Title
Evolution and Function of Drososphila melanogaster cis-regulatory Sequences

Author
Hardin, Aaron

Publication Date
2013

Peer reviewed|Thesis/dissertation

Evolution and Function of *Drososphila melanogaster cis*-regulatory Sequences

By

Aaron Hardin

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael Eisen, Chair
Professor Doris Bachtrog
Professor Gary Karpen
Professor Lior Pachter

Fall 2013

Evolution and Function of *Drososphila melanogaster cis*-regulatory Sequences

Abstract

Evolution and Function of *Drososphila melanogaster cis*-regulatory Sequences

by

Aaron Hardin

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Michael Eisen, Chair

In this work, I describe my doctoral work studying the regulation of transcription with both computational and experimental methods on the natural genetic variation in a population. This works integrates an investigation of the consequences of polymorphisms at three stages of gene regulation in the developing fly embryo: the diversity at *cis*-regulatory modules, the integration of transcription factor binding into changes in chromatin state and the effects of these inputs on the final phenotype of embryonic gene expression.

I dedicate this dissertation to Mela Hardin who has been here for me at all times, even when we were apart.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to acknowledge the following individuals for their invaluable contributions to this work and my personal growth in and out of the lab: Matt Davis, Xiao-Yong Li, Tommy Kaplan, Jacqueline Villalta, Mike Eisen, Devin Scannell, Doris Bachtrog, John Pool, Ryan Shultzaberger, Colin Brown, Susan Lott, Rich Lusk, Kelly Schiabor, Peter Combs, Dan Richter, Matt Taliaferro, Oh-Kyu Yoon, Tera Levin, Heather Bruce, Adam Session, Lior Pachter, Adam Roberts, Christopher Villalta, Eric Buhlis, Gary Karpen, Craig Miller, Robert Fowler, and many other members of the Eisen Lab, the Department of Molecular and Cell Biology, and the community of the University of California.

Chapter 1

Introduction

> *The power of Selection, whether exercised by man or brought into play under nature through the struggle for existence and the consequent survival of the fittest, absolutely depends on the variability of organic beings. Without variability, nothing can be effected; slight individual differences, however, suffice for the work, and are probably the chief or sole means in the production of new species.*
>
> (Charles Darwin 1868[1])

When Charles Darwin described the mechanisms of evolution[2], he had no knowledge of the genetic mechanisms of inheritance. He proposed gemmules[1], small packages carrying information for each part of the organism to the germ cells where they were collected and integrated forming the next generation. Darwin's erroneous solution was motivated by the puzzle as to how quantitative and discrete heritable variation could function. We now know that both result from the interaction of a multitude of genes and their variants [3]. These networks of interactions have explained several of Darwin's original observations of variability in and between species. The heritable variance that lead to differences in finch beaks[4] and dog varieties[5] have been traced down to the level of individual nucleotides through the study of the molecule of inheritance we now know to be DNA[6].

DNA contains the information for producing enzymatic and structural proteins and ultimately all metabolic processes and products are encoded in DNA. DNA is heritable and is passed down from parents to offspring over generations, accumulating changes in both the regions which encode proteins and the regulatory regions that control when, where, how much, and what form of protein is produced. DNA can be modified in the short term through the addition of chemical modifications such as methylation[7] but these effects do not persist over evolutionary time scales[8], only the changes of the sequence itself. Understanding how the primary sequence of DNA is interpreted and translated into the complex form of life has been one of the major scientific efforts since the 1950s.

As time and resources were put into determining the genomes of species, we collected genomes on larger scales starting with small viruses with only 4 genes[9], bacteria[10], and finally model eukaryotes such as yeast[11], worms[12], and flies which have around 14,000 genes[13]. Based on comparisons of the protein content between eukaryotes and the observed

and apparent difference in complexity between these model organisms and humans it was predicted that there should be at least 100,000 genes in the human genome[14]. It was a great surprise to observe that in fact there is only a small amount of additional genes in the human genome[15, 16] compared to worms and far less than the 60,000 genes found in Trichomonas vaginalis, a single celled infectious protozoa[17]. The amount of alternative splicing and variety in the expression regulation of the genes in the human genome suggested that gene regulation plays a major role in the interpretation of the genome into an organism. Humans were not unique in this regard and careful examination of gene expression within other species confirms that the majority of differences in tissues are due to regulatory control and modulation[18, 19].

Regulation of gene expression is a complex process with a network of positive and negative feedback[20] that act on DNA before and during gene expression and after expression on the transcribed RNA[21]. DNA is rarely found without any interacting protein and is packaged within the nucleus by wrapping around histones complexes known as nucleosomes[22]. These nucleosomes can be further organized into heterochromatin structures so dense that they sequester DNA away from other proteins[23]. Entire chromosomes can be prevented from expressing by this mechanism[24, 25]. DNA in an euchromatin state consists of nucleosomes that are compacted but not so much that the DNA is completely inaccessible. Histone acetyltransferase can initiate the process of opening up these DNA-protein complexes by adding chemical groups to the histones, decreasing their affinity for the DNA and allowing the DNA more freedom to move[26]. This exposes sequences which broadly fall into two classes: the promotor region immediately proximal to the start of transcription and distal *cis*-regulatory modules.

The promotor region is predominantly a scaffold for the organization of generic transcriptional machinery by the relative positioning of a short self-descriptive DNA motif called the TATA box to the beginning of the gene[27]. Eukaryotic promotors contain a core set of binding sites for transcription factor TFIID recognition and binding[28]. TFIID itself is part of and recruits the RNA polymerase II holozyme complex including a general transcription factor TATA-binding protein (TBP)[29]. While the promotor region contains transcription factor binding sites, they recruit a common set of transcription factors to all transcribed genes. It is the *cis*-regulatory modules that contain combinations of transcription factor binding sites for transcription factors that are expressed at specific developmental times, in spatially restricted domains, or tissues[30]. These transcription factors recruited to *cis*-regulatory modules contain additional domains for recruiting combinations of other proteins, coactivators and corepressors, through protein-protein interactions[31]. Through these combinations of transcription factors, a cell can gain spatial and temporal information or respond to external conditions[32]. In contrast to the number of genes, the number of transcriptions factors found in the genome increases as the organism increases in complexity resulting in multiple layers of regulation[33].

cis-regulatory modules are found at almost every gene that is not expressed at all times[34]. Genes with a specific expression patterns often have multiple *cis*-regulatory modules allowing a single gene to be expressed in multiple tissues or times[35]. Despite the

importance of *cis*-regulatory modules, only a few have been studied in great detail. Dissecting these regions and discovering the connections between sequence and transduction factor binding has yielded few general principles. This is largely due to the large range of conditions *cis*-regulatory module are exposed to. Describing *cis*-regulatory module effects by modulating the *trans* environment of the simultaneous combination of transcription factors is difficult to precisely control in an experiment. The majority of experiments on *cis*-regulatory modules modify or disrupt individual transaction factor binding sites or combinations of transcription factor binding sites revealing two major classes of organization. In some *cis*-regulatory modules, the disruption of a single transcription factor binding site or modification of the spatial arrangement is sufficient to ablate the activity of the entire enhancer or silencer[36, 37, 38]. These *cis*-regulatory modules are likely to function through the coordination of several transcription factors, all of which are required for proper function of the *cis*-regulatory module. In other *cis*-regulatory modules, there is remarkable flexibility in both the affinity of the transcription factor binding sites and also the spatial organization[30, 39]. These regions likely function through a combination of transcription factors that compensate and actively coordinate with each other to maintain a consistent output[40]. Classifying *cis*-regulatory modules into these two groups is crucial for the interpretation of their function. However manipulating individual *cis*-regulatory modules is extremely laborious and genome wide comparative approaches that found conserved and functional protein coding regions yield less definitive information in comparisons of *cis*-regulatory modules between and within species[41]. The conservation of physical binding is observed in the absence of sequence conservation between species and variability in transcription factor binding site sequence may represent a conserved tuning to species specific inputs and not a lack of functional conservation[41, 42]. These comparisons are also all limited by by our ability to describe and predict which sequences are bound by transcription factors.

Prediction of transcription factor binding sites has been informed by the experiment that defined biological interactions. The first steps in defining a transcription factor binding site were to discover which bases are physically associated with the transcription factor through DNA footprinting[43]. In this assay a DNA fragment of interest is incubated with the transcription factor and then digested with an endonuclease, DNaseI. The DNA that is bound by the transcription factor will be protected from the catalytic activity and by separating the fragments of digested DNA by length, the protected bases can be found. Repeating this process on potential *cis*-regulatory modules reveals a handful of locations with an affinity to the transcription factor of interest. However, this is a low throughput method and is restricted to relatively small DNA fragments. The handful of binding sites collected are examined for similarities and a first approximation of the properties of the binding sites is made from the consensus sequence[44, 45]. These sequences are short, usually between six and 11 base pairs long and are found at thousands to millions of locations within a genome, both within *cis*-regulatory modules and outside of presumptive *cis*-regulatory modules[44]. Additionally, a single transcription factor can bind to a range of transcription factor binding sites that may be related by only a handful of nucleotides that are consistent between them yielding a degenerate consensus sequence[45]. A single transcription factor binding site does

not necessarily only recruit a single transcription factor but may bind to several transcription factors[46, 47] leading to an explosion in possible binding events predicted from consensus sequences.

Efforts to more precisely characterize the binding site of a transcription factor have taken the collection of short sequences and instead of reducing them to only of the most common nucleotides, collated them into a matrix which describes the frequency of observing each nucleotide at each position[44]. This is called a position weight matrix (PWM) and can be used to predict the location of a binding site by the scoring the match of the input sequence against the values in the matrix[44]. Now instead of a binary match to a nucleotide polymer of length K (kmer), a p-value can be calculated taking into consideration aspects such as expected frequencies of each nucleotide in the genome. For these PWM to be useful, a larger number of inputs is required, more than the handful collected with DNaseI footprints. SELEX is an *in vitro* method of exposing a transcription factor binding domain to a large pool of random sequences and purifying those that bind[48]. The advantage of SELEX methods is that relative affinities to be measured, the PWM can be refined by the thousands of input sequences and the predictive power the PWM is increased[49]. The drawback of these PWMs are that they measuring an affinity that only exists in isolation and not in the context of larger sequences or other bound transcription factor interactions[50].

While comparative genomes has been highly informative for protein sequence inference, proteins are well defined in their primary encoding and understood well enough in their secondary structures[51, 52] in ways that are not true for regulatory codes. It is not to say that no useful evolutionary inferences can be made on transcription factor binding sites individually and together as modules. Position specific rates of evolution have been found in motifs defined by the PWM. Columns of this matrix in which a single nucleotide is strongly represented are under stronger purifying selection to maintain that representation than columns in which there is no specific requirement for particular nucleotide [53]. Comparing enhancers that drive identical expression patterns in different species of *Drosophila*[54] and other Dipterans[55] has shown that very divergent sequences at the nucleotide level can define identical expression patterns through the conservation of the total composition of binding sites. Comparative protein analysis has been successful because large numbers of proteins with conserved functions in divergent species have been identified. The function of *cis*-regulatory modules is far less understood and there are far fewer examples of homologous *cis*-regulatory modules between species.

Chromatin immunoprecipitation (ChIP) methods have avoided predictions and directly measured the position and strength of the binding[56, 57]. Briefly, this method isolates protein DNA complexes for a transcription factor recognized by an antibody. The protein-DNA interface is fixed with a chemical cross-link in vivo and then separated from non-bound DNA by fragmentation. The mixture of fragmented protein-DNA is then exposed to the antibody which is fixed to a column so that DNA which is bound by the transcription factor of interest is recovered. These fragments of DNA can be mapped back to the genome using chip hybridization or direct sequencing[58] and the genome wide position of the transcript factor can be inferred to within hundreds or tens of base pairs depending on the method.

This method has been applied to hundreds of transcription factors, creating a complete map of the regulatory landscape in yeast[59, 60], and to a lesser extent in humans[61, 62] and *Drosophila*[63, 64]. In particular, the use of high throughput sequencing of highly fragmented short pieces of bound DNA to the point that the bound region is near the end of the fragment leads to stand specific peaks, flanking the binding site. With sufficient sequencing depth, the position of the binding site can be inferred[65]. While this method has been very successful in creating genome wide graphs of binding activity, it has illuminated an feature of transcription factor bind that was previously unappreciated: pervasive genome wide clusters of weak binding[63]. In *D. melanogaster* clusters of weak binding were observed in ChIP-chip assays of 21 transcription factors[66] known from genetic screens to regulate the progression of anterior-posterior and dorsal-ventral patterning in early embryogenesis. These overlapping positions of multiple weakly bound transcription factors were found away from known targets of these factors.

What is the function of a transcription factor binding event? What is the impact? What kind of binding thresholds are biologically relevant? The presence of clusters of weak binding puts a new perspective on these questions. If transcription factor binding to *cis*-regulatory modules does not necessarily have any impact on the regulation of nearby genes then new models need to be constructed. Previously the assumption was that a bound protein was bound because it was functioning in gene regulation. However we do not have a good understanding of function and therefore computationally thresholding weak versus strong binding and linking this to function is fraught with error. It is possible that weak binding as defined by the graph of signal strength over a background is functional in a way that we do not understand due to the architecture of the *cis*-regulatory module. Indeed in many cases of weak binding the cognate transcription factor binding site is absent and our knowledge of the of these sequences is insufficient to determine whether this is due to random protein DNA interaction without any functional consequence or that it represents an unknown but functional protein DNA interaction.

To solve the translation of genotype to phenotype in *cis*-regulatory modules, we need a large set of divergent sequence that drive the same expression pattern so that conserved components and interactions can be identified, or a large set of highly similar sequences with informative differences at key positions so that inferences can be made from the disruptions. Here I show progress in interpreting the within species variation in *cis*-regulatory modules with particular focus on weakly bound regions by:

1. Classifying *cis*-regulatory modules by diversity (chapter two)
2. Exploring chromatin state changes between individuals (chapter three)
3. Linking chromatin variation to expression variation (chapter four)

Chapter 2

Within Species Diversity in *cis*-Regulatory Modules

The early embryo of *Drosophila melanogaster* has several of the best studied gene regulatory networks in animals. In particular, the factors that regulate anterior-posterior patterning were identified by genetic methods several decades ago, and we now well understand many of their activities, expression patterns and targets. Previous work in the Eisen lab has shown that these factors bind to thousands of regions across the genome. The most strongly bound regions contain most of the sequences known to be involved in gene regulation, while the function - if any - of the more abundant weakly bound regions is unclear. Our limited understanding of the molecular forces that shape transcription factor binding limit our ability to predict how altering these sequences will affect binding and activity. Examination of these regions shows that the underlying *cis*-regulatory modules are under strong purifying selection regardless of the strength of the *trans*-acting factor binding, suggesting that these regions are under selection to be in an open chromatin state.

2.1 Introduction

Gene expression is modulated by the biochemical interaction of *trans*-acting proteins known as transcription factors with *cis*-regulatory sequences spatially linked to the effected gene. These *cis*-regulatory sequences contain nucleotide sequences that are bound by a DNA binding domain in the transcription factor. Each *cis*-regulatory sequences can contain several transcription factor binding sites for the same transcription factor and also for combinations of transcription factors. These combinations of transcription factor binding sites can regulate genes in a unique pattern across tissues and development. Identifying these regions, the transcription factors that bind to them and how they interact to regulate nearby genes is a major goal for the understanding of expression. Changes within the *cis*-regulatory sequences lead to changes in the transcription factor binding and thus the expression of the regulated gene. These gene expression changes has been linked adaptive phenotypes[67, 68]

and diseases[69] both between and within species.

Although these transcription factor binding sites play an important regulatory role, we have a very poor understanding how the combination and composition into *cis*-regulatory modules effect gene regulation. Defining the gross location of the *cis*-regulatory sequences is assisted by the conservation of these regions [70] but experimentally isolating, verifying and manipulating a single region is a time consuming and expensive process. The projects that have put the time into testing *cis*-regulatory sequences conserved across distant species showed that massive sequence changes are tolerated and integrated to conserve function[55]. Examination of the conserved sequence features of the *eve* locus in *Drosophila melanogaster* shows that the conserved combinatorial composition of transcription factor binding site explains the majority of the functional conservation although the expression domain of *eve* is not perfectly conserved[55].

Since it is conserved binding of transcription factors that drives conservation of gene expression patterns, knowing where these transcription factors bind genome-wide is crucial for developing a global model of *cis*-regulatory sequence regulation of gene expression. By making DNA-protein crosslinks and using antibodies against transcription factors known to be present and active in the cell, and then recovering the DNA, we can capture the locations of these transcription factor - DNA interactions by hybridization or sequencing. This technique is called ChIP-seq or ChIP-chip (chromatin immuno-precipitation followed by sequencing or chip hybridization) and the Eisen lab has used ChIP-chip for six transcription factors (BCD, CAD, GT, HB, KNI, KR)[63] regulating anterior-posterior patterning during early embryogenesis of *D. melanogaster* helping to expand the gene regulatory network of the early embryo of *D. melanogaster*, one of the best studied gene regulatory networks in animals.

These transcription factors bound with a range of strengths to thousands of regions, including near genes not previous thought to be regulated by these factors[63] such as anterior-posterior regulators binding at genes that are expressed in a dorsal-ventral pattern. The unexpectedly bound regions also tended to be less strongly bound than at known anterior-posterior patterning genes. Explanations for this observation are that they might regulate expression at another stage in development, they might have a structural function within the chromatin, or they might not be functional at all. Several HB bound regions were known to be regulated by HB at later stages but there are only a handful of examples that link these transcription factors, which are at their peak expression now, to later regulation and the majority of regions weakly bound by KR showed no activity during development[71].

These bound regions were enriched for the presence of the corresponding transcription factor binding sites, including the weakly bound regions, but only BCD binding sites were more conserved within bound regions than without when compared to the sister species *D. simulans*. However, comparing under 200 regions of *Drosophila* species X chromosome diversity with a McDonald-Kreitman approach[72] results in estimates of over 50% of intergenic sequence, including *cis*-regulatory modules, is under constraint and functional[73].

Here, I leverage the advent of next-generation genome sequencing to obtain a whole genome perspective of *D. melanogaster* diversity and specifically the diversity in regions

bound by transcription factors. I sequence several whole genomes of *D. melanogaster* and use additional public resequencing *D. melanogaster* projects to create a high quality set of variants. Using previously collected ChIP data and predictions of transcription factor binding sites, I show that bound regions are under strong purifying selection, including the weakly bound regions.

2.2 Results

2.2.1 Genome wide diversity in transcription factor binding sites

The primary set of *D. melanogaster* genetic lines used in this work are the Drosophila Genetic Resource Panel, a set of 165 inbred and sequenced lines from Raleigh, North Carolina, USA led by Dr. Trudy Mackay[74]. A subset (37) of these lines were independently genotyped as part of the Drosophila Population Genomics Project led by Dr. Charles Langley [75]. These lines were originally collected from gravid females used to create individual lines from 20 generations of full-sib inbreeding. These strains were one of the first large scale resequencing efforts after yeast[76] to use next generation high throughput short read sequencing. These line were sequenced to at least 15x coverage with 36bp single-end reads with first generation Illumina GAI machines[74, 75]. This coverage was not sufficient to fully sequence single nucleotide polymorphisms (SNPs) with low error rates[77, 78, 79] and and the length of the reads provided very low ability to find small insertions and deletions (INDELs). To supplement this set, 100bp paired-end genomic libraries were made for four strains and resequenced with an Illumina HiSeq 2000.

Additional high quality, high depth sequencing libraries were also available for 20 strains *D. melanogaster* of isolated from Europe [80] and 139 strains from African, the ancestral home of *D. melanogaster*. The French strains are isofemale full-sib inbred lines from Montpellier, France and were sequenced to 50x coverage with an Illumina HiSeq 2000 by the BGI. The African isofemale strains were collected from 20 populations across Africa and while not inbred through full-sib matings, libraries were made from haploid embryos[81, 82] and sequenced with an Illumina GA2[83].

