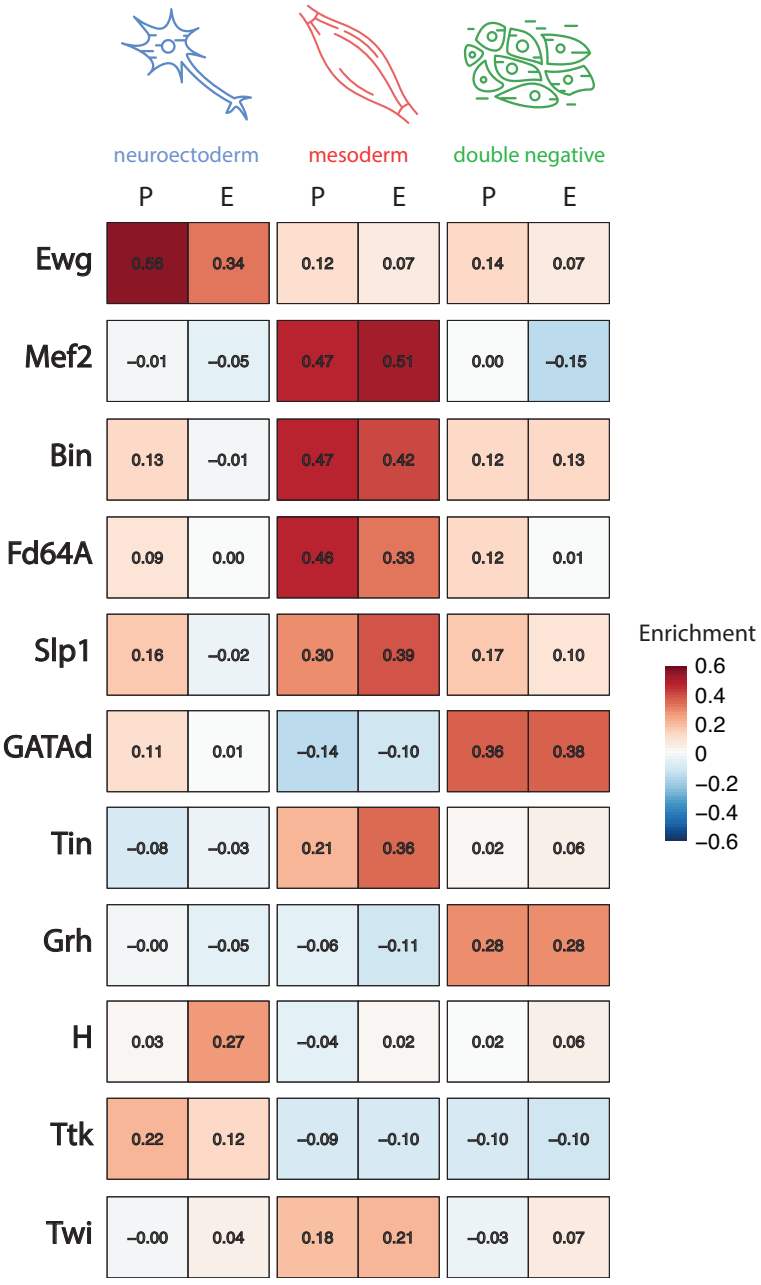


Figure 43 – Motif enrichment within the k-mers from the 6 conditions. The plot shows the median score of k-mers matching Transcription Factors PWMs. Positive values indicate that k-mers that match the PWM are associated with the positive training set. For example, the top k-mers in the mesoderm tissue match the Mef2 PWM well, but the same PWM is not found in the top k-mers of neuroectoderm and Double Negative tissue. The enrichment scores correspond to the median LS-GKM score of the top 10-mers matching each PWM, scaled by the range of the distribution. Enrichment score can range between -1 and 1. P: promoter-proximal DHS; E: enhancer. Ewg: Erect wing; Mef2: Myocyte enhancer factor 2; Bin: Binou; Fd64A: Forkhead box L1; Slp1: Sloppy paired 1; Tin: Tinman; Grh: Grainy head; H: Hairy; Ttk: Tramtrack; Twi: Twist.

2.4 - deltaSVM scores give insight into the impact of variants on chromatin accessibility

LS-GKM provides deltaSVM scores that correspond to the likelihood of the non-reference allele to increase or decrease chromatin accessibility. These scores represent a useful resource to prioritize variants that are more likely to have an impact on accessibility and to predict the direction of the effect. In this section, I will discuss how deltaSVM scores can provide global information about the impact of variants on accessibility.

2.4.1 - Variants in enhancers have a larger impact compared to promoter-proximal DHS

The deltaSVM scores follow a normal distribution: variants with high absolute deltaSVM values are more likely to have an effect on chromatin accessibility (Supplementary Figure 6). The distribution of deltaSVM is shifted towards negative values indicating that non-reference variants are more likely to reduce accessibility. In addition, promoter-proximal DHS and enhancers have very different deltaSVM ranges, with the distribution on enhancers being two times broader than on promoter-proximal DHS. This observation suggests that enhancer variants have a larger effect on accessibility, while promoter-proximal DHS are more robust to variation.

2.4.2 - Rare variants have a larger impact on chromatin accessibility

By combining the allele frequency and deltaSVM information, we observe that rare variants have larger absolute deltaSVM values. Figure 44 shows the comparison of deltaSVM scores between rare and common variants. Rare variants, within the DGRP population, were identified as having a non-reference frequency smaller than 0.01; common variants had a non-reference allele frequency greater than 0.5. In all cases, rare variants had higher deltaSVM than common variants, except for the mesodermal promoter-proximal DHS (the training condition with the smallest positive set and lowest AUC). Rare variants were previously observed to have larger

effects[131], and therefore it was suggested that they are more likely to be under negative selection. In conclusion, the deltaSVM scores indicate that enhancer variants generally have larger effects compared to variants in promoter-proximal DHS, with rare variants have larger impact on chromatin accessibility compared to common variants.

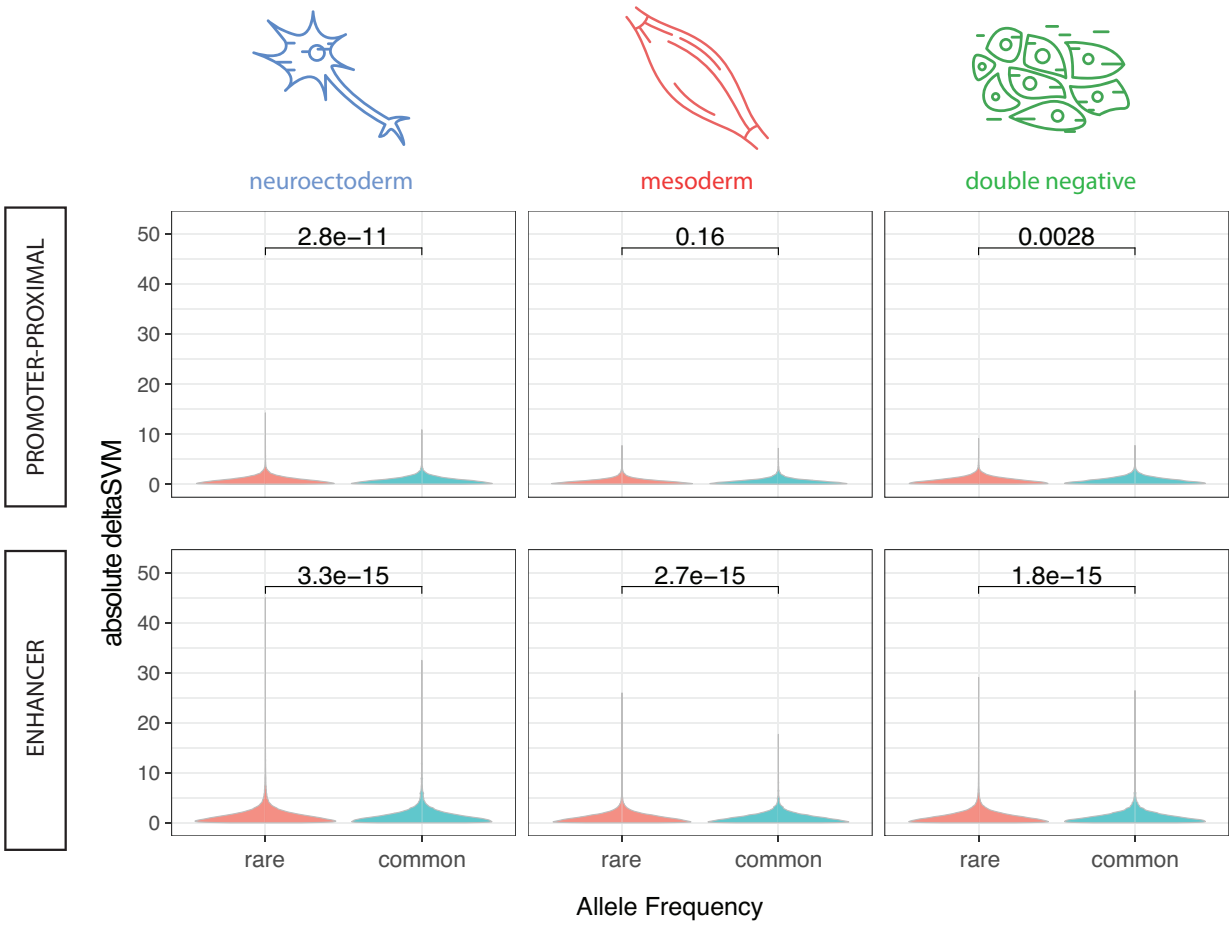


Figure 44 – Rare variants have a larger impact on chromatin accessibility. The plot compares the distribution of absolute deltaSVM between rare and common variants. The absolute deltaSVM scores are divided by tissue and promoter-proximal DHS and enhancer. The number comparing the distributions corresponds to the Wilcoxon p-value. Rare variants have a non-reference allele frequency smaller than 0.01 while common variants have a non-reference allele frequency greater than 0.5.

