**TITLE:** PATTERNS OF CHROMATIN ACCESSIBILITY VARIATION DIFFER ACROSS THE *DROSOPHILA* GENOME.

**INTRO**

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*-factors bind to *cis­*-regulatory elements to recruit transcriptional machinery (Stern and Orgogozo 2008), although this finding may in part be biased due to our historically lagging knowledge of gene regulation outside of transcriptional activation. Our rapidly growing understanding of gene regulation in its entirety allows us to now survey the role and/or contribution of other mechanistic steps to gene expression evolution.

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 2019). The extent to which regions are made and maintained accessible is known as chromatin accessibility. The molecular mechanisms that give rise to chromatin accessibility are varied and depend on the regulatory function of a given region (e.g. promoters vs enhancers) (Floc’hlay et al. 2021), but can also vary within each of 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 which vary in transcription initiation dynamics – often classified as either “broad” or “peaked” architectures – have different nucleosome affinities and epigenetic mark profiles, suggesting varied chromatin remodeling mechanisms (Nozaki et al. 2011; Hoskins et al. 2011) (Nozaki et al. 2011). As we begin to understand chromatin accessibility variation and how it relates to gene expression variation, we must also integrate our mechanistic understanding of chromatin remodeling to ask: do evolutionary patterns of chromatin accessibility divergence vary across genomic regions remodeled by different mechanisms, and how may this impact the relationship between chromatin accessibility and gene expression variation?

One observation is common across all prior work of chromatin accessibility variation: the majority of chromatin accessibility variation is due to changes in *cis* (Connelly, Wakefield, and Akey 2014; Floc’hlay et al. 2021). However, the extent to which this chromatin accessibility variation is predictive of gene expression variation is unclear, seemingly depending on the organism and tissue being assayed. As for how these patterns may vary across the genome, when distal enhancer and proximal promoter regions are separated and compared, 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). Here, we integrate our prior knowledge and data on remodeling and regulatory mechanisms to measure how evolutionary patterns of chromatin accessibility variation and its relationship to gene expression variation may vary across the genome depending on the regulatory mechanism. We collected parental and hybrid 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.

**RESULTS**

***Quantifying gene expression and chromatin accessibility across the genome***

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). These two populations of *D. melanogaster* have, on average, of 1% and X% sequence divergence in coding regions and non-coding regions (Supp Fig), respectively, and specifically dissecting L3 imaginal wing discs allowed us to draw on prior functional work on chromatin remodeling mechanisms in these tissue s(Bravo González‐Blas et al. 2020; Jacobs et al. 2018). 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 reduced correlation coefficients between genotypes, as expected (Supp Fig).

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 defined these regions as +/- 500bp from the annotated predominant TSS and TSS coordinates (cite UCSC browser). For intergenic and intragenic regions, we obtained regions with a peak calling software (cite macs2) and filtered 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. For both intergenic and intragenic regions, the final coordinates were defined as +/- 500bp from the center of the called peak region. Reads of each sample were then counted within region coordinates for each class. Processing the data in this way results in the highest average read counts falling at the center of these coordinates, suggesting we are accurately defining coordinates of interest (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 (Figure 1D).

**Intergenic regions exhibit the most chromatin accessibility divergence and percentage contribution of *cis* changes**

To quantify the extent and mode of chromatin accessibility variation, we estimated the difference in chromatin accessibility between parental strains as well the difference in chromatin accessibility between hybrid alleles (Figure 2A). Since the chromatin accessibility variation between hybrid alleles is measured in a controlled *trans* environment, we can infer the contribution of *cis* and *trans* changes to chromatin accessibility variation between *D. melanogaster* strains. To do so for each region while normalizing for the magnitude of variation, we calculated the contribution the percentage of chromatin accessibility variation due to *cis* changes (Figure 2B, top). In agreement with all prior studies measuring the genetic architecture of chromatin accessibility variation, for most regions in the genome, chromatin accessibility variation is largely due to changes in *cis* (Figure 2B, bottom).