The largest high quality set of variants from a N. American population of *D. melanogaster* is from a collection of 96 inbred strains derived from a population in Winters, CA in which a 148kb *bab* locus has been Sanger sequenced [84]. Comparing the alleles and frequencies from the DGRP and Winters populations in Figure 2.1 shows that nearly all variants at the locus are shared and at similar frequencies. In total, my final variant collection contains 5,594,892 variants: 3,274,246 substitutions and 2,320,646 indels. This is 1,398,051 less than the 4,672,297 substitutions than were found using the DGRP project's Joint Genotyper for Inbred Lines[85] variant caller which does not take pervasive sequencing technology errors into account and reflects the benefit of using the global strain collection while jointly variant calling [86]

2.2.2 Genome wide purifying selection on *cis*-regulatory modules

Conservation and scans for selective pressure may allow us to classify the thousands of transcription factor binding sites[87]. My approach was to first define the general selective pressures on the *D. melanogaster* genome and then to narrow down my focus to the bound regions and the variations in predicted binding sites. Previous studies of isolated loci on the X chromosome suggested that most of the genome is under selection, including most of the noncoding sequence [73]. This observation is supported by measuring the overabundance of rare alleles across the entire genome, summarized by Tajima's D in Figure 2.2. If particular bound regions or regulatory sequences are under stronger selective pressures then we should observe further excess in rare alleles in these regions when compared to other noncoding regions. However, for most bound regions we do not observe a significantly larger selective pressure. To further characterize the effects of selection on bound regions, I measured the relationship between selection and binding strength (Figure 2.4) as well as the relationship between the binding strength and the conservation of the predicted binding sites in the bound regions (Figure 2.3). In both cases there is no significant relationship, which indicates that while these regions are under a general purifying selection, there is no simple relationship between the strength of binding and the selective pressures on that bound region. However, there is an enrichment of bound regions that contain a small number of polymorphic predicted binding sites in Figure 2.3.

To test if this enrichment of polymorphic predicted binding sites was significant, I modified the Tajima's D test so that instead of measuring the skew in the allele frequency of all single nucleotide polymorphisms in the bound regions, I measured the skew in only the polymorphisms that disrupted a predicted binding site in a bound region (Figure 2.5), essentially focusing the test to be most sensitive to only the features we are interested in and removing possible confounding background polymorphisms when looking at regions in different contexts across the genome. This skew was not significant when compared to an analysis of the same regions but using a randomly permuted matrix of the binding motif ($p > 0.35$, Wilcoxon rank-sum test).

2.3 Discussion

The diversity within the North American population of *D. melanogaster* from which the common lab strain *Oregon-R* was derived[88] was screened for polymorphism which disrupt transcription factor binding sites and compared to the binding strength of the cognizant transcription factor in those regions. These data suggest that even though the strength of binding of the transcription factors for anterior - posterior patterning has a very dynamic range, even low levels of binding are under purifying selection. However several factors may be obscuring the analysis.

First, the low levels of polymorphism in general reduces the number of polymorphic binding sites I have to measure. The average bound region contains 0.0491 segregating sites/bp but only 0.0083 segregating binding site motifs/bp. In spite of this, the mean

binding site D is within the range of intergenic D (Figures 2.2,2.5). The general trend of an excess of rare alleles within all bound regions including those in binding sites indicates that although there is a deficit of informative variants, the general trend is consistent. In addition, these strains were also largely purged of rare strongly deleterious alleles through the generations of inbreeding[74] so the tests are conservative and if unbiased population samples were collected it is possible that a subset of these regions would be found to be under differential selective pressures.

Related to the biases of the DGRP strains is the demographics of the N. American strains in general. *D. melanogaster* has experienced an bottleneck and expansion into N. American within the last 10,000 years[89, 90, 91] which will lead to an excess of rare alleles if the population has not reached equilibrium[92]. The impact of demographics has been considered by contrasting the allele distributions of disrupted binding sites with genomic regions, introns (Figure 2.2), that are undergoing less selection and should more sensitive to demographic effects. Compared to these regions, bound *cis*-regulatory sequences appear to be under stronger purifying selection. An additional assumption is the independence of the each segregating site. In contrast to humans[93], this assumption is largely justified since *D. melanogaster* has a recombination rate that, while not completely uniform, is largely free of dominating recombination hotspots[94] and much smaller linkage between loci[74], r^2 decaying to less than 0.1 within 100bp.

Lastly, the motifs predictions themselves are subject to some inaccuracy due to the stochastic binding and context dependency of transcription factor binding[95]. The motifs themselves were defined by a variety of methods, SELEX, Y1H and DNase footprints, which measure slightly difference features of the binding motifs. These motifs do have reason to be relied upon with some confidence though since occupancy can be predicted from these motifs once other genomic contexts have been accounted for[96].

In total, these analysis show that weakly bound regions are under purifying selection as strong as the highly bound regions. Whether these the selection acting on these regions is actually a consequence of the weak binding of transcription factors is not clear. Transcription factors are known to transitively bind in a nonspecific manner and the presence of these factors could be of no biological consequence[97]. Indeed, the complete lack of differential selection on the binding sites and correlation with occupancy suggests that the binding is not the phenotype under selection. This supports a hypotheses that the it is a general feature of the region that is under selection and the binding is a spandrel[98]. A possible explanation is that the chromatin in all bound regions is under selection to be in an open state, perhaps so that the region will be accessible for regulation at a later time in development. However, there could also be an unknown transcription factor that is actively regulating during this stage and primarily responsible for the open chromatin. If this is case, variants that effect the binding of these unknown actors could causes observable changes in the chromatin state and I explore this possibility in Chapter 3.

2.4 Methods for finding polymorphisms

Raw short Illumina reads for public data sets DGRP and DPGP2 were downloaded from the NCBI SRA, accessions SRP000694 and SRP005599 respectively. European *D. melanogaster* data courtesy of Casey Bergman and retrieved from Bergman lab webpage[80]. Additional high quality paired-end libraries were generated and sequenced on a Illumina HiSeq to over 50x coverage for four strains: DGRP208, DGRP324, DGRP437, DGRP505. Libraries were made from adult females and the reads are available on request. Each read set was filtered for reads that failed standard Illumina quality controls and adapters removed with Timmomattic [99] using settings 'ILLUMINACLIP:TruSeq2-SE.fa:2:30:10'.

True positive data sets: Odorant Receptor loci [100] from the same DGRP lines were retrieved from NCBI Genbank [GQ919302 - GQ920615] and *bab* locus sequences[84] from an unrelated North American population from Winters, CA, USA were obtained through personal communication with Ryan Bickel. Each loci was aligned to the reference genome loci with FSA[101], ambiguous alignment regions were fixed by manual inspection.

Quality controlled reads for each genomic library were aligned separately to the dm5 genome from FlyBase [102] with stampy v1.0.21[103, 104] with '–substitutionrate=0.01' using bwa 0.5.9-r16 [105] as a prealigner. Strains with multiple libraries were merged with picard[106] MergeSamFiles into single bam files. Duplicate reads were removed from each bam file with picard MarkDuplicates. Each strain bam file was then preprocessed with GATK IndelRealigner and ReduceReads and all strains were jointly genotyped with GATK UnifiedGenotyper [107]. A preliminary high quality variant set was generated with the high coverage strains using the intersection of GATK and bcftools[108] whole genome hard filtered variant calls; SNP filters QD $<$ 2.0, MQ $<$ 40.0, FS $>$ 60.0, MQRankSum $<$ -12.5, ReadPosRankSum $<$ -8.0 and INDEL filters QD $<$ 2.0, ReadPosRankSum $<$ -20.0, FS $>$ 200.0. The raw variant calls were recalibrated with GATK VariantRecalibrator using the Sanger validated variants as truth and preliminary variants with high confidence. Final variants were accepted if they fit the GATK true positive model and were in the highest specificity category.

2.5 Methods for inferring selection on binding sites

Bound regions were defined using the ChIP-seq footprints from previous experiments [109]. Transcription factor binding sites used matrices from in vitro footprinting [110], one-hybrid assays [111], and SELEX [63]. Thresholds were set to match *in vitro* binding [55] using patser-v3e [112]

Tests for neutrality of binding motifs were performed by encoding the presence and absence of motifs in bound windows and a neutrality score was calculated with custom scripts available on request. The score is defined as:

$$\frac{\pi - \frac{S}{a_i}}{\sqrt{[e_1 S + e_2 S(S-1)]}} \tag{2.1}$$

where S is the number of segregating motifs, π is the mean number of motif differences, a_i is the sum of the harmonic series of i, the number of samples, and e_1, e_2 are normalizing parameters defined by Tajima[113].

2.6 Figures

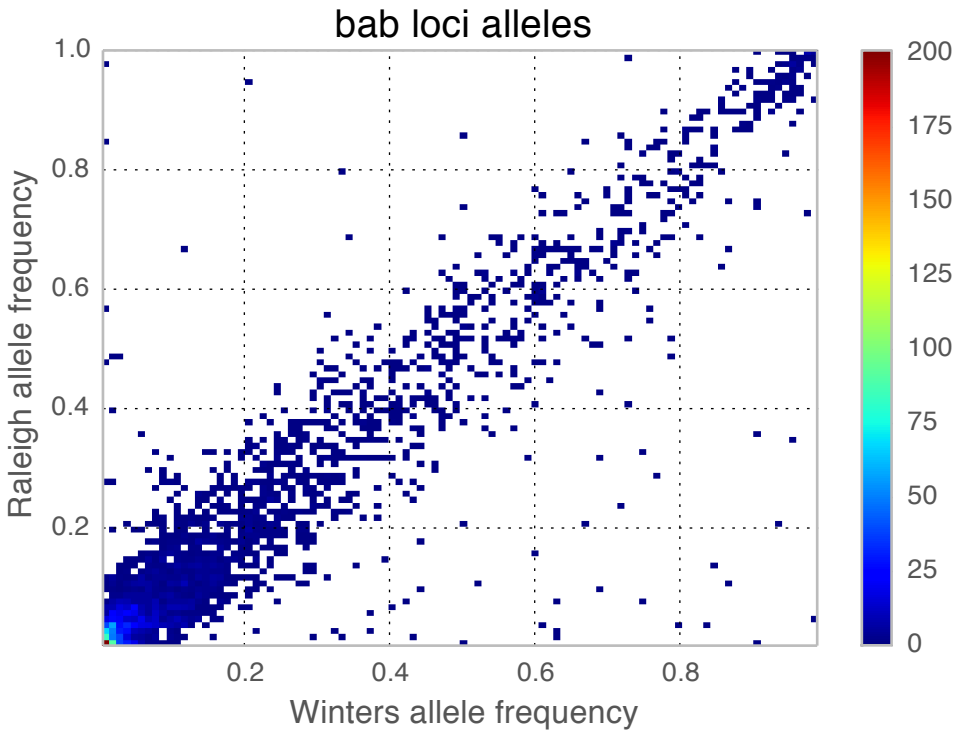

Figure 2.1: Histogram of allele frequencies from two populations of N. American strains. The alleles found from the Raleigh population measured with short reads correlate highly with alleles in the Winters population from Sanger sequencing

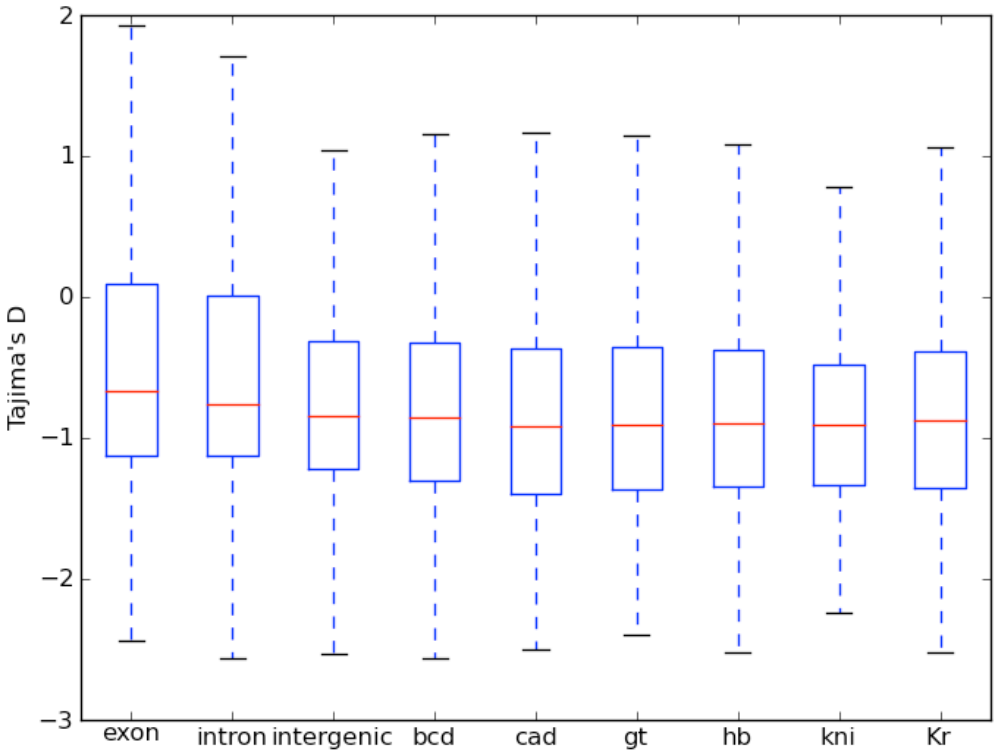

Figure 2.2: Distribution of Tajima's D values for classes of *D. melanogaster* DNA. Bar are given with whiskers indicating two standard errors. Transcription factors regions for each factor were appended together.

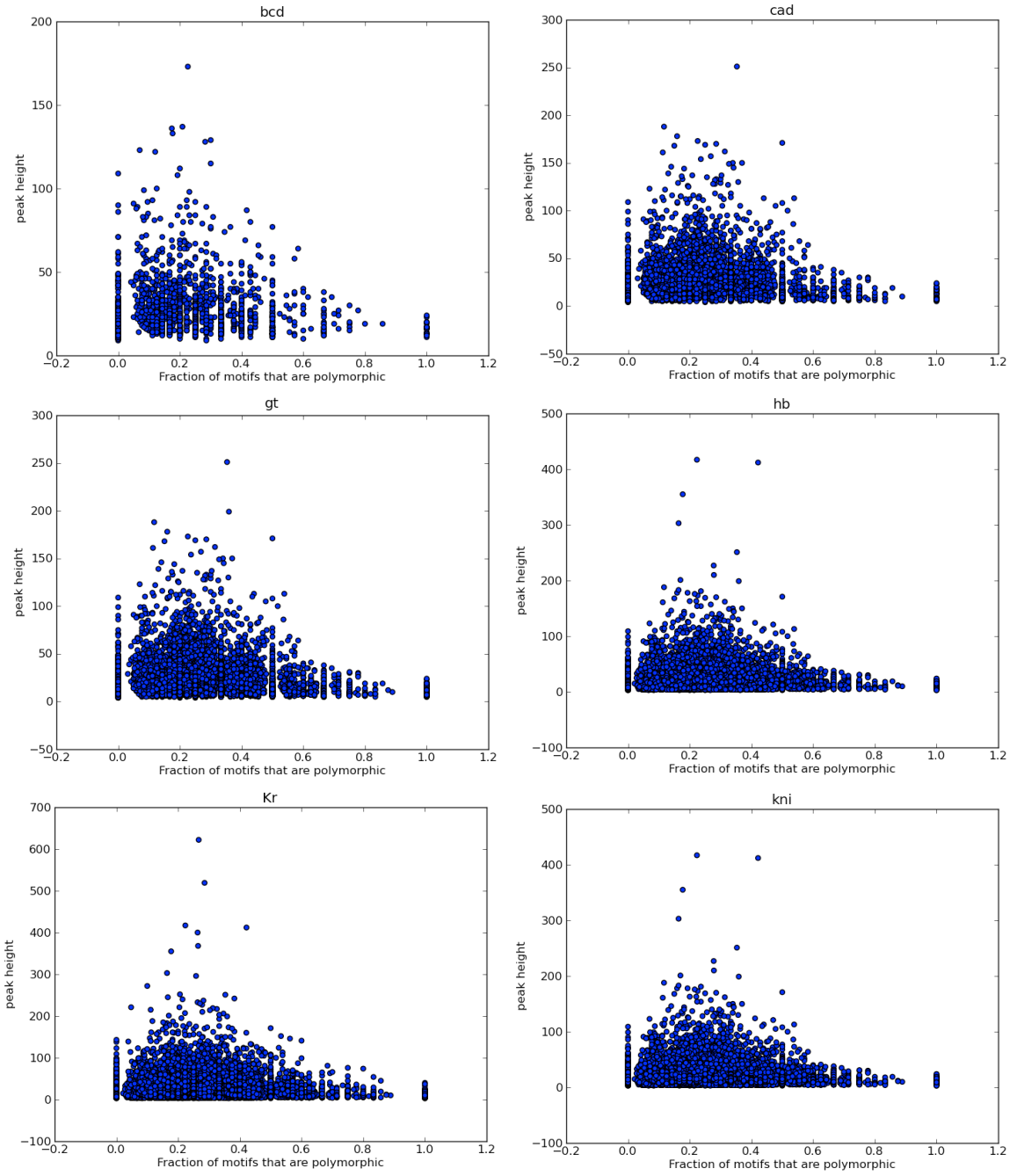

Figure 2.3: Strength of AP transcription factors binding in relationship to population motif diversity in bound regions. Strength of binding is measured by the maximum fragment depth of the region and motifs were considered as binary absent or present if the motif passed a p-value threshold.

