Title
Characterize and classify genetic variation in chromatin state in Drosophila melanogaster

Author
Huynh, Khoi

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,

IRVINE

Characterize and classify genetic variation in chromatin state in *Drosophila melanogaster*

DISSERTATION

submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Ecology and Evolutionary Biology

by

Khoi Huynh

Dissertation Committee:

Professor Anthony Douglas Long, Chair

Professor Kevin R Thornton

Associate Professor James Jordan Emerson

2023

DEDICATION

To

my dear mother
who
gives me life,
supports me tirelessly in the pursue of my career,
means the world to me

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

As a cancer focused biologist after BS degree, PhD graduate study in bioinformatic has been a long and tough road for me since I never coded nor focused in genetic before. However, today is the end of the road for graduate study, without my advisor, my colleagues, friends, family, and google, it would have been much more challenging. The years I spent learning to code and learning genetic field have been some of the toughest years, but also some of the best years. UC Irvine has granted me with wonderful opportunity to pursue my career as a scientist, filled with new knowledge, and challenges.

First, I would like to express the deepest appreciation to my committee chair, Professor Anthony D. Long. You took a huge change on accepting me to your lab knowing my background of no genetic, and no coding. It is fascinating to listen to your teaching and advice so that I can grow as a bioinformatic scientist. You continually and convincingly conveyed a spirit of importance in conducting research the right way with ample quality control steps. It was not easy, and I still make a few of the same old mistakes, albeit not much anymore. However, I would not be the scientist I am today without your guidance, and I am grateful for your teaching and patience.

To my mother, Hanh. With your love and guidance, I can pursue my dream of becoming a scientist. You have been the guiding light for the importance of perseverance. Without your push for me to get education at every levels despite the financial hardship in my

childhood, I wouldn’t be finishing my dissertation. For your love, your sacrifices, and your hard work, I am eternally grateful.

I would like to thank all the graduate students, post-doc, and other professors whom I have asked for guidance. Without all your help, I would definitely be buried under the avalanche of google search results.

Chapter 1 is (or "Portions of chapter 1 are") a reprint of the material as it appears in https://doi.org/10.1371/journal.pgen.1010439, published CC BY 4.0 and used with permission from PLOS Genetics. The co-authors listed in this publication are Brittny R. Smith, Stuart J. Macdonald, Anthony D. Long.

Lastly, I would like to thank all my co-workers in Dendreon Pharmaceutical LLC. Your advices were what finally pushed me toward the PhD route. Without you, I would still be working in biotech companies with my BS in biological sciences instead of achieving my goal as a scientist today.

VITA

Khoi Huynh

EDUCATION

2018-2023 University of California, Irvine, CA

Doctor of Philosophy in Ecology and Evolutionary Biology

2021 University of California, Irvine, CA

Master of Science in Biological Sciences

2015 University of California, Irvine, CA

Bachelor of Science in Biological Sciences

RESEARCH EXPERIENCE

2017-2023 Graduate Eco/Evo 200: Quantitative Genetic

Dr Anthony Long Laboratory, UC Irvine

2016 Research Assistant: Photodynamic Therapy

Dr Henry Hirschberg Laboratory, UCI Beckman Laser Institute, UC Irvine

2015 Undergraduate Bio 199: Photodynamic Therapy

Dr Henry Hirschberg Laboratory, UCI Beckman Laser Institute, UC Irvine

2014 Undergraduate Bio 199: Biology of Aging

Rose and Mueller Laboratories, UC Irvine

2014 Undergraduate Bio 198: Directed Group Studies

Rose and Mueller Laboratories, UC Irvine

PUBLICATIONS

PUBLISHED:

Huynh K, Smith BR, Macdonald SJ, Long AD. Genetic variation in chromatin state across multiple tissues in Drosophila melanogaster. PLoS Genet. 2023 May 5;19(5):e1010439. doi: 10.1371/journal.pgen.1010439. PMID: 37146087; PMCID: PMC10191298.

Madsen SJ, Christie C, **Huynh K**, Peng Q, Uzal FA, Krasieva TB, Hirschberg H. Limiting glioma development by photodynamic therapy-generated macrophage vaccine and allo-stimulation: an in vivo histological study in rats. J Biomed Opt. 2018 Feb;23(2):1-7. doi: 10.1117/1.JBO.23.2.028001. PMID: 29417766; PMCID: PMC5802332.

Christie C, Molina S, Gonzales J, Berg K, Nair RK, **Huynh K**, Madsen SJ, Hirschberg H. Synergistic chemotherapy by combined moderate hyperthermia and photochemical

internalization. Biomed Opt Express. 2016 Mar 14;7(4):1240-50. doi: 10.1364/BOE.7.001240. PMID: 27446650; PMCID: PMC4929636.

PROJECT NEAR PUBLICATION:

Huynh K, Long AD. Classification of *cis*- and *trans*- nature for genetic variation in chromatin state in Drosophila melanogaster.

TEACHING EXPERIENCE:

2018-2023 Teaching Assistant; University of California, Irvine

Bio 99: Molecular Biology

Bio Sci 100: Scientific Writing

Bio M116L: Molecular Biology Lab

Bio 94: Organisms to Ecosystems

Bio 97: Genetics

ABSTRACT OF DISSERTATION

Characterize and classify genetic variation in chromatin state in Drosophila melanogaster

by

Khoi Huynh

Doctor of Philosophy in Ecology and Evolutionary Biology

University of California, Irvine, 2023

Professor Anthony Douglas Long, Chair

There are two types of genetic traits. The first is monogenic traits which are caused by rare variants that disrupt the function of a single gene. Monogenic traits typically follow the classic Mendelian inheritance, and are rare in nature. In contrast, the second type of genetic traits involve heritable traits that do not follow the classic Mendelian inheritance. These traits are classified as complex traits, and are thought to involve multiple genes. As a result, many studies have spent great efforts to elucidate the nature of these complex traits. However, an appreciable fraction of heritable variation remains unexplained, and is referred to as "missing heritability" [1]. It is widely believed that these missing heritable variations are variations in gene expression due to the binding of transcription factors to enhancers [2-4]. These binding events can be identified by the local chromatin configuration which should be open in particular tissue or timepoint necessary for a trait [5,6]. Therefore, I argue that a genome-wide landscape

of variation in chromatin accessibility in a large number of tissues would be valuable for complex trait studies.

Thus, my first chapter is to utilize ATAC-seq to assess chromatin accessibility across multiple genotypes and tissues from Drosophila melanogaster. In this first chapter, I performed ATAC-seq to study chromatin accessibility for four different tissues: adult female brain, ovaries, wing and eye-antennal imaginal discs. Each sample is also collected from eight different inbred strains. I have identified 44099 ATAC-seq peaks-regions with high ATAC-seq fragment coverage. Furthermore, since the eight inbred founder strains have reference quality genome assemblies, I also performed structural variant correction on my ATAC-seq data. These structural variants contributed to an elevated rate (55%) of the identification of false positive differences in chromatin state between genotypes. After structural variant correction, I have found 1050, 30383, and 4508 regions whose peak heights are polymorphic among genotypes, tissues, or for genotype by tissue interactions respectively. Finally, I identified 249 SNPs and 3 SVs candidate causative variants that explained 100% of the variation at nearby chromatin profiles varying among genotypes.

While having a completely characterized open chromatin landscape is helpful for complex trait communities, the question of whether those polymorphic regulatory elements are in *cis* or in *trans* remain unanswered. Thus, my second chapter aim is to elucidate the *cis* and *trans* nature of the identified regulatory elements from the first study. Therefore, I performed ATAC-seq, utilized our developed quantile normalization of ATAC-seq data,SV-correction, ANOVA-based statistical analysis, and haplotype phasing to examine chromatin accessibility and its *cis*, and *trans* nature in *Drosophila*

melanogaster ovaries collected from two parental strains (A4, B6) and their F1 offspring. We identified 3006 ATAC-seq peaks that are significantly different between parental genotypes. Out of those ATAC-seq peaks, 106 and 45 peaks are identified to be *cis* and *trans* regulatory respectively using cis-*trans* value.

INTRODUCTION

I. Complex trait overview:

All variances in phenotypic traits have genetic components [7,8]. For example, according to Online Mendelian Inheritance in Man resource (OMIMR), there are at least 8930 diseases (traits) that are Mendelian in nature, which are caused by an inheritable alteration of one gene or abnormality in the genomes [9]. Despite those 8930 disease traits on OMIMR and another 12279 recruiting and completing clinical studies on genetic disease as described on ClinicalTrials.gov, the extent of causal genes' contributions to genetic diseases or trait variation remains unclear [10]. In fact, the contributions of genetic variances to the complex traits have been shown to not only be linked directly to causal genes but also to be linked to multiple single nucleotide polymorphism (SNP) loci which are in linkage disequilibrium to causal genes or regulatory elements linked directly to genetic traits [11]. Furthermore, these complex traits, which are affected by variations in genetic component-SNPs or variations in environmental effect, can be disease traits (cardiovascular disease [12], psychological disorders [13], and type 2 diabetes [14] in humans), or non-disease traits (milk yield in dairy cattle [15], rice yield [16], and human height [17]). Thus, complex traits are undoubtedly important targets for studies in multiple disciplines such as medicine, agriculture, and evolution to name a few [18].

As such, complex traits have been studied rigorously for more than 100 years yet the longest-standing question on how the genetic variation contributes to the phenotypic variation remains unclear. Historically, there was a debate between Mendelians, who

believed in large effect causal gene contribution to monogenic phenotypes, and biometricians, who argued that such contribution couldn't explain the continuous variation observed in many phenotypic traits [19]. This was largely resolved by the "infinitesimal model" published by Fisher in 1918 [20]. In this model, normal distribution variation of each phenotypic trait was attributed to multiple causal genes instead of a single causal gene as previously believed by the Mendelians. Thus, the contributions of each gene became infinitesimally small [21]. However, while the infinitesimal model had been successful in accounting for multiple causal genes contributing to a complex trait [22–24], the actual number of causal genes per traits, and their effect size remains unclear until now [19].

II. Complex trait study methods and missing heritability:

In order to study complex traits and to identify the actual number of causal genes, there have been two predominant methodologies: Quantitative trait locus (QTL) mapping and Genome-Wide Association Studies (GWAS). In principle, QTL mapping studies need two isogenic strains which have different alleles at loci affecting the trait of interest, and polymorphic marker linkage map. Then, backcross F2, or recombinant inbred lines (RILs) are created to identify QTL affecting the complex traits [25]. In contrast, GWAS starts with selecting an appropriate outbred population for a complex trait of interest. Individuals from a collection of case/control individuals from that population are phenotypes and genotypes using whole-genome sequencing for SNPchips. Association tests then identify regions of the genome at which alleles differ in

frequency between cases and controls [26]. GWAS is not limited to case/control panels and can also employ a large cohort of individuals with genotypic state regressed on phenotype [27,28].

Given their powerful nature as genetic variation study methods, it is not surprising that both GWAS and QTL mapping have been dominating the field of complex trait study. However, both of these methods have employed the view that quantitative traits are under additive effect of identified variations. This can be observed as GWAS and QTL mapping traditionally identify causative loci by screening the entire genome for loci where alternative genotypes differ significantly [51]. Then, the variation of a complex trait is ,then, defined by these hundred of identified causative loci. This viewpoint, however, completely ignores the multi-allelism nature of genetics. In fact, a different way that genetic variants can contribute to the expression of complex traits is to have different alleles at the same locus affecting the trait in addition to variations found in other loci affecting the same trait. The locus with such polymorphic genotypes is said to have genetic variance-heterogeneity and is identified as vQTL [52,53].