2.4.3 - Many variants have tissue-specific effects

deltaSVM scores can be visualized and compared across tissues (Figure 45). The heatmap shows the deltaSVM across tissues for the 1,000 variants with the highest scores, each row represents a variant. Promoter-proximal DHS show variable scores across tissues, with many variants having tissue-specific effects: a high score is

observed in one tissue and a near zero score in other tissues. There are a few examples of discordant scores at the top and at the bottom of the heatmap. Enhancers show more uniform behavior with roughly 60% (top of the heatmap) of the variants behaving in the same way across tissues. The remaining 40% (bottom of the heatmap) have discordant behaviors across tissues. The latter cases are especially interesting to examine because they provide examples of variants increasing accessibility in one tissue and decreasing it in another.

The high variability of deltaSVM scores between tissues for the promoter-proximal DHS is unexpected given that promoters tend to function in a constitutive way. It is important to notice that the heatmap only shows the 1,000 variants with the highest deltaSVM scores out of 342,760 scored variants within enhancers and 193,554 variants within promoter-proximal DHS. At the global level, deltaSVM scores have higher correlation across tissues in promoter-proximal DHS (Person correlation between deltaSVM scores of the three tissues range between 0.83 and 0.93) than in enhancers (Pearson correlation between deltaSVM scores of the three tissues range between 0.08 and 0.29). This indicates that globally variants on enhancers tend to act more often in a tissue-specific way than on promoter-proximal DHS.

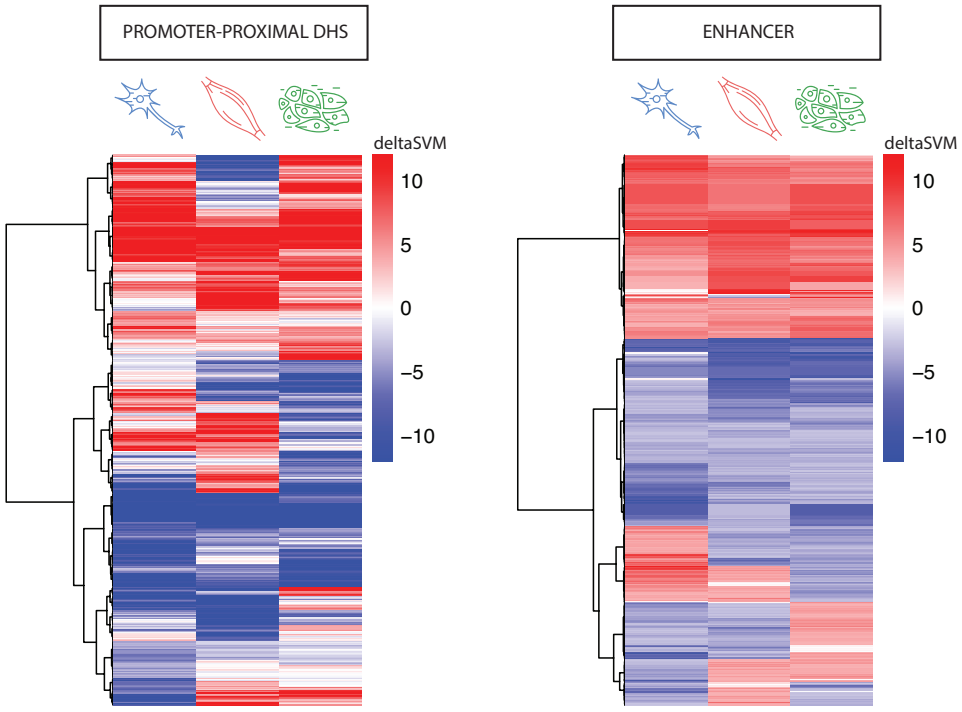


Figure 45 – Heatmap of tissue-specific variant scores divided by promoter-proximal DHS and enhancers. The heatmaps report the top 1,000 variants with highest deltaSVM scores for promoter-proximal DHS and enhancers (the variants are not the same across the two heatmaps). Heatmap colors correspond to deltaSVM values.

2.5 - Merging of variant calls from different populations

To provide a unified set of variant scores, we merged the variant calls (vcf files[92]) for the DGRP and GDL. When considering both variant files, 72% of variants are uniquely called in one of the vcf files, while the remaining 28% are common between the two panels. Specifically, the DGRP vcf includes 6,131,648 variants and the GDL vcf includes 6,752,029 variants, 2,828,011 of which are in common. The vcf files were merged using two strategies:

- All variants from the two files were merged. If a variant was genotyped only in one file, we assumed that all lines in the other file were harboring the reference allele.
- Only the variants genotyped in both vcf files were joined.

The first strategy does not cause any loss of data but it assumes the genotype for 72% of the variants. The assumption would hold true only if the vcf files were including all true positive variants. The second strategy is more conservative and does not require any assumption, but it causes a significant loss of data.

To assess the quality of the two merged vcf files, we retrieved the population structure of the lines from them. Following the protocol in Grenier *et al*.[93], we performed a PCA on neutrally evolving variants selected from the two merged vcf files. The expected clustering of the 5 populations within the GDL is shown in Figure 30 from Grenier *et al*.[93]. In addition, the DGRP population should cluster with the Ithaca population, given their common origin in the same geographic location.

I then performed a PCA using the merged vcf file containing all variants and assuming reference genotypes (Figure 46, top) and the merged vcf file including only variants genotyped in both original vcf files (Figure 46, bottom). The top plots show a strong separation between the DGRP and the GDL lines, only the second PC separates the Zimbabwe lines from the others. This indicates that the assumption that ungenotyped variants are similar to the reference allele biases the results by including an incorrect structure in the data, meaning that both vcf files are missing variants within the populations. The bottom plots show the expected separation between populations: the first PC splits the Zimbabwe lines and the second one the Beijing lines. The DGRP lines cluster with the Ithaca as expected.

Taken together these results show that it is not possible to merge all variants from two separate vcf files without performing a *de novo* genotyping. Following this result, we provided separate deltaSVM scored for the DGRP and GDL vcf files.

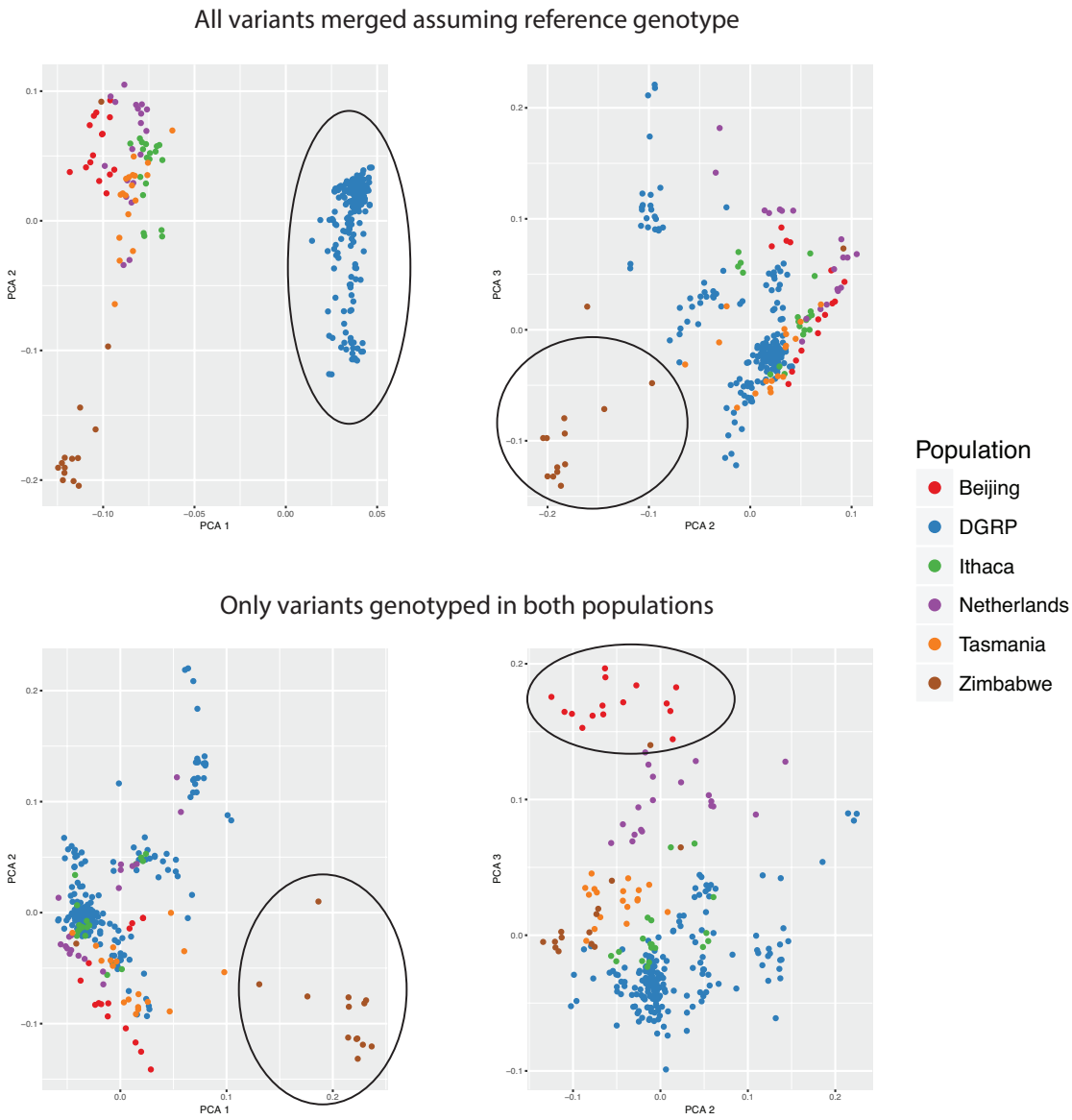


Figure 46 – Merging vcf files from two different populations. Principal Component Analysis of variants from two independent vcf files. The two populations are the GDL and DGRP. (a) The plots show the results of merging all variants from the two vcf files. When a variant is genotyped only in one file, it is assumed that the other file has a reference allele. The first PC separates the DGRP lines from the GDL showing strong batch effects. (b) The plots show the results of merging only the variants genotyped in both vcf files. The first PC clearly separates the Zimbabwe lines as in Grenier *et al.*[93], the third separates the Beijing lines. The DGRP lines cluster with the Ithaca lines that come from the same geographic location.