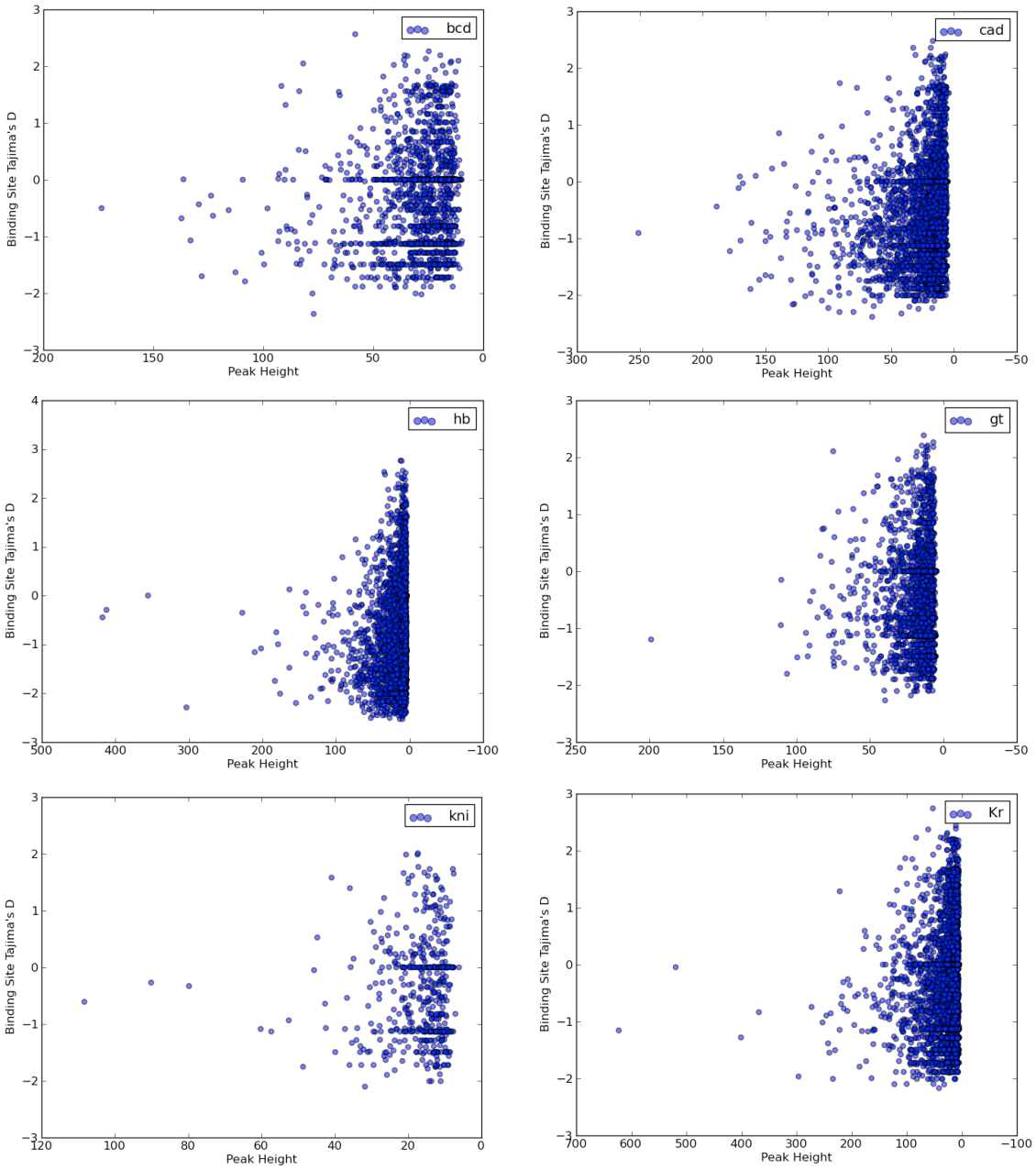

Figure 2.4: Binding of AP transcription factors in relationship of excess low frequency motifs D as calculated in equation 2.1

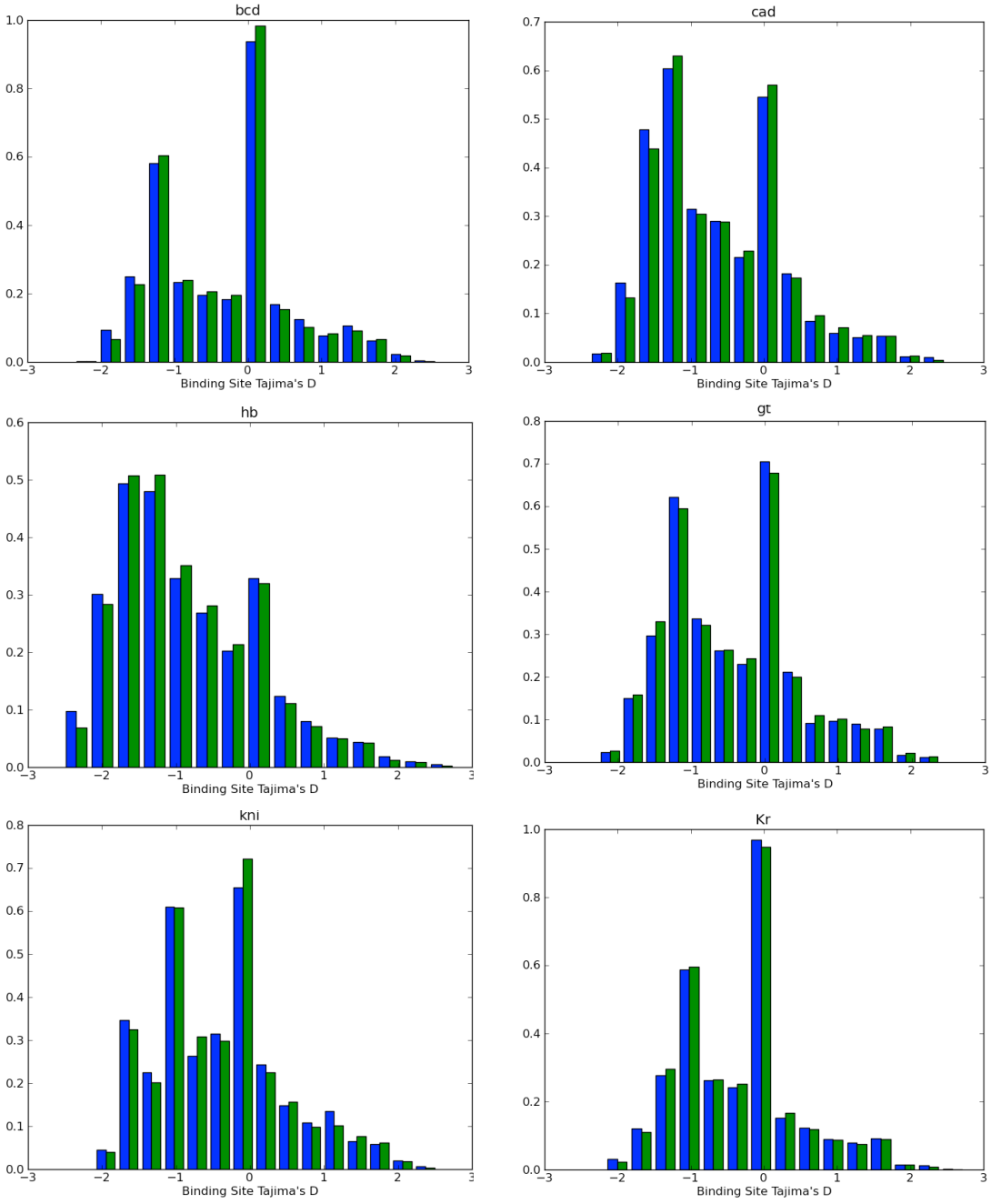

Figure 2.5: Transcription factor binding sites are under similar selective pressures as the entire *cis*-regulatory module. Each motif was shuffled by column followed by row and scored against the bound regions using the same thresholds as the canonical motif. The score distributions were compared with Wilcoxon rank-sum test and no significant differences were found.

Chapter 3

Early Embryo Chromatin Landscape

The foundations of Metazoan body plans are laid during early embryogenesis. Combinatorial expression of transcription factors over time progressively refine cellular domains and impart eventual tissue specificity. Expression of these genes is proceeded by the integration of regulatory proteins binding to and/or creating regions of DNA released from the tightly wrapped chromatin. Regulation of all but a few well studied genes is poorly understood and the variation in the accessible regions of DNA within *Drosophila melanogaster* has not been explored previously. I used DNaseI hypersensitivity to define this accessibility landscape and find that is it highly constrained between strains during embryogenesis. Within the limits of this constraint, there are DNA motifs that are associated with quantitative variation in the level of accessibility which are linked in transcription factors. The overall stability of the accessible region combined with sequence variation within those regions leads to a model of chromatin regulation defined by the combination of transcription factors.

3.1 Introduction

DNA packaging and nucleosomes Metazoan development is process in which a single cell divides and differentiates into a diversity of cell types mediated by a set of interactions between proteins and DNA that leads to a complex network of gene expression patterns[114, 115]. The genetic regulatory elements of this process include short DNA sequences both proximal and distal to the regulated gene that will bind to proteins and affect a multitude of processes such as to enhance[116, 117], repress or silence gene expression[118, 119, 120]. Additional regulatory elements act at the promoters of genes[121], at the far edges to insulate inappropriate regulation[122], or to tether multiple regulatory elements together[123, 124]. These regulatory elements are regulated themselves by sequestering these regions in tightly wrapped and inaccessible protein complexes called nucleosomes or displaying these sequences in nucleosome free regions[125, 126, 127]. In addition, these regulatory regions are often found together allowing for combinatorial regulation and interactions between the bound proteins[128, 129].

During development, the chromatin DNA/nucleosome complexes at the promoter are

remodeled to regulate gene expression. The histones in the nucleosome are modified at the promoter to include the H2A.Z variant [130] which assists in recruiting RNA polymerase[131]. Chromatin modifications also occur in other regions as transcription factors are recruited to *cis*-regulatory regions. Several examples in embryogenesis and cellular differentiation show that transcriptions factors bind in an orderly fashion, starting with a pioneering transcription factor[132] which recruits histone modifying complexes[133, 126] and decreases the affinity of the nucleosomes to DNA[134]. The relaxed chromatin to then accessible to additional transcription factors which can bind and enhance gene expression. Conversely, repressors and insulators of gene regulation can recruit Polycomb complexes[135, 136] that tighten the chromatin and prevent further transcription factor binding[137, 138].

The two dominant models of transcription factor binding in *cis*-regulatory modules are the "billboard" [39, 30, 139] model and the enhanceosome [38, 140] which differ in the interactions of the bound transcription factors. The enhanceosome model fits situations where a collection of transcription factor binding sites regulate synergistically and a single mutation in one the binding sites will obliterate the functional activity as seen in the IFN-*beta* gene regulation[38] and some dorsal-ventral regulators[140]. The billboard model on the other hand describes the display of a collection of semi-independent binding sites where regulation is mediated the binding of a subset of these site, allowing for some flexibility in the composition of the CRM, a model which more accurately describes the enhancers of *eve*. Both models require the depletion of nucleosomes for transcription factor binding.

Previously Bradley et al. examined the binding of a set of transcription factors regulating anterior-posterior patterning in the closely related species *D. melanogaster* and *D. yakuba*[109]. The evolutionary distance between these two Dipteran species is small enough that 95-98% of the transcription factor binding sites are conserved between them in contrast to the 11-59% conserved between humans and mice[141] or the 22-34%[142] and 7-42%[143] conserved between yeast species. While these transcription factor binding sites contained divergent positions, the predictive value of these changes where limited and the dominate predictor quantitative changes in transcription factor binding was the presence or absence of a CAGGTAG kmer known to be a binding site of Zelda[144], an activator of the early zygotic genome[145]. This kmer was associated with the joint increase or decrease in binding of several transcription factors simultaneously suggesting that the changes in the accessibility of the chromatin mediated divergence between species.

When using comparative genomics, we need to consider the amount and context of the divergence we observe. Observations of conservation in comparisons between distantly related species in which genomic sequences have accumulated large amount of divergence allows for identification of the core elements required for a particular process. However, when trying to determine the consequences of variations in a process, comparisons between divergent species is complicated by the large number of variations, leading to difficulty determining which variations are of interest. A further consideration is that the comparison between species that are diverged could result in erroneous conclusions about variation if the process has diverged in function. In contrast, variations between very closely related species or even variation within a species occur in a more highly conserved background, increasing

the power to link the consequences of a polymorphism to a change in the phenotype. The trade-off is that there are much fewer number of variations which we can observe. While *D. melanogaster* is relatively closely related to *D. yakuba* and contains extensive changes in binding but we are unable to predict the underlying sequence changes that influence these changes due to the large number of polymorphisms between species. I will avoid this issue by using a within species comparison.

Ideally, I would like to characterize all DNA-protein interactions in all of the sequenced *D. melanogaster* strains, and thereby directly identify specific protein-DNA interactions that vary in the population. However, such an experiment is not practical in either time or resources. It would be feasible to carry out a more limited version of this experiment - looking at a few factors in a subset of strains. But this is likely to have low yield. So I have designed an alternative strategy.

In this work, I identified regions of the *D. melanogaster* genome whose chromatin accessibility in blastoderm embryos varies among the sequenced strains. There is an emerging body of evidence that nucleosome occupancy and transcription factor binding are inversely correlated[146], and that changes in transcription factor binding are associated with changed in chromatin accessibility[109]. Thus, if a particular protein-DNA interaction varies among strains, it will be associated with variable nucleosome occupancy and chromatin accessibility. A genome-wide screen for variable chromatin accessibility revealed polymorphic chromatin accessibility generically for all factors active in the embryo. From these regions of differential accessibility I identified drivers of chromatin state change within *D. melanogaster*. These results will inform future models of transcriptional regulation and evolution.

3.2 Results

An embryo undergoing early development is a very dynamic system, with rapid cellular divisions and the activation and refinement of zygotic gene expression. Measurements of the embryo during this active stage must take special care to ensure that samples of embryos are of comparable states. For example, the contamination of 5% stage 11 embryos with 40,000 nuclei per embryo [147] into a otherwise pure collection of stage 5 embryos with 6,000 nuclei per embryo [148] will represent contamination of 25.6% of the nuclei. Taking extreme measures such as hand sorting embryos to ensure purity has been used to discover fine stage specific differences in transcription factor binding at a mitotic cycle resolution [96]. The chromatin accessibility landscape is tightly correlated with the proximal transcription factor binding and exclusion of nucleosomes [149, 150, 151, 132] and therefore should mirror the dynamic binding of transcription factors. Previous surveys of DNaseI hypersensitivity (DHS) across longer time windows in *D. melanogaster* development sampling stages 5 to 14, found that many DHS regions are stage specific, are linked to stage specific expression of nearby genes, and are associated with motifs presumed to also be bound in a stage specific manner[152].

DNaseI hypersensitivity is an assay based on the steric blockage of cleavage capacities

of the active site of the deoxyribonulcease by DNA-protein complexes [153, 154, 155]. DNA wrapped up and bound by nucleosomes are protected from the cleavage of DNases while regions that are nucleosome free are susceptible to fragmentation [154]. Chromatin that has been digested to a degree that results in several hundred base-pair fragments followed by next-generation sequencing results in the ability to determine at base-pair resolution the accessibility of the chromatin [156], a substantial improvement over previous DNaseI microarrays [157]. Since both ends of the fragment should be cleaved by DNaseI and cleavage is a stochastic process, the degree of the digestion and therefore the average fragment size will lead to different distributions of cleavage sites [158].

For those reasons, the chromatin and DNaseI treatment comparisons were collected in a pairwise fashion where each strain was mixed at the embryo stage as shown in Figure 3.1. This design attempts to equalize all DNaseI treatments so that any differences in recovered fragments reflects the genetic differences between strains and sample input instead of staging problems or variations in fragmentation. When collecting embryos for each strain two at a time, I was able to ensure that each fly population was fed at the same time, retained embryos were cleared at the same time and embryos were collected and aged simultaneously. After collecting embryos, equal weights of each strain were pooled and fragmented in the same tube. By comparing the combinations of sequencing depths of pairwise pooled alleles, accurate effects of each pair of alleles can be determined [159].

The chromatin accessibility landscape is very consistent between samples as seen at the *eve* locus in Figure 3.2. In fact, quantifying the similarity by finding windows of increased DNaseI accessibility with MACS2 for each sample showed that 93-97% of the top 20% of regions in any sample are also are accessible in the top 20% of all other samples and those that are not found in the top 20% are all found in the top 60%. These highly conserved regions are also those with the highest sensitivity. The q-values in these variable regions are much weaker than the conserved regions (mean conserved q-value = 62.02, mean variable q-values = 39.59, $p = 1.04 \times 10^{-11}$).

This shared conservation of chromatin accessibility between strains of *D. melanogaster* is linked to the overlapping binding of early embryonic transcription factors and other signals of chromatin accessibility such as RNA PolII and TFIIB. Comparing the DNase accessibility regions in these strains to the binding of 21 early embryonic transcription factors shows that 67% of the accessible regions overlap at least of the these factor by 50%, comparable to previous findings[152]. While general accessibility of a region can be classified by the amount of DNaseI activity, with nucleosome free regions allowing access to the DNA, DNaseI is not the only protein that will be binding to the free DNA. These open regions are biologically interesting due to the binding of these factor and they too have an influence on the accessibility of DNaseI but on a much smaller scale, often only blocking DNaseI access at the short DNA binding site. With high enough sensitivity, these DNase footprints will be discoverable on a genome scale just as DNaseI footprints were found by isolating single fragments with gel assays[160]. The signal of these footprints in surrounding accessible regions will the steep and local decrease in accessibility corresponding the binding site. The clarity of the these digital footprints will be mediated by the amount of occupancy of the bound protein and

homogeneity of the sample. Complex mixtures of cell types with diversity of expressed transcription factors or concentration will obscure the signature of these footprints. While the *D. melanogaster* stage 5 embryo is indeed a complex mixture of transcription factors without cellular compartmentalization, there are several transcription factors and coactivators that are ubiquitously expressed. Zelda is a maternally deposited protein that is distributed uniformly throughout the embryo and I used the bound regions defined by ChIP-seq [96] to find ZLD footprints in the overlap of ZLD bound and DNase accessible sequences. Figure 3.4 shows the aggregate footprint in these regions.

Although the genome wide accessibility landscape is highly constrained between strain, it is possible that quantitative variation in the accessibility can be explained by variation at the sequence level. With a small sample size, finding individual quantitative trait loci that explain any of the variance in accessibility is infeasible but by examining each DHS for genome wide trends, general trends can be elucidated. In particular, comparisons between species suggests that kmer composition in CRMs can be linked to quantitative variation in transcription factor binding and DNaseI hypersensitivity [109]. I examined each DNaseI hypersensitive region for variations and the differential accessibility was scored for each kmer overlapping a variant. The effect on scores were thresholded at a 0.05 FDR yielding 14 kmers with at least an 1.1 absolute mean differential effect on accessibility. A complete list of the kmers is shown in Table 3.1. These kmers are observed to have log_2 fold effects on accessibility (Figure 3.6) that span several orders of magnitude.

The kmers that are associated with a change in DNaseI accessibility are nearly all binding sites for known transcriptional regulators. The kmers associated with an increase is accessibility include the binding sites for Ci, an transcription activator required for segment polarity[161], Broad which is primarily known for its role in metamorphosis[162] but is also an activator male specific genes in embryos[163], tramtrack represses several pair-rule genes[164] and is also found proximal to promoters in RNA PolII complexes[165], sugarbabe is a repressor of the fat catabolism pathway[166], Blimp-1 is a master regulator of cells destined for the trachial system [167], HKB is a terminal gap activator in the tail and repressor in the head[168], and Zelda is a well known activator of embryonic gene expression[169, 170, 171, 172, 173, 174, 96, 175, 145]. The kmers associated with decreases in accessibility include the motif for the gap repressor Krupple[], lola ,a repressor of Spire[176, 177], tinman, the activator of mesoderm fates[178] and ovo which is also an activator and repressor[179, 180, 181] but the remainder are less well characterized. Ets98B is a regulator of germ cell migration with an ETS DNA binding domain but the role it plays in regulating gene expression is unknown.[182, 183] while the kmer CGTAGTA/TACTACG which has the second highest association with a decrease in accessibility has no known protein partner.

Measuring the difference in expected accessibility between strains conditioned on the effect of a SNP which is differentially accessible shows a local effect on the surrounding chromatin, Figure 3.7. There is a difference between the regions that are increasing in accessibility verses decreasing in that regions around a SNP which increased accessibility have an consistently similar reaction while there is no significant deviation from expected

accessibility around SNP with decreased accessibility.

The accessible regions are enriched for function during this stage of the life cycle and targets of selection. The function of regions with variable accessibility is unknown but examination of the population diversity, π, in regions with conserved accessible verses those that are polymorphic with in the samples may distinguish these states. A reduction in diversity is found in variations in all accessible regions, regardless of conservation of chromatin state within the population as shown in Figure 3.8. The reduction in diversity for both classes is significant compared to the diversity in matching windows in the adjacent regions, $p = 1.57 \times 10^{-6}$ using a Wilcoxon rank-sum test. These diversity rates were calculated using combined substitutions and indels and separately with equivalent results.

3.3 Discussion

This study is the first exploration of DNaseI hypersensitivity variation in developing tissues from a fully genotyped strains. While extensive resources have been invested in the ENCODE project to determine chromatin states in hundreds of cell lines derived from a variety of tissues, very few of these have been genotyped. A collection of 70 Yoruba lymphoblastoid cell lines have been genotyped and assayed for chromatin accessibility[127] but these are immortalized lines and have idiosyncrasies that are not found in primary tissues[184]. The only previous experiment to reach this level of closeness to *in vivo* measurements is a study of terminally differentiated erythroblasts collected directly from eight inbred and genotyped strains of mice[185]. Other experiments include yeast[155], seedling and callus tissues of rice [186] and leaf and flower tissues in Arabidopsis[187] and a *D. melanogaster* time series of early embryogenesis[152].

The overall conservation of DHS regions is higher than any other study thus far. The mean DNaseI correlation between the 70 Yoruban lines is 0.7 with a range of 0.35 to 0.94[127] , while the correlation between any two *D. melanogaster* samples was never less that 0.86. Comparing *D. melanogaster* peak variability which is completely conserved at a qualitative level to those found in mouse mature primary erythroid cells where 14% showed gains and losses between stains again shows that that embryonic accessibility landscape is under far more regulation than terminally differentiated cells[185]. The erythroid comparison may not be completely fair due to a large amount of the changes found that many variable DHS regions were due to nearby RNA expression differences and not due to any *cis* variation with the region [185]. Regardless of expression induced changes, the qualitative variability in the accessibility in non-embryonic cells is far higher than embryonic.

This early embryonic accessibility landscape is also more conserved within a stage than between stages of development. Previous comparisons of chromatin accessibility at five stages of *D. melanogaster* (stages 5, 9, 10, 11, and 14) spanning 8 hours of development found a dynamic reorganization. Even between stage 5 and the closest stage 9, there was a gain or loss of 20% of the accessible regions and the correlation between the youngest and oldest stages dropped to 0.52[152]. Comparisons of chromatin accessibility during *D. melanogaster*

larval stages using FAIRE-seq[188] do find that imaginal discs of thorasic appendages at the same stage are more similar to each other than to the central nervous system. The majority of the larval tissues are differentiated at this stage and if differences between strains were examines, they may be more similar in character to the mouse erythroids.