III. MPP, and RIL for QTL mapping:

Despite the contradicting results between QTL mapping and GWAS studies, QTL mapping remains a powerful method to identify loci that co-segregate with a varied phenotypic traits [54], and has been widely used with improving strategies [49]. Throughout the 1990s, QTL mapping was low power with mapping being done on only the F2 generation of a pair of inbred parents following the outlined procedures from two

landmark papers [55,56]. However, QTL mapping has some major disadvantages. The first is the requirement of a large sample size [57]. The second disadvantage-given how QTL mapping was done historically [55,56]- is the fact that QTL mapping can't detect segregating alleles if they are not present in the selected parent genomes. As a result, many segregating alleles that are not represented in the parents would remain hidden. This raises a question as to how relevant the identified QTL is to the actual genotypic variance in the population of multiple varied genotypes. The third disadvantage of QTL mapping is that it is only accurate to within 2cM or less [58]. While the first disadvantage can't be easily solved, the improving high throughput technologies and genomics has improved the utility of QTL mapping techniques by providing adequately dense markers map [59]. Furthermore, with the improving PCR technique and GWAS, QTL mapping studies have improved to include hundreds of crosses instead of just F1, and to incorporate GWAS into the mapping studies as well [49]. As a result, QTL mapping technique resolution has significantly improved since QTL and GWAS can identify linked genes and unlinked genes respectively. As a result, the number of hidden segregating alleles can be reduced.

Among the advanced strategies for QTL mapping studies, there are two major recently developed strategies which are the usage of recombinant inbred lines (RIL), and the usage of Multi-Parent Population (MPP). Both are extremely valuable for different reasons. RIL is formed by crossing two isogenic parents for F1 and F2, and by crossing several brother-sister F2 pairs. This inbred crossing of F2 would be continued for many generations which would result in RIL that are genetically identical to one or the other progenitor's alleles [60]. Thus, RIL has one huge advantage which is the

greater mapping resolution. This is due to a denser breakpoints compared to any cross that has only one meiosis event [61]. On the other hand, MPP starts with k highly-inbred founder strains, and creates individuals who are genetic mosaics of the founders after n generation cross [49]. As a result, with crossing of multiple founder strains, MPP allows for a more complete view of genetic variations as the number of natural haplotypes segregating at any given gene is greatly expanded, improving the resolution of QTL mapping studies [49]. Thus, MPP has solved the second disadvantage of QTL mapping discussed above.

Interestingly, Drosophila Synthetic Population Resource (DSPR) mapping panel, which is the brainchild of Dr Stuart Macdonald and Dr Anthony Long, have incorporated both RIL and MPP [62]. Each of the two synthetic populations of DSPR was created by intercrossing 8 founder lines -with 1 founder line shared between two populations- through 50 generations [62]. More than 1600 recombinant inbred lines (RIL) were created by 25 generation inbred crossing. Such a large number of generations of recombinants have no doubt increased mapping resolution due to the average genomic segment being 3cM in size [49,63]. Furthermore, the usage of 8 founder lines collected from across the world is also a great strength of DSPR. This allows for greater variation in segregating alleles to be captured which will lower the chance of missing any alleles due to them not being found in founder lines. As a result, the identified QTL using DSPR would be more representative for populations instead of just for individuals seen in traditional QTL mapping studies. Furthermore, since DSPR contains 8 different founder strains, it is also a perfect resource to address the genetic variance-heterogeneity of complex traits.

IV. Non-coding regulatory elements contribution to complex traits:

In addition to causal genes' direct contribution to complex traits variation through changing the proteins [64], non-coding genetic variants can also participate in driving the variations of these traits as regulatory elements [11,65,66]. In fact, a widely held current belief is that variation in complex traits is often due to variation in the gene regulation machinery [19,67–71], and especially cis-acting factors that control gene expression in specific tissues or developmental time points that determine traits [2]. These varied regulatory elements control gene expression by binding transcription factors to enhancers, but those binding events can only take place if the local chromatin configuration is open in the particular tissue or timepoint important for that trait. Thus, we argue that knowing the genome-wide landscape of variation in chromatin accessibility in a large number of tissues would be valuable for complex trait studies.

Until recently, non-coding regions with regulatory function have been difficult to identify at scale, but genome-wide profiling of open chromatin regions using experimentally straightforward ATAC-seq (Assay for Transposase Accessible Chromatin) approach [72] have allowed characterization of chromatin state in large panels of genotypes [73,74]. ATAC-seq uses Tn5 transposase to insert sequencing adapters , and to cut the DNA at regions of accessible chromatin. In contrast, if the DNA regions are not accessible, steric hindrance would prevent the binding of Tn5. As a result, only open chromatin regions are probable allowing for amplification and high-throughput sequencing of DNA fragments located at open chromatin regions [72]. An

example schematic of ATAC-seq is shown in figure intro.1 . Then, ATAC-seq "peaks" are called using the ATAC-seq DNA sequence reads pileup. These chromatin accessibility “peaks" which vary across tissue and genotypes are of potentially even great value, as these polymorphic chromatin configurations could presage variation in complex traits.

V. Chromatin accessibility in cis and in trans:

Normally, 75-90% of genomic DNA exists as densely compacted nucleosome arrays, which are bent sequences of DNA wrapping tightly around histones [75,76]. These nucleosome structures act as a barrier to prevent the binding of RNA polymerase[77–80], and most transcription factors [81,82]. DNA wrapped in a nucleosome also prevents repair, recombination complexes [76]. Furthermore, nucleosomes can also recruit other proteins through interaction with the histone tail domains [83]. Thus, the nucleosome must be evicted resulting in a nucleosome free DNA region (open chromatin regions) so that they can participate in regulation of gene expression allowing the binding of polymerase or transcription factors. These open chromatin regions are the regulatory elements that I am interested in for the reasons discussed in the previous section.

Given the importance of nucleosome free DNA regions, great effort has been spent to elucidate the mechanism of nucleosome eviction. It has been shown that DNA sequences can bend differently depending on their nucleotide sequences [84–86]. Consequently, nucleosome stability is greatly dependent on histone affinity to specific

DNA sequences [87,88] which can be 1000 fold or greater [89]. Thus, the less affinity there is between histone and DNA sequence, the less stable a nucleosome is. Therefore, the resulting openness of chromatin due to nucleosome eviction can be in cis due to substantial DNA sequence reference. However, others argue that the histone DNA sequence reference might not be too meaningful [76]. It has been shown that nucleosome positions can be regulated in trans by ATP-dependent nucleosome remodeling complexes [90,91]. Thus, it is important to elucidate the cis and trans nature of open chromatin regions since they regulate gene expression differently. An example for nucleosome eviction in cis and in trans is shown in the figure intro.2 left using an allele of isogenic A4 genotype.

VI. Major shortcomings in the field:

Among the model organisms, Drosophila melanogaster is one of the most widely used model organisms in complex trait fields due to their fast life cycle (~2 weeks), well-characterized reference genome, and a vast-array of specific developed genetic tools. Furthermore, given the ease of use of ATAC-seq and the importance of characterizing complete complex trait loci, there have been at least 14 papers that utilize the ATAC-seq to characterize open chromatin regions in Drosophila melanogaster. However, there are three major shortcomings of these studies. The first is that they primarily focus on embryo samples with only 5 studies utilizing third instar larvae [92–94], adult gut [95], and adult testes [96]. Thus, the utility of identified open chromatin regions are limited, and may not apply to adult tissues and/or phenotypes.

One example is the study on distinct expression characteristics found in genes with promoters that have a TATA box or pausing elements for RNA polymerase II (Pol II) [97]. Both of these promoters play an active role in regulation of gene expression [97–103]. In fact, Pol II pausing has been shown to be involved in regulating transcriptional activation [104]. As a result, genes with Pol II pausing elements are highly regulated during development [98,99,105], and mediate synchronous gene expression between cells [106,107]. In contrast, TATA box is found to be a core player in the gene promoter which directly controls transcription [108]. Therefore, genes containing TATA promoters are often associated with high expression variability [109–111], and are highly enriched among effector genes [103,112,113], which have been shown to be associated with complex traits [114]. Thus, the study on distinct expression characteristics between genes with Pol II pausing and TATA promoters is an important first step to further elucidate the mechanisms behind polymorphic gene expression in complex traits as both promoters are involved directly with gene expression regulation. However, in the study [115], the authors only utilized embryos to characterize differential chromatin accessibility between two promoter groups. Furthermore, tissue specificity of chromatin accessibility between two promoter groups is inferred using only embryonic tissue (tissue collected from embryo). Thus, such a study, while showing valuable polymorphic chromatin accessibility between two promoter groups, can't illustrate the complete picture on differential chromatin state landscape between these two promoter groups given the differences between embryo stage and adult stage, which is skipped.

The second major shortcoming is that the majority of Drosophila studies have focused on a single genotype (or cell line), have used different mutant backgrounds, or have employed a small number of wildtype strains that lack a high-quality genome sequence [92–95,116–118]. Furthermore, none of these genotypes are of the Drosophila melanogaster reference strain, against which the resulting ATACseq reads are subsequently aligned. As a result, they completely ignore the considerable number of SNPs, short insertion/deletion variants, and a wide array of structural variants (SVs) distinguishing any pair of Drosophila strains [119]. This has two major issues. The first is the possible misalignment caused by the effect of SVs on read coverage which has been well-documented leading to incorrect inference of coverage [120]. As outlined in Mahmoud et al., structural variants cause complication in read mapping due to their size compared to the read size, repeated patterns in copy number variant, or the overlapping of multiple SVs in the same regions [120]. All of these may cause the mapped reads to be of low quality and to be thrown out subsequently as the result which would cause incorrect calculation of read coverage. In fact, this issue was observed with RNAseq data [121]. Another issue of using non-reference strains which are aligned only to the reference genome is that it is not possible to analyze genetic and phenotypic variation in chromatin accessibility (as measured by ATAC-seq). This can be observed in the results of hundreds of QTL mapping and GWAS studies which explains a small portion of trait heritability [1,122]. One hypothesis for this is that many hidden variants, such as SNPs, or SVs, make significant contributions to complex trait variation despite being rare in nature [122,123]. Thus, using non-reference strains which don't have known genome assemblies will not be as valuable in elucidating the actual genotypic and

phenotypic variants observed in complex traits due to missing information on genome variations.

In summary, while those studies do produce ATAC-seq landscapes for their respective samples, I would argue that those are not representative and potentially incorrect. This is, firstly, due to their lack of representative usage of tissues from different developmental cycles. Furthermore, they fail to account for the false positive effect of structural variant on inference of open chromatin regions due to SV effect on read mapping errors. These false positive rates can sometimes reach 50% or more as will be shown in my first paper. Lastly, the lack of consideration toward diploid nature of species effectively disallows characterization of trans-acting open chromatin regions. Thus, while the current studies which utilizes ATAC-seq to characterize open chromatin regions do produce a landscape of TFBS for their respective samples. Such landscapes are nowhere near being representative to the species.