Next, we contrasted both metrics – parental chromatin accessibility variation and percent *cis* – across different functional/spatial groups. In accordance with prior work, intergenic regions have greater parental chromatin accessibility variation than TSS regions (Floc’hlay et al. 2021) but are also greater than both TES and intragenic regions (Figure 2C). Furthermore, intergenic regions also exhibit the greatest percentage of chromatin accessibility due to changes in *cis* as compared to the other functional/spatial groups (Figure 2D). Taken together, these findings are consistent with our expectations based on intergenic regions having the greatest amount of sequence variation, underscoring the relationship between local sequence and chromatin accessibility variation (Supp figure).

**Narrow promoter architectures are more accessible and exhibit different accessibility evolutionary patterns than broad promoter architectures**

Generally, broad promoters tend to have TATA box motifs and are found to regulate ubiquitously expressed genes, whereas peaked promoters often lack TATA box motifs and regulate tissue-specific genes (Hoskins et al. 2011). More importantly, prior work has shown differences in nucleosome affinity and epigenetic marking, suggesting evolutionary patterns of chromatin accessibility may differ between promoter architectures. To measure any potential differences in our dataset, we using a CAGE-seq dataset from the modENCODE project (Hoskins et al. 2011) to label our TSS regions as either broad or peaked promoters and compared various metrics between the two groups.

Interestingly, we found that the TSS regions of peaked promoters are on average more accessible than those of broad promoters (Figure 3A), which is consistent with prior work showing broad promoters to have greater nucleosome affinity (Hoskins et al. 2011). As for how these different promoter architectures may result in varied evolutionary patterns of chromatin accessibility, we found that peaked promoters are less variable (Figure 3B) but better explained by *cis*-regulatory divergence (Figure 3C) than broad promoters. Taken together, these results demonstrate that chromatin accessibility measurements can capture the previously characterized varied chromatin states of broad and peaked promoters, and these varied chromatin states are associated with different evolutionary patterns of chromatin accessibility variation.

**Sequence variation in pioneer factor-mediated regions has a disproportionately large effect on chromatin accessibility variation**

Although it is well-known that different regions are remodeled and made accessible through different mechanisms, we have little functional data speaking to *which* regions are remodeled by a given mechanism. Epithelial cells of *Drosophila* imaginal discs are one of the few tissue types for which we do have some functional data on remodeling mechanism. We used this data from prior work to appropriately characterize our regions as pioneer factor-mediated based on whether pioneer factor functionality is necessary for wild-type accessibility. We then asked whether patterns of chromatin accessibility evolution are distinct for regions made accessible by a pioneer factor-mediated mechanism.

In contrast to the rest of the genome, the 230 regions classified as pioneer factor-mediated exhibit a greater magnitude of chromatin accessibility variation (Figure 4A) and are more often due to changes in *cis* (Figure 4B). However, we found notably less sequence variation in the pioneer factor-mediated regions, meaning that the average effect size per nucleotide variant is greater for pioneer factor-mediated regions than the rest of the genome (Figure 4C). *These findings prompted us to scan for motifs with variation.. [need to play around with this analysis – still unsure what to do to provide some explanation or hypothesis for this pattern].*

**Promoter architecture influences the extent to which chromatin accessibility variation predicts gene expression divergence**

Given that regions that differ by function, space, architecture, and remodeling mechanism exhibit different patterns of chromatin accessibility variation, we next asked whether any of these parameters have an influence on the extent to which chromatin accessibility variation can predict gene expression variation. 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 5A). Overall, gene expression and chromatin accessibility are weakly correlated (Pearson correlation coefficient = 0.30), which is consistent with observations in yeast and, to some extent, fly embryos (Figure 5B).

To statistically test whether the predictability of gene expression variation from chromatin accessibility variation was influenced by categorical identity, we built an analysis of covariance based generalized linear model (ANCOVA) to contrast regression lines within each category. We tested this for functional/spatial category, promoter architecture, and remodeling mechanism and found that only promoter architecture has a significant effect on the extent to which gene expression variation can be predicted by chromatin accessibility (Figure 5C, top). Furthermore, if we contrast the Pearson correlation coefficients generated from gene expression – chromatin accessibility variation pairs for both broad and peaked promoters, we see that this result is driven by the peaked promoters, for which there is a near zero relationship between gene expression and chromatin accessibility variation (Figure 5C, bottom).