3 - Perspectives

LS-GKM is able to learn tissue-specific features and can predict the effect of variants on chromatin accessibility. We are currently performing two experiments to validate the predictions of LS-GKM. In addition, we hope to set meaningful cutoffs for the deltaSVM scores, that correspond to a high chance of differential accessibility. In particular, we are performing tissue-specific ATAC-Seq to measure accessibility and SuRe to quantify promoter and enhancer function.

3.1 Tissue-specific ATAC-Seq

ATAC-Seq will be performed for two DGRP lines (DGRP-57 and DGRP-714) to validate the deltaSVM scores *in vivo*. This experiment is currently being performed in collaboration with Rebecca Rodrigues Viales from the Furlong laboratory. We collected DGRP-57 and DGRP-714 embryos staged at 10-12h and FACS-sorted the cells with the same procedure as for the DNase hypersensitivity assay (see "II - Genetic variation as a tool to associate *cis* Regulatory Modules with their target genes: 5.2.1 - DHS identification"). ATAC-seq is currently being performed on sorted populations of muscle and neuronal cells in both genetic backgrounds. After identifying differentially accessible peaks, I will compare the differential accessibility between the two lines with the deltaSVM scores predicted by LS-GKM.

3.2 SuRe to measure variants impact on CRM function

CRMs exert their function by recruiting Transcription Factors to the DNA. In this project, we used chromatin accessibility as a proxy for TF binding. By measuring TF binding, chromatin accessibility also correlates with CRM activity. A more direct way to assess promoter and enhancer activity[78] is by quantifying self-transcription. SuRe is an *in vitro* technique capable of measuring self-transcription genome-wide in the chosen cell line[20]. We are currently performing SuRe in *Drosophila* Kc cells on genomic fragments from 6 DGRP and 6 GDL lines to measure differential activity of promoters and enhancers genome wide. This experiment is currently being performed in collaboration with Matteo Perino from the Furlong laboratory and Ludo Pagie and Marcel De Haas from the laboratory of Bas van Steensel. We plan to

measure differential activity across the genomic DNA for the two lines and identify the causal variants. We will then compare differential activity with the deltaSVM scores.

4 - Discussion

LS-GKM is an effective method to capture the predictive regulatory motifs of *Drosophila melanogaster's cis*-regulatory elements. We identified tissue-specific DHS and separated them into promoter-proximal and promoter-distal (putative enhancer elements) for a total of six training sets. LS-GKM shows good performance on the small training sets provided in this study. After training the model on tissue-specific DHS, we could retrieve the expected motifs for transcription factors that are important regulators in the relevant tissues. deltaSVM scores also provide insights into functional effect of variants on accessibility. We observed that variants have larger predicted effects on enhancers compared to promoter-proximal accessibility. In addition, rare variants have a larger effect on accessibility compared to common variants.

In order to increase the relevance and usefulness of this resource, we will further validate the deltaSVM scores. We are currently performing tissue-specific ATAC-Seq on two DGRP lines to directly correlate deltaSVM scores with accessibility measures. We hope to identify a cutoff for deltaSVM scores over which the variants are very likely to have an effect on accessibility. The caQTL validation suggests that this is possible. In fact, Figure 42 shows that all deltaSVM scores with an absolute value larger than 6 correctly predict the direction of the caQTL. We will also try to assess the specificity and sensitivity of the deltaSVM scores. Finally, we will test if deltaSVM scores can predict promoter and enhancer activity measured by SuRe-seq. Predicting enhancer and promoter function will represent a greater challenge than predicting accessibility. In fact, deltaSVM scores are poor predictors of gene expression. Previous observation estimate that in roughly 30% of the cases accessibility and gene expression are negatively correlated[123] suggesting that about 30% of CRMs have a negative effect on transcription and might represent silencers.

5 - Methods

5.1 - LS-GKM training

LS-GKM is a machine learning approach that can classify DNA sequences. The training is a crucial step that requires a careful selection of positive and negative (background) sequences. In this section, I will discuss the steps that were taken to train and assess the model. The pipeline is summarized in Figure 40.

5.1.1 - Identification of tissue-specific DHS

The tissue-specific DHS were identified following the pipeline in "II - Genetic variation as a tool *to* associate cis Regulatory Modules with their target genes: 5.2.3 - Tissue-specific DHS" independently for all time points (4-6 hpf, 6-8 hpf, 8-10 hpf, 10-12 hpf). If a DHS was defined as tissue-specific in one time point, then it was considered as being tissue-specific. DHS were then separated in promoter-proximal DHS and enhancers based on vicinity to TSS annotated in Flybase 6.13. If a DHS was closer than 500 bp to a known TSS it was annotated as promoter-proximal, otherwise, it was annotated as an enhancer.

5.1.2 - Positive set

We obtained the DHS sequences from BDGP6 genome assembly using bedtools[88] getfasta. We excluded the DHS that contained missing nucleotides (Ns) because they are not handled by LS-GKM. The DHS were divided into six sets corresponding to neuroectoderm promoter-proximal, neuroectoderm enhancer, mesoderm promoter-proximal, mesoderm enhancer, Double Negative promoter-proximal and Double Negative enhancer.

5.1.3 - Negative set selection

For each positive set, we selected five negative sets, each including the same number of sequences as the positive set. We removed known exon sequences and DHS from the BDGP6 *Drosophila* genome assembly and tiled the remaining genome in sequences of 300 bp. To match the sequence composition and increase the complexity of the most discriminative k-mers, we used the R package MatchIt[132]. MatchIt was employed so that the background sequence would match the nucleotide and di-nucleotide composition of the positive set. For each positive set, we selected five times more sequences from the non-exon non-DHS tiles that best matched the positive set. The matched sequences were then randomly divided in five batches to form five independent negative sets.

5.1.4 - LS-GKM training

We downloaded LS-GKM[122] from GitHub (https://github.com/Dongwon-Lee/lsgkm/). Five independent replicate trainings were performed for each of the six DHS set using different background sequences. We ran LS-GKM with the options "-t 2 -l 10 -k 6" to use gapped k-mers of total length 10 with 4 gaps. LS-GKM can be set to give more relevance to the k-mers in the center of the sequence. While this feature is valuable for ChIP-Seq data, we noticed a reduction of performance when using it with DHS data. Our pipeline was then run without this option (corresponding to "-t 2").

We performed 10 cross-validations for each independent training. LS-GKM performs cross-validations by excluding a random 10% of the positive and negative sequences form the training.

All unique 10-mers were generated using nrkmers.py script provided with LS-GKM. We assigned SVM weights to k-mers using gkmpredic. We obtained final k-mer scores by averaging the scores across the 5 replicates.

5.2 - Variants scoring

We analyzed variants from two independent populations: the Global Diversity Lines[93] (GDL) and *Drosophila* Genetic Reference Panel[63] (DGRP). The variants overlapping a DHS were divided into two groups, depending if the DHS was promoter-proximal or enhancer. deltaSVM scores were computed using deltasvm.pl from LS-GKM package for the six trainings. Each variant overlapping a DHS was then associated to three deltaSVM scores corresponding to the tissue-specific effect on chromatin accessibility.

We also computed DHS level scores for each line by summing the deltaSVM at the variant level. Delta SVM scores are computed by comparing the alternative allele to the reference allele: variants that have the reference allele have a deltaSVM of 0. Unknown genotypes were imputed by averaging the deltaSVM at the population level. For heterozygous variants, we averaged the deltaSVM of the parental alleles. The DHS level scores have proven to be less predictive than using the top score variant for the caQTL validation ("2.2 - Prediction of chromatin accessibility QTLs"). We will further test this approach to predict the tissue-specific ATAC-Seq and SuRe-Seq data.

5.3 - Validation of caQTLs

We received the ATAC-Seq peaks and caQTL files described in Jacobs *et al*.[124] from the laboratory of Stein Aerts. The coordinates were moved from BDGP5 to BDGP6 using liftOver. The ATAC-Seq peaks overlapping embryonic DHS or exons were excluded. A total of 14,682 eye-antennal imaginal disc specific ATAC-Seq peaks were used in the training. We then obtained the three positive sets: all peaks, promoter-proximal peaks and enhancers (following as in "5.1.2 - Positive set"). The background sequences were selected following "5.1.3 - Negative set selection" with tiles of length 455 bp, corresponding to the median length of the ATAC-seq peaks. The training was performed for the three training sets separately following the same procedure as in "5.1.4 - LS-GKM training". Finally, we scored the DGRP variants

using the three LS-GKM model trained on the eye-antennal imaginal disc specific ATAC-Seq peaks. Each variant was assigned two scores corresponding: either promoter-proximal or enhancer and all DHS.

Jacobs *et al*.[124] identify 4,288 caQTLs on 2,048 unique ATAC peaks and report the GLM fit statistics. We repeated the caQTL fit with a simple linear regression and used the Pearson r as a measure of direction and size of the QTL effect. If more caQTLs were overlapping the same peak they were excluded, since it was not possible to discriminate the causal variant. caQTLs with low effect size were excluded from the correlation (p-value > 0.01 or Pearson r < 0.5). We then performed three correlations between the three training sets scores and the corresponding caQTL Pearson r.