VII. Thesis aims and methods:

As an effort to elucidate the "missing heritability", the first aim of this thesis is to characterize a genome-wide landscape of regulatory elements by identifying open chromatin regions using ATAC-seq. Furthermore, we also address the first two major shortcomings of the field by performing ATAC-seq on multiple tissue samples collected from multiple genotypes from DSPR [119] which have reference genome quality assembly and identified structural variants. Here, we carry out a biologically-replicated ATACseq experiment to characterize chromatin accessibility in four adult tissues in

seven highly-characterized isogenic genotypes of D. melanogaster [119]. Similar to the discussion above on how ATAC-seq is used to identify open chromatin regions, we expect to identify different open chromatin peaks. Furthermore, inspired by the quantile normalization method used in microarray studies, we perform genome wide normalization for the ATAC-seq reads. Statistical analysis will be carried out to identify polymorphic peaks that differ in coverage by tissue, by genotype, and by genotype:tissue interaction. As the seven founder lines that we use are highly isogenic and contain fully characterized structural variants [119], we can correct for SVs effect on coverage due to mis-mapped read pairs in order to correctly infer open chromatin configurations. Lastly, as all seven founder strains also have complete SNP profiles, we can also identify a set of causal cis-acting SNPs/SVs that are linked to polymorphic peaks. In conclusion, I believe that our results would be valuable representative profiles of polymorphic open chromatin regions as we have included tissues from two developmental stages of D. melanogaster (embryo imaginal disc and adult tissues), and have corrected for SVs effect on reads mapping coverage.

Furthermore, as the first study fails to include data from F1 hybrid individuals which is necessary to characterize trans-acting open chromatin regions, the question of whether those polymorphic regulatory elements are in cis- or in trans- remain unanswered. As a result, it is not as helpful in elucidating the mechanisms underlying transcriptional regulation. Thus, my second aim for the thesis would be to address the third shortcomings and to elucidate the cis- and trans- nature of the identified regulatory elements from the first study. This is done by fully characterizing cis-acting and trans-acting open chromatin regions using ATAC-seq for ovary tissues collected from F1

hybrid which is the result of a F0 cross between A4 genotype and B6 genotype- both of which are founders of the DSPR. This second study is inspired by the alternative method to identify trans-acting elements as described in yeast [126,127], maize [128], or fruit fly [128–130]. Inspired by this alternative method, we expect that cis- and trans-effect variations in chromatin accessibility can be dissected in the same manner using our ATAC-seq data. If the ratio of coverage between F1 haplotypes is different from the ratio of coverage between the two parents at the same polymorphic open chromatin regions, such difference can be attributed to the trans-acting variations. As a result, we would expect to identify cis-acting and trans-acting polymorphic open chromatin regions. These results can be valuable to the complex trait community despite the limited number of tissues and genotypes in use.

VIII. Figures:

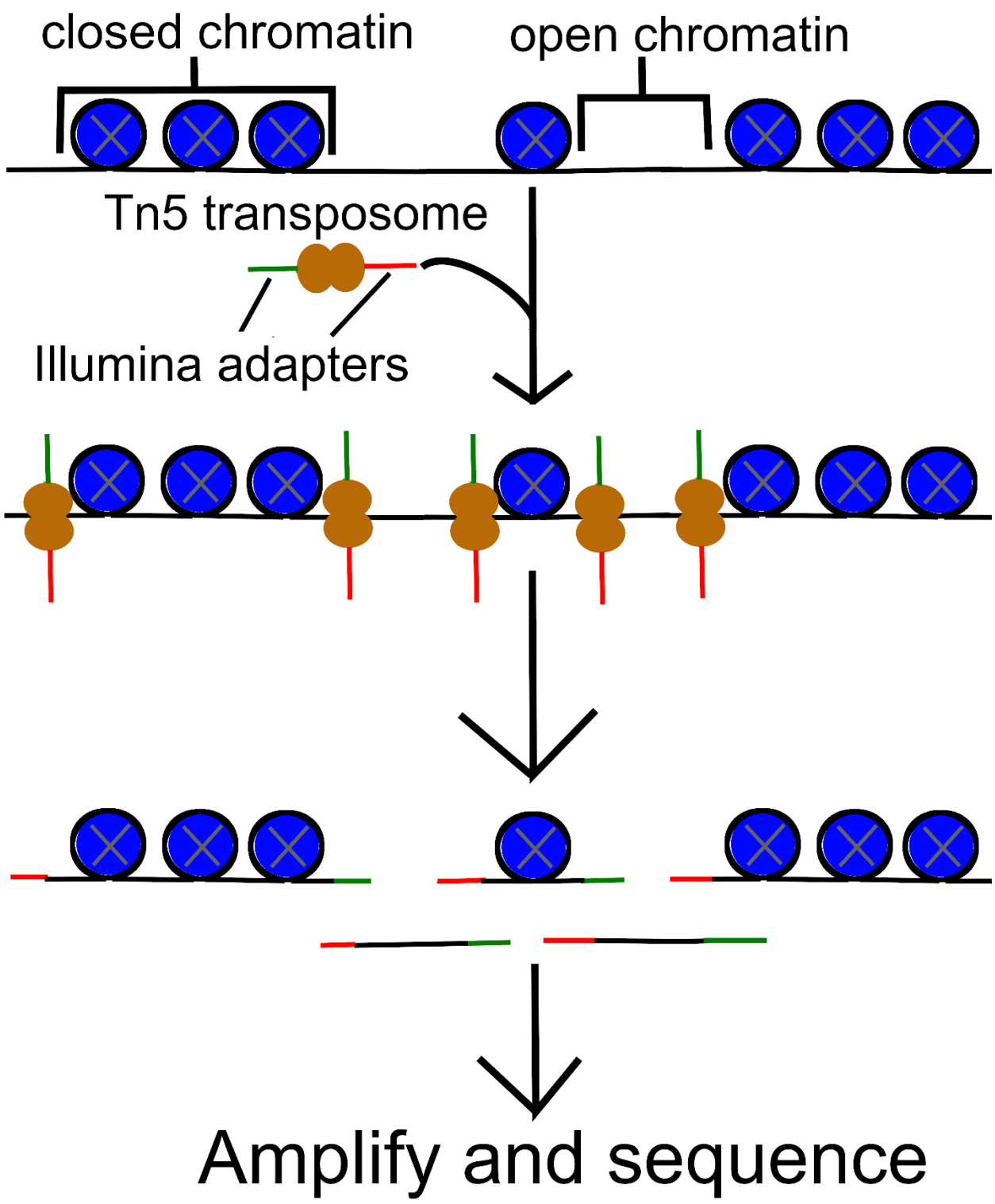


Fig intro.1: Schematic of ATAC-seq showing Tn5 transposome

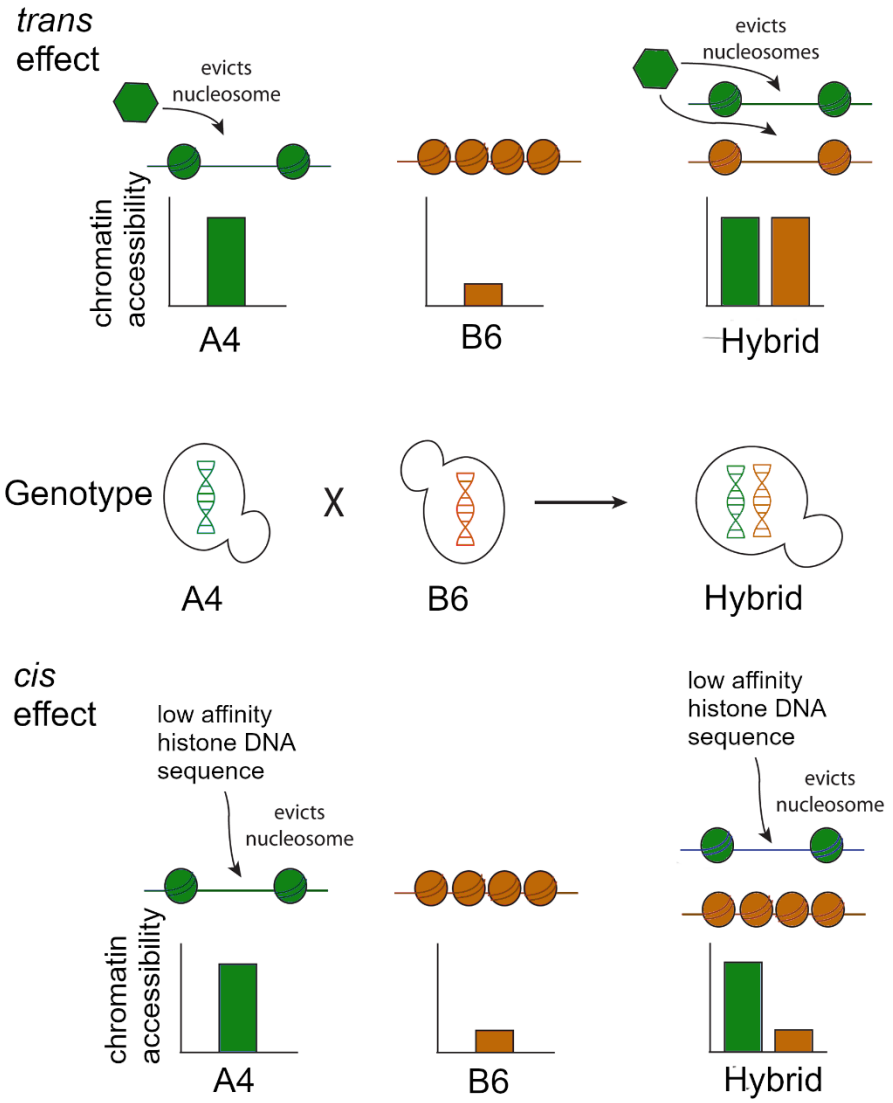


Fig intro.2: Schematic of approach to detect cis and trans effects on chromatin accessibility adapted from Connelly et al., 2014 *[131].*

(Left) examples of nucleosome evictions in trans by ATP-dependent nucleosome remodeling complexes (shown as a hexagon), and in cis by low affinity histone DNA sequence for one allele in A4 isogenic genotype. (Middle) an example of one allele from B6 isogenic genotype without any nucleosome modification. (Right) example of open chromatin region in trans and in cis. An open chromatin is in trans when there is a difference in chromatin accessibility in parental haploids, but there is no such difference between the two alleles in the diploid hybrid. An open chromatin is in cis when there is a difference in chromatin accessibility in parental haploids, and the cis effect is shown by the same difference in accessibility detected between the two alleles in diploid hybrids.

CHAPTER 1

Genetic Variation in Chromatin State Across Multiple Tissues in *Drosophila melanogaster*

1.1 ABSTRACT

We use ATAC-seq to examine chromatin accessibility for four different tissues in *Drosophila melanogaster*: adult female brain, ovaries, and both wing and eye-antennal imaginal discs from males. Each tissue is assayed in eight different inbred strain genetic backgrounds, seven associated with a reference quality genome assembly. We develop a method for the quantile normalization of ATAC-seq fragments and test for differences in coverage among genotypes, tissues, and their interaction at 44099 peaks throughout the euchromatic genome. For the strains with reference quality genome assemblies, we correct ATAC-seq profiles for read mis-mapping due to nearby polymorphic structural variants (SVs). Comparing coverage among genotypes without accounting for SVs results in a highly elevated rate (55%) of identifying false positive differences in chromatin state between genotypes. After SV correction, we identify 1050, 30383, and 4508 regions whose peak heights are polymorphic among genotypes, among tissues, or exhibit genotype-by-tissue interactions, respectively. Finally, we identify 3988 candidate causative variants that explain at least 80% of the variance in chromatin state at nearby ATAC-seq peaks.