While the overall accessibility landscape is relatively stable, the quantitative changes at alleles between strains are reproducible and represent subtle but consistent regulatory effects. With a small sample, I was able to find general genome wide trends that implicate a combination of transcription factors, including the pioneering factor ZLD, in enhancing and repressing DNaseI accessibility. While ZLD was expected in this set of motifs, it striking what is missing in this set. There are no motifs associated with insulator factors such as CTCF, CP190, BEAF-32, Su(Hw), Mod-(mdg4) and GAF which previous studies have found to be present at the boundaries of *cis*-regulatory modules[189]. Also missing are kmers associated with the Male Specific Lethal (MSL) complex[190] that acetylates histones and increases transcription on the X chromosome in males.This suggests that there is strong selection against the variable presence of these motifs, again even in the weakly accessible regions.

The number of kmers significantly associated with a change in chromatin accessibility is surprisingly low considering potential effects that the loss of a transcription factor binding site can have [191] and supports the predominance of the billboard model of enhancer activity that maintains open enhancers and buffers combinations of variable affinity transcription factor binding sites. The strong conservation of DNase accessible regions also lends support to the interpretation of weakly bound regions having at least a neutral role in development and the suppression of diversity in accessible regions suggests that these regions are under selection for a role at some other point in development. This adds further evidence that weak transcription factor binding are likely to be a byproduct of setting up open chromatin and are nonfunctional at this stage. There are several caveats to these conclusions including the unknown effect that these weakly bound and variable accessible regions have on expression and this is examined in Chapter 4

3.4 Methods of Embryo Collection and Treatment

Each strain of *D. melanogaster* was maintained at in vials until expansion for collection. During expansion, 25-30 flies were transferred into fresh bottles and pushed every two days for six days and then discarded. This process was repeated for each generation until the population reached approximately 150ml of flies at which point the newly hatched flies were transferred to population cages and maintained until maturity. Embryos were collected on yeast/molasses plates for 50min and aged for 2 hours and 15-25 min in order to obtain stage 5 embryos, verified by microscopic inspection.

DNase I Treatment of Nuclei Isolated from Drosophila Embryos

Modified from Xiao-yong Li/Eisen lab

Stock Buffers and Solutions

3 M KCl

1M CaCl2
0.5 M EDTA
0.5 M EGTA
1M TrisCl, pH 8.0
10% NP40
0.5 M spermine
0.5 M spermidine
10 mg/ml RNase
20% SDS
10 U/ul DNase I

Buffer A (500 ml)

Final Concentration	Stock conc.	Amount to add
M15 mM Tris-Cl, pH 8.0	1 M	7.5 mL
15 mM NaCl	5 M	1.5 mL
60 mM KCl	3 M	10 mL
1 mM EDTA, pH 8.0	0.5 M	1 mL
0.5 mM EGTA, pH 8.0	0.25 M	1 mL
0.5 mM Spermidine	1 M	250 μl
H2O, sterile		478.75 ml

Filter (0.22 um), store at store at 4 °C

Stop Buffer (500 ml)

Final Concentration	Stock conc.	Amount to add
50 mM Tris-Cl, pH 8.0	1 M	25 mL
100 mM NaCl	5 M	10 mL
0.10 % SDS	20%	2.5 mL
100 mM EDTA, pH 8.0	0.5 M	100 mL
H2O, sterile		362.5 ml

Filter (0.22 um), store at room temperature (or 4 °C).

Working solution

Buffer A + spermine: 5 ml/g embryos

Amount to add	Stock
25.0 ml	buffer A
7.5 μl	0.5 M spermine(final 0.15 mM)
25 ml	

Set on ice. Add DTT to 0.5 mM, and PMSF to 1 mM right before use

Buffer A: need 10 ml/g embryos
Chill on ice
DNase I 1X Digestion Buffer: 5 mL/g stage 5 embryos

Amount to add	Stock	Final
25 ml		buffer A
0.15 mL	1M CaCl2	6 mM
0.375 mL	5M NaCl	75 mM
25 ml		

Stop Buffer*: 2.5 ml/g stage 5 embryos

Amount to add	Stock	Final
25 ml	Stop solution	
25 μl	10 mg/ml RNase A	10 ug/ml
25 μl	1 M Spermidine	1 mM
15 μl	0.5 M Spermine	0.3 mM
25 ml		

*: if precipitate forms, heat to 55 °C to bring back to solution
10% NP40
DNase I aliquot
Other setups Tubes for aliquoting DNase I solution, 1 per reaction Tubes for Nuclei digestion, 1 per reaction Dounce homogenizer (5 ml), with pestle A, and B Mira-cloth wet in sterile ddH2O.

I. Nuclei Prep

Prior to Isolation of Nuclei:
1. Prepare stop solution, as above, prewar it in a 37 °C water bath.
2. Prepare fresh 1XDNase I Buffer. Transfer 0.5 ml/0.1g embryo to a new 15 ml tube, and put in a 37 °C water bath, after known amount of embryos to be treated.

Embryo collection and nuclei isolation (stage 5: for each 0.5 g, carry out 2.5 ml digestion; stage 9, 0.3g in 2.5 ml digestion)
1. Collect embryos from population cages, allow the embryos to develop to desired developmental stage; harvest and dechorinate the embryos.
2. Resuspend in 5 ml cold (buffer A + spermine), + 0.5 mM DTT, + 1 mM PMSF) per gram of embryos
3. Homogenize with a dounce homogenizer, pestle A, 2-3 strokes
4. Pass the homogenate through Miracloth pre-wetted with cold H2O, into another clean dounce homogenizer.
5. Further homogenize using a dounce homogenizer, pestle B, 5-6 strokes. Transfer homogenate to microcentrifuge tubes.
6. Add 10% NP-40 drop-wise to final concentration of 0.5%, mix well.

7. Spin 1.5 ml aliquots in microcentrifuge at 3 krpm at 4 °C for 3 min.
8. Re-suspend gently each nuclei pellet in 1 ml of fresh buffer A. (if not sure about nuclei concentration, take a small aliquot and set aside for nuclei count).
9. Spin the nuclei down at 3 krpm at 4 °C for 3 min.
10. Repeat steps 8-9.
11. Re-suspend the nuclei in each tube into 25 ul remaining buffer.
12. Transfer nuclei to a 15 ml tube

Nuclei Count
1. dilute nuclei 1:10, Count nuclei on the hemacytometer while pelleting nuclei.(count can be done in buffer A)

II. DNase I treatment
Use Roche DNase I at 10,20,40, or 60 units/ml (non-Mg++ buffer). With small amount of embryos, carry out just one DNase I digestion a time, repeat with different amount of DNase I from different harvest. (most likely the 1st 2nd concentration will produce the right digestion). Now, Stop Buffer and 1x DNase I buffer (already aliquoted) should have been equilibrated to 37 °C, (>30 minutes.)
1.For each 0.1 g embryos, or 0.5 ml DNase I digestion buffer, add 0.5, 1, 2, 3 ul (10 units/μl) DNase I enzymes to each of the equilibrated tube containing DNase I buffer.
2. Mix thoroughly by pipetting.
3. Place tubes with nuclei pellets in 37 °C water bath and allow temperature to equilibrate for 1 minute at 37 °C.
4. Re-suspend nuclei pellet with 1x DNase I buffer plus enzyme, pipette several times gently to ensure homogenous suspension.
5. Incubate for 3 minutes at 37 °C.
6. Add equal volume of stop buffer to each reaction tube, mix by inverting several times, transfer tubes to 55 °C water bath.
7. Allow digestion tubes to incubate at 55 °C for 15 minutes, then for each ml of sample, add 2.5 *mu*l 10 mg/ml proteinase K.
8. Allow digestion with Proteinase K to continue overnight (min. 16 hr.) at 55 °C.
9. Store at 4 °C.

3.5 Methods of Comparing DNase-seq

Mapping and snp calling
DNaseI libraries were prepared using the Illumina V2 protocols and sequenced at the QB3 Vincent J. Coates Genomics Sequencing Laboratory. The resulting fragments were trimmed with Timmomattic [99] with settings 'illuminaclip:TruSeq2-SE.fa:2:30:10'. The trimmed reads were then aligned to the dm5.54 genome from FlyBase [102] with stampy v1.0.21 (r1654) [103] with '–substitutionrate=0.01' using bwa 0.5.9-r16 [105] as a prealigner. The

mapped reads were then realigned against the known variants from Chapter 2 with GATK 2.4-9-g532efad [107]. For each pairwise mixture, the allelic depths were extracted using GATK against the known variants for those strains and compared to the expected mixture proportions from the genotypes of the mixed strains with a binomial test. Mixtures proportions were estimated from genotype ratios of private homozygous alelles with coverage between 10-25 to avoid random biases in low coverage regions and DNase specific biases in high coverage regions.

DNaseI Hypersensitive Regions

DNaseI hypersensitive regions were called using MACS2[192] with a q-value threshhold of 0.02. Other methods for determining accessible regions such as ZINBA[193] and fseq[194] gave similar results.

Kmer association

Kmers associations were calculated assigning a score to each instance of a kmer overlapping variants A and a with read coverage D_A and D_a where the kmers containing the overrepresented variant are given a positive score:

$$S = \log_2 \frac{max(D_A, D_a) + 1}{min(D_A, D_a) + 1}$$

and the kmers containing the underrepresented variant are assigned a negative score:

$$S = -\log_2 \frac{max(D_A, D_a) + 1}{min(D_A, D_a) + 1}$$

The list of kmers was consolidated by collapsing each kmer with its reverse complement and assigning pair of values to the lexicographically larger kmer. An FDR was calculated by permuting the depths of the alleles across the genome [195]. Kmers were matched to known *Drosophila* transcription factor motifs with the TOMTOM motif comparison server version 4.9.1 [196]

3.6 Figures

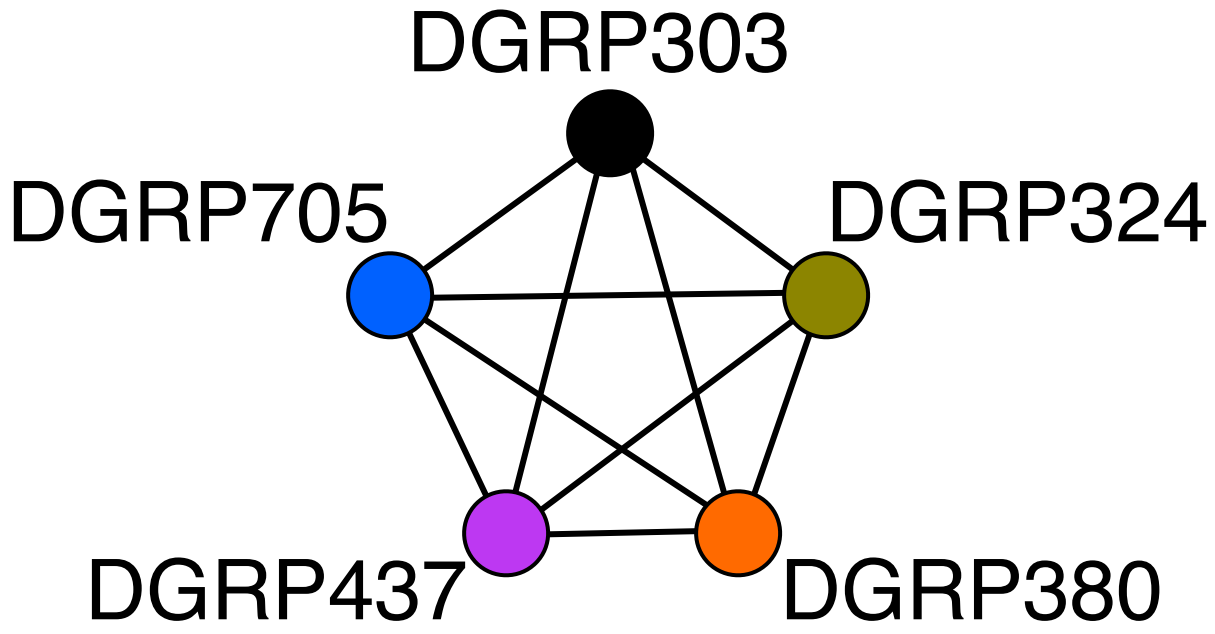

Figure 3.1: DGRP strains used for mixture combinations. Each strain was mixed with all others at least once and biological replicates for DGRP705 mixed with DGRP380 and DGRP437 were collected and sequenced.

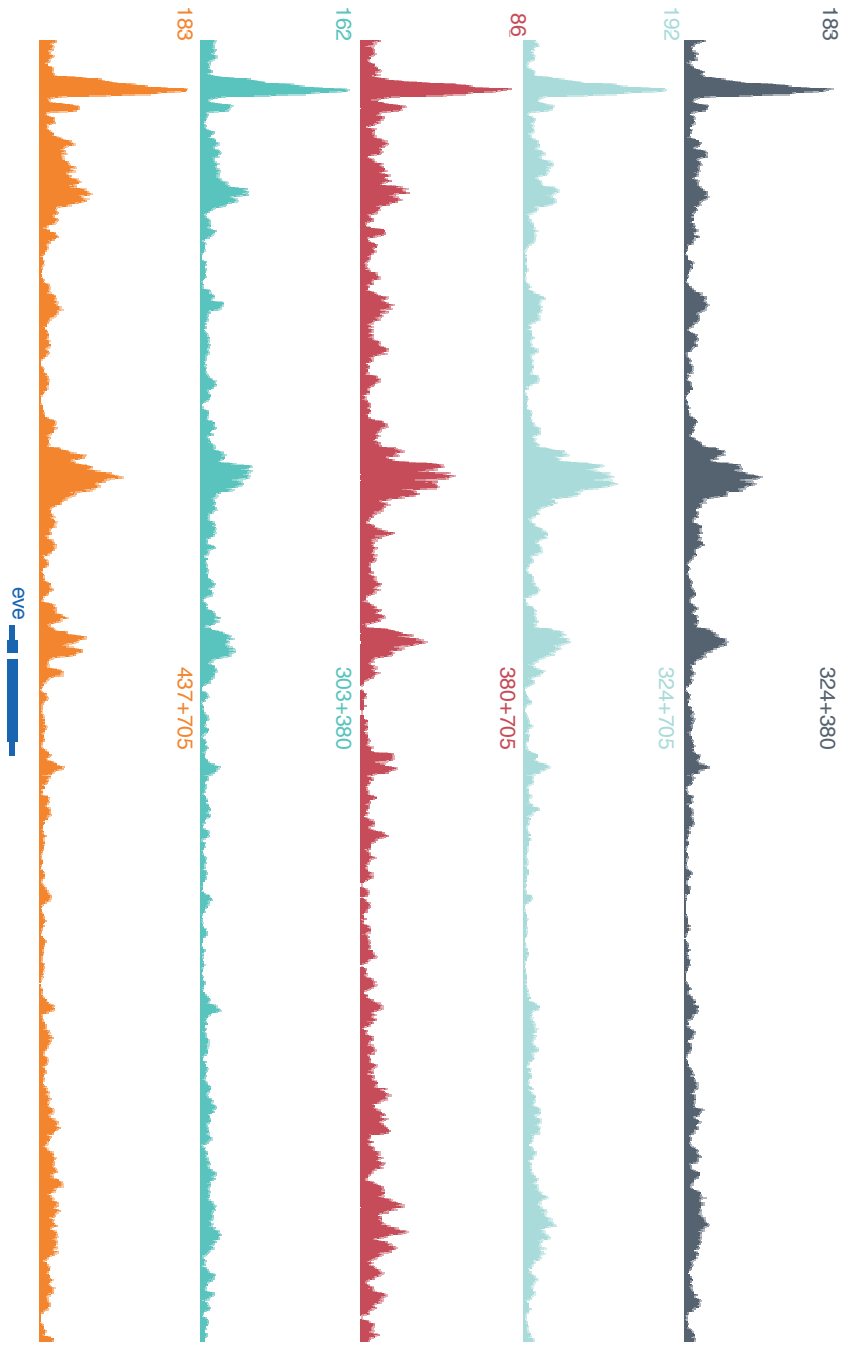

Figure 3.2: DNase landscape at *eve*. A subset of the strain mixtures shows the range of accessibilities and general consistency between strains.

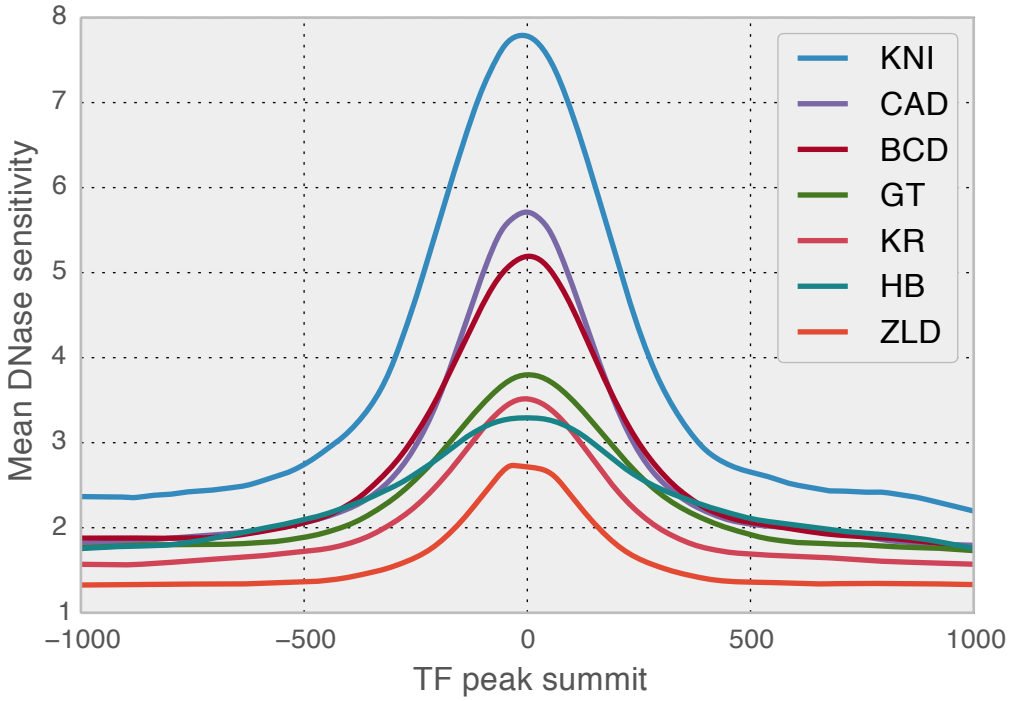

Figure 3.3: DNase signal at TF peaks from stage 5 embryos.