5.4 - Identification of TF motifs enrichment from k-mers

The enrichment plots in Figure 45 were obtained by comparing the distribution of the top matching k-mers best matching to the PWM to the global k-mers distribution. We collected a total of 1,796 high quality motifs for *Drosophila* from the following sources:

- CIS-BP database[133] (downloaded on 20 June 2018)
- Fly Factor Survey (downloaded on 20 June 2018)
- Jaspar Core[134] Insecta (downloaded on 20 June 2018)
- On the Fly[135] (downloaded on 20 June 2018)
- *de novo* motif call performed by Olga Sigalova in the Furlong laboratory from ModERN ChIP-Seq database[136].
- High quality mesodermal Transcription Factor motifs from ChIP-chip experiments (Zinzen *et al*.[27])
- Grainy head motif from Yao *et al*.[137]

We matched each motif with all 10-mers (obtained in “5.1.4 - LS-GKM training”) using R Biostrings package[138]. Each motif was associated with the top 100 k-mers having a match score of at least 0.8. The k-mer enrichment score was obtained by

subtracting the median SVM weight of all the k-mer from the median SVM weight of matched k-mers, scaled by half of the range of the k-mer SVM weight:

$$\frac{median(SVM_{matched}) - median(SVM_{all})}{max(SVM_{all}) - min(SVM_{all})} \times 2$$

V - Conclusions

In this thesis, I have presented three projects developed during my Ph.D. They are complementary since they each explore the relationship between the effects of natural sequence variation on the regulation of gene expression during *Drosophila melanogaster* embryonic development.

In the first project, we built on the eQTL framework to specifically associate CRMs (using DNase hypersensitivity as a proxy) with their target genes. I identified 2,967 DHS-eQTLs and in particular 2,005 promoter-proximal DHS to gene associations and 962 enhancers to gene associations. This represents, to my knowledge, the largest functional CRM-to-gene map in *Drosophila*. We validated the results *in silico* by enrichment of eQTL signal on DHS, enrichment of tissue concordance between DHS and target genes and Hi-C contact enrichment, and experimentally by RT-qPCR. The results show extensive CRM sharing between genes. We also observe frequent long range gene regulation from both enhancers and promoter-proximal DHS. It will be crucial to assess if the promoter-proximal distal activity is classic *cis* regulation or if it can be attributed to *trans* effects. The predicted function of the regulated genes does not suggest that genes close to promoter-proximal DHS with distal activity are enriched for transcription factors. The ongoing CRISPR deletions of a number of selected promoter-proximal DHS-QTLs should shed light on this issue, especially if analyzed in an F_1 context.

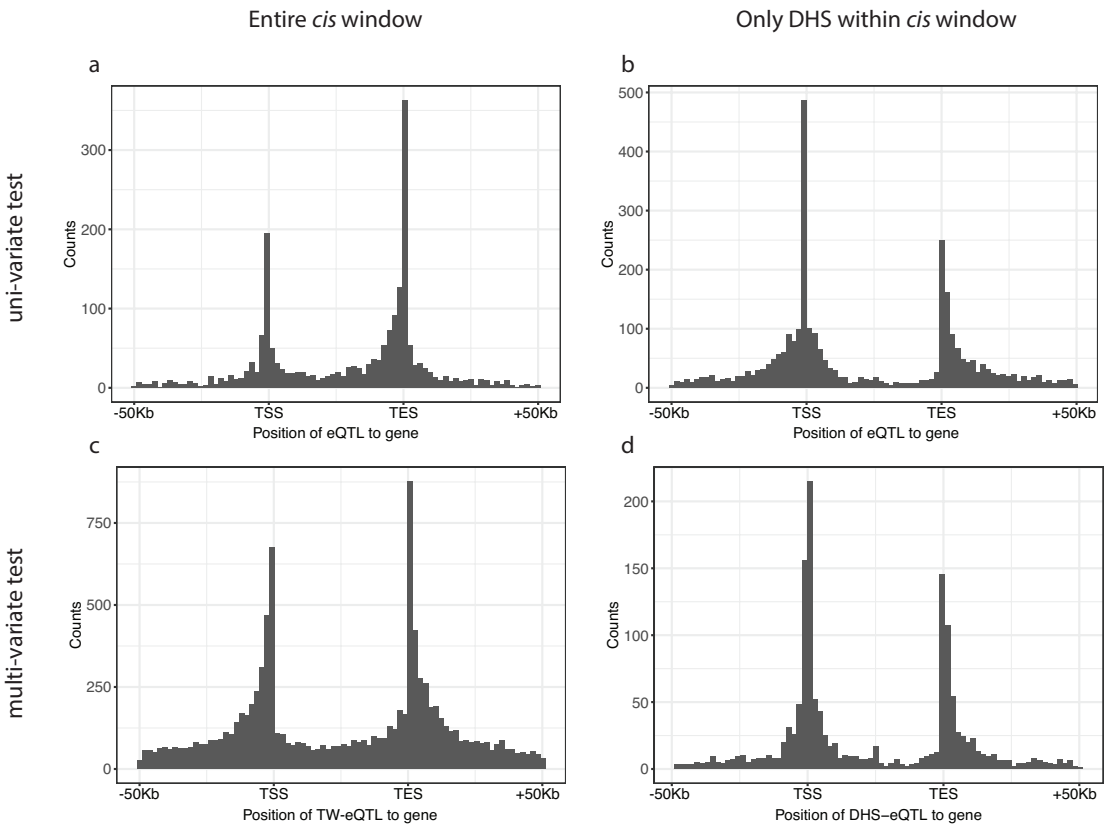
In the second project, we performed RNA-Seq on 80 samples from the Global Diversity Lines, which were collected at 10-12 hpf of embryogenesis. We performed multiple quality-control tests to ensure that the gene expression dataset is of high quality. The GDL come from five continents and show transcriptional diversity. In particular, the African population is the most separated, confirming the observations at the genetic level while the lines of European descent have similar transcriptomes. The Netherlands lines overexpress genes involved in the cuticle formation at 10-12 hpf, indicating an adaptation to the environment. I used this data to perform and eQTL analysis, which identified 903 gene and 2,021 exon eQTLs. This is, to my knowledge, the first map of splicing related eQTLs in *Drosophila* development.

In the third project, we applied LS-GKM (an SVM approach based on gapped k-mers) to score variants for their predicted impact on chromatin accessibility. We trained LS-GKM on six tissue-specific training sets: neuroectodermal, mesodermal and double negative DHS divided in promoter-proximal and promoter-distal. The method shows very good performance despite the small training sets. We could retrieve tissue-specific TFBS from the scored k-mers validating the training. We then scored the genetic variants from the DGRP and GDL populations to provide the population genetics community with a resource for variant prioritization. Rare variants generally show higher absolute deltaSVM scores indicating a larger impact on chromatin accessibility. To confirm these results, and thereby increase the usability of this resource, we are performing tissue-specific ATAC-Seq on two DGRP lines. This will enable us to assess LS-GKM predictions and associate the deltaSVM scores with a measure of statistical confidence. The resulting resource will then provide predictions for the functional impact of genetic variants on open chromatin (i.e. on enhancer and promoter occupancy).

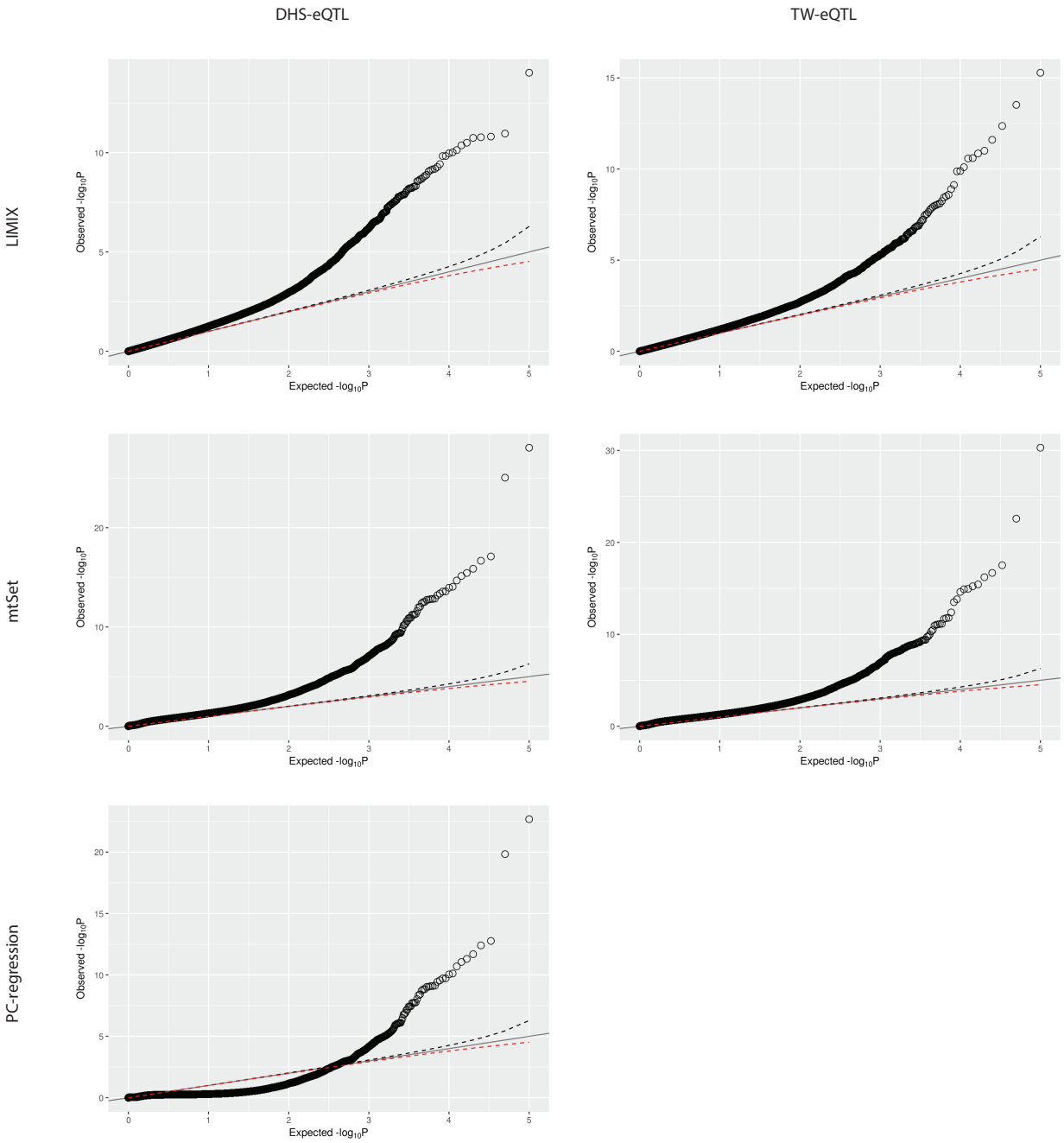
In summary, my Ph.D. has used population genetics as a tool to assign a function to regulatory elements at different levels. First, by dissecting the functional impact of genetic variants on open-chromatin at enhancers and promoters (Chapter IV), by analyzing transcriptional diversity among flies from five continents (Chapter III), and then by functionally linking enhancers and promoter-proximal elements to their target genes (Chapter II), while uncovering an unexpected level of complexity and distal regulation, and potential enhancer sharing.