1.2 INTRODUCTION

Many human complex diseases, such as heart disease and diabetes, are highly heritable [1] . Large high-powered Genome-Wide Association Studies (GWAS) have dominated the study of such diseases over the last decade, but despite thousands of associations between markers and traits, the exact causative variants underlying risk typically remain hidden [2,3], and an appreciable fraction of heritable variation remains unexplained [4]. Recent papers propose that variation in human complex traits is due to thousands of mostly intermediate-frequency, tiny effect variants [5–7]. In contrast, QTL mapping studies in yeast [8,9], mouse[10–12] , and *Drosophila* [13] consistently map factors of much larger effect, with mapped QTL collectively explaining a considerable fraction of heritability in a cross. Efforts to fully characterize complex trait loci in model systems may hold the most promise for "lifting the statistical fog" [14] associated with genetic mapping, and point to causative, functional alleles.

A promising strategy for identifying causative variants at candidate genes identified via GWAS or QTL mapping is to focus on regions near those genes that act as *cis*-regulators of gene expression. There is now a preponderance of evidence that the bulk of variation in complex traits is due to regulatory variants [6,15–18] , with little evidence that amino acid variants explain human GWAS hits [19] . Yet, so little is actually known about complex traits that even this claim is debated [19,20]. Until recently, non-coding regions with *cis*-regulatory function have been difficult to identify at scale, but genome-wide profiling of open chromatin regions using DNase-I HS (DNase-I hypersensitive sites) sequencing [21] and/or the more experimentally straightforward ATAC-seq (Assay for Transposase Accessible Chromatin) approach [22] have allowed

characterization of chromatin state in large panels of genotypes [23,24]. ATAC-seq employs the Tn5 transposase sequencing chemistry to make an Illumina-compatible paired-end sequencing library using nucleosome-bound DNA as template for the transposition reaction. Regions of DNA bound by transcription factors or nucleosomes are protected from Tn5 insertion, whereas more open chromatin regions - likely harboring active *cis*-regulatory features - are associated with higher levels of sequence coverage. Much like RNA-seq data, open chromatin regions identified by ATAC-seq can vary among tissues, developmental timepoints, and genotypes [23,25–27] . In terms of the genetics of complex traits, chromatin features displaying variation among genotypes, especially in a tissue-specific manner, are of considerable interest as potential contributors to trait variation.

Multiple DNase1-HS-seq and ATAC-seq studies have been carried out in *Drosophila melanogaster* [26,28–41] as well as other insects such as *Anopheles gambiae* [42]. The majority of *Drosophila* studies have focused on a single genotype (or cell line), have compared different mutant backgrounds, or have employed a small number of wildtype strains that lack a high-quality genome sequence (*c.f.* [26,29,33,35,36,38,39]). In no case has the genotype queried been the *Drosophila melanogaster* reference strain (*i.e.*, Bloomington stock 2057 or “iso1”), the strain ATAC-seq reads are generally aligned to. There are routinely a considerable number of SNPs, short insertion/deletion variants, and a wide array of structural variants (SVs) distinguishing any pair of *Drosophila* strains [43] , and such events – if they are effectively “hidden” due to the absence of high-quality genomes for the target strains – may complicate the analysis of chromatin state, as has been observed with RNAseq

data [44] . Furthermore, chromatin accessibility studies in *Drosophila* have focused principally on early embryonic stages [26,30–32,34–37,41], cell lines [28], or whole adults, with only five studies examining specific adult tissues or imaginal discs [29,33,38–40]. In terms of the complex traits that tend to be studied in the *Drosophila* research community, which are skewed towards traits measured in adults and larvae (*c.f.* Table 3 of [45]), *cis*-regulatory elements active in imaginal discs or adult tissues are likely of broad interest.

Here we carry out a biologically-replicated ATAC-seq experiment to characterize chromatin accessibility in four adult tissues in several highly-characterized isogenic genotypes of *D. melanogaster* [43] (throughout this paper we use genotype to refer to a genome wide genotype or isogenic strain). We identify a set of peaks with evidence for an open chromatin configuration in at least one of the tissues. Unlike previous studies, and inspired by the quantile normalization method deployed in microarray research [46] , we develop a method for normalizing ATAC-seq reads across tissues, genotypes, and biological replicates. We carry out statistical tests to identify ATAC-seq peaks that differ in coverage as a function of tissue, genotype, or that display a tissue-by-genotype interaction. By virtue of studying highly isogenic genotypes with reference quality *de novo* assemblies, we correct for artifacts in peak coverage due to hidden SVs. We show that a failure to correct for the impact of SVs can result in a high rate of peaks inferred as differing between genotypes, which are in fact due to mis-mapped read pairs. We finally identify a set of SNPs near to variable ATAC-seq peaks that potentially represent candidate causal *cis*-acting factors.

1.3 RESULTS

Workflow and samples: We dissected wing and eye-antennal imaginal discs from male third instar larvae, and brains and ovaries from adult females, for eight *Drosophila* Synthetic Population Resource [47] founder strains. All eight have been re-sequenced using short-read sequencing [48], while seven have extremely well characterized, reference-quality genomes [43]. For each tissue and genotype combination we obtained three biological replicates. The eight genotypes chosen are highly inbred and represent a world-wide sampling of variation within the species (see S1.1 Fig and S1.1 Table). Dissected samples were immediately processed to make indexed ATAC-seq libraries [22] and sequenced to obtain 20-147 million Illumina paired-end reads per sample (mean=73M, SD=21M). Reads were aligned to the *D. melanogaster* reference genome (dm6) and pooled across genotypes, but within tissues, to identify open chromatin “peaks" located throughout the euchromatic genome using MACS2 [49]. Individual replicate/genotype/tissue samples were separately normalized to obtain a weighted coverage at each identified peak. Finally, we utilized reference quality assemblies for the seven assembled strains to correct read coverage statistics for the presence of nearby polymorphic structural variants (SVs) and carried out statistical tests at peaks to identify chromatin structures that varied among the four tissues, the seven genotypes, or exhibited a tissue-by-genotype interaction. Our general workflow is depicted in Supplementary Figure S1.2 and read mapping statistics for each sample are given in Supplementary Table 1.4.

ATAC-seq identifies open chromatin regions across four tissues in the *Drosophila* genome: Peaks were filtered to only include those in euchromatin regions (see S1.2 Table) that were also significantly enriched above background at $p<0.01$ as defined by MACS2. After filtering, we identified 25464, 18111, 18496, and 17413 euchromatic peaks for adult female brain, ovary, eye-antennal imaginal disc and wing imaginal disc tissue, respectively. Venn diagrams showing peaks shared among tissues for the set of peaks enriched at $p<0.01$ and at $p<0.001$ (Fig 1.1A) are qualitatively similar, supporting the idea that the significance threshold for enrichment that is employed only subtly impacts the collection of peaks we consider. The Venn diagram at $p<0.01$ (Fig 1.1A) shows that although peaks shared among tissues are not uncommon – 9.8%, 7.5%, and 17.2% of the total collection of peaks are shared by all four, three, or two tissues, respectively – 65.6% of the peaks are private to a single tissue. Brain tissue exhibits the highest number of private ATAC-seq peaks, but even the pair of disc tissues – which one might naively think would be the most similar of our target tissue types – have appreciable numbers of private peaks, highlighting the value of tissue-specific chromatin characterization.

For each peak, within each tissue, we characterized fold enrichment (a measure of peak “height” based on read count in the peak relative to the local background [49]) to explore whether the properties of the peaks we identify resemble those observed in previous studies. Figure 1.1B, depicting fold enrichment for the brain, shows that the vast majority of peaks (>90%) have fold enrichments of less than 5. We observed the same trends for the four other tissues (see S1.3A, S1.4A, and S1.5A Figs). This observation is consistent with results from DNase1-HS-seq experiments [50,51] and

other ATAC-seq datasets [52,53] . We further examined the distribution of fold enrichment as a function of distance from transcription start sites (TSSs) for brain peaks, as TSSs often exhibit strong enrichment patterns [54,55] . Figure 1.1C shows fold-enrichment as a function of distance from the TSS for the female brain (S3B, S4B, and S1.5B Figs for the other tissues). The patterns we see largely mirror other studies [52,54,55] . We finally examined average fold enrichment as a function of HOMER annotation type (Fig 1.1D depicts female brain). Fold enrichments are strongest for 5'UTR, TSS, and perhaps transcription termination sites (TTS). Enrichments are more subtle for other feature types, although for all feature types there was clearly a subset of peaks with strong fold enrichment scores. The same trend in peak enrichment with regard to feature types can also be observed in other tissues (S1.3C, S1.4C, and S1.5C Figs). Overall, properties of the ATAC-seq peaks observed for our four target tissues are comparable to those observed in the Drosophila literature [56] , giving us confidence that the peaks of this study are robustly inferred. Finally, there is some suggestion that more highly enriched peaks (e.g., those near TSSs) tend to be more likely to be shared among tissues. Supplementary Figure 1.6 shows the degree of peak sharing among tissues as a function of the feature type that peak is located in.

Our next goal was to obtain a common set of genomic locations (or loci) at which statistical tests to evaluate variation in chromatin accessibility over genotypes and/or tissues could be carried out. To do this we merged peaks (*i.e.,* the single base position where coverage peaked) over all tissues and genotypes that were within 200-bp of one another, and whose MACS2-defined boundaries overlapped. In contrast, peaks that were separated by more than 200-bp were not merged even if their MACS2 boundaries

overlapped. To illustrate the merging procedure Figure 1.2 (top panel) depicts a representative ~30kb region centered on the gene *hairy* (a gene contributing to embryonic segmentation and peripheral neurogenesis) showing peaks called separately for each of the four tissues, as well as the consensus set of peaks with adjacent peaks merged (the “all tissues” track; see methods). Red hashes show the location of each peak, and horizontal black bars depict the entire peak interval from MACS2. The lower panel zooms in on a smaller 10kb region with a more detailed depiction of the raw coverage data (the y-axis is fold enrichment). As with typical ATAC-seq datasets we often see a strong peak near the TSS that is consistently identified across tissues. In contrast, for non-TSS peaks, MACS2 boundaries may only sometimes overlap depending on the tissue. The lower panel illustrates how our heuristic merges peaks close to one another across tissues to define a single peak location (red hashes). The heuristic gives a single “all tissues” location for the peak associated with the TSS of *hairy*, despite the peak position varying slightly among tissues. Furthermore, consider the region downstream of the 3’ UTR of *hairy*, the MACS2 boundaries (indicated by the black bars) for two peaks overlap for ovaries, but not for the two disc tissues, and the six peaks each have different locations. Despite the MACS2 boundaries overlapping in ovaries, the raw coverage clearly suggests two peaks. As those peaks are greater than 200bp apart, the heuristic calls two peaks and further merges the positions of those two peaks across tissues. An algorithm that merges peaks based on overlapping boundaries, especially when data is collected from multiple tissues, would merge these two peaks (since their boundaries overlap), despite evidence they are separate. Based on visual inspections of the fold-enrichment profiles for many other regions (not shown)

we observe many such instances where merging peaks based on overlapping MACS2 boundaries, especially those observed only in a subset of tissues, seems misleading, whereas keeping the peaks separate appears correct.