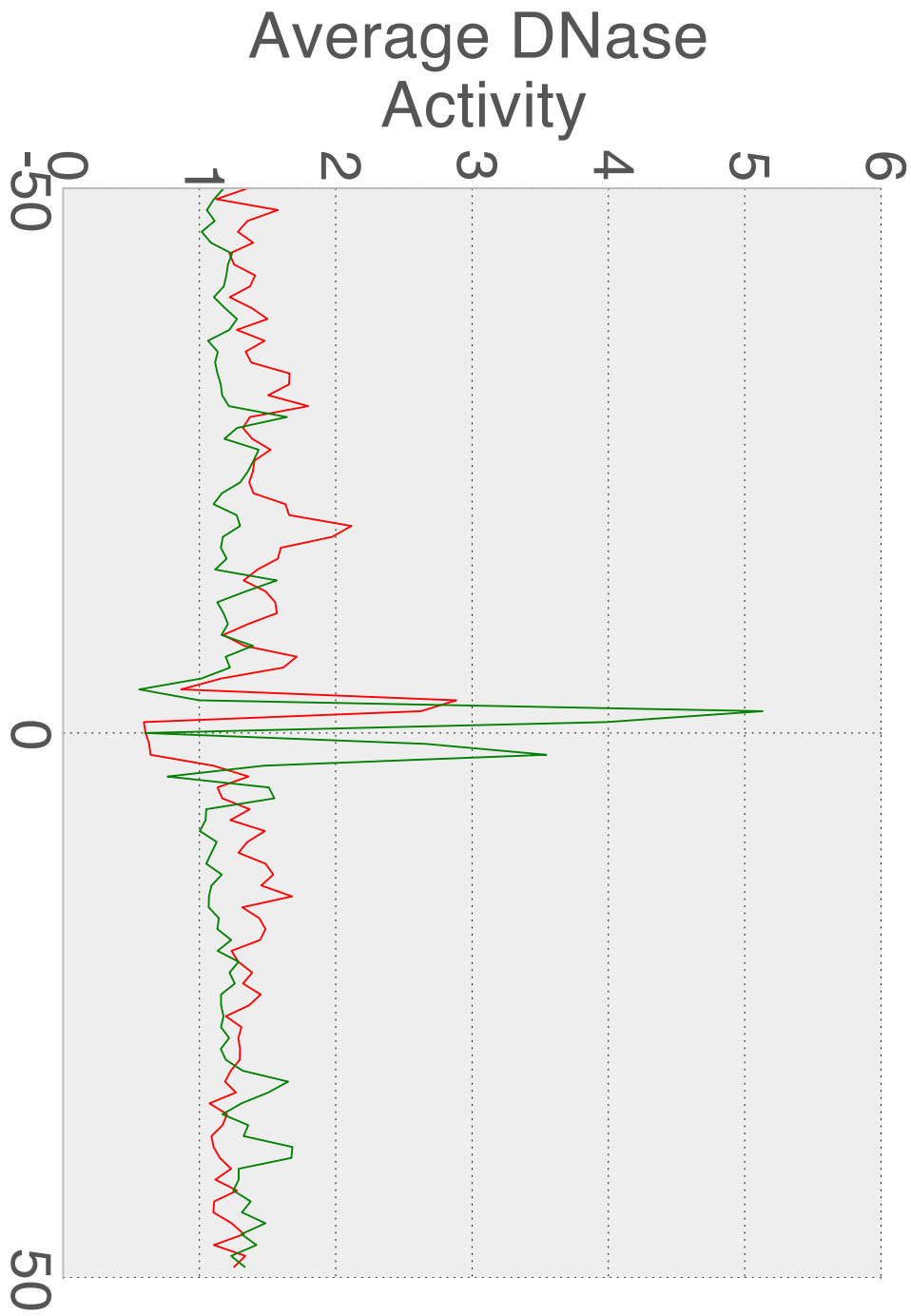

Figure 3.4: DNase I cleavage (per nucleotide per 1M reads) at ChIP-seq identified ZLD binding sites. Red line is the average cuts on the positive strand and green represents DNase cuts mapping to the negative strand.

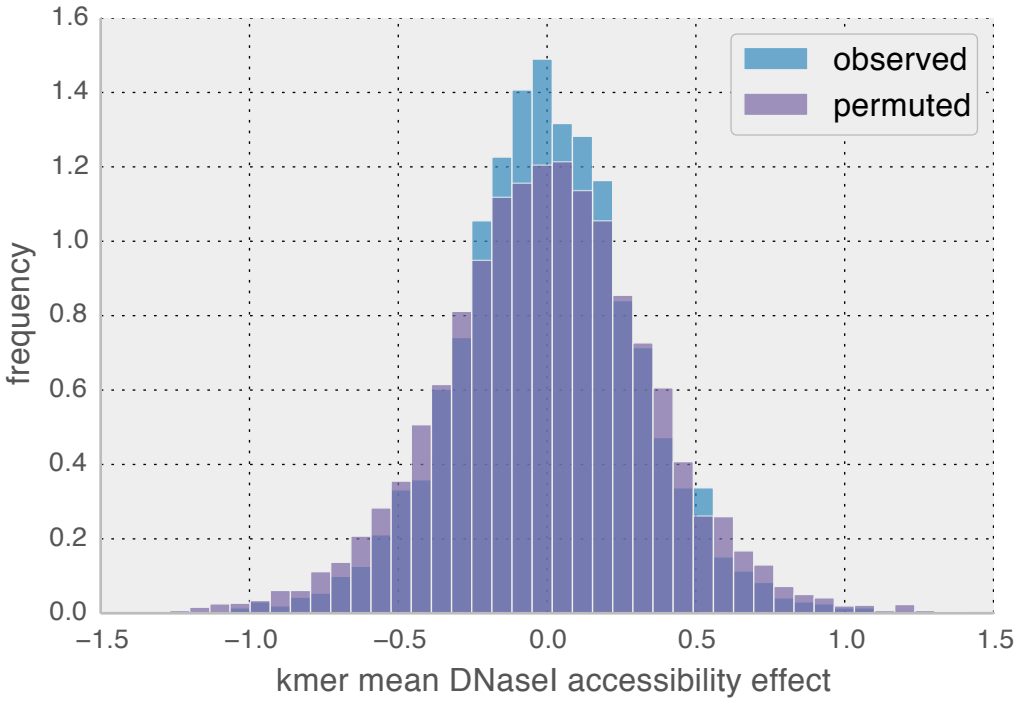

Figure 3.5: Kmer effects on chromatin accessibility

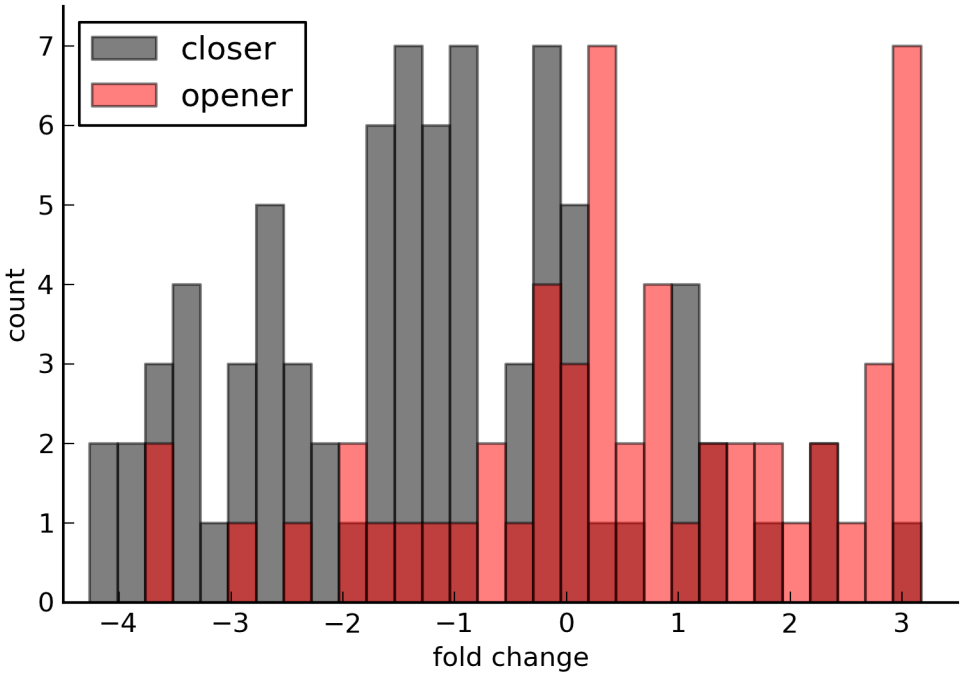

Figure 3.6: Fold change in accessibility at kmers with largest mean effects on chromatin state.

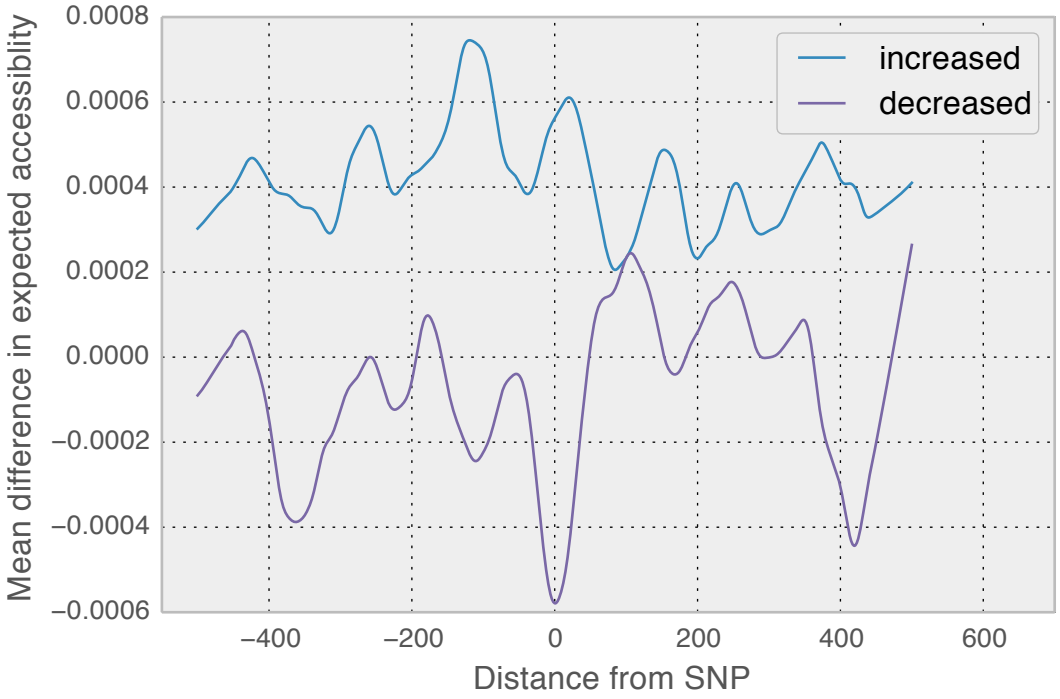

Figure 3.7: Change in accessibility around differentially accessibility SNPs. The difference in expected accessibility corrected for mixture ratios around SNPs which deviated significantly from expected ratios.

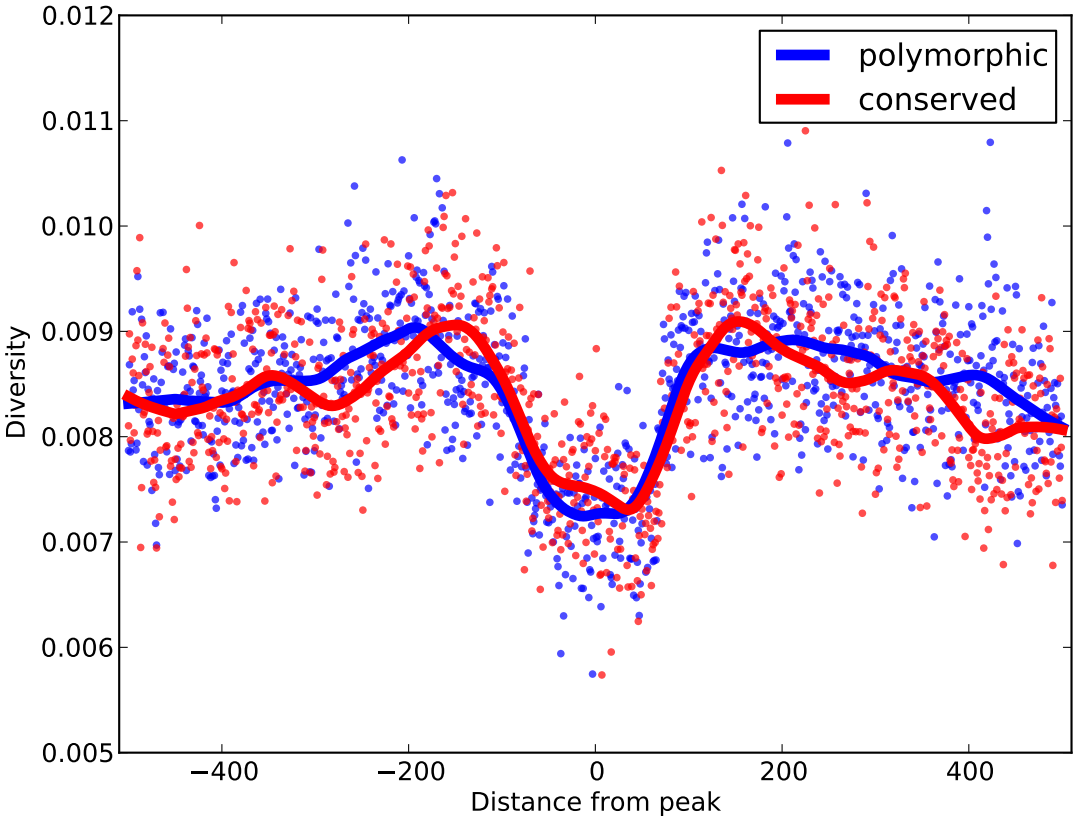

Figure 3.8: Regions with conserved and variable chromatin accessibility are both under purifying selection. The mean pairwise diversity around the peak of chromatin accessible regions were smoothed with LOWESS for each class of chromatin accessible regions, shared within all strains in red and polymorphic in blue.

3.7 Tables

Kmer	Mean DNase Effect	Transcription Factor
CCCCGGA/TCCGGGG	-1.67	Ets98B
CGTAGTA/TACTACG	-1.24	no match
CCTGAGA/TCTCAGG	-1.22	tin
CTGTAAC/GTTACAG	-1.21	ovo
CGGGTTA/TAACCCG	-1.13	Kr
GACCCTC/GAGGGTC	-1.13	lola
GGCCACC/GGTGGCC	1.12	Ci
CAGGTCG/CGACCTG	1.18	Zelda
ATAGACC/GGTCTAT	1.21	br
GGCAGGA/TCCTGCC	1.21	ttk
CCGCGGA/TCCGCGG	1.24	sug
CTTACCC/GGGTAAG	1.28	Blimp-1
CCCCACG/CGTGGGG	1.37	sug
AGGCCCG/CGGGCCT	1.71	hkb

Table 3.1: Kmer effects on chromatin accessibility

Chapter 4

Interaction of Chromatin and Expression in Early Embryogenesis

The map of genotype to phenotype for a single trait is mediated by layers of regulation and a network of interactions with other phenotypes. A change in a transcription factor binding site can increase the phenotype of protein binding promoting expression but compensatory regulatory processes can maintain the steady state levels of the transcript through increasing mRNA decay. In *D. melanogaster* embryos, a wide range of binding strengths are observed in the potential regulatory regions of developmental genes within a single strain. The chromatin accessibility landscape is also observed to be quantitatively variable between strains but the effects of these changes on gene expression is unknown. I examine the functional consequences of the *in vivo* variation in chromatin accessibility on the expression of nearby genes. Comparisons of differential gene expression and chromatin accessibility establish that while there is substantial expression variability between strains, expression has no relationship to the variability of the chromatin state of the strain. The provides addition evidence that while weak binding of transcription factors are conserved within a population, they likely have no impact of the regulation of nearby genes.

4.1 Introduction

The search for genetic regulators of gene expression variability has shown that while there are both *cis* and *trans* loci, the bulk of the regulation variability is linked to nearby variation. Joint examinations of segregating traits in F1 crosses in yeast[197] found that quantitative trait loci effecting expression (eQTLs) act in *cis* for over a third of the variably expressed genes. Searches for eQTLs in other organisms has found thousands of loci in humans[198], mice[], rats, insects[199], and plants[200]. While the genetic architecture of each species can obscure the causal variants, in cases where they have been precisely identified, the variants often lay in predicted regulatory regions. A few variants have been exhaustively studied and found to disrupt transcription factor binding sites.

Chromatin accessibility has also been treated as a quantitative trait and the genetic bases of this variation have been found to play an important role in tissue specific regulation and differentiation. In a survey of human cell lines collected from genetically diverse humans, *cis* acting dsQTLs (DNaseI sensitivity QTL) were found to regulate thousands of region of chromatin. Examining the context of these dsQTLs showed again many were directly binding sites for regulators of chromatin such as CTFT, an insulator. In addition to the direct actors on chromatin accessibility, a subset of the dsQTLs were found to influence transcription factors that have a more direct effect of gene regulation, such as NF-κB. This overlap in the targets of QTLs that are linked to the same pathway of gene expression highlights the role that chromatin accessibility as an important factor in this process. The tight linkage is further supported by the joint role of over 55% of known eQTLs for human lymphoblastoid expression are also dsQTLs, while 16% of dsQTLs also effected expression.

While this connection of eQTL to dsQTL in immortal human cells cultures is an important discovery, the broader application and causal relationship is unknown. Mouse chromatin landscapes in primary and terminally differentiated erythroblasts found that the chromatin landscape is heavily influenced by expression differences and exhibits quantitative variation with no underlying genetic variation, suggesting that it is the expression that is driving chromatin change and not the converse [185].

An advantage in using stable cell line and terminally differentiated cells is that chromatin and gene expression are relatively constant over time. In these cells, an enhancer is held open and genes are turned on in stable state. It is not clear that the lessons learned in these settings have much to do with the dynamic in vivo state that is the developing embryo. In addition, human and mice have a lower rate of diversity and a larger spread in the spatial arrangement of their *cis*-regulatory models. *D. melanogaster cis*-regulatory modules have a higher density of transcription factor binding sites and there is substantial evidence that these binding sites act in a billboard type of action where several transcription factors all colocalize and act in a threshold manner that is less sensitive to the particular binding sites composition.

In this work, I examine the functional consequences of the *in vivo* variation in chromatin accessibility on the expression of nearby genes. Gene expression phenotypes play a closer and more direct role in the fitness of the organism and may be under converging levels of regulation that dampen the effects of chromatin accessibility. Here I use the same pairwise mixture approach to increase sensitivity of gene expression comparisons and establish that while there is substantial expression variability between strains, expression has no relationship to the variability of the chromatin state of the strain. The provides addition evidence that while weak binding of transcription factors are conserved within a population, they likely have no impact of the regulation of nearby genes.

4.2 Results

Using variants described in Chapter 2, I constructed isoform haplotypes for each strain mixture with an average of 23744 out of 36200 FlyBase annotated transcripts containing segregating variants between strains for qualification of allele specific expression. The fragments were also mapped to the reference sequence for comparison. Figure 4.1 shows that the fragments mapped to genes regions in a consistent but $5'$ biased coverage. Using eXpress to estimate the read counts for each transcript model corrected these biases and allowed for less biased measurements of ASE using EBseq. The differentially expressed transcripts in each mixture illuminate relative changes and are not polarized relative to an ancestral state. Increases and decreases were defined relative to the comparisons in the mixture with the highest sequencing coverage, 303-324. Using these definitions, each pairwise mixture of RNA was examined for differential expression and an average of 9.25% of transcripts were differentially expressed at FDR=0.05 out of 11161 expressed transcripts across all samples.

These deferentially expressed transcripts shown in Figure 4.2 were not enriched for any particular developmental pathway and GO term analysis found one pathway, negative regulation of cell cycle, was enriched representing two genes, *grapes* and *altered disjunction*. I then looked at the potential targets of active transcription factors using ChIP-chip[66] and ChIP-seq[109] binding footprints 21 embryonic transcription factors within 2kb upstream of the expressed genes. The differentially expressed genes that are potential targets of anterior-posterior (A-P) and dorsal-ventral (D-V) patterning regulation in Figure 4.3 are less likely to be differentially expressed (3.98% of 7357 transcripts, $p = 2.36 \times 10^{-6}$ by permuting annotations). When these transcription factor targeted gene are differentially expressed, the magnitude of difference is smaller than genes without nearby A-P or D-V binding ($p = 9.37 \times 10^{-9}$).

Collecting mRNA from the same biological samples for which DNase hypersensitive landscapes have been measured enables a direct comparison of the chromatin changes to expression changes. For each differentially expressed transcript, I determined how much change there was in the accessibility of the chromatin for that strain. Increases and decreases in chromatin accessibility have been found to be linked to expression changes in human tissues and cell lines and to a lesser degree in mouse cells. In Figure 4.4 I examine the change in chromatin state between two strains within 2Kb upstream of all expressed transcripts and find no significant correlation between an increase or decrease in chromatin accessibility and the expression of nearby genes (Spearman $\rho = -0.01, p = 0.06$). The correlation between DNase accessibility and only differentially expressed transcripts instead of all expressed transcripts is even less significant (Spearman $\rho = -0.01, p = 0.6$).