VI - Appendix

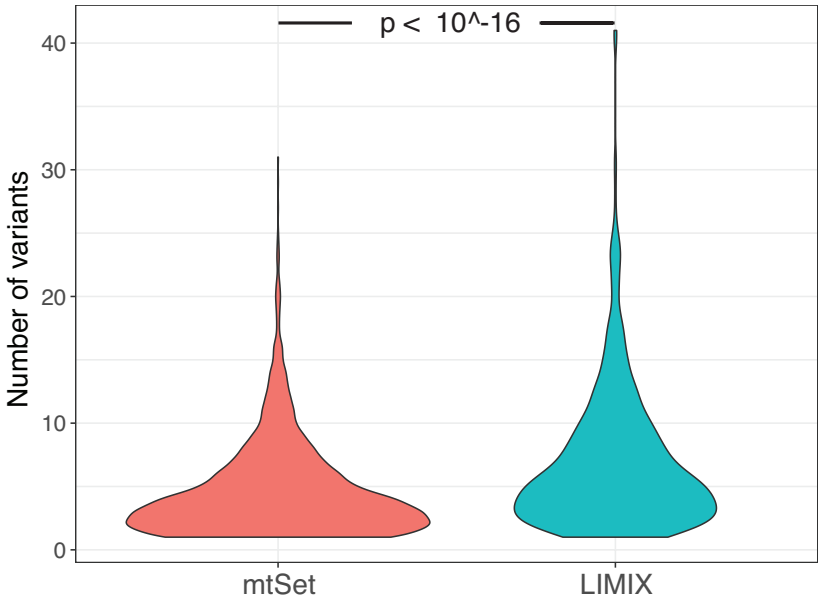
1 - Supplementary Figures



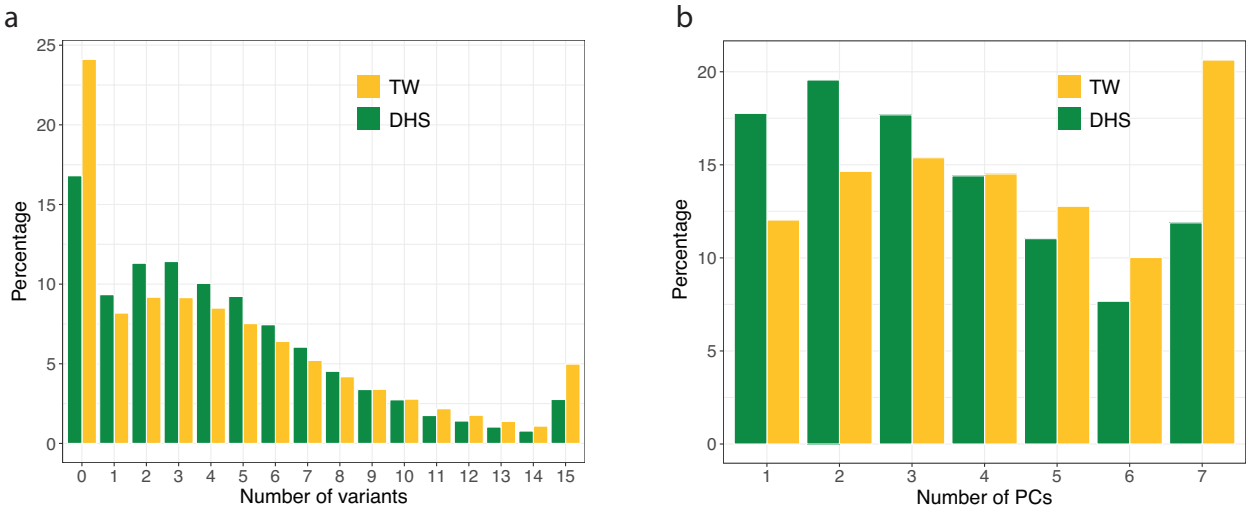
Supplementary Figure 1 – eQTL distribution around target genes. The plot shows the distribution of different types of eQTLs around their target genes. From top to bottom and from left to right: LIMIX eQTLs on the entire *cis* window; LIMIX eQTLs on DHS; TW-eQTLs; DHS-eQTLs. TSS: Transcription Start Site, TES: Transcription End Site.



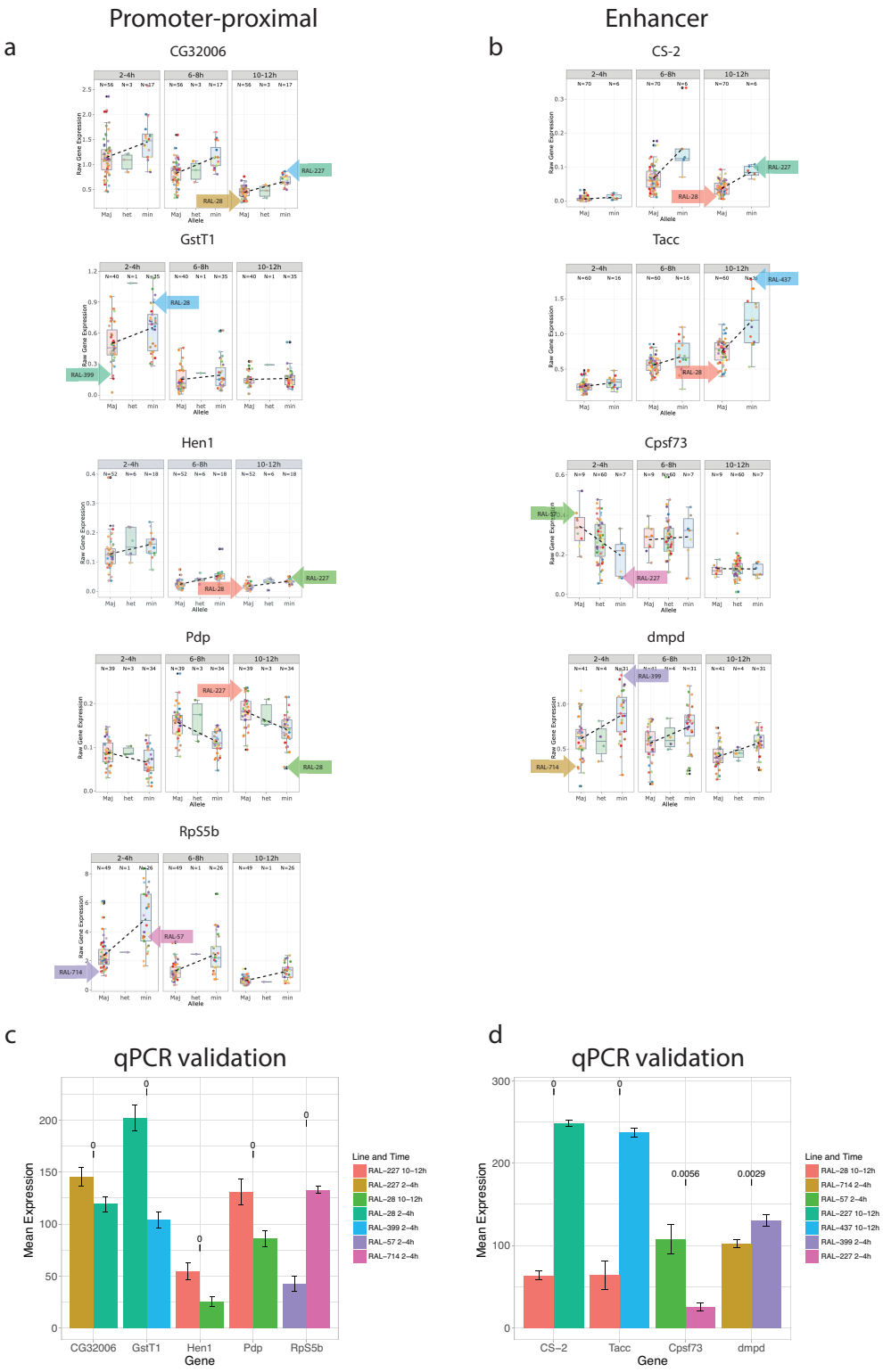
Supplementary Figure 2 - qqplots for different eQTL methods and tests. The quantile-quantile plots are shown by black points. The expected p-value distributions are shown as a solid black line with ±95% confidence intervals as dashed lines.