Normalizing coverage corrects for sample-to-sample variation and the presence of structural variants: Different samples yield different numbers of raw reads. Additionally, histograms of ATAC-seq fragment lengths show a characteristic periodicity representing nucleosome free DNA, mono-nucleosome bound DNA, di-nucleosome bound DNA, and so on (see Figure 1.2 of [22] and S1.7 Fig for convenience). Figure 1.3, depicts the distribution of raw fragment lengths for two biological replicates of brain tissue ATAC-seq from the A4 strain in red (*i.e.*, independent tissue dissections and library preps). It is evident that replicate 2 has more nucleosome bound DNA than replicate 1. We hypothesize that such differences might arise from subtle differences in sample prep that result in different rates of disassociation of nucleosomes from DNA, and this sample-to-sample variation is likely challenging to experimentally control for. To allow comparisons across tissues and genotypes we normalized each sample so that the genome-wide distribution of fragment sizes are identical (see methods) using an approach akin to the quantile normalization technique used extensively in the context of gene expression [57]. Our normalization results in a weight being assigned to each fragment and by working with those weights, as opposed to raw fragment counts, histograms have identical fragment size distributions across all samples (Fig 1.3, blue curves). This normalization allows for straightforward statistical testing between tissues and genotypes. Supplementary Figure 1.8 depicts the distribution of fragment lengths

across the 96 samples of this study prior to normalization and the removal of one sample due to low data quality.

A second concern often ignored in ATAC-seq analysis, that can make it difficult to compare samples across genotypes, is the presence of structural variants that could masquerade as polymorphisms in chromatin structure. ATAC-seq data obtained from different genotypes are generally aligned to a single reference genome, and a polymorphic structural variant near an ATAC-seq peak can result in unaligned reads, which will present as a local drop in coverage, and lead to the incorrect inference of more closed chromatin in that region of the genome. The eight genotypes examined in this study are highly isogenic and seven are associated with reference quality *de novo* assemblies, putting us in the unique position of being able to correct for polymorphic structural variants. We correct for SVs by excluding all fragments across *all* samples that span a structural variant present in *any* of the several assembled samples.

We illustrate the impact of correcting for SVs on wing disc ATAC-seq data for a 10kb region around the *rpr* gene (a gene important in programmed cell death) for two genotypes (B6 in brown, A4 in green), B6 harbors a ~17kb *mdg1* transposable element ~5kb upstream of the TSS of *rpr* (Fig 1.4A). In the uncorrected for SVs wing disc dataset, there is an apparent difference in chromatin configuration near the *mdg1* insertion. But after correcting for reads mis-mapped due to the *mdg1* insertion it appears that such an inference is incorrect and the lower coverage in the B6 genotype is largely due to mis-mapped reads associated with the *mdg1* TE. Although, even after correcting for the *mdg1* insertion, there does appear that there is a subtle difference in chromatin structure to the right of its location. Interestingly, this region contains two

other SVs (a 2.8 kb *F* insertion in A4 and a 111bp deletion in B6) that do not impact chromatin structure inferences.

Figure 1.4B depicts a second example of an ATAC-seq peak in the first intron of the *Mef2* gene, whose product is crucial in myogenesis [58] . The top panel shows a 45kb region centered on the *Mef2* gene, while the bottom panel zooms in on a 550bp region entirely contained within the first intron showing coverage for brain samples with and without SV correction. In the SV-uncorrected data, this peak significantly varies by genotype with a $-\log_{10}$(FDR p-value) of 3.7. Seven of the genotypes exhibit a relatively open chromatin configuration, while A4 (in dark green) exhibits lower coverage in a region that contains a TE insertion. In a typical experiment, where the existence of the TE insertion would be unknown, and "hidden" from short read callers, the effect on read mapping of the TE would not have been corrected for, and we would have incorrectly inferred a genetic difference in chromatin accessibility. After SV correction, the ATAC-seq peak is not identified as being polymorphic. It is important to note that our correction acts by masking regions close to SVs in non-SV containing samples, so our proposed solution is far from perfect. But uncharacterized structural variants in non-reference genotypes can clearly cloud the interpretation of ATAC-seq datasets (as we show below).

Although both panels of Figure 1.4 illustrate transposable elements, other structural variants can impact read mapping. Supplemental Figure 1.9 depicts a polymorphic 1.9kb deletion relative to the reference in strain A4 in the first intron of the *Abl* gene. The deletion knocks out two ATACseq peaks, but if only mapping short reads to the reference strain, A4 would appear to have a closed chromatin configuration.

ANOVA identifies polymorphic chromatin structures: For every merged-peak in the euchromatic genome we carried out an ANOVA to determine if chromatin accessibility varies across Tissues, Genotype, or their interaction (T:G). We carried out this analysis for data either corrected or uncorrected for polymorphic structural variants for the seven genotypes with reference genomes. As the statistical analysis involved roughly sixty-eight thousand peaks and three *p*-values for each peak (Tissue, Genotype, and their interaction) we convert *p*-values to a false discovery rate and consider a test significant if the FDR is less than 0.5%. Table 1.1 gives the number of significant chromatin profile differences by factor, and Figure 1.5A shows tissue overlap using a Venn diagram. A robust observation is that for the SV-corrected data, close to 100% of all peaks display differences in chromatin features among the four tissues we examine. Of the peaks showing differences between tissues ~84% are not significant for a genotype or tissue by genotype interaction (Fig 1.5A). Thus, chromatin features are far more likely to vary among tissues than genotypes. Although differences between genotypes are far less frequent than differences between tissues, we still identify roughly 1000 such peaks (Table 1.1). Interestingly we identify roughly four times as many tissue by genotype interactions than simple genotype specific peaks. Finally, we created Manhattan plots for all ANOVA tests, and observe that SV-corrected “hits” are largely uniformly distributed throughout the euchromatic genome with no evidence for “hotspots” (S1.10 Fig), although perhaps there is a tendency for an increased rate of significant genotype hits nearer centromeric regions (despite aggressive filtering for euchromatin only regions).

Since we carried out ANOVA on both SV corrected and uncorrected data, we can assess the impact of failing to correct for SVs on inference. There is considerable overlap in those peaks showing tissue-only effects between the uncorrected and SV-corrected datasets (Fig 1.5C). In contrast, we observe many fewer ATAC-seq peaks following SV correction in the genotype and tissue-by-genotype peak sets (Fig 1.5B & 1.5D): Of the peaks identified in the uncorrected analysis, 55% for the genotype-only set, and 21% for the tissue-by-genotype set are eliminated by correcting for SVs. We more carefully examined the peaks eliminated by SV correction (*n*=1441 genotype-only, *n*=4041 tissue-only, *n*=1382 interaction) to determine what might be driving their disappearance (Table 1.2). The vast majority of these peaks - 89%, 99%, and 90% for genotype, tissue, or the interaction, respectively - are either completely contained within an SV or are within 800bp of an SV boundary (Table 1.2). The location of these peaks suggests that they are purely the result of incorrect mapping of short sequencing reads from a non-reference genotype to a common reference genome. The remaining genotype- and interaction-only peaks that disappear following SV correction, but that are greater than 800bp from an SV, appear to be excluded by just failing to survive thresholds. Either they are eliminated by having their average coverage drop just below our threshold of 50 following SV-correction, or by just failing to reach our 0.5% FDR threshold in the SV corrected dataset (S1.11 Fig). Failing to correct for SVs during ATAC-seq peak calling - as is the norm when *de novo* genome assemblies are not available for the target strains - will generate large numbers of false positive peaks that do not, in truth, impact chromatin accessibility.

Supplementary Figure 1.12 depicts false positive differences between genotype, tissues, or a genotype by tissue interaction as a function of the SV-type corrected for, and if the ATACseq peak is inside the SV or instead within 800-bp of an SV. In the case of indels, for example, an ATACseq peak could be contained within a deletion present only in one of the non-reference strains. In contrast, for a peak to be within a TE, that TE would need to be present in the reference strain and absent in the other strains examined, due to the way mapping to a reference genome works. Chakraborty *et al.* [43] observed 7347 TE insertions, 1178 duplication CNVs, 4347 indels, and 62 inversions in the euchromatin genomes of DSPR strains based on de novo sequencing. As expected, TEs dominate false positives due to SVs within 800bp of an ATACseq peak, whereas INDELs dominate the landscape for peaks contained within an SV. In general, the likelihood of a false positive is a complex function of the type of event, its population frequency, and how that event presents to short read mappers relative to the reference genome.

Examples of polymorphic chromatin structures: Figure 1.6A depicts SV-corrected coverage brain and ovary samples centered between the TTS of the *Npc2f* gene, whose human ortholog (*NPC2* gene) is implicated in Niemann-Pick disease and Niemann-Pick disease type C2 due to its involvement in regulating sterol transport [59] , and the TSS of *Kal1* gene, whose human ortholog (*Anosmin-1* gene) is responsible for the X-linked Kallmann's syndrome [60]. We observe a genotype polymorphism in chromatin state with the B2 genotype (light green) exhibiting a more closed chromatin state compared to the other genotypes for ovary tissue ($-\log_{10}$(FDR p-value) = 3.6). This ATAC-seq peak is

further polymorphic by tissue with brain tissue exhibiting a generally closed chromatin state. We speculate that the B2 genotype has lower expression of *Kal1* in ovaries, with the chromatin structure impacting its TSS. Figure 1.6B depicts a polymorphic ATAC-seq peak located within the intron or near the TSS of the *eIF4A* gene (depending on isoform), with *eIF4A* acting as RNA-dependent ATPase and ATP-dependent RNA helicase that facilitates attachment of the 40S ribosomal subunit [61] . Coverages are higher for A5, A7, and B3, than the other genotypes in wing disc tissue ($-\log_{10}$(FDR p-value) = 3.5) with other tissues (not shown) showing similar trends in coverage. The location of the peak suggests a role in mediating isoform usage between genotypes via an alternative TSS, with the peak heights suggesting an allelic series. Figure 1.6C is an example of two adjacent peaks exhibiting a genotype:tissue interaction ($-\log_{10}$(FDR p-value) = 4.5 and 4.4 respectively) located in intron 1 of the *Eip75B* gene isoform F, and near TSS of *Eip75B* gene isoform E. This gene has been shown to regulate the complex traits of feeding behavior, fat deposition, and developmental timing [62–64]. As with the example of *EIF4A* we speculate that this polymorphism impacts isoform usage via alternative TSSs. Figure 1.6D depicts four peaks polymorphic by tissue, or by genotype:tissue interaction, for an interesting 14kb region directly upstream of TSS of *hairy*. *hairy* is well studied in the context of developmental biology [65–67] and the genetics of complex traits [68–70] , with several *cis*-regulatory enhancers in this region playing a role in regulating the seven stripes formed in the blastoderm stage [71,72] . The four ATAC-seq peaks exhibit chromatin configurations that vary among tissues ($-\log_{10}$(FDR p-value) = 14.7, 17.9,10.9, and 14.9 left to right). Finally, the peak at

chr3L:8672906 is polymorphic for a genotype:tissue interaction ($-\log_{10}$(FDR p-value) = 2.3).

Candidate causative SNP identification: For each of the SV-corrected peaks significant for a genotype or genotype:tissue interaction we estimate the proportion of variation in peak height explained by each SNP (or marker) within 250bp of the peak (Fig 1.7). We speculate that such SNPs are strong candidates for *cis*-regulatory factors that control chromatin configuration. Two caveats are that we are only examining seven reference genomes so our models may be over-fitted, and truly implicating events as causative would require gene replacement experiments. We identify and test 6707 and 33570 SNPs located within 250bp of genotype or genotype: tissue interaction specific ATAC-seq peaks respectively. Out of those, there are 1253 (18.7%), and 2735 (8.1%) SNPs that explain greater than 80% of the variation in peak heights due to Genotype or G:T respectively, an average of 6 nearby SNPs per significant peak (Table 1.3). We further annotate all SNPs that explain 100% of variance for functional impact. Out of 687 SNPs that explain 100% of variance in peak height (by genotype or for a genotype: tissue interaction), there are a total of 22 SNPs annotated as having a high functional impact (i.e., missense, premature start codon, or splice variant), which is odd given that there is no reason to think a mutation of high functional impact on a transcribed protein is likely to impact a nearby chromatin configuration. The potential functional impact of the remaining 665 SNPs is more difficult to discern (S1.13 Fig).

