INTRO

**Gene expression important for phenotypic evolution**

Gene regulation guides spatiotemporal gene expression throughout development and organismal life. As such, changes in gene regulation that produce gene expression variation often underlie the evolution of higher-order phenotypes, such as morphology and behavior (Carroll 2008; Stern 2000; Stern and Orgogozo 2008; Carroll 2005). These mechanistic changes often take place specifically during transcription activation, during which *trans*-acting elements bind to *cis­*-regulatory elements to recruit transcriptional machinery (Stern and Orgogozo 2008), although we may be in part biased towards this step due to our lagging knowledge of the other steps. However, our rapidly growing understanding of gene regulation in its entirety now allows us to survey the role and/or contribution of other mechanistic steps to gene expression evolution.

**Chromatin remodeling increasingly well studied and being integrated into our understanding of evolutionary mechanisms**

Chromatin remodeling is an essential early step of gene regulation. To allow for regulatory machinery to activate transcription, DNA must first be made accessible by evicting nucleosomes followed by epigenetic modification of regions to maintain accessibility (Klemm, Shipony, and Greenleaf n.d.). The extent to which regions are made and maintained accessible is known as chromatin accessibility (Klemm, Shipony, and Greenleaf n.d.). The molecular mechanisms that give rise to chromatin accessibility are varied and can depend on the functional category (e.g. promoter or enhancer) of a given region, and can also vary within these categories. For example, the accessibility of some *Drosophila* enhancers seems to be primarily determined by the binding of a single transcription factor with the ability to bind nucleosome-bound DNA (known as a pioneer factor), whereas other enhancers’ accessibility is determined by the collective binding of numerous transcription factors(Jacobs et al. 2018; Bravo González‐Blas et al. 2020). Additionally, promoters with different architectures – commonly classified into either “broad” or “narrow” – are known to differ in nucleosome affinity (Nozaki et al. 2011). Generally, broad promoters tend to have TATA box motifs and are found to regulate ubiquitously expressed genes, whereas narrow promoters often lack TATA box motifs and regulate tissue-specific genes (Hoskins et al. 2011). The diverse mechanisms by which chromatin accessibility is established and maintained across the genome creates two interesting evolutionary questions: 1) how does chromatin accessibility variation differ across the genome and 2) does it affect the extent to which this variation relates to gene expression?

Previous work has found that, importantly, chromatin accessibility variation can explain TF binding variation unaccounted for by motif sequence variation, demonstrating that chromatin variation can causally lead to gene expression variation (Peng et al. 2019). Additionally, when regions are parsed into proximal/promoter versus enhancer/distal regions, distal enhancer regions are more variable than proximal promoter elements, although variation in the latter better correlates with gene expression variation (Floc’hlay et al. 2021). Finally, across both promoter/proximal and enhancer/distal elements, the chromatin accessibility variation is largely due to changes in *cis*, which may be binding sites for chromatin remodeling factors, transcription factors, etc (Floc’hlay et al. 2021). Here, we collect parental and allelic chromatin accessibility and gene expression data from *Drosophila* imaginal wing discs, which allows us to further delimit both promoter and enhancer regions by chromatin remodeling mechanism and test the relationship to gene expression variation. These data describe chromatin accessibility variation at a more finely resolved resolution, as well as incorporate current molecular biology into our understanding of how variation across mechanistic layers is related.

*Experimental schematic and classification – Figure 1*

To measure the extent and mode of chromatin accessibility variation, as well as its relationship to gene expression variation, we collected ATAC-seq (accessible chromatin) and RNA-seq (gene expression) data from the L3 imaginal wing discs of two populations of *D. melanogaster* (United States and Zimbabwe) as well as the F1 hybrid (Figure 1A/B). We used these two populations of *D. melanogaster* to yield more genetic variation than population samples, and specifically dissected L3 imaginal wing discs because of prior work on chromatin remodeling of enhancers in this tissue. For each genotype, we collected three biological replicates for the ATAC-seq and six biological replicates of RNA-seq. Read counts across regions were highly correlated across biological replicates with, as expected, reduced correlation coefficients between genotypes (Figure S1).