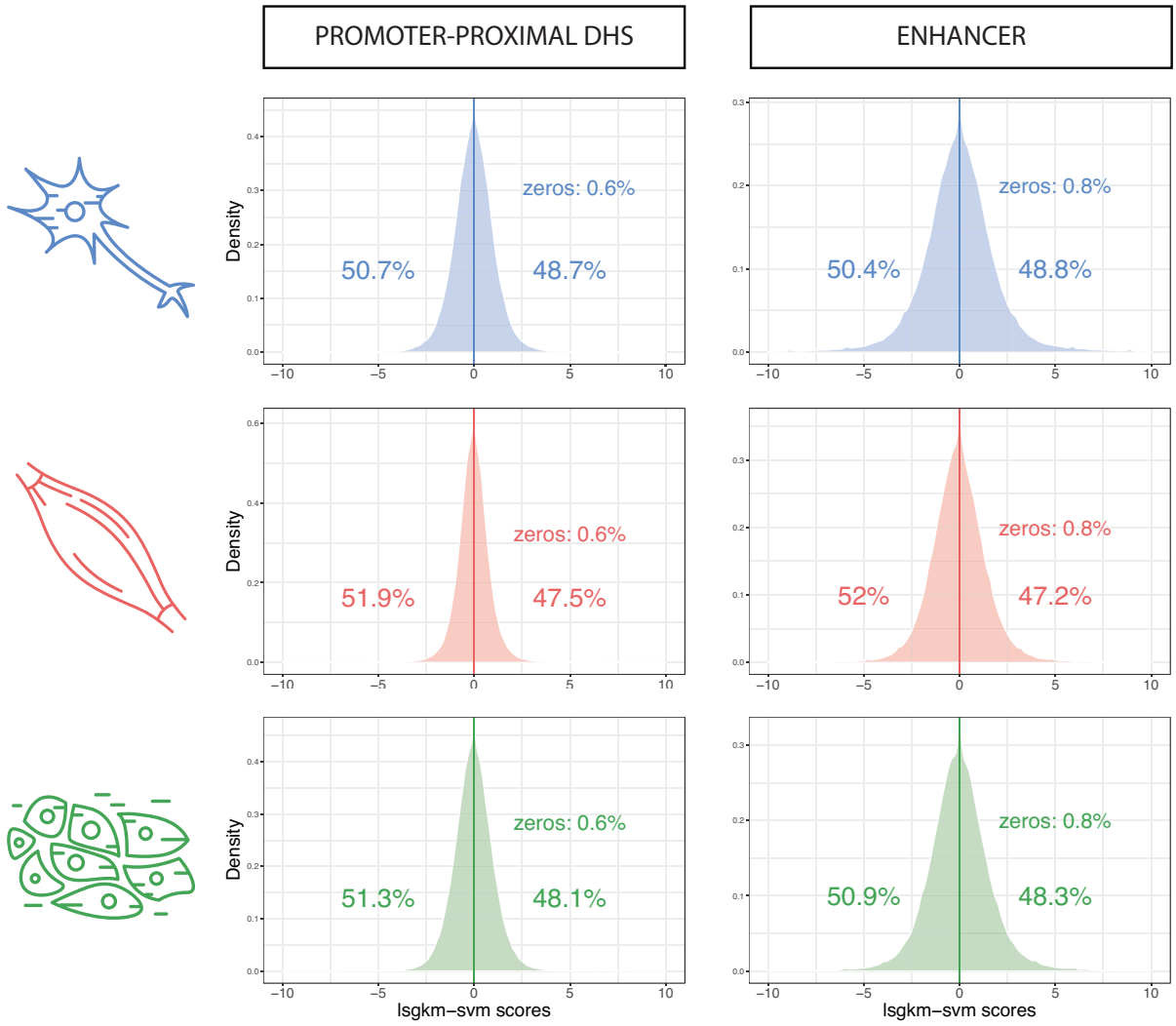
Supplementary Figure 3 - Number of variants on DHS-eQTLs significant for mtSet and LIMIX.



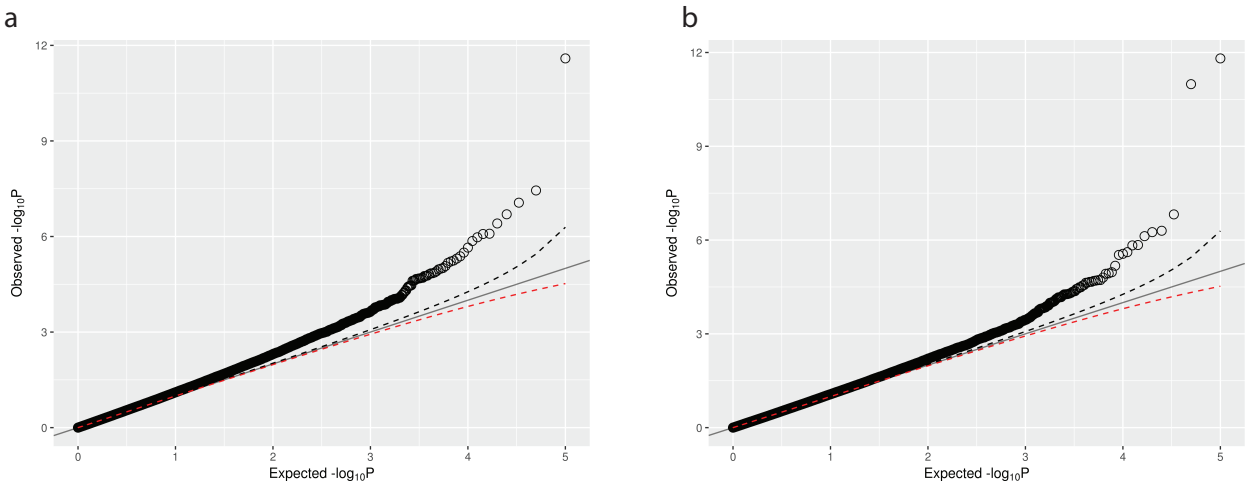
Supplementary Figure 4 – Number of variants and PCs necessary to explain them on DHS and Tiling Windows. (a) Number of variants (b) Number of Principal Components. DHS are shown in green while TW in yellow.



Supplementary Figure 5 – qPCR setup. (a) The plots show the distribution of gene expression values for promoter-proximal-eQTLs target genes divided by Maj and min alleles (based on the lowest p-value variant in the DHS). Two lines (one with the Maj, the other with the min allele) have been selected for RT-qPCR testing. The selected lines are indicated with an arrow. (b) Same as in (a) but for enhancer-eQTLs. (c) Gene expression measured with RT-qPCR for the lines and the genes involved in promoter-proximal-eQTLs indicated in (a). (d) Gene expression measured with RT-qPCR for the lines and the genes involved in enhancer-eQTLs indicated in (b).



Supplementary Figure 6 – Distribution of deltaSVM scores in 6 conditions. The figure shows the distribution of deltaSVM scores across the six training sets.



Supplementary Figure 7 – GDL gene-eQTL and exon-eQTL qqplots. (a) qqplot for gene-eQTLs called on the Global Diversity Lines. (b) qqplot for the exon-eQTLs.

Supplementary Figure 8 – PCA for batch effects on the Global Diversity Lines RNA-Seq. PCA of GDL samples. All plots show the first three Principal Components. The points represent samples and are colored to distinguish potential sources of batch effects.

2 - Supplementary Tables

Table 1 – Tissue-specific numbers for enhancer-eQTLs.

enhancer-eQTLs		DHS-eQTL	Number of DHS	Characterized enhancers	Number of target genes	Same gene / assigned
6-8h	**Neuro**	55	44	10	50	0 / 2
	Mesoderm	4	4	2	4	0 / 1
	Double Negative	23	21	2	23	1 / 4
	Multiple / all tissues	880	819	123	625	13 / 52
10-12h	**Neuro**	157	142	33	135	2 / 8
	Mesoderm	40	38	5	37	1 / 3
	Double Negative	78	68	5	70	0 / 4
	Multiple / all tissues	687	640	94	514	12 / 45

Table 2 – Enhancers associated to two genes.

enhancers associated to 2 genes	Gene expression correlation	BDGP concordance perfect / overlap / all
58	34	2 / 4 / 4

Table 3 – Promoter-proximal DHS and relationship with the target gene TSS.