Examples of potentially causative SNPs: Figure 1.8A depicts an ATAC-seq peak downstream of TTS of *Bre1* isoform A and exon 4 of isoform B, a gene involved in regulation of *Notch* signaling [73–75]. Genotypes B2 and B3 are more closed in the eye disc and brain compared to all other genotypes (Fig 1.8A). There is a potential causal SNP almost centered on the peak explaining 100% and 57% of the variation in eye disc and brain respectively. Figure 1.8B depicts a polymorphic peak for brain and ovary in which the A6 (purple) genotype appears more open than the others ($-\log_{10}$(FDR p-values) = 6.1 and 10.2 for genotype and G:T respectively). The peak is located in a intron of *Ptpmeg* (involved in the maintenance of axon projection [76] and inhibition of EGFR/Ras/mitogen-activated protein kinase signaling pathway during wing morphogenesis [77]), as well as ~400bp downstream of TTS of *mthl9* (whose gene subfamily plays important role in *Drosophila* development, stress response, and regulation of life span [78]). Two nearby SNPs each explain 100% of variation in genotypes, and both are private to the A6. Figure 1.8C depicts a peak polymorphic by genotype that appears largely brain specific with A7 (pink) being more closed relative to other genotypes, and B2 (light-green) perhaps more open slightly downstream but not associated with a called peak. A nearby SNP private to A7 in the 5'-UTR (and 51bp downstream of a TSS) of a *Nna1* isoform explains 99% variance in genotype in the brain. Figure 1.8D depicts potentially causal SNPs exhibiting a genotype:tissue interaction located upstream of two TSSs for the gene *stv* (involved in the chaperone pathway essential for muscle maintenance [78,79]). For both peaks and tissues the A4 (green) genotype exhibits a more closed configuration especially in the wing disc and to

a lesser extent the brain. A SNP private to A4 explains 81%, 95%, 55%, and 97% of the variance in coverage for brain and wing disc at left and right peaks respectively.

1.4 DISCUSSION

Previous ATAC-seq/DNase1-HS-seq experiments in *Drosophila* have focused almost exclusively on embryos, whole adult bodies, or cell lines, and only rarely have compared multiple genotypes. We carried out a replicated ATAC-seq experiment on two adult tissues (female brains and ovaries) and two imaginal disc tissues (wing and eye-antennal imaginal discs) from which adult tissues are ultimately derived. It is widely believed that the sites that contribute to complex trait variation are likely to be regulatory in nature, thus chromatin features expressed in adult tissues are strong candidates to harbor such causative sites. Thus, we expect the data collected as part of this experiment will be of utility to the *Drosophila* complex trait community, who tend to study traits that manifest in adult or larval flies (c.f. Table 3 of Mackay and Huang 2018)[45], and see utility in our distributing coverage as a function of genotype and tissue as a series of Santa Cruz Genome Browser tracks (http://goo.gl/LLpoNH). We characterize eight highly isogenic strains of *Drosophila* that are a subset of the strains used to found the *Drosophila* Synthetic Population Resource [48] , with seven of those strains having reference quality genome assemblies levels [43] . We largely employ a standard ATAC-seq peak calling pipeline, apart from our strategy for merging peaks within and between tissues, to obtain a union dataset consisting of 44099 open chromatin peaks.

Our analyses identified approximately thirty thousand peaks that differed in coverage between tissues, highlighting the future need for tissue specific chromatin

maps. We further identified on the order of one thousand chromatin peaks that differ among genotypes and five thousand that vary among genotypes in a tissue specific manner. Chromatin peaks that differ among genotypes associated with candidate genes identified via QTL mapping in DSPR [13] or GWAS using DGRP [80] are strong candidates for contributing to differences in gene expression levels. Surprisingly, peaks displaying genotype by tissue interaction are more frequent than the genotype specific peaks. Such peaks represent candidates for modulating gene expression in a genotype dependent manner in a small subset of tissues that gene impacts. It is reasonable to speculate that ATACseq peaks displaying tissue by genotype interactions underlie QTL that appear to be tissue or complex trait specific and do not show a great deal of pleiotropy.

We carried out statistical testing to identify chromatin states that vary among tissues, seven of the eight wild-type genotypes, and/or exhibit a tissue by genotype interaction (*i.e.*, differences among genotypes varying in the tissue dependent manner). To facilitate statistical testing, we carried out two important data normalization steps unique to this study. We first developed a per sample normalization procedure that creates per fragment weights that control for differences between samples in the total number of reads, and the percentage of read pairs that are nucleosome-free, mononucleosomic, binucleosomic, etc. The degree to which normalization impacts inference depends on how similar the fragments distributions are between samples. Some tissues seem easily amenable to ATACseq preps, especially cell lines, in which case perhaps no correction is necessary. On the other hand, more difficult tissues, will result in larger differences in fragment size distributions, and the correction is more

likely to be beneficial. Any normalization method is likely to be most useful when comparing genotypes within tissue, where subtle differences in ATACseq peak heights could be biologically meaningful. There is some risk that differences in fragment size distributions between tissues could be biological in origin, a problem shared among all between tissue normalization methods. These caveats acknowledged, the observation of fragment size distribution differences between biological replicates suggests that normalization may be beneficial.

A second important normalization step attempted to control for false positive inferences due to hidden structural variants. By virtue of seven of the eight isogenic strains being associated with reference quality *de novo* assemblies [43], we control for the potential artifact of polymorphic structural variants creating read-alignment differences that in turn could masquerade as differences in chromatin configuration. We accomplish this by masking regions in all strains harboring a nearby SV present in any strain. We carried out statistical testing on datasets either ignoring or following correction for polymorphic SVs and estimated the potential to identify false positives in data sets where SVs are hidden. Failure to account for SVs does not strongly impact the inference of differences in coverage between tissues, but it can have a huge impact in terms of detecting difference in chromatin accessibility between genotypes or those showing a genotype by tissue interaction, where we estimate potential false positive rates of 48% and 19% respectively. Our method of correcting for SVs is conservative and consists of masking regions associated with polymorphic SVs.

A shortcoming of our masking SVs is that we cannot perform in depth analyses of possible biological effects of the SVs themselves (unless they exert those effects

over distances longer than ~800bp). A potential solution would be to align reads to a genome private to each strain, followed by lifting those alignments over to a universal coordinate system to compare genotypes. Although this approach works well for SNP-based variation in well-behaved genomic regions, we find that lift-overs tend to break down when structural variants distinguish strains [81], and this is especially problematic for events like duplications where there is not even a 1:1 mapping between genomes. While our method of masking SV is not perfect, it is simple to implement and can remove upward of 50% of false positive peaks.

For ATAC-seq peaks that vary significantly in coverage among genotypes or that show a tissue by genotype interaction, we attempted to identify nearby SNPs (or markers) that may control that variation. It is both reasonable to suggest, and supported by experiments (c.f., [24,81–85]), that alleles that control chromatin accessibility peaks are likely to be in *cis* and physically close to the peak. We identify several thousand such SNPs that explain more than 80% of the variation due to genotype or a genotype by tissue interaction for coverage, a collection likely enriched for causative polymorphisms, despite our over-fitting of the data. It would be of value to extend this approach to a much larger collection of genotypes, although such work may necessitate focusing on a single tissue and require more *de novo* genome sequences to control for hidden structural variants. As Crispr/Cas9/allele swapping methods continue to come of age in *Drosophila* [86–90] medium-throughput functional assays capable of confirming specific allele chromatin peaks interactions could alternatively be used to characterize alleles regulating nearby chromatin states and gene expression levels.

1.5 MATERIALS & METHODS

Strains: We employed 8 of the 15 strains that serve as founders of the *Drosophila* Synthetic Population Resource (DSPR), a multiparental, advanced generation QTL mapping population consisting of hundreds of recombinant inbred lines [48]. These highly-inbred strains - A4, A5, A6, A7, B2, B3, B6, and B7 (S1.1 Table) - are a worldwide sample of genotypes (S1.1 Fig), and seven of the eight (excluding B7) have reference quality assemblies such that virtually all SVs are known [43,48] .

Tissue dissection and ATAC-seq library preparation: The 8 inbred strains were raised and maintained in regular narrow fly vials on a standard cornmeal-yeast-molasses media in an incubator set to 25°C, 50% relative humidity, and a 12 hour Light : 12 hour Dark cycle. We isolated nuclei from four different tissues for our 8 target strains. (1) Adult brains (central brain + optic lobes) were dissected and pooled from ten 1-4 day old females per replicate. (2) Ovaries were dissected and pooled from five 1-5 day old females per replicate. (3) Wing imaginal discs were dissected and pooled from 3-7 male wandering third instar larvae per replicate. (4) Eye-antennal imaginal discs were dissected and pooled from 4-7 male wandering third instar larvae per replicate. For each strain/tissue combination we generated 3 replicates. All dissections were carried out 1-9 hours after lights on, and following dissection all samples were immediately subjected to nuclei isolation.

Our full protocol for ATAC-seq library construction is provided in Supplementary Text 1, but in brief: Animals were dissected in nuclei lysis buffer under a standard

stereoscope, and dissected tissue for a given replicate pooled into 200-µl of nuclei lysis buffer on ice. Each sample was then subjected to manual grinding, passed through 30-µM filter cloth, spun down, and the supernatant removed. Subsequently, 25-µl of tagmentation reaction mix was added to the pellet, and incubated at 37°C for 30-min before freezing at −20°C. After thawing, the sample was cleaned using a MinElute PCR purification column (Qiagen, 28004), and an aliquot was subjected to PCR to add on custom, Illumina-compatible indexing oligos. Finally, samples were cleaned using a standard bead-based approach, quantified using a Qubit dsDNA BR kit (ThermoFisher, Q32850), and examined via a TapeStation 2200 using genomic DNA ScreenTapes (Agilent Technologies, 5067-5365 / 5067-5366).

All 96 libraries (8 strains × 4 tissues × 3 replicates) were pooled at equal amounts - along with a series of other libraries that are not part of the project - and run over 16 lanes of an Illumina HiSeq4000 sequencer at the UCI Genomics High-Throughput Facility collecting PE50 reads.

Read processing: Adapters were trimmed from the raw reads using Trimgalore-0.4.5 [91,92], and trimmed reads were aligned to the dm6 *D. melanogaster* reference genome [93] using bwa 0.7.8 [94] . Unmapped reads, and reads with unmapped mates were removed with samtools 1.3 (option -F 524 -f 2) [95] , and all non-primary reads and improperly aligned reads were also removed with samtools 1.3 (option fixmate -r and option -F 1084 -f 2). Following this, duplicate reads are removed using picard 2.18.27 [95,96] via MarkDuplicates and REMOVE_DUPLICATES=TRUE. Only reads aligning to the five major chromosome arms - X, 2R, 2L, 3R, and 3L - were retained for analysis.

BAM files were corrected to reflect the actual insertion points of the Tn5 transposase acting as a dimer by having plus strand reads shifted +4bp and minus strand reads shifted −5bp as suggested in [22] . We refer to these as “corrected BAM files”. Paired end BED files reflecting mapped fragments were generated using bedtools 2.25.0 [97] . The same process was carried out for all 96 samples (8 genotypes, 4 tissues, and 3 replicates).