4.3 Discussion

A major goal of modern biology is to define and understand the entirety of the process in which a single cell embryo integrates the information in the genome and develops into

a multi celled organism with complex patterns of gene expressions, protein synthesis and modifications, regulating and feeding back on each other. Methods for creating lists of the functional elements have been fraught with problems in how to define what is functional, ranging from any physical interaction in which by definition every atom in a genome could be considered functional to various thresholds set by measurements of selective constrain or consistent interactions in many observations. While some projects such as ENCODE set a very low threshold for including a region of the human genome as functional by including all regions that are bound at some point to some protein as functional, these standards have been criticized for ignoring any evolutionary information such as the lack of sequence constraint or nonadaptive processes. In *D. melanogaster* a similar problem occurs when attempting to define the functional regulatory elements in embryogenesis. Simply defining all regions bound by transcription factors as functional leads to the inclusion of many sequences that have no nearby genes, that are weakly bound, and do not contain canonical transcription factor binding sites. Only by carefully considering functional directly or by an impact on a series of processes with a phenotype can be understood in an evolutionary context can function be determined.

Studying the role of chromatin accessibility in the relative to gene expression has been done in human cell lines. These QTLs studies with large numbers of samples found that changes in chromatin accessibility are linked to changes in gene expression and account for 80% of the variance for a handful of genes. In these *D. melanogaster* embryonic experiments though the there are fewer variable regions overall and those that are variable are not consistently associated with changes in expression.

A trivial explanation is that there is not enough power in this experiment to detect associations. It is indeed underpowered to detect individual dsQTLs and eQTLs but as seen in Chapter 3, genome wide trends in chromatin accessibility can be found and there is no reason to assume that if those variable regions had an effect on expression, we would fail to observe it. The features of the human dsQTLs point to several viable reasons why chromatin accessibility is linked in those cell lines and not in embryos. The most significant lymphoblast dsQTLs are found in the binding sites of NF-κB. NF-κB binding is known in some cases to be part of a complex of tightly interacting transcription factors such that the change in binding in one factor strong influences the binding of the interacting partners and turns genes on/off in a switch like manner. The known partners of NF-κB , PU.1[201], SP1[202], and others are significantly enriched for dsQTLs in binding sites for these proteins.

In contrast to the enhanceosome regulation of the lymphoblast cells, the enhancers in *D. melanogaster* development appear to be dominated by the billboard model of Kulkarni and Arnosti[30]. Chapter 2 and previous studies[55, 63] presented evidence that the regions with strongest transcription factor binding are as plastic in the composition of the cogent sites as the weakest regions. Additionally, many *cis*-regulatory module in *D. melanogaster* development integrate combinations of activators and repressors at several *cis*-regulatory modules[203], so that even if a single minimal *cis*-regulatory module is rendered inactive through the modification of an individual binding site, the overall regulation in maintained.

Since the magnitude of chromatin accessibility at *cis*-regulatory modules is correlated

with the strength of transcription factor binding[149] and the weakest regions of chromatin accessibility are the most variable but still consistently accessible as summarized in Chapter 3, modest changes in the transcription factor binding are likely to be common in these regions. As we did not see any correlation in changes in gene expression linked to these variably accessible regions, this supports the conclusion that the thousands of weakly bound regions are not regulating gene expression. Future studies of functional elements in the *D. melanogaster* genome will be able to use an expanded set of annotations in order to distinguish what characteristics define a functional element.

4.4 Methods

RNA was collected from mixtures of two strains of *D. melanogaster* during homogenization for DNaseI treatment. Immediately after the dechoniated embryos were homogenized with a dounce, a sample of the lysate was fixed with trizol. Sequencing libraries were made from total RNA using the mRNA TruSeq kit from Illumina, following the manufacturer's instructions. Libraries were sequenced on a HiSeq 2000 Illumina sequencer, yielding 17M-22M reads per library.

Reads were mapped to FlyBase transcripts generated from the r5.47 annotation. Each mixture was mapped to a custom reference sequences in strain specific references were made from the homozygous alleles found in the annotated transcripts. These paired transcript haplotype references were then used to create mapping indexes with STAR 2.3.0e[204] genomeGenerate. Each collection of reads was preprocessed as in 3 and mapped with STAR with default parameter except increasing the number of allowed multimapping positions to 20.

Transcript proportions were then estimated with eXpress v1.5.0[205] with custom haplotypes for allele specific estimates. The counts per transcript were then examined with EBSeq[206] for differential expression. Pairs of transcripts were accepted as differentially expressed if the posterior probability was greater than 0.95. Genes with differential expression were examined for GO term enrichment with GOrilla[207].

Comparisons of expression and differential DNase expression are all relative to the mixture of the individual strains. To take advantage of the replication provided by the combination of mixtures, alleles were considered if they occurred at least twice in any strain and arbitrarily labeled as 303 or 324 if the more frequent allele was found in that strain.

4.5 Figures

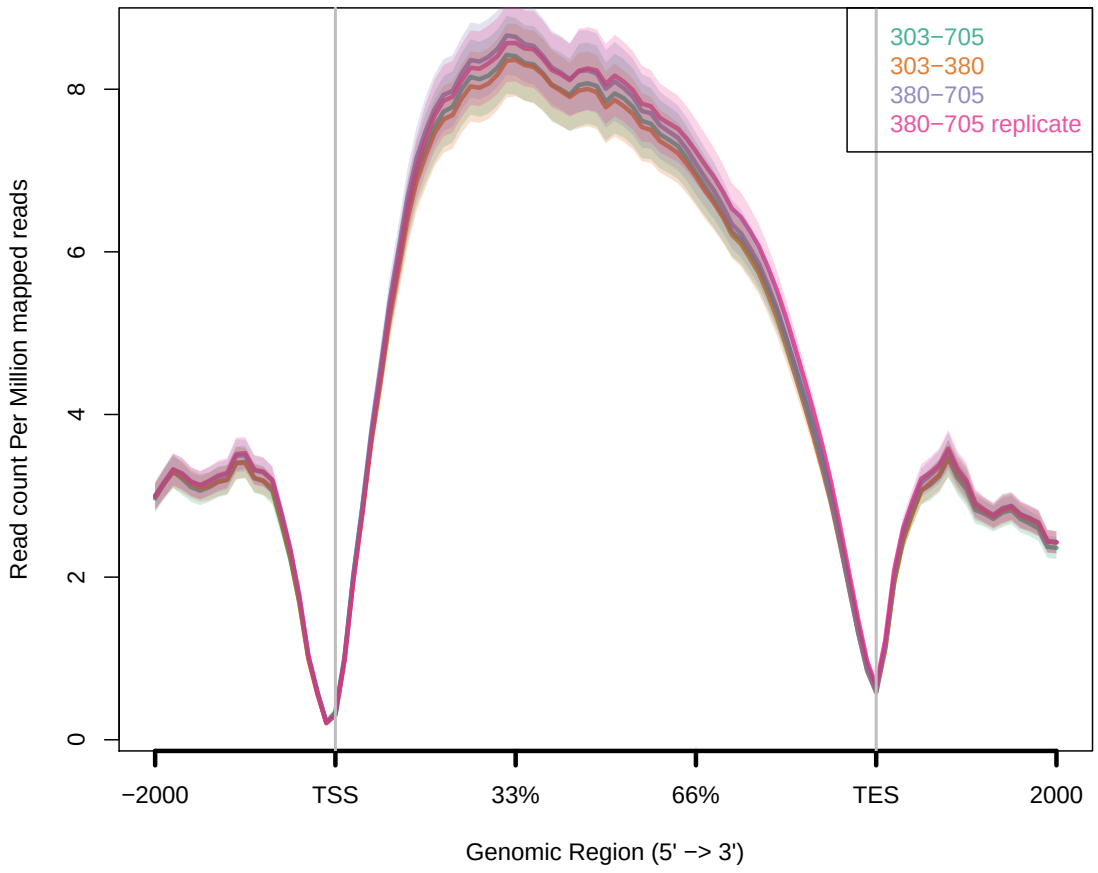

Figure 4.1: RNA fragment distribution between stain mixtures, including biological replicates.

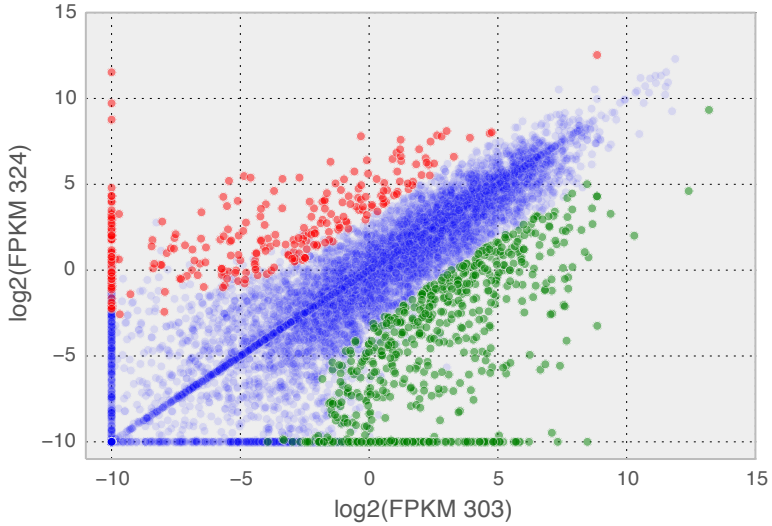

Figure 4.2: Differential Expression in mixture of strains with alleles classified into matching strains 324 and 303. Red transcripts are expressed higher in 324, green transcripts are expressed higher in 303 with an FDR of 0.05 calculated with EBSeq.

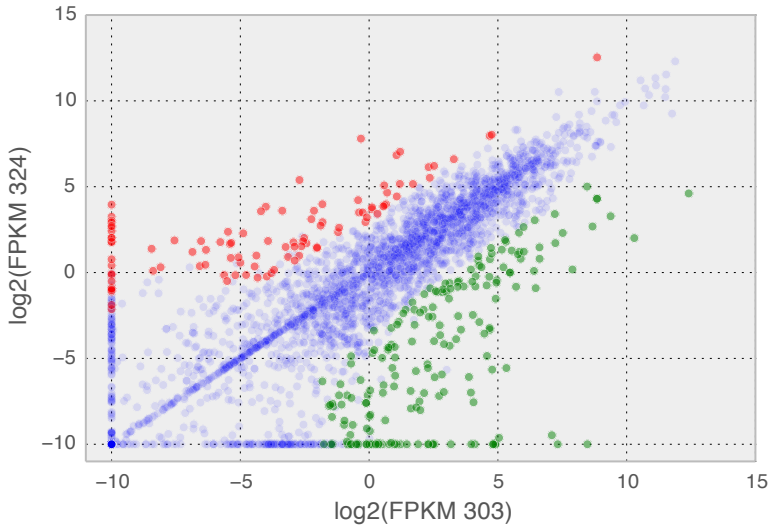

Figure 4.3: Expression of AP and DV target genes. Of the 7064 expressed genes, 4% show differential expression

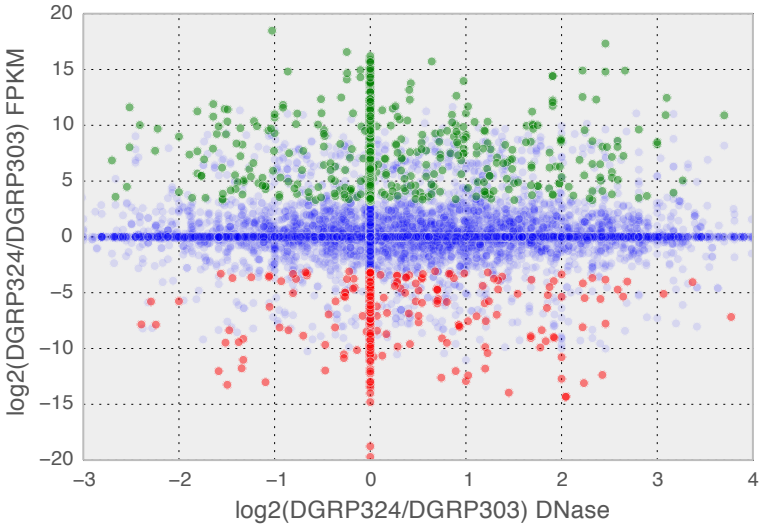

Figure 4.4: Differential chromatin accessibility within 2Kb of TSS and expression per transcript. The colors of each transcript match Figure 4.2. All correlations of expression and DNase accessibility per transcript partitioned into increased expression, decreased expression, no change, and all transcripts were non-significant, p = 0.06 to 0.90

Bibliography

[1] Charles Darwin. *The variation of animals and plants under domestication. Volume I.* John Murray, 1868.

[2] Charles Darwin. *On the origins of species by means of natural selection.* London: Murray, 1859.

[3] RA Fisher. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[4] R Mallarino, O Campàs, and J A Fritz. Closely related bird species demonstrate flexibility between beak morphology and underlying developmental programs. In *Proceedings of the ...*, 2012.

[5] Adam R Boyko, Pascale Quignon, Lin Li, Jeffrey J Schoenebeck, Jeremiah D Degenhardt, Kirk E Lohmueller, Keyan Zhao, Abra Brisbin, Heidi G Parker, Bridgett M vonHoldt, Michele Cargill, Adam Auton, Andy Reynolds, Abdel G Elkahloun, Marta Castelhano, Dana S Mosher, Nathan B Sutter, Gary S Johnson, John Novembre, Melissa J Hubisz, Adam Siepel, Robert K Wayne, Carlos D Bustamante, and Elaine A Ostrander. A simple genetic architecture underlies morphological variation in dogs. *PLoS Biology*, 8(8):e1000451, August 2010.

[6] Alfred D Hershey and Martha Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of general physiology*, 36(1):39–56, 1952.

[7] K D Robertson and P A Jones. DNA methylation: past, present and future directions. *Carcinogenesis*, 21(3):461–467, March 2000.

[8] Neil A Youngson and Emma Whitelaw. Transgenerational epigenetic effects. *Annual Review of Genomics and Human Genetics*, 9:233–257, 2008.

[9] W Fiers, R Contreras, F Duerinck, G Haegeman, D Iserentant, J Merregaert, W Min Jou, F Molemans, A Raeymaekers, A Van den Berghe, G Volckaert, and M Ysebaert. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507, April 1976.

[10] F R Blattner. The Complete Genome Sequence of Escherichia coli K-12. *Science (New York, NY)*, 277(5331):1453–1462, September 1997.

[11] A Goffeau, B G Barrell, H Bussey, R W Davis, B Dujon, H Feldmann, F Galibert, J D Hoheisel, C Jacq, M Johnston, E J Louis, H W Mewes, Y Murakami, P Philippsen, H Tettelin, and S G Oliver. Life with 6000 Genes. *Science (New York, NY)*, 274(5287):546–567, October 1996.

[12] C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science (New York, NY)*, 282(5396):2012–2018, December 1998.

[13] M D Adams, S E Celniker, R A Holt, C A Evans, J D Gocayne, P G Amanatides, S E Scherer, P W Li, R A Hoskins, R F Galle, R A George, S E Lewis, S Richards, M Ashburner, S N Henderson, G G Sutton, J R Wortman, M D Yandell, Q Zhang, L X Chen, R C Brandon, Y H Rogers, R G Blazej, M Champe, B D Pfeiffer, K H Wan, C Doyle, E G Baxter, G Helt, C R Nelson, G L Gabor, J F Abril, A Agbayani, H J An, C Andrews-Pfannkoch, D Baldwin, R M Ballew, A Basu, J Baxendale, L Bayraktaroglu, E M Beasley, K Y Beeson, P V Benos, B P Berman, D Bhandari, S Bolshakov, D Borkova, M R Botchan, J Bouck, P Brokstein, P Brottier, K C Burtis, D A Busam, H Butler, E Cadieu, A Center, I Chandra, J M Cherry, S Cawley, C Dahlke, L B Davenport, P Davies, B de Pablos, A Delcher, Z Deng, A D Mays, I Dew, S M Dietz, K Dodson, L E Doup, M Downes, S Dugan-Rocha, B C Dunkov, P Dunn, K J Durbin, C C Evangelista, C Ferraz, S Ferriera, W Fleischmann, C Fosler, A E Gabrielian, N S Garg, W M Gelbart, K Glasser, A Glodek, F Gong, J H Gorrell, Z Gu, P Guan, M Harris, N L Harris, D Harvey, T J Heiman, J R Hernandez, J Houck, D Hostin, K A Houston, T J Howland, M H Wei, C Ibegwam, M Jalali, F Kalush, G H Karpen, Z Ke, J A Kennison, K A Ketchum, B E Kimmel, C D Kodira, C Kraft, S Kravitz, D Kulp, Z Lai, P Lasko, Y Lei, A A Levitsky, J Li, Z Li, Y Liang, X Lin, X Liu, B Mattei, T C McIntosh, M P McLeod, D McPherson, G Merkulov, N V Milshina, C Mobarry, J Morris, A Moshrefi, S M Mount, M Moy, B Murphy, L Murphy, D M Muzny, D L Nelson, D R Nelson, K A Nelson, K Nixon, D R Nusskern, J M Pacleb, M Palazzolo, G S Pittman, S Pan, J Pollard, V Puri, M G Reese, K Reinert, K Remington, R D Saunders, F Scheeler, H Shen, B C Shue, I Sidén-Kiamos, M Simpson, M P Skupski, T Smith, E Spier, A C Spradling, M Stapleton, R Strong, E Sun, R Svirskas, C Tector, R Turner, E Venter, A H Wang, X Wang, Z Y Wang, D A Wassarman, G M Weinstock, J Weissenbach, S M Williams, WoodageT, K C Worley, D Wu, S Yang, Q A Yao, J Ye, R F Yeh, J S Zaveri, M Zhan, G Zhang, Q Zhao, L Zheng, X H Zheng, F N Zhong, W Zhong, X Zhou, S Zhu, X Zhu, H O Smith, R A Gibbs, E W Myers, G M Rubin, and J C Venter. The genome sequence of Drosophila melanogaster. *Science (New York, NY)*, 287(5461):2185–2195, March 2000.

[14] P Goodfellow. A big book of the human genome. Complementary endeavours. *Nature*, 377(6547):285–286, September 1995.

[15] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee,

Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H Waterston, Richard K Wilson, LaDeana W Hillier, John D McPherson, Marco A Marra, Elaine R Mardis, Lucinda A Fulton, Asif T Chinwalla, Kymberlie H Pepin, Warren R Gish, Stephanie L Chissoe, Michael C Wendl, Kim D Delehaunty, Tracie L Miner, Andrew Delehaunty, Jason B Kramer, Lisa L Cook, Robert S Fulton, Douglas L Johnson, Patrick J Minx, Sandra W Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A Gibbs, Donna M Muzny, Steven E Scherer, John B Bouck, Erica J Sodergren, Kim C Worley, Catherine M Rives, James H Gorrell, Michael L Metzker, Susan L Naylor, Raju S Kucherlapati, David L Nelson, George M Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, François Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W Davis, Nancy A Federspiel, A Pia Abola, Michael J Proctor, Bruce A Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W Richard Mccombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L Aravind, Jeffrey A Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G Brown, Christopher B Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R Copley, Tobias Doerks, Sean R Eddy, Evan E Eichler, Terrence S Furey, James Galagan, James G R Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L Steven Johnson, Thomas A Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W James Kent, Paul Kitts, Eugene V Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V Moran, Nicola Mulder, Victor J Pollara, Chris P Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F A Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I Wolf, Kenneth H Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S Guyer, Jane Peterson, Adam Felsenfeld, Kris A Wetterstrand, Richard M Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R Cox, Maynard V Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J

Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[16] J C Venter. The Sequence of the Human Genome. *Science (New York, NY)*, 291(5507):1304–1351, February 2001.