Number of targets for promoter-proximal DHS	Closest TSS	Distal TSS 1	Distal TSS 2	Gene expression correlation	BDGP concordance perfect / overlap / all
1	615				
1		1004			
2	96			56	8 / 19 / 19
2		77		57	2 / 11 / 11

Table 4 – BDGP expression tissue to FACS sorted tissue

BDGP general tissue term	FACS sorted tissues
Maternal	none
Ubiquitous	Neuro / Meso / Double Negative
Gonad pole cells	Double Negative
Blastoderm	Double Negative
Mesoderm derivatives	Meso
Gut	Double Negative
Ectoderm	Double Negative
Nervous System	Neuro
Malphyngian tubule	Double Negative
Tracheal system	Double Negative
No staining	none
Sense organs	ambiguous
Amnioserosa	Double Negative
Hemolymph	Double Negative
Fat body	Double Negative
Mesectoderm	ambiguous

3 - List of abbreviations

ATAC-Seq	Assay for Transposase Accessible Chromatin Sequencing
AUC	Area Under the Curve
BAM	Binary Alignment Map
BDGP	Berkeley *Drosophila* Genome Project
BED	Browser Extensible Data
CAGE	Cap Analysis Gene Expression
ChIP-Seq	Chromatin immunoprecipitation and sequencing
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRM	*cis* Regulatory Module
DGRP	*Drosophila* Genetic Reference Panel
DNA	Deoxyribonucleic Acid
DNase I	Deoxyribonuclease I
DHS	DNase Hypersensitive Site
DPE	Downstream Promoter Element
DRE	DNA Recognition Motif
FACS	Fluorescence-Activated Cell Sorting

FDR	False Discovery Rate
GDL	Global Diversity Lines
GO	Gene Ontology
GWAS	Genome Wide Association Studies
Hi-C	High-throughput chromosome conformation capture
hpf	hours post fertilization
IDR	Irreproducible Discovery Rate
LD	Linkage Disequilibrium
LIMIX	Linear mixed model
LS-GKM	Large Sample gapped k-mer support vector machine
mtSet	multi trait Set test
MQ	Mapping Quality
pA	polyadenylation
PCA	Principal Component Analysis
PEER	Probabilistic Estimation of Expression Residuals
PIC	Pre-Initiation Complex
PRO-Seq	Precision nuclear Run-On Sequencing
PWM	Position Weight Matrix
QTL	Quantitative Trait Locus

eQTL	expression QTL
caQTL	chromatin accessibility QTL
hQTL	histone QTL
DHS-eQTL	DNase Hypersensitivity expression QTL
TW-eQTL	Tiling Window expression QTL
RNA	Ribonucleic Acid
mRNA	messenger RNA
rRNA	ribosomal RNA
RNA-Seq	RNA Sequencing
RT-qPCR	Real Time quantitative Polymerase Chain Reaction
SAM	Sequence Alignment Map
STAR	Spliced Transcripts Alignment to a Reference
STARR-Seq	Self-transcribing active regulatory region sequencing
SuRe	Survey of Regulatory Elements
TAD	Topologically Associating Domain
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSS	Transcription Start Site
VCF	Variance Call File

VII - References

1. Alberts, B. *et al. Molecular Biology of the Cell*. (W. W. Norton & Company, 2014).

2. Nüsslein-Volhard, C. & Wieschaus, E. Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795–801 (1980).

3. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287**, 2185–95 (2000).

4. Turner, F. R. & Mahowald, A. P. Scanning electron microscopy of Drosophila embryogenesis. 1. The structure of the egg envelopes and the formation of the cellular blastoderm. *Dev. Biol.* **50**, 95–108 (1976).

5. Petschek, J. P., Perrimon, N. & Mahowald, A. P. Region-specific defects in l(1)giant embryos of Drosophila melanogaster. *Dev. Biol.* **119**, 175–189 (1987).

6. Kosman, D. *et al.* Multiplex Detection of RNA Expression in Drosophila Embryos. *Science (80-.).* **305**, 846–846 (2004).

7. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **19**, 621–637 (2018).

8. Long, H. K. *et al.* Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170–1187 (2016).

9. Ayer, S. & Benyajati, C. Conserved enhancer and silencer elements responsible for differential Adh transcription in Drosophila cell lines. *Mol. Cell. Biol.* **10**, 3512–23 (1990).

10. Schwartz, Y. B. & Cavalli, G. Three-Dimensional Genome Organization and Function in Drosophila. *Genetics* **205**, (2017).

11. Hampsey, M. Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol. Mol. Biol. Rev.* **62**, 465–503 (1998).

12. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).

13. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).

14. Roider, H. G., Lenhard, B., Kanhere, A., Haas, S. A. & Vingron, M. CpG-depleted promoters harbor tissue-specific transcription factor binding signals—implications for motif overrepresentation analyses. *Nucleic Acids Res.* **37**,

6305–6315 (2009).

15. Hoskins, R. A. *et al.* Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res.* **21**, 182–192 (2011).

16. Engstrom, P. G., Ho Sui, S. J., Drivenes, O., Becker, T. S. & Lenhard, B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* **17**, 1898–1908 (2007).

17. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15776–81 (2003).

18. Schor, I. E. *et al.* Promoter shape varies across populations and affects promoter evolution and expression noise. *Nat. Genet.* (2017). doi:10.1038/ng.3791

19. Mahat, D. B. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–76 (2016).

20. van Arensbergen, J. *et al.* Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* **35**, 145–153 (2017).

21. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).

22. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).

23. Fujioka, M., Emi-Sarker, Y., Yusibova, G. L., Goto, T. & Jaynes, J. B. Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. *Development* **126**, 2527–38 (1999).

24. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–22 (2008).

25. 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**, 1213–1218 (2013).

26. Carroll, J. S. *et al.* Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.* **38**, 1289–1297 (2006).

27. 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).

28. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–7 (2013).

29. Arunachalam, M., Jayasurya, K., Tomancak, P. & Ohler, U. An alignment-free method to identify candidate orthologous enhancers in multiple Drosophila genomes. *Bioinformatics* **26**, 2109–2115 (2010).

30. Harju-Baker, S., Costa, F. C., Fedosyuk, H., Neades, R. & Peterson, K. R. Silencing of Agamma-globin gene expression during adult definitive erythropoiesis mediated by GATA-1-FOG-1-Mi2 complex binding at the -566 GATA site. *Mol. Cell. Biol.* **28**, 3101–13 (2008).

31. Bushey, A. M., Dorman, E. R. & Corces, V. G. Chromatin Insulators: Regulatory Mechanisms and Epigenetic Inheritance. *Mol. Cell* **32**, 1–9 (2008).

32. Gurudatta, B. V. & Corces, V. G. Chromatin insulators: lessons from the fly. *Briefings Funct. Genomics Proteomics* **8**, 276–282 (2009).

33. Valton, A.-L. & Dekker, J. TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* **36**, 34–40 (2016).

34. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–93 (2009).

35. Ghavi-Helm, Y. *et al.* Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* (2014). doi:10.1038/nature13417

36. Grosveld, F., van Assendelft, G. B., Greaves, D. R. & Kollias, G. Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell* **51**, 975–985 (1987).

37. Brooker, R. J. *Genetics : analysis & principles*.

38. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).

39. Macgregor, S., Cornes, B. K., Martin, N. G. & Visscher, P. M. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum. Genet.* **120**, 571–580 (2006).

40. Dubois, L. *et al.* Genetic and environmental contributions to weight, height, and BMI from birth to 19 years of age: an international study of over 12,000 twin pairs. *PLoS One* **7**, e30153 (2012).

41. Kimura, M. Evolutionary Rate at the Molecular Level. *Nature* **217**, 624–626 (1968).

42. Kimura, M. *The neutral theory of molecular evolution*. (Cambridge University Press, 1983).

43. Félix, M.-A. & Barkoulas, M. Pervasive robustness in biological systems. *Nat. Rev. Genet.* **16**, 483–496 (2015).

44. Perry, M. W., Boettiger, A. N., Bothma, J. P. & Levine, M. Shadow enhancers foster robustness of Drosophila gastrulation. *Curr. Biol.* **20**, 1562–7 (2010).

45. Mackay, T. F. The genetic architecture of quantitative traits: lessons from Drosophila. *Curr. Opin. Genet. Dev.* **14**, 253–257 (2004).

46. Nielsen, R. & Slatkin, M. *An introduction to population genetics : theory and applications*. (Sinauer Associates, 2013).

47. Corbett-Detig, R. B., Zhou, J., Clark, A. G., Hartl, D. L. & Ayroles, J. F. Genetic incompatibilities are widespread within species. *Nature* **504**, 135–137 (2013).

48. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

49. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).

50. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 2941 (2018).

51. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–4 (2012).

52. Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051–1065 (2015).

53. Cannavò, E. *et al.* Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature* **541**, 402–406 (2016).

54. Huang, W. *et al.* Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res.* **24**, 1193–208 (2014).

55. Nelson, J. D., Denisenko, O., Sova, P. & Bomsztyk, K. Fast chromatin immunoprecipitation assay. *Nucleic Acids Res.* **34**, e2 (2006).

56. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

57. Gallo, S. M. *et al.* REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res.* **39**, D118–D123 (2011).

58. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).

59. Kvon, E. Z. *et al.* Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512**, 91–95 (2014).

60. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858-871.e8 (2018).

61. Mackay, T. F. C. The Genetic Architecture of Quantitative Traits. *Annu. Rev.*

Genet. **35**, 303–339 (2001).

62. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7**, 862–872 (2006).

63. Mackay, T. F. C. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**, 173–178 (2012).

64. Jiang, N., Emberly, E., Cuvier, O. & Hart, C. M. Genome-Wide Mapping of Boundary Element-Associated Factor (BEAF) Binding Sites in Drosophila melanogaster Links BEAF to Transcription. *Mol. Cell. Biol.* **29**, 3556–3568 (2009).

65. Kim, T. H. *et al.* Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell* **128**, 1231–1245 (2007).

66. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*. (Springer Berlin Heidelberg, 2005). doi:10.1007/978-3-540-27752-1

67. Lippert, C., Casale, F. P., Rakitsch, B. & Stegle, O. LIMIX: genetic analysis of multiple traits. *bioRxiv* 003905 (2014). doi:10.1101/003905

68. Casale, F. P., Rakitsch, B., Lippert, C. & Stegle, O. Efficient set tests for the genetic analysis of correlated traits. *Nat. Methods* **12**, 755–8 (2015).

69. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–7 (2012).

70. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).

71. Mohrs, M. *et al.* Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nat. Immunol.* **2**, 842–847 (2001).

72. Tomancak, P. *et al.* Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.* **3**, research0088.1 (2002).

73. 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).

74. Spies, N. *et al.* Constraint and divergence of global gene expression in the mammalian embryo. *Elife* **4**, e05538 (2015).

75. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

76. Oliver, B., Parisi, M. & Clark, D. Gene expression neighborhoods. *J. Biol.* **1**, 4 (2002).

77. Spellman, P. T. & Rubin, G. M. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J. Biol.* **1**, 5 (2002).

78. 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).

79. Muerdter, F., Boryń, Ł. M. & Arnold, C. D. STARR-seq — Principles and applications. *Genomics* **106**, 145–150 (2015).