ATAC-seq peak calling: Corrected BAM files from all 96 samples were merged by tissue across replicates and genotypes, and MACS2 [49] was used to call peaks separately on the ovary, brain, wing disc, and eye disc. MACS2 options were -f, -p 0.01 to set cut-off p-value for peaks to be considered significant, -B --SPMR, --no-model to skip any read shifting as we were using corrected BAM files, and -g was set to 142573017, the summed length of the major chromosome arms in the dm6 genome release. The peak calling resulted in four ENCODE “tissue NarrowPeak files”, one for each tissue.

Merging of peaks across tissues: Tissue NarrowPeak files were concatenated, sorted by chromosome and peak summit, then a custom python script grouped and averaged peak summit locations that were within 200 bp of one another, but greater than 200 bp from the nearest adjacent peak summit. Each averaged peak summit is associated with a minimum left interval boundary and maximum right interval boundary obtained from all the summit peaks contributing to an average peak. The resulting file was converted to ENCODE NarrowPeak format for viewing using the UCSC genome browser with "peak"

as peak name,"1000" as peak score, "." as peak strand, "10" as peak enrichment, and "-1" as q-value and p-value to accommodate the ENCODE NarrowPeak format, referred to as the “all tissue” track/peak file. Only the chromosome and the mean peak summit columns are used in downstream statistical analysis steps.

Euchromatin peak filter and peak annotation: We choose to focus solely on euchromatic regions of the genome since heterochromatic regions are gene poor, poorly annotated, and enriched for structural variants and transposable elements. The euchromatin region boundaries we employ are given in Supplementary Table 1.2 and come from [98]. All peaks in the all tissue peak file, and the four tissue NarrowPeak files used in downstream analyses, include only euchromatin located peaks.

We used HOMER v 4.11.1 [99] and the tissue NarrowPeak files separately for each of the four tissues to annotate each peak summit as belonging to one of eight exclusive groups based on their location relative to features annotated in the dm6 reference genome: (1) transcription start site (TSS: −1000 to +100bp from the transcription start site), (2) transcription termination site (TTS: −100 to +1000bp from transcription termination site), (3) coding exons, (4) 5' UTR exon, (5) 3' UTR exon, (6) intronic, (7) intergenic, and (8) non-coding (referring to non-protein-coding, but nonetheless transcribed DNA). In the case of a peak belonging to more than one feature type it is assigned to a single feature type with priority according to the numeric order of the features in the previous sentence (*i.e.*, TSS has priority over 5’ UTR, etc). Since we focus in this work on peak summits, whereas HOMER annotates a peak as being at the mid-points of an interval, we edited the interval associated with each peak

to be the peak summit +/− 1bp. The percentages of peaks falling into each feature type by tissue are given in the Supplementary Table 1.3. These percentages are compared to the annotation types associated with one million randomly assigned peak locations. Comparing these percentages between tissues and/or to other studies is a measure of quality control.

Quality control of ATAC-seq peaks: We carried out manual quality control steps on the dataset. First, we generated Venn diagrams for each tissue to compare the number of peaks using two different cut-offs for significance in the MAC2 peak calling. We compared cut-offs of -log(p-value) >= 2 and >=3 (MACS2 p-value cutoff suggestions) using R package VennDiagram version 1.6.20. The number of overlapping peaks were calculated via the mergePeak function in HOMER with option -venn on the tissue bed files. We forced the maximum distance between peak centers to be <=100bp for two peaks to be considered "overlapping". We observed the degree of overlap to be qualitatively similar for -log(p-value) cutoffs of either 2 or 3, as a result we employ a cutoff of 2 for peaks called by MACS2.

We further created several plots using the peak fold-enrichment profiles obtained from MACS2. Peak fold-enrichments are a measure of read counts at peaks relative to the local random Poisson distribution of reads [49]. ATAC-seq peaks are typically highly enriched in transcription related genomic regions, such as TSS, TTS, 5' UTR, or exons [52,100] . We similarly examined fold-enrichment as a function of annotated region type to confirm our data is consistent with previous work. We similarly examined fold-enrichment profiles as a function of distance from the nearest TSS to ensure our peaks

were consistent with prior work. We also looked at the distribution of fragments lengths for each library to make sure libraries were not dominated by naked DNA and exhibited peaks associated with nucleosome bound DNA (c.f. S1.7 and S1.8 Figs). Lastly, we generated Manhattan plots of peak locations to ensure that they are not spatially clustered at a gross genomic scale.

Normalization for differences between tissues and genotypes: For the j^{th} sample (*i.e.*, replicate/tissue/genotype combination) we have a "fragment file" generated from the corrected BAM file that is a 3-column BED file with the chromosome, corrected start and corrected stop base of each fragment defined by a set of paired reads. In order to carry out statistical tests at peaks using our replicated ATAC-seq data we normalized the 95 different fragment files associated with each tissue/genotype/replicate combination (as one sample failed a visual quality control check). Our normalization procedure is based on the observation that both the number of reads and the distribution of fragment lengths varies between samples (see Fig 1.3A and 1.3B). The former just reflects variation in the number of reads obtained per library, and we believe the latter is due to subtle differences in sample preparation that inadvertently selects for differing fractions of nucleosome free DNA. Our normalization consists of adding a 4th column to the fragment file that can be thought of as a "weight" used in all downstream analyses, where that weight normalizes the fragment files across the J samples. The weight is inspired by the "quantile normalization" method used in the field of gene expression [57] and is simply: $w_{ij} = N_{i\cdot}/N_{ij}$, where N_{ij} is the number of fragments of length i in the j^{th} sample, and the "$\cdot$" is the average over samples. As can be seen from

the unweighted and weighted histograms of Figure 1.3 these weights result in a distribution of fragment lengths that are identical between samples.

With weights in hand, we calculated the weighted Coverage for each sample at any given position in the genome, C, as the sum of the weights of all fragments covering that position. We finally averaged coverage over replicates (within tissues and genotypes) to generate UCSC Genome browser tracks [101–104] , although the biological replicates were retained for statistical testing.

Accounting for structural variants: In a typical ATAC-seq experiment raw reads are aligned to a reference genome, fragment files are derived from those alignments, and the resulting fragment files are perhaps normalized. However, in the case of the seven strains (A4,A5,A6,A7,B2,B3,B6) of this study we have complete *de novo* reference quality genome assemblies [43]. The genomes of these strains are distinguished from the dm6 reference by thousands of SVs, such as mobile element insertions, smaller insertions or deletions of DNA sequences (large enough to generally not be identified by standard pipelines), tandem duplications, and inversions. These “hidden” structural variants can impact inferences regarding chromatin structure obtained from ATAC-seq data assembled to a standard reference. To illustrate this issue we simulated 50bp PE reads from a 30kb or 32.5 kb genomic region using samtools::wgsim, with the two sequences being identical aside from a 2.5kb insertion of DNA sequence derived from a transposable element. Simulated short reads are then obtained from each region with an average fragment length of 400bp (standard deviation of 100 bp), similar to the fragment length distribution of ATAC-seq reads. Reads were mapped back to the

shorter region (akin to a “reference genome”) and coverage is depicted in Supplementary Figure 1.14. The figure clearly shows the potential for mis-mapped reads associated with a polymorphic SV to create a large localized dip in sequence coverage (that could be interpreted as a closed chromatin structure), with the footprint of this phenomena likely restricted to +/- 800bp (i.e., 2 standard deviation in read length) around the SV.

In the work of this paper, by virtue of seven of the genotypes examined having reference quality *de novo* assemblies, we can control for the effect of unmapped reads due to structural variants by removing fragments *across all genotypes* that span an SV *in any given genotype*. This correction is done using bedtools intersect to remove all reads from all fragment files that span insertion or deletion variants. For duplication variants, we first calculated duplicated regions by adding and subtracting the total length of duplication from the insertion site. Then, all reads spanning duplicated regions are removed. We then calculate new weights as described above, and recalculate coverage.

Statistical testing: We carry out two ANOVA statistical tests for seven strains with reference qualify assemblies (A4, A5, A6, A7, B2, B3, B6) at peaks to identify those that differ among genotypes, tissues, or their interaction for weighted log transformed Coverage (lnC = ln(C+5)) after peaks with a weighted average coverage < 50 were dropped as: *lnC ~ geno + tissue + geno:tissue***.** Adding 5 to the number of counts makes the rare case of counts near zero less extreme relative to other strains. A False Discovery Rate (FDR) associated with each p-value was calculated using the p.adjust

function in R [105,106]. Tests with FDR adjusted p-values < 0.005 (or $-\log_{10}$(FDR p-value) > 2.3) are considered significant. QQ plots and Manhattan plots were generated for the ANOVA results.

We carried out statistical tests on both SV corrected and uncorrected fragment files. Loci significant in the SV-uncorrected data but not significant in the SV corrected data potentially represent false positives. We define hits unique to the SV-uncorrected dataset as false positives and estimate the rate of such false positives in experiments that do not correct for hidden SVs. Results are also represented as Venn diagrams. We further examined each potential false positive to determine if the ATAC-seq peak was actually contained within a SV (e.g., a deletion relative to the reference), was within +/-800bp of an SV boundary, or was >800bp from an SV (for peaks >800bp from a SV both F&R reads are expected to map to the reference genome and thus such peaks are not expected to be impacted by the correction). We finally compared the p-values between SV-corrected and uncorrected data for peaks >800bp from an SV to determine if any remaining differences were due to simple sampling error in p-values near significant cut-offs.

Causative SNP and SV identification by random effect model: For peaks significant for genotype or genotype:tissue we attempted to identify SNPs within 250bp of the peak that could potentially explain the significant result. We accomplished this via the following random effects model in R::lme4 (version 1.1-23): *lnC ~ (1|marker) + (1|marker:tis) + (1|tis) + (1|geno:marker) + (1|tis:geno:marker)*

We estimate the proportion of variance explained by a marker as $var_m/[var_m + var_{g:m}]$ or marker:tissue as $var_{m:t}/[var_{m:t} + var_{g:m:t}]$ respectively. Here marker refers to a state of a SNP, thus several genotypes could share the same marker state. Furthermore, since the strains of this study are isogenic, markers are either REF or ALT, and never heterozygous. In both cases a ratio close to 100% identifies a SNP that explains all the variation associated with a significant peak. We similarly estimate the proportion of variation explained for each tissue by dropping terms involve tissue. These SNPs are strong candidates for being causative, with the strong caveat that only 7 genotypes are examined in this study, so we are almost certainly over-fitting and confirmatory experiments are necessary. We examine the distributions of these marker tests and maintain a list of polymorphisms explaining 100% of the variation associated with peaks. We finally annotate SNPs explaining 100% variance using SnpEff [107]and HOMER. In addition, a list of SNPs/SVs which individually explain less than 80% variance of polymorphic peaks is also provided. These SNPs/SVs potentially explain only a fraction of the variation in peak height, with the remaining due to other cis-acting or trans-acting variants. Future confirmatory experiments are even more necessary to confirm the causal effect of these SNPs/SVs.

1.6 Data and script availability: The raw fastq files are submitted to NCBI as BioProject: PRJNA761571. A github containing the codes used in this work is here: https://github.com/Kh0iHuynh/ATAC-seq-Project. Several intermediate data tables resulting from different analyses are hosted here: https://wfitch.bio.uci.edu/~tdlong/sandvox/publications.html or

https://doi.org/10.7280/D1FM5F. Many of the results, such as ATAC-seq coverage tracks, SNPs/SVs tracks, can also be visualized as Santa Cruz Genome Browser (SCGB) tracks here: http://goo.gl/LLpoNH.