To define genomic regions to measure accessibility, we took two approaches to establish coordinates for 1) transcription start sites (TSSs) and transcription end sites (TESs) and 2) intergenic and intragenic regions. For TSSs and TESs, we used annotated TSS and TSS coordinates and defined these regions as +/- 500bp from this point. For intergenic and intragenic regions, we took only peaks non-overlapping with the TSS and TES sites; if the peak fell within a coding region it was classified as intragenic; if the peak fell outside of a coding region it was classified as intergenic. To normalize region length across the four categories, for both intergenic and intragenic regions, the center of the peaks were taken +/- 500bp. Reads were then counted for region coordinates for each class (Figure 1C). As expected, the highest average read counts fall at the center of these coordinates for each class (Figure 1C). Lastly, to gain better insight into the intragenic class, we calculated the frequency of these regions falling in introns, exons, or spanning both, and found that the majority fell into exons, one-third into introns, and few into both.

*Extent and mode of chromatin accessibility across categories – Figure 2*

To quantify the extent and mode of chromatin accessibility variation, we used a Bayesian approach to estimate the difference between parental strains and hybrid alleles. Using the estimate of parental divergence (Figure 2A), we found that, in agreement with previous work and expectation, chromatin accessibility divergence is greatest for intergenic regions (Figure 2B). Next, we used both the parental and hybrid allele estimates to calculate the contribution of *cis* changes (percent *cis*) to the observed chromatin accessibility variation for each region (Figure 2C). Consistent with prior work, we find that for most regions the majority of chromatin accessibility variation is due to changes in *cis* (Figure 2D), and also find that this pattern is most extreme for intergenic genic regions. In summary, these findings are in accordance with previous findings of chromatin accessibility variation in that 1) accessibility is due to cis and 2)

*Extent and mode of chromatin accessibility divergence differs for TSS regions with different promoter types – Figure 3*

Since prior work has shown varied nucleosome affinities based on promoter architecture, we wanted to test whether this may impact patterns of chromatin accessibility variation. To characterize the promoter architecture of our TSS regions, we used a CAGE-seq dataset from modENCODE to annotate our regions as having either broad or narrow promoters. These data were collected in embryos but concluded that the promoter profiles were highly conserved between embryos and adults, suggesting the annotations may be confidently used for imaginal discs as well.

Interestingly, by visualizing and comparing the average chromatin accessibility across these two categories, peaked regions seem to be more accessible at the main TSS than broad regions (Figure 3A), consistent with our understanding that 1) peaked regions have fewer stably bound nucleosomes and 2) the location of transcription initiation is less variable. As for patterns of variation, narrow promoters are less divergent than broad promoters (Figure 3B), but *cis* changes explain a larger percentage of the divergence (Figure 3C). These results demonstrate that different promoter architectures exhibit different chromatin accessibility profiles and patterns of variation, suggesting that the relationship between chromatin accessibility and gene expression may be different for different types of promoters.

*Extent and mode of chromatin accessibility divergence differs for intergenic regions made accessible by pioneer factor or non-pioneer factor mechanisms – Figure 4*

To contrast patterns of chromatin accessibility divergence by regions made accessible by known different molecular mechanisms, such as pioneer factor-mediated, we used prior functional work to characterize regions for which pioneer factor activity is necessary to establish accessibility in this tissue. With this approach, we identified 105 regions which is likely an under-estimate, but a conservative characterization. In contrast to the rest of the regions, pioneer factor-mediated regions are more divergent and more often due to *cis*-regulatory changes. However, we found less sequence variation in these regions, meaning that the average effect size per nucleotide variant is greater for pioneer factor-mediated regions than the rest of the genome. One explanation for this finding is that these regions contain more TF binding sites with the capacity of altering chromatin accessibility. Indeed, we [ ].

*The relationship between gene expression and chromatin accessibility divergence based on chromatin remodeling mechanism – Figure 5*

We next sought to contrast the relationship between gene expression and chromatin accessibility divergence between regions, as well as test how this relationship may vary for different functional categories. To pair accessible regions with gene expression values, we either 1) matched TSS and TES regions with the annotated gene or 2) used a proximity approach in which each intergenic and intragenic region was paired with the closest expressed gene (Figure 1A). Overall, gene expression and chromatin accessibility are weakly correlated (pearson = 0.34), which is consistent with observations in yeast and, to some extent, fly embryos (Figure 1B). We calculated pearson correlation coefficients for each subset of annotated pair (i.e. functional region, promoter type, and pioneer or not) to get a quantitative comparison of the relationship, and then built a linear models to determine whether the interaction between a given functional category and accessibility divergence impacted the correlation with gene expression divergence.