[17] Jane M Carlton, Robert P Hirt, Joana C Silva, Arthur L Delcher, Michael Schatz, Qi Zhao, Jennifer R Wortman, Shelby L Bidwell, U Cecilia M Alsmark, Sébastien Besteiro, Thomas Sicheritz-Pontén, Christophe J Noel, Joel B Dacks, Peter G Foster, Cedric Simillion, Yves Van De Peer, Diego Miranda-Saavedra, Geoffrey J Barton, Gareth D Westrop, Sylke Müller, Daniele Dessi, Pier Luigi Fiori, Qinghu Ren, Ian Paulsen, Hanbang Zhang, Felix D Bastida-Corcuera, Augusto Simoes-Barbosa, Mark T Brown, Richard D Hayes, Mandira Mukherjee, Cheryl Y Okumura, Rachel Schneider, Alias J Smith, Stepanka Vanacova, Maria Villalvazo, Brian J Haas, Mihaela Pertea, Tamara V Feldblyum, Terry R Utterback, Chung-Li Shu, Kazutoyo Osoegawa, Pieter J De Jong, Ivan Hrdy, Lenka Horvathova, Zuzana Zubacova, Pavel Dolezal, Shehre-Banoo Malik, John M Logsdon, Katrin Henze, Arti Gupta, Ching C Wang, Rebecca L Dunne, Jacqueline A Upcroft, Peter Upcroft, Owen White, Steven L Salzberg, Petrus Tang, Cheng-Hsun Chiu, Ying-Shiung Lee, T Martin Embley, Graham H Coombs, Jeremy C Mottram, Jan Tachezy, Claire M Fraser-Liggett, and Patricia J Johnson. Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science (New York, NY)*, 315(5809):207–212, January 2007.

[18] Douglas L Black. Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry*, 72:291–336, 2003.

[19] Gene Yeo, Dirk Holste, Gabriel Kreiman, and Christopher B Burge. Variation in alternative splicing across human tissues. *Genome Biology*, 5(10):R74, 2004.

[20] P Smolen, D A Baxter, and J H Byrne. Modeling circadian oscillations with interlocking positive and negative feedback loops. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 21(17):6644–6656, September 2001.

[21] D A Day and M F Tuite. Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. *The Journal of endocrinology*, 157(3):361–371, June 1998.

[22] Jeffrey J Hayes, Thomas D Tullius, and Alan P Wolffe. The structure of DNA in a nucleosome. *Proceedings of the National Academy of Sciences of the United States of America*, 87(19):7405–7409, 1990.

[23] K S Weiler and B T Wakimoto. Heterochromatin and gene expression in Drosophila. *Annual review of genetics*, 1995.

[24] M Gatti, S Pimpinelli, and G Santini. Characterization of Drosophila heterochromatin. *Chromosoma*, 1976.

[25] M F LYON. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature*, 190:372–373, April 1961.

[26] Sharon Y Roth, John M Denu, and C David Allis. Histone acetyltransferases. *Annual review of biochemistry*, 70(1):81–120, 2001.

[27] Stephen T Smale and James T Kadonaga. The RNA polymerase II core promoter. *Annual review of biochemistry*, 72:449–479, 2003.

[28] D B Starr and D K Hawley. TFIID binds in the minor groove of the TATA box. *Cell*, 67(6):1231–1240, December 1991.

[29] Andrew KP Taggart, Timothy S Fisher, and B Franklin Pugh. The TATA-binding protein and associated factors are components of pol III transcription factor TFIIIB. *Cell*, 71(6):1015–1028, 1992.

[30] David N Arnosti and Meghana M Kulkarni. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry*, 94(5):890–898, 2005.

[31] Scott Barolo and James W Posakony. Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. *Genes & Development*, 16(10):1167–1181, May 2002.

[32] A M Turing. The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 237(641):37–72, August 1952.

[33] Erik van Nimwegen. Scaling laws in the functional content of genomes. *Trends in Genetics*, 19(9):479–484, September 2003.

[34] Stefan Bonn and Eileen E M Furlong. cis-Regulatory networks during development: a view of Drosophila. *Current Opinion in Genetics & Development*, 18(6):513–520, December 2008.

[35] Long Li, Qianqian Zhu, Xin He, Saurabh Sinha, and Marc S Halfon. Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. *Genome Biology*, 8(6):R101, 2007.

[36] M Hoch and H Jäckle. Transcriptional regulation and spatial patterning in Drosophila. *Current Opinion in Genetics & Development*, 3(4):566–573, August 1993.

[37] A H Brivanlou. Signal Transduction and the Control of Gene Expression. *Science (New York, NY)*, 295(5556):813–818, February 2002.

[38] Daniel Panne, Tom Maniatis, and Stephen C Harrison. An atomic model of the interferon-beta enhanceosome. *Cell*, 129(6):1111–1123, June 2007.

[39] David N Arnosti. Analysis and function of transcriptional regulatory elements: insights from Drosophila. *Annual review of entomology*, 48:579–602, 2003.

[40] Patricia J Wittkopp and Gizem Kalay. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1):59–69, December 2011.

[41] Matthew T Weirauch and Timothy R Hughes. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends in Genetics*, 26(2):66–74, February 2010.

[42] K L Gordon and I Ruvinsky. PLOS Genetics: Tempo and Mode in Evolution of Transcriptional Regulation. *PLoS Genetics*, 2012.

[43] D J Galas and A Schmitz. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research*, 5(9):3157–3170, September 1978.

[44] G D Stormo. DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, 16(1):16–23, January 2000.

[45] D Pribnow. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences of the United States of America*, 72(3):784–788, March 1975.

[46] Rutger Hermsen, Sander Tans, and Pieter Rein ten Wolde. Transcriptional regulation by competing transcription factor modules. *PLoS Computational Biology*, 2(12):e164, December 2006.

[47] J Massague. Smad transcription factors. *Genes & Development*, 19(23):2783–2810, December 2005.

[48] C Tuerk and L Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (New York, NY)*, 249(4968):505–510, August 1990.

[49] Christina Lorenz, Frederike von Pelchrzim, and Renée Schroeder. Genomic systematic evolution of ligands by exponential enrichment (Genomic SELEX) for the identification of protein-binding RNAs independent of their expression levels. *Nature Protocols*, 1(5):2204–2212, December 2006.

[50] Istvan Ladunga. An Overview of the Computational Analyses and Discovery of Transcription Factor Binding Sites. In *Computational Biology of Transcription Factor Binding*, pages 1–22. Humana Press, Totowa, NJ, August 2010.

[51] Linus Pauling and Emile Zuckerkandl. Chemical paleogenetics molecular restoration studies of extinct forms of life. *Acta Chem Scand*, 17(suppl 1):59–516, 1963.

[52] J A Eisen, D Kaiser, and R M Myers. Gastrogenomic delights: a movable feast. *Nature medicine*, 3(10):1076–1078, October 1997.

[53] Alan M Moses, Derek Y Chiang, Daniel A Pollard, Venky N Iyer, and Michael B Eisen. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biology*, 5(12):R98, 2004.

[54] Alan M Moses, Daniel A Pollard, David A Nix, Venky N Iyer, Xiao-Yong Li, Mark D Biggin, and Michael B Eisen. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Computational Biology*, 2(10):e130, October 2006.

[55] Emily E Hare, Brant K Peterson, Venky N Iyer, Rudolf Meier, and Michael B Eisen. Sepsid even-skipped Enhancers Are Functionally Conserved in Drosophila Despite Lack of Sequence Conservation. *PLoS Genetics*, 4(6):e1000106, June 2008.

[56] D S Gilmour and J T Lis. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proceedings of the National Academy of Sciences of the United States of America*, 81(14):4275–4279, July 1984.

[57] Oscar Aparicio, Joseph V Geisberg, Edward Sekinger, Annie Yang, Zarmik Moqtaderi, and Kevin Struhl. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Current protocols in molecular biology / edited by Frederick M Ausubel [et al]*, Chapter 21:Unit 21.3, February 2005.

[58] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, NY)*, 316(5830):1497–1502, June 2007.

[59] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D MacIsaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, Ezra G Jennings, Julia Zeitlinger, Dmitry K Pokholok, Manolis Kellis, P Alex Rolfe, Ken T Takusagawa, Eric S Lander, David K Gifford, Ernest Fraenkel, and Richard A Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, September 2004.

[60] Bryan J Venters, Shinichiro Wachi, Travis N Mavrich, Barbara E Andersen, Peony Jena, Andrew J Sinnamon, Priyanka Jain, Noah S Rolleri, Cizhong Jiang, Christine Hemeryck-Walsh, and B Franklin Pugh. A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces. *Molecular Cell*, 41(4):480–492, February 2011.

[61] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, Renqiang Min, Pedro Alves, Alexej Abyzov, Nick Addleman, Nitin Bhardwaj, Alan P Boyle, Philip Cayting, Alexandra Charos, David Z Chen, Yong Cheng, Declan Clarke, Catharine Eastman, Ghia Euskirchen, Seth Frietze, Yao Fu, Jason Gertz, Fabian Grubert, Arif Harmanci, Preti Jain, Maya Kasowski, Phil Lacroute, Jing Leng, Jin Lian, Hannah Monahan, Henriette O'Geen, Zhengqing Ouyang, E Christopher Partridge, Dorrelyn Patacsil, Florencia Pauli, Debasish Raha, Lucia Ramirez, Timothy E Reddy, Brian Reed, Minyi Shi, Teri Slifer, Jing Wang, Linfeng Wu, Xinqiong Yang, Kevin Y Yip, Gili Zilberman-Schapira, Serafim Batzoglou, Arend Sidow, Peggy J Farnham, Richard M Myers, Sherman M Weissman, and Michael Snyder. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 488(7414):91–100, April 2013.

[62] Shane Neph, Jeff Vierstra, Andrew B Stergachis, Alex P Reynolds, Eric Haugen, Benjamin Vernot, Robert E Thurman, Sam John, Richard Sandstrom, Audra K Johnson, Matthew T Maurano, Richard Humbert, Eric Rynes, Hao Wang, Shinny Vong, Kristen Lee, Daniel Bates, Morgan Diegel, Vaughn Roach, Douglas Dunn, Jun Neri, Anthony Schafer, R Scott Hansen, Tanya Kutyavin, Erika Giste, Molly Weaver, Theresa Canfield, Peter Sabo, Miaohua Zhang, Gayathri Balasundaram, Rachel Byron, Michael J MacCoss, Joshua M Akey, M A Bender, Mark Groudine, Rajinder Kaul, and John A Stamatoyannopoulos. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 488(7414):83–90, April 2013.

[63] Xiao-Yong Li, Stewart MacArthur, Richard Bourgon, David Nix, Daniel A Pollard, Venky N Iyer, Aaron Hechmer, Lisa Simirenko, Mark Stapleton, Cris L Luengo Hendriks, Hou Cheng Chu, Nobuo Ogawa, William Inwood, Victor Sementchenko, Amy Beaton, Richard Weiszmann, Susan E Celniker, David W Knowles, Tom Gingeras, Terence P Speed, Michael B Eisen, and Mark D Biggin. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biology*, 6(2):e27, February 2008.

[64] Mathilde Paris, Tommy Kaplan, Xiao-Yong Li, Jacqueline E Villalta, Susan E Lott, and Michael B Eisen. Extensive Divergence of Transcription Factor Binding

in Drosophila Embryos with Highly Conserved Gene Expression. *PLoS Genetics*, 9(9):e1003748, September 2013.

[65] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, 5(9):829–834, September 2008.

[66] Stewart MacArthur, Xiao-Yong Li, Jingyi Li, James B Brown, Hou Cheng Chu, Lucy Zeng, Brandi P Grondona, Aaron Hechmer, Lisa Simirenko, Soile VE Keränen, David W Knowles, Mark Stapleton, Peter Bickel, Mark D Biggin, and Michael B Eisen. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biology*, 10(7):R80, 2009.

[67] Kevin Chen, Erik van Nimwegen, Nikolaus Rajewsky, and Mark L Siegal. Correlating gene expression variation with cis-regulatory polymorphism in Saccharomyces cerevisiae. *Genome Biology and Evolution*, 2:697–707, 2010.

[68] Jacek Majewski and Tomi Pastinen. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends in Genetics*, 27(2):72–79, February 2011.

[69] Teri A Manolio. Genomewide association studies and assessment of the risk of disease. *The New England journal of medicine*, 363(2):166–176, July 2010.

[70] Thomas Dickmeis and Ferenc Müller. The identification and functional characterisation of conserved regulatory elements in developmental genes. *Briefings in Functional Genomics and Proteomics*, 3(4):332–350, February 2005.

[71] William W Fisher, Jingyi Jessica Li, Ann S Hammonds, James B Brown, Barret D Pfeiffer, Richard Weiszmann, Stewart MacArthur, Sean Thomas, John A Stamatoyannopoulos, and Michael B Eisen. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 109(52):21330–21335, 2012.

[72] Nick G C Smith and Adam Eyre-Walker. Adaptive protein evolution in Drosophila. *Nature*, 415(6875):1022–1024, February 2002.

[73] Peter Andolfatto. Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437(7062):1149–1152, October 2005.

[74] Trudy F C Mackay, Stephen Richards, Eric A Stone, Antonio Barbadilla, Julien F Ayroles, Dianhui Zhu, Sònia Casillas, Yi Han, Michael M Magwire, Julie M Cridland, Mark F Richardson, Robert R H Anholt, Maite Barrón, Crystal Bess, Kerstin Petra Blankenburg, Mary Anna Carbone, David Castellano, Lesley Chaboub, Laura Duncan, Zeke Harris, Mehwish Javaid, Joy Christina Jayaseelan, Shalini N Jhangiani, Katherine W Jordan, Fremiet Lara, Faye Lawrence, Sandra L Lee, Pablo Librado, Raquel S Linheiro, Richard F Lyman, Aaron J Mackey, Mala Munidasa, Donna Marie Muzny, Lynne Nazareth, Irene Newsham, Lora Perales, Ling-Ling Pu, Carson Qu, Miquel Ràmia, Jeffrey G Reid, Stephanie M Rollmann, Julio Rozas, Nehad Saada, Lavanya Turlapati, Kim C Worley, Yuan-Qing Wu, Akihiko Yamamoto, Yiming Zhu, Casey M

Bergman, Kevin R Thornton, David Mittelman, and Richard A Gibbs. The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482(7384):173–178, January 2012.

[75] C H Langley, K Stevens, C Cardeno, Y C G Lee, D R Schrider, J E Pool, S A Langley, C Suarez, R B Corbett-Detig, B Kolaczkowski, S Fang, P M Nista, A K Holloway, A D Kern, C N Dewey, Y S Song, M W Hahn, and D J Begun. Genomic Variation in Natural Populations of Drosophila melanogaster. *Genetics*, 192(2):533–598, October 2012.

[76] Gianni Liti, David M Carter, Alan M Moses, Jonas Warringer, Leopold Parts, Stephen A James, Robert P Davey, Ian N Roberts, Austin Burt, Vassiliki Koufopanou, Isheng J Tsai, Casey M Bergman, Douda Bensasson, Michael J T O'Kelly, Alexander van Oudenaarden, David B H Barton, Elizabeth Bailes, Alex N Nguyen, Matthew Jones, Michael A Quail, Ian Goodhead, Sarah Sims, Frances Smith, Anders Blomberg, Richard Durbin, and Edward J Louis. Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–341, March 2009.

[77] Subramanian S Ajay, Stephen C J Parker, Hatice Ozel Abaan, Karin V Fuentes Fajardo, and Elliott H Margulies. Accurate and comprehensive sequencing of personal genomes. *Genome Research*, 21(9):1498–1505, September 2011.

[78] K Nakamura, T Oshima, T Morimoto, S Ikeda, H Yoshikawa, Y Shiwa, S Ishikawa, M C Linak, A Hirai, H Takahashi, M Altaf-Ul-Amin, N Ogasawara, and S Kanaya. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13):e90–e90, July 2011.

[79] Frazer Meacham, Dario Boffelli, Joseph Dhahbi, David IK Martin, Meromit Singer, and Lior Pachter. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, 12(1):451, 2011.

[80] PR Haddrill and CM Bergman, http://bergmanlab.smith.man.ac.uk/?p=1685.

[81] Y FUYAMA. Gynogenesis in Drosophila melanogaster. *Idengaku Zasshi*, 59(1):91–96, 1984.

[82] C H Langley, M Crepeau, C Cardeno, R Corbett-Detig, and K Stevens. Circumventing Heterozygosity: Sequencing the Amplified Genome of a Single Haploid Drosophila melanogaster Embryo. *Genetics*, 188(2):239–246, June 2011.

[83] John E Pool, Russell B Corbett-Detig, Ryuichi P Sugino, Kristian A Stevens, Charis M Cardeno, Marc W Crepeau, Pablo Duchen, J J Emerson, Perot Saelao, David J Begun, and Charles H Langley. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. *PLoS Genetics*, 8(12):e1003080, December 2012.

[84] Ryan D Bickel, Wendy S Schackwitz, Len A Pennacchio, Sergey V Nuzhdin, and Artyom Kopp. Contrasting Patterns of Sequence Evolution at the Functionally Redundant bric à brac Paralogs in Drosophila melanogaster. *Journal of Molecular Evolution*, 69(2):194–202, August 2009.

[85] Eric A Stone. Joint genotyping on the fly: identifying variation among a sequenced panel of inbred lines. *Genome Research*, 22(5):966–974, May 2012.

[86] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype

and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, June 2011.

[87] Sònia Casillas, Antonio Barbadilla, and Casey M Bergman. Purifying selection maintains highly conserved noncoding sequences in Drosophila. *Molecular Biology and Evolution*, 24(10):2222–2234, October 2007.

[88] Dan L Lindsley and E H Grell. *Genetic variations of Drosophila melanogaster*. Carnegie, 1967.

[89] P T Ives. GENETIC CHANGES IN AMERICAN POPULATIONS OF DROSOPHILA MELANOGASTER. ... *Academy of Sciences of the United States of America*, 1954.

[90] M O Kauer, D Dieringer, and C Schlötterer. A microsatellite variability screen for positive selection associated with the "out of Africa" habitat expansion of Drosophila melanogaster. *Genetics*, 165(3):1137–1148, November 2003.

[91] John E Pool and Charles F Aquadro. History and structure of sub-Saharan populations of Drosophila melanogaster. *Genetics*, 174(2):915–929, October 2006.

[92] S Aris-Brosou and L Excoffier. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Molecular Biology and Evolution*, 13(3):494–504, March 1996.

[93] G A T Mcvean. The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. *Science (New York, NY)*, 304(5670):581–584, April 2004.

[94] Andrew H Chan, Paul A Jenkins, and Yun S Song. Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLoS Genetics*, 8(12):e1003090, December 2012.

[95] Amin Zia and Alan M Moses. Towards a theoretical understanding of false positives in DNA motif finding. *BMC Bioinformatics*, 13:151, 2012.

[96] Melissa M Harrison, Xiao-Yong Li, Tommy Kaplan, Michael R Botchan, and Michael B Eisen. Zelda Binding in the Early Drosophila melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition. *PLoS Genetics*, 7(10):e1002266, October 2011.

[97] L Teytelman, D M Thurtle, and J Rine. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. In *Proceedings of the ...*, November 2013.

[98] S J Gould and R C Lewontin. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London Series B, Containing papers of a Biological character Royal Society (Great Britain)*, 205(1161):581–598, September 1979.

[99] Marc Lohse, Anthony M Bolger, Axel Nagel, Alisdair R Fernie, John E Lunn, Mark Stitt, and Björn Usadel. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40(Web Server issue):W622–7, July 2012.

[100] Ping Wang, Richard F Lyman, Svetlana A Shabalina, Trudy F C Mackay, and Robert R H Anholt. Association of polymorphisms in odorant-binding protein genes with variation in olfactory response to benzaldehyde in Drosophila. *Genetics*, 177(3):1655–

1665, November 2007.

[101] Robert K Bradley, Adam Roberts, Michael Smoot, Sudeep Juvekar, Jaeyoung Do, Colin Dewey, Ian Holmes, and Lior Pachter. Fast statistical alignment. *PLoS Computational Biology*, 5(5):e1000392, May 2009.

[102] Steven J Marygold, Paul C Leyland, Ruth L Seal, Joshua L Goodman, Jim Thurmond, Victor B Strelets, Robert J Wilson, and FlyBase consortium. FlyBase: improvements to the bibliography. *Nucleic Acids Research*, 41(Database issue):D751–7, January 2013.

[103] G Lunter and M Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, pages 1–5, October 2010.

[104] Swetansu Pattnaik, Srividya Vaidyanathan, Durgad G Pooja, Sa Deepak, and Binay Panda. Customisation of the Exome Data Analysis Pipeline Using a Combinatorial Approach. *PLoS ONE*, 7(1):e30080, January 2012.

[105] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–595, March 2010.

[106] http://picard.sourceforge.net/.

[107] Mark A Depristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, Aaron Mckenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, pages 1–11, April 2011.