1.7 FIGURES

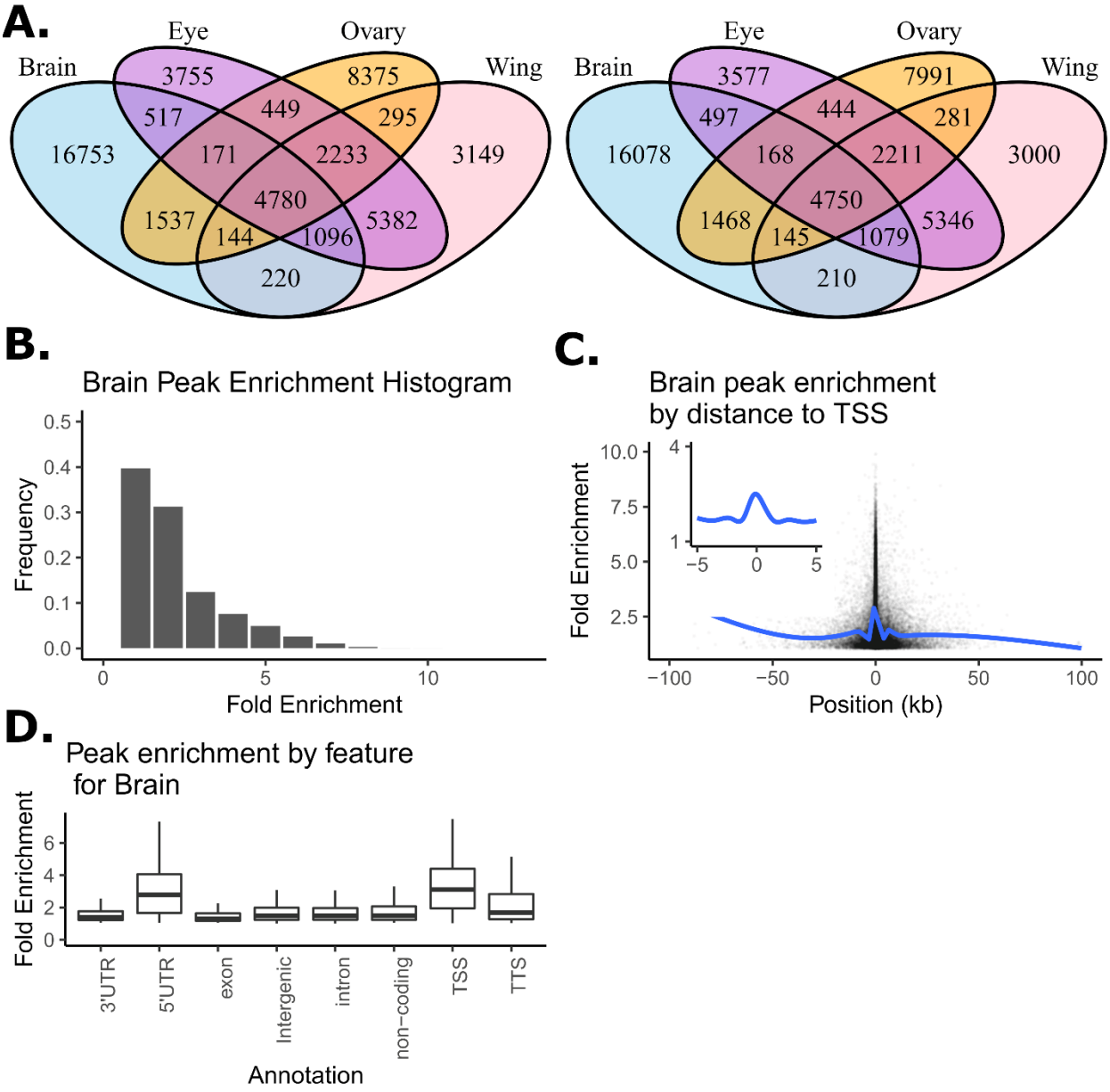


Fig 1.1. Summary of open chromatin peaks identified across four tissues. *(A) Venn diagram showing overlap in peak calls across tissues as a function of the p-value cut-offs of 0.01 (left) and 0.001 (right). (B) Distribution of peak enrichment scores for the brain samples. (C) Peak enrichment scores as a function of distance to the nearest transcription start site with a smoothing line for brain samples. Insert focues on peaks within 10kb of the TSS showing only the smoothing line. (D) Peak enrichment distribution as a function of genomic feature for brain samples (TSS, transcription start site; TTS, transcription termination site).*

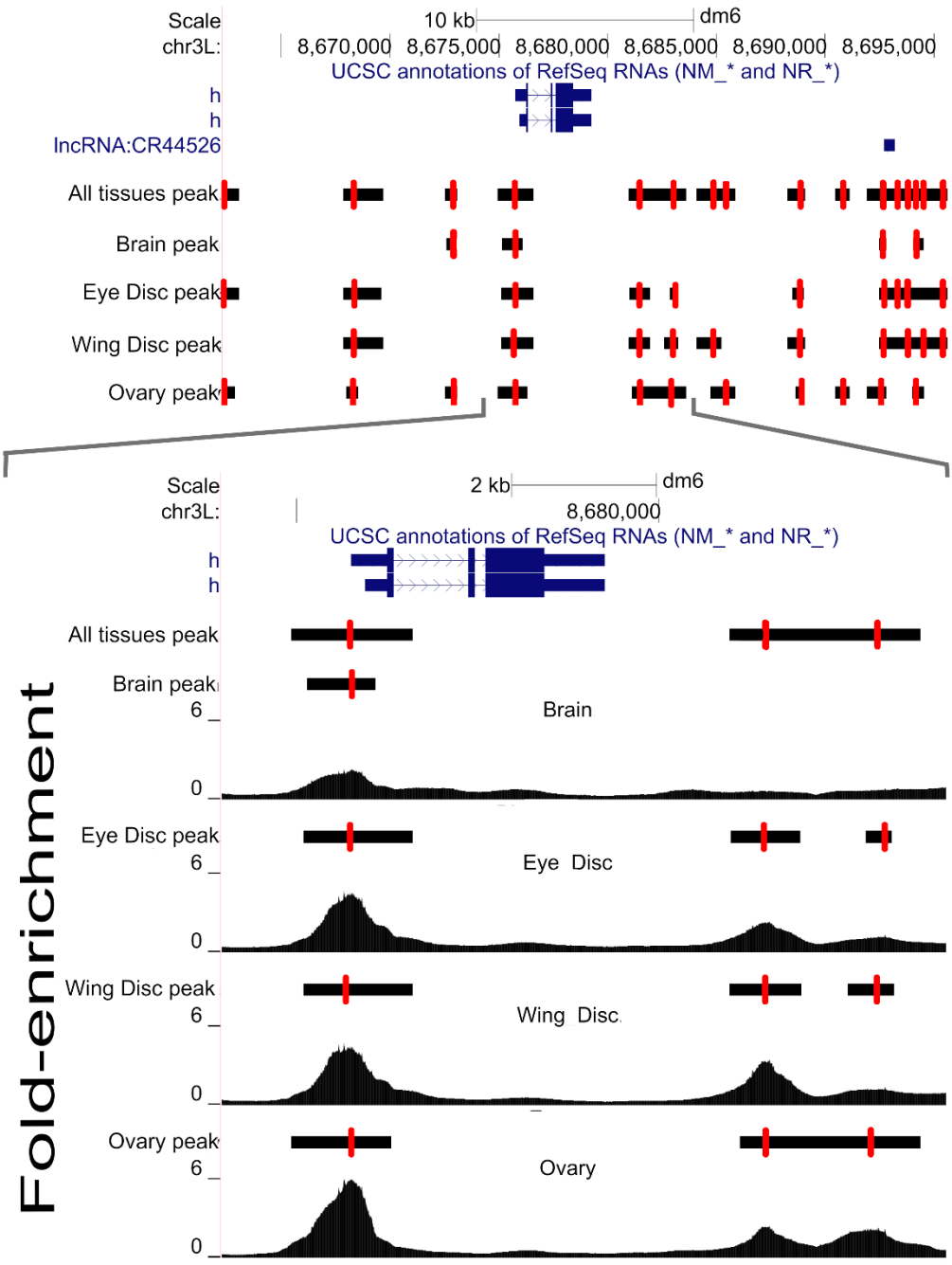


Fig 1.2. An illustrative example of peak calling results near the gene hairy. *(Top panel) Peaks called separately by tissue as well as a consensus set of peaks calls (labeled "all tissues"). Single base peaks are indicated with red hash lines with black bars representing uncertainty. (Bottom panel) a zoomed region showing peaks and raw read coverage.*

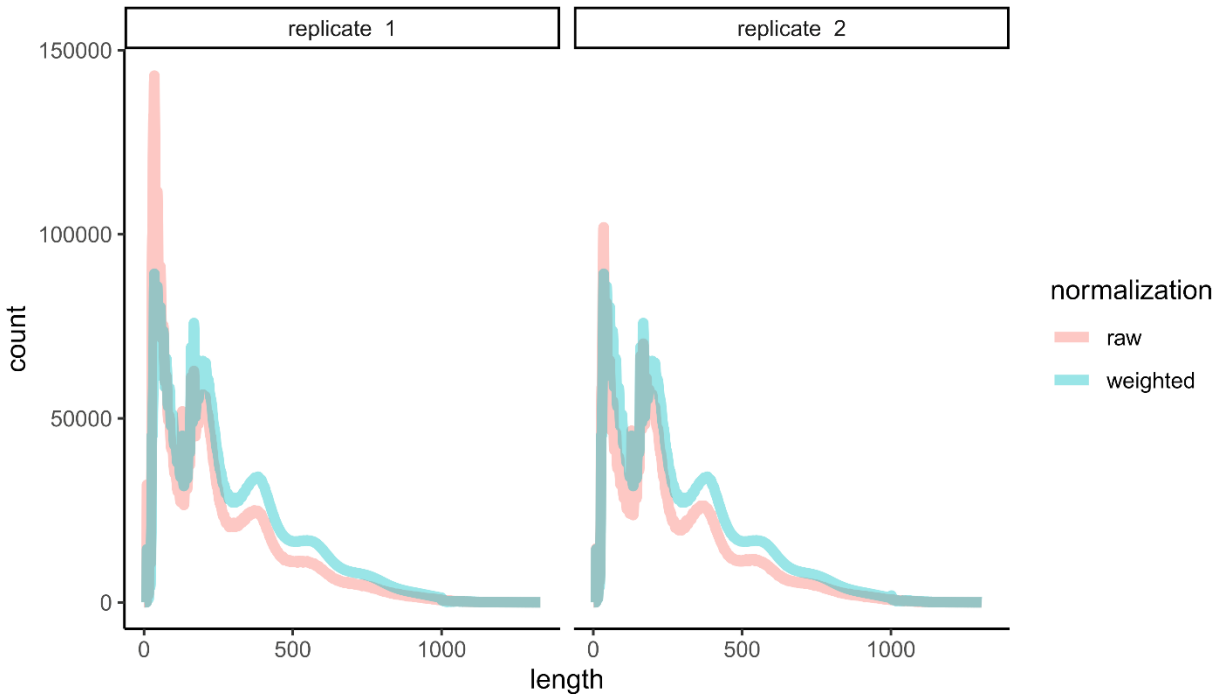


Fig 1.3. Distribution fragment lengths before and after normalization. *Representative examples of the raw fragment size distribution for genotype A4 and brain tissue for two replicates in red. The same two samples are depicted in blue following normalization.*