80. Bailey, T. L. *et al.* MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **37**, (2009).

81. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

82. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).

83. dos Santos, G. *et al.* FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.* **43**, D690–D697 (2015).

84. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–9 (2015).

85. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).

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

87. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).

88. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1-11.12.34 (2014).

89. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-5 (2016).

90. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47–e47 (2019).

91. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

92. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8 (2011).

93. Grenier, J. K. *et al.* Global Diversity Lines-A Five-Continent Reference Panel

of Sequenced Drosophila melanogaster Strains. *G3: Genes|Genomes|Genetics* **5**, 593–603 (2015).

94. 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**, 289–300 (1995).

95. Gratz, S. J., Rubinstein, C. D., Harrison, M. M., Wildonger, J. & O'Connor-Giles, K. M. CRISPR-Cas9 genome editing in Drosophila. *Curr. Protoc. Mol. Biol.* **111**, 31.2.1 (2015).

96. Keller, A. Drosophila melanogaster's history as a human commensal. *Curr. Biol.* **17**, R77–R81 (2007).

97. Lachaise, D. *et al.* Historical Biogeography of the Drosophila melanogaster Species Subgroup. in *Evolutionary Biology* 159–225 (Springer US, 1988). doi:10.1007/978-1-4613-0931-4_4

98. Thornton, K., Genetics, P. A.- & 2006, undefined. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. *Genet. Soc Am.*

99. Arguello, J. R., Laurent, S. & Clark, A. G. Demographic History of the Human Commensal *Drosophila melanogaster*. *Genome Biol. Evol.* **11**, 844–854 (2019).

100. Broad Institute. Picard Tools (http://picard.sourceforge.net). *Broad Institute, GitHub repository* (2015). doi:http://broadinstitute.github.io/picard.

101. Brown, J. B. *et al.* Diversity and dynamics of the Drosophila transcriptome. *Nature* **512**, 393–399 (2014).

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

103. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

104. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

105. Graveley, B. R. *et al.* The developmental transcriptome of Drosophila melanogaster. *Nature* **471**, 473–479 (2011).

106. Ostrowski, S., Dierick, H. A. & Bejsovec, A. Genetic control of cuticle formation during embryonic development of Drosophila melanogaster. *Genetics* **161**, 171–82 (2002).

107. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–11 (2013).

108. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term

association. *Bioinformatics* **23**, 257–258 (2007).

109. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

110. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* **6**, 80–92 (2012).

111. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–106 (2012).

112. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17**, 93–108 (2016).

113. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

114. Roman, T. S. *et al.* A Type 2 Diabetes–Associated Functional Regulatory Variant in a Pancreatic Islet Enhancer at the *ADCY5* Locus. *Diabetes* **66**, 2521–2530 (2017).

115. Lupiáñez, D. G., Spielmann, M. & Mundlos, S. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends Genet.* **32**, 225–237 (2016).

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

117. Sonawane, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. *Cell Rep.* **21**, 1077–1088 (2017).

118. Consortium, Gte. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

119. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS Comput. Biol.* **10**, e1003711 (2014).

120. Svetlichnyy, D., Imrichova, H., Fiers, M., Kalender Atak, Z. & Aerts, S. Identification of High-Impact cis-Regulatory Mutations Using Transcription Factor Specific Random Forest Models. *PLoS Comput. Biol.* **11**, e1004590 (2015).

121. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–9 (2016).

122. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).

123. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA

sequence. *Nat. Genet.* **47**, 955–961 (2015).

124. Jacobs, J. *et al.* The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. *Nat. Genet.* 1 (2018). doi:10.1038/s41588-018-0140-x

125. Haussmann, I. U., White, K. & Soller, M. Erect wing regulates synaptic growth in Drosophila by integration of multiple signaling pathways. *Genome Biol.* **9**, R73 (2008).

126. Lilly, B. *et al.* Requirement of MADS domain transcription factor D-MEF2 for muscle formation in Drosophila. *Science* **267**, 688–93 (1995).

127. Okumura, T., Matsumoto, A., Tanimura, T. & Murakami, R. An endoderm-specific GATA factor gene, dGATAe, is required for the terminal differentiation of the Drosophila endoderm. *Dev. Biol.* **278**, 576–586 (2005).

128. Narasimha, M., Uv, A., Krejci, A., Brown, N. H. & Bray, S. J. Grainy head promotes expression of septate junction proteins and influences epithelial morphogenesis. *J. Cell Sci.* **121**, 747–752 (2008).

129. Jin, H. *et al.* Genome-Wide Screens for In Vivo Tinman Binding Sites Identify Cardiac Enhancers with Diverse Functional Architectures. *PLoS Genet.* **9**, e1003195 (2013).

130. Gebelein, B., McKay, D. J. & Mann, R. S. Direct integration of Hox and segmentation gene inputs during Drosophila development. *Nature* **431**, 653–659 (2004).

131. Gorlov, I. P., Gorlova, O. Y., Frazier, M. L., Spitz, M. R. & Amos, C. I. Evolutionary evidence of the effect of rare variants on disease etiology. *Clin. Genet.* **79**, 199–206 (2011).

132. Ho, D. E., Imai, K., King, G. & Stuart, E. A. **MatchIt**: Nonparametric Preprocessing for Parametric Causal Inference. *J. Stat. Softw.* **42**, 1–28 (2011).

133. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**, 1431–1443 (2014).

134. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).

135. Shazman, S., Lee, H., Socol, Y., Mann, R. S. & Honig, B. OnTheFly: a database of Drosophila melanogaster transcription factors and their binding sites. *Nucleic Acids Res.* **42**, D167-71 (2014).

136. Kudron, M. M. *et al.* The ModERN Resource: Genome-Wide Binding Profiles for Hundreds of *Drosophila* and *Caenorhabditis elegans* Transcription Factors. *Genetics* **208**, 937–949 (2018).

137. Yao, L. *et al.* Genome-wide identification of Grainy head targets in *Drosophila*

reveals regulatory interactions with the POU domain transcription factor Vvl. *Development* **144**, 3145–3155 (2017).

138. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. (2018).

VIII - Acknowledgments

I would like to acknowledge the many people I had the pleasure to meet during my Ph.D. and that made this work possible. I want to start thanking Eileen for giving me the chance to do my Ph.D. in such an amazing lab. Thank you for the guidance, advice and encouragement. It has been great to work on stimulating projects and to collaborate with so many people, I feel like I learned a lot.

I want to thank every past and present member of the Furlong lab. Thank you, Fede. You have been the best student I could ask for. I wish you all the best for your Ph.D. Working with you has been fun and inspiring. It has been incredibly productive too: half of the content of this thesis we did it together. I also want to thank Enrico Cannavò for the fruitful discussions and the 3'-Tagged Sequencing data that he generated and I had the chance to analyze. The same goes for James, David and Charles for the DHS dataset. Big thanks go to the lab technicians that have all been involved in the projects presented here. In particular, Lucia for the Global Diversity Lines RNA extraction and library preparation, Rebecca for the RT-qPCR and the ATAC-Seq, Katharina and Songjie for the CRISPR and the *in situ* hybridization and Raquel for following me during my first step at the bench. Thank you, Matteo for the SuRe experiments and good advice. I also want to thank Alek for the Hi-C analysis and David for the helpful discussions that helped to shape the projects at the beginning of my Ph.D. I want to express my gratitude to Marijn for proof-reading this thesis and all the valuable advice, together with Aleksander, James, Federica, Anna and Tim. A big thank you to all the "monkey room" colleagues that I had the pleasure to share so much time with: Olga, Alek, Adam, Fede, David, Dermot, Matthias and Swann.

Many thanks to everyone who helped me during my Ph.D. In particular to Oliver Stegle, Paolo Casale and Nils Kölling for their precious advice on eQTL pipelines, LIMIX and mtSet. A thank you to Michael Beer for the insightful discussion about LS-GKM and Andrew Clark for providing the GDL samples and helpful suggestions. I would also like to thank my TAC members – Oliver Stegle, Jan Korbel, Henrik Kaessmann and Paolo Provero – for their guidance during my Ph.D.

Grazie a chi è stato al mio fianco durante il mio percorso di studi. Senza di voi non sarei qui. Grazie a Elena, per avermi insegnato le gioie e i dolori della programmazione ed a Paolo per avermi guidato durante i primi passi nel mondo della ricerca. Grazie a tutti i compagni del corso di biotecnologie che col loro entusiasmo mi hanno ispirato ogni giorno. In particolare grazie a Francesco, Nicola, Simona, Marta, Gabriele, Elisa, Riccardo, Martina. Grazie anche ai nuovi amici di Heidelberg per le cene, le bevute e le chiacchierate inseme. In particolare grazie a Michael, Enric, Elisa, Floriana, Stefano, Alice, Ilaria, Matteo e Matteo.

Voglio ringraziare i miei genitori per il supporto che mi hanno sempre dato, per aver creduto in me ed avermi aiutato ad ogni mio passo. Grazie per avermi comprato qualsiasi libro vi abbia mai chiesto. Se sono arrivato fin qui è soprattutto grazie a voi. Grazie a nonno per essere stato il primo ad ispirare in me l'amore per la natura e a nonna per tutto l'affetto che mi ha dato. Grazie anche a Laura per essermi sempre stata vicina ed in bocca al lupo per il tuo futuro da biologa.

Grazie a tutti i miei amici di sempre, in particolare a Gimmo per le chiacchierate in questi anni nonostante la distanza e per i consigli, sempre validi. Grazie anche per avermi aiutato a conoscermi meglio e a capire quali sono le cose più importanti.

Grazie Roby per tutti questi anni insieme e per aver sopportato la distanza. Non so come abbiamo fatto, ma siamo riusciti a crescere insieme. Grazie per avermi incoraggiato nei momenti difficili ed aver condiviso quelli belli. Non vedo l'ora di poterti avere vicina tutti i giorni.