FIGURE LEGENDS

**Figure 1. Quantifying gene expression and chromatin accessibility across the genome.** **(A)** Imaginal wing discs were dissected from L3 larvae of two *D. melanogaster* strains separated by 10,000 years of divergence and their F1 hybrid. RNA-seq and ATAC-seq data was collected in replicates of 6 and 3 per genotype, respectively. **(B)** Genome browser for one sample of each genotype illustrating transcript abundance and chromatin accessibility at the X locus, a gene known to be active in imaginal wing discs. **(C, top)** Density plots for a representative sample showing ATAC-seq mean read count signal *across* regions of each functional category. **(C, bottom)** Heatmap showing read count signal for *each* region in all functional categories. **(D)** Pie chart showing the proportion of intragenic regions that overlapped with an intron, exon, or both.

**Figure 2. Intergenic regions exhibit the most chromatin accessibility divergence and percentage contribution of *cis* changes. (A)** Chromatin accessibility divergence was estimated for parental strain and hybrid allele comparisons and plotted on the x- and y-axes, respectively. **(B)** Smoothed histograms of the calculated metric: percentage of accessibility divergence due to *cis* changes, showing that for most regions the majority of chromatin accessibility divergence is due to *cis* changes. **(C)** Boxplots comparing the distribution of estimated parental accessibility divergence (absolute value) across functional categories. **(D)** Boxplots comparing the distribution of percentage of accessibility divergence due to *cis* changes access across functional categories. Boxplot notches indicate 95% confidence of median.

**Figure 3. Narrow promoter architectures are more accessible and exhibit different accessibility evolutionary patterns than broad promoter architectures. (A, top)** Density plot for a representative sample showing ATAC-seq mean read count signal for TSS regions of broad (blue) and peaked (green) promoters. **(A, bottom)** Heatmap showing read count signal for TSS regions of each broad (top heatmap) and peaked (bottom heatmap) promoter region, showing that narrow promoters tend to have more accessible chromatin. **(B)** Boxplots comparing the distribution of estimated parental accessibility divergence (absolute value) between TSS regions of broad and narrow promoters. **(C)** Boxplots comparing the distribution of percentage of accessibility divergence due to *cis* changes access between TSS regions of broad and narrow promoters. Boxplot notches indicate 95% confidence of median.

**Figure 4. Sequence variation in pioneer factor-mediated regions has a larger effect on chromatin accessibility divergence.** For regions classified as pioneer factor-mediated and other mechanism, boxplots comparing the distribution of **(A)** estimated parental accessibility divergence (absolute value), **(B)** percentage of accessibility divergence due to *cis* changes, and **(C)** the average amount of cis-regulatory divergence per nucleotide variant in each region, calculated by log transforming the from the absolute value of estimated hybrid allelic difference (i.e. chromatin accessibility divergence due to *cis* regulatory changes) by the number of nucleotide variants for each region. Boxplot notches indicate 95% confidence of median. **(D)** *Something related to predicted TFBS changes???*

**Figure 5. Promoter architecture influences the extent to which chromatin accessibility divergence can predict gene expression divergence. (A)** Schematic depicting the bioinformatic approach used to pair accessible regions with expressed genes. **(B)** The chromatin accessibility (x-axis) and gene expression (y-axis) divergence estimated values are shown as a scatterplot for each of the region – gene pair. Points are colored by density as indicated by the legend. Pearson correlation coefficient is provided at the bottom right of the plot, indicating the weak correlation between chromatin accessibility and gene expression divergence. **(C, top)** Table providing Pearson correlation coefficients relating chromatin accessibility and gene expression divergence for each class of region (top), **(C, bottom)** as well as results of an analysis of covariance (ACOVA), in which the regression lines of the predictor and response variables (chromatin accessibility and gene expression divergence) are compared for the different levels within each categorical variable (functional region, promoter architecture, and remodeling mechanism).