[108] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, August 2009.

[109] Robert K Bradley, Xiao-Yong Li, Cole Trapnell, Stuart Davidson, Lior Pachter, Hou Cheng Chu, Leath A Tonkin, Mark D Biggin, and Michael B Eisen. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biology*, 8(3):e1000343, March 2010.

[110] Thomas A Down, Casey M Bergman, Jing Su, and Tim JP Hubbard. Large-scale discovery of promoter motifs in Drosophila melanogaster. *PLoS Computational Biology*, 3(1):e7, 2007.

[111] M B Noyes, X Meng, A Wakabayashi, S Sinha, M H Brodsky, and S A Wolfe. A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Research*, 36(8):2547–2560, February 2008.

[112] Jacques van Helden. Regulatory sequence analysis tools. *Nucleic Acids Research*, 31(13):3593–3596, 2003.

[113] Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, 1989.

[114] Joel Smith, Christina Theodoris, and Eric H Davidson. A gene regulatory network subcircuit drives a dynamic pattern of gene expression. *Science (New York, NY)*, 318(5851):794–797, November 2007.

[115] Eric H Davidson. Emerging properties of animal gene regulatory networks. *Nature*, 468(7326):911–920, December 2010.

[116] Mike Levine. Transcriptional enhancers in animal development and evolution. *Current biology : CB*, 20(17):R754–63, September 2010.

[117] Michael Bulger and Mark Groudine. Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144(3):327–339, February 2011.

[118] Hanna M Petrykowska, Christopher M Vockley, and Laura Elnitski. Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. *Genome Research*, 18(8):1238–1246, August 2008.

[119] Steven A Vokes, Hongkai Ji, Wing H Wong, and Andrew P McMahon. A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes & Development*, 22(19):2651–2663, October 2008.

[120] S Ayer and C Benyajati. Conserved enhancer and silencer elements responsible for differential Adh transcription in Drosophila cell lines. *Molecular and cellular biology*, 10(7):3512–3523, July 1990.

[121] Boris Lenhard, Albin Sandelin, and Piero Carninci. Regulatory elements: Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, 13(4):233–245, March 2012.

[122] Miklos Gaszner and Gary Felsenfeld. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics*, 7(9):703–713, September 2006.

[123] S Ohtsuki, M Levine, and H N Cai. Different core promoters possess distinct regulatory activities in the Drosophila embryo. *Genes & Development*, 12(4):547–556, February 1998.

[124] Vincent C Calhoun, Angelike Stathopoulos, and Michael Levine. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. *Proceedings of the National Academy of Sciences of the United States of America*, 99(14):9243–9247, July 2002.

[125] Xiao Liu, Cheol-Koo Lee, Joshua A Granek, Neil D Clarke, and Jason D Lieb. Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Research*, 16(12):1517–1528, December 2006.

[126] Sam John, Peter J Sabo, Robert E Thurman, Myong-Hee Sung, Simon C Biddie, Thomas A Johnson, Gordon L Hager, and John A Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43(3):264–268, March 2011.

[127] Jacob F Degner, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, Matthew Stephens, Hai Wang, and Jonathan K Pritchard. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, pages 1–5, February 2012.

[128] Robert P Zinzen, Kate Senger, Mike Levine, and Dmitri Papatsenko. Computational models for neurogenic gene expression in the Drosophila embryo. *Current Biology*,

16(13):1358–1365, July 2006.

[129] Benjamin P Berman, Yutaka Nibu, Barret D Pfeiffer, Pavel Tomancak, Susan E Celniker, Michael Levine, Gerald M Rubin, and Michael B Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):757–762, January 2002.

[130] Travis N Mavrich, Cizhong Jiang, Ilya P Ioshikhes, Xiaoyong Li, Bryan J Venters, Sara J Zanton, Lynn P Tomsho, Ji Qi, Robert L Glaser, Stephan C Schuster, David S Gilmour, Istvan Albert, and B Franklin Pugh. Nucleosome organization in the Drosophila genome. *Nature*, 453(7193):358–362, May 2008.

[131] Zhaoyu Li, Paul Gadue, Kaifu Chen, Yang Jiao, Geetu Tuteja, Jonathan Schug, Wei Li, and Klaus H Kaestner. Foxa2 and H2A.Z Mediate Nucleosome Depletion during Embryonic Stem Cell Differentiation. *Cell*, 151(7):1608–1616, December 2012.

[132] Rasmus Siersbæk, Ronni Nielsen, Sam John, Myong-Hee Sung, Songjoon Baek, Anne Loft, Gordon L Hager, and Susanne Mandrup. Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. *The EMBO Journal*, 30(8):1459–1472, April 2011.

[133] Gary Felsenfeld and Mark Groudine. Controlling the double helix. *Nature*, 421(6921):448–453, January 2003.

[134] Jiang I Wu, Julie Lessard, and Gerald R Crabtree. Understanding the words of chromatin regulation. *Cell*, 136(2):200–206, January 2009.

[135] P Arnold, A Scholer, M Pachkov, P Balwierz, H Jorgensen, M B Stadler, E van Nimwegen, and D Schubeler. Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Research*, September 2012.

[136] Miki Fujioka, Guizhi Sun, and James B Jaynes. The Drosophila eve Insulator Homie Promotes eve Expression and Protects the Adjacent Gene from Repression by Polycomb Spreading. *PLoS Genetics*, 9(10):e1003883, October 2013.

[137] Daniel Enderle, Christian Beisel, Michael B Stadler, Moritz Gerstung, Prashanth Athri, and Renato Paro. Polycomb preferentially targets stalled promoters of coding and noncoding transcripts. *Genome Research*, January 2011.

[138] K L Reddy, J M Zullo, E Bertolino, and H Singh. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature*, 452(7184):243–247, March 2008.

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

[140] Dmitri Papatsenko and Mike Levine. A rationale for the enhanceosome and other evolutionarily constrained enhancers. *Current biology : CB*, 17(22):R955–7, November 2007.

[141] Duncan T Odom, Robin D Dowell, Elizabeth S Jacobsen, William Gordon, Timothy W Danford, Kenzie D MacIsaac, P Alexander Rolfe, Caitlin M Conboy, David K Gifford, and Ernest Fraenkel. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics*, 39(6):730–732, June 2007.

[142] Anthony R Borneman, Tara A Gianoulis, Zhengdong D Zhang, Haiyuan Yu, Joel Rozowsky, Michael R Seringhaus, Lu Yong Wang, Mark Gerstein, and Michael Snyder. Divergence of Transcription Factor Binding Sites Across Related Yeast Species. *Science (New York, NY)*, 317(5839):815–819, August 2007.

[143] Brian B Tuch, David J Galgoczy, Aaron D Hernday, Hao Li, and Alexander D Johnson. The evolution of combinatorial gene regulation in fungi. *PLoS Biology*, 6(2):e38, February 2008.

[144] John R ten Bosch, Joseph A Benavides, and Thomas W Cline. The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription. *Development (Cambridge, England)*, 133(10):1967–1977, May 2006.

[145] Hsiao-Lan Liang, Chung-Yi Nien, Hsiao-Yun Liu, Mark M Metzstein, Nikolai Kirov, and Christine Rushlow. The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature*, 456(7220):400–403, October 2008.

[146] Noam Kaplan, Irene K Moore, Yvonne Fondufe-Mittendorf, Andrea J Gossett, Desiree Tillo, Yair Field, Emily M LeProust, Timothy R Hughes, Jason D Lieb, Jonathan Widom, and Eran Segal. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366, March 2009.

[147] Katrin Weigmann, Robert Klapper, Thomas Strasser, Christof Rickert, Gerd Technau, Herbert Jäckle, Wilfried Janning, and Christian Klämbt. FlyMove–a new way to look at development of Drosophila. *Trends in Genetics*, 19(6):310–311, June 2003.

[148] José Antonio Campos-Ortéga and Volker Hartenstein. *The embryonic development of Drosophila melanogaster*. Springer Verlag, 1997.

[149] Xiao-Yong Li, Sean Thomas, Peter J Sabo, Michael B Eisen, John A Stamatoyannopoulos, and Mark D Biggin. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biology*, 12(4):R34, April 2011.

[150] T N Mavrich, I P Ioshikhes, B J Venters, C Jiang, L P Tomsho, J Qi, S C Schuster, I Albert, and B F Pugh. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research*, 18(7):1073–1083, July 2008.

[151] Cheol-Koo Lee, Yoichiro Shibata, Bhargavi Rao, Brian D Strahl, and Jason D Lieb. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics*, 36(8):900–905, July 2004.

[152] Sean Thomas, Xiao-Yong Li, Peter J Sabo, Richard Sandstrom, Robert E Thurman, Theresa K Canfield, Erika Giste, William Fisher, Ann Hammonds, Susan E Celniker, Mark D Biggin, and John A Stamatoyannopoulos. Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biology*, 12(5):R43, May 2011.

[153] M E Hogan, M W Roberson, and R H Austin. DNA flexibility variation may dominate DNase I cleavage. *Proceedings of the National Academy of Sciences of the United States of America*, 86(23):9273–9277, December 1989.

[154] C Wu, P M Bingham, K J Livak, R Holmgren, and S C Elgin. The chromatin structure

of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell*, 16(4):797–806, April 1979.

[155] Jay R Hesselberth, Xiaoyu Chen, Zhihong Zhang, Peter J Sabo, Richard Sandstrom, Alex P Reynolds, Robert E Thurman, Shane Neph, Michael S Kuehn, William S Noble, Stanley Fields, and John A Stamatoyannopoulos. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, 6(4):283–289, April 2009.

[156] Alan P Boyle, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, January 2008.

[157] Gregory E Crawford, Sean Davis, Peter C Scacheri, Gabriel Renaud, Mohamad J Halawi, Michael R Erdos, Roland Green, Paul S Meltzer, Tyra G Wolfsberg, and Francis S Collins. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature Methods*, 3(7):503–509, July 2006.

[158] Jorja G Henikoff, Jason A Belsky, Kristina Krassovsky, David M Macalpine, and Steven Henikoff. Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences*, 108(45):18318–18323, November 2011.

[159] M K Kerr and G A Churchill. Experimental design for gene expression microarrays. *Biostatistics*, 2(2):183–201, June 2001.

[160] Alan P Boyle, Lingyun Song, Bum-Kyu Lee, Darin London, Damian Keefe, Ewan Birney, Vishwanath R Iyer, Gregory E Crawford, and Terrence S Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, 21(3):456–464, March 2011.

[161] Cynthia Kelsey Motzny and Robert Holmgren. The Drosophila cubitus interruptus protein and its role in the wingless and hedgehog signal transduction pathways. *Mechanisms of Development*, 52(1):137–150, July 1995.

[162] G Tzolovsky, W M Deng, T Schlitt, and M Bownes. The function of the broad-complex during Drosophila melanogaster oogenesis. *Genetics*, 153(3):1371–1383, November 1999.

[163] Thomas G Wilson, Yoram Yerushalmi, David M Donnell, and Linda L Restifo. Interaction between hormonal signaling pathways in Drosophila melanogaster as revealed by genetic interaction between methoprene-tolerant and broad-complex. *Genetics*, 172(1):253–264, January 2006.

[164] D Read and J L Manley. Alternatively spliced transcripts of the Drosophila tramtrack gene encode zinc finger proteins with distinct DNA binding specificities. *The EMBO Journal*, 11(3):1035–1044, March 1992.

[165] Julia E Dallman, Janet Allopenna, Andrew Bassett, Andrew Travers, and Gail Mandel. A conserved role but different partners for the transcriptional corepressor CoREST in fly and mammalian nervous system formation. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 24(32):7186–7193, August 2004.

[166] Ingo Zinke, Christina S Schütz, Jörg D Katzenberger, Matthias Bauer, and Michael J

Pankratz. Nutrient control of gene expression in Drosophila: microarray analysis of starvation and sugar-dependent response. *The EMBO Journal*, 21(22):6162–6173, November 2002.

[167] Teclise Ng, Fengwei Yu, and Sudipto Roy. A homologue of the vertebrate SET domain and zinc finger protein Blimp-1 regulates terminal differentiation of the tracheal system in the Drosophila embryo. *Development genes and evolution*, 216(5):243–252, February 2006.

[168] Luiz Paulo Andrioli, Luciano Antonio Digiampietri, Lilian Ponce de Barros, and Ariane Machado-Lima. Huckebein is part of a combinatorial repression code in the anterior blastoderm. *Developmental Biology*, October 2011.

[169] Panagiotis Giannios and Sonia G Tsitilou. The embryonic transcription factor Zelda of Drosophila melanogaster is also expressed in larvae and may regulate developmentally important genes. *Biochemical and biophysical research communications*, 438(2):329–333, August 2013.

[170] Abbie Saunders, Leighton J Core, Catherine Sutcliffe, John T Lis, and Hilary L Ashe. Extensive polymerase pausing during Drosophila axis patterning enables high-level and pliable transcription. *Genes & Development*, 27(10):1146–1158, May 2013.

[171] Elodie Darbo, Carl Herrmann, Thomas Lecuit, Denis Thieffry, and Jacques van Helden. Transcriptional and epigenetic signatures of zygotic genome activation during early drosophila embryogenesis. *BMC Genomics*, 14(1):226, 2013.

[172] Joseph C Pearson, Joseph D Watson, and Stephen T Crews. Drosophila melanogaster Zelda and Single-minded collaborate to regulate an evolutionarily dynamic CNS midline cell enhancer. *Developmental Biology*, 366(2):420–432, June 2012.

[173] R Satija and R K Bradley. The TAGteam motif facilitates binding of 21 sequence-specific transcription factors in the Drosophila embryo. *Genome Research*, January 2012.

[174] Chung-Yi Nien, Hsiao-Lan Liang, Stephen Butcher, Yujia Sun, Shengbo Fu, Tenzin Gocha, Nikolai Kirov, J Robert Manak, and Christine Rushlow. Temporal coordination of gene networks by Zelda in the early Drosophila embryo. *PLoS Genetics*, 7(10):e1002339, October 2011.

[175] Amy Tsurumi, Fan Xia, Jinghong Li, Kimberly Larson, Russell LaFrance, and Willis X Li. STAT is an essential activator of the zygotic genome in the early Drosophila embryo. *PLoS Genetics*, 7(5):e1002086, May 2011.

[176] S Goeke, E A Greene, P K Grant, and M A Gates. Alternative splicing of lola generates 19 transcription factors controlling axon guidance in Drosophila. *Nature*, 2003.

[177] Michael A Gates, Ramakrishnan Kannan, and Edward Giniger. A genome-wide analysis reveals that the Drosophila transcription factor Lola promotes axon growth in part by suppressing expression of the actin nucleation factor Spire. *Neural development*, 6:37, 2011.

[178] K Jagla, M Bellard, and M Frasch. A cluster of Drosophila homeobox genes involved in mesoderm differentiation programs. *BioEssays*, 23(2):125–133, February 2001.

[179] Élio Sucena, Isabelle Delon, Isaac Jones, François Payre, and David L Stern. Regula-

tory evolution of shavenbaby/ovo underlies multiple cases of morphological parallelism. *Nature*, 424(6951):935–938, August 2003.

[180] D Pauli and A P Mahowald. Germ-line sex determination in Drosophila melanogaster. *Trends in Genetics*, 1990.

[181] J Andrews, D Garcia-Estefania, I Delon, J Lu, M Mevel-Ninio, A Spierer, F Payre, D Pauli, and B Oliver. OVO transcription factors function antagonistically in the Drosophila female germline. *Development (Cambridge, England)*, 127(4):881–892, February 2000.

[182] D J Montell. Moving right along: regulation of cell migration during Drosophila development. *Trends in Genetics*, 1994.

[183] T Chen, M Bunting, F D Karim, and C S Thummel. Isolation and characterization of five Drosophila genes that encode an ets-related DNA binding domain. *Developmental Biology*, 151(1):176–191, May 1992.

[184] Richard M Eglen, Annette Gilchrist, and Terry Reisine. The use of immortalized cell lines in GPCR screening: the good, bad and ugly. *Combinatorial chemistry & high throughput screening*, 11(7):560–565, August 2008.

[185] Mona Hosseini, Leo Goodstadt, Jim R Hughes, Monika S Kowalczyk, Marco De Gobbi, Georg W Otto, Richard R Copley, Richard Mott, Douglas R Higgs, and Jonathan Flint. Causes and Consequences of Chromatin Variation between Inbred Mice. *PLoS Genetics*, 9(6):e1003570, June 2013.

[186] W Zhang, Y Wu, J C Schnable, Z Zeng, M Freeling, G E Crawford, and J Jiang. High-resolution mapping of open chromatin in the rice genome. *Genome Research*, November 2011.

[187] Wenli Zhang, Tao Zhang, Yufeng Wu, and Jiming Jiang. Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *The Plant cell*, 24(7):2719–2731, July 2012.

[188] Daniel J McKay and Jason D Lieb. A common set of DNA regulatory elements shapes Drosophila appendages. *Developmental cell*, 27(3):306–318, November 2013.

[189] Nicolas Nègre, Christopher D Brown, Parantu K Shah, Pouya Kheradpour, Carolyn A Morrison, Jorja G Henikoff, Xin Feng, Kami Ahmad, Steven Russell, Robert A H White, Lincoln Stein, Steven Henikoff, Manolis Kellis, and Kevin P White. A comprehensive map of insulator elements for the Drosophila genome. *PLoS Genetics*, 6(1):e1000814, 2010.

[190] Artyom A Alekseyenko, Joshua W K Ho, Shouyong Peng, Marnie Gelbart, Michael Y Tolstorukov, Annette Plachetka, Peter V Kharchenko, Youngsook L Jung, Andrey A Gorchakov, Erica Larschan, Tingting Gu, Aki Minoda, Nicole C Riddle, Yuri B Schwartz, Sarah C R Elgin, Gary H Karpen, Vincenzo Pirrotta, Mitzi I Kuroda, and Peter J Park. Sequence-Specific Targeting of Dosage Compensation in Drosophila Favors an Active Chromatin Context. *PLoS Genetics*, 8(4):e1002646, April 2012.

[191] Wei Zheng, Hongyu Zhao, Eugenio Mancera, Lars M Steinmetz, and Michael Snyder. Genetic analysis of variation in transcription factor binding in yeast. *Nature*, 464(7292):1187–1191, April 2010.

[192] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nussbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008.

[193] Naim U Rashid, Paul G Giresi, Joseph G Ibrahim, Wei Sun, and Jason D Lieb. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology*, 12(7):R67, July 2011.

[194] Alan P Boyle, Justin Guinney, Gregory E Crawford, and Terrence S Furey. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics (Oxford, England)*, 24(21):2537–2538, November 2008.

[195] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[196] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, 2007.

[197] R B Brem. Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science (New York, NY)*, 296(5568):752–755, March 2002.

[198] Eric E Schadt, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, Stephen B Milligan, John R Lamb, Guy Cavet, Peter S Linsley, Mao Mao, Roland B Stoughton, and Stephen H Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, March 2003.

[199] W Jin, R M Riley, R D Wolfinger, K P White, G Passador-Gurgel, and G Gibson. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nature Genetics*, 29(4):389–395, December 2001.

[200] M A L West, K Kim, D J Kliebenstein, H van Leeuwen, R W Michelmore, R W Doerge, and D A St Clair. Global eQTL Mapping Reveals the Complex Genetic Architecture of Transcript-Level Variation in Arabidopsis. *Genetics*, 175(3):1441–1450, December 2006.

[201] Fulai Jin, Yan Li, Bing Ren, and Rama Natarajan. PU.1 and C/EBP(alpha) synergistically program distinct response to NF-kappaB activation through establishing monocyte specific enhancers. *Proceedings of the National Academy of Sciences*, 108(13):5290–5295, March 2011.

[202] F Hirano, H Tanaka, Y Hirano, M Hiramoto, H Handa, I Makino, and C Scheidereit. Functional interference of Sp1 and NF-kappaB through the same DNA binding site. *Molecular and cellular biology*, 18(3):1266–1274, March 1998.

[203] Michael W Perry, Alistair N Boettiger, and Michael Levine. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proceedings of the National Academy of Sciences*, 108(33):13570–13575, August 2011.

[204] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, January

2013.

[205] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73, November 2012.

[206] Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart M G Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics (Oxford, England)*, 29(8):1035–1043, April 2013.

[207] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48, 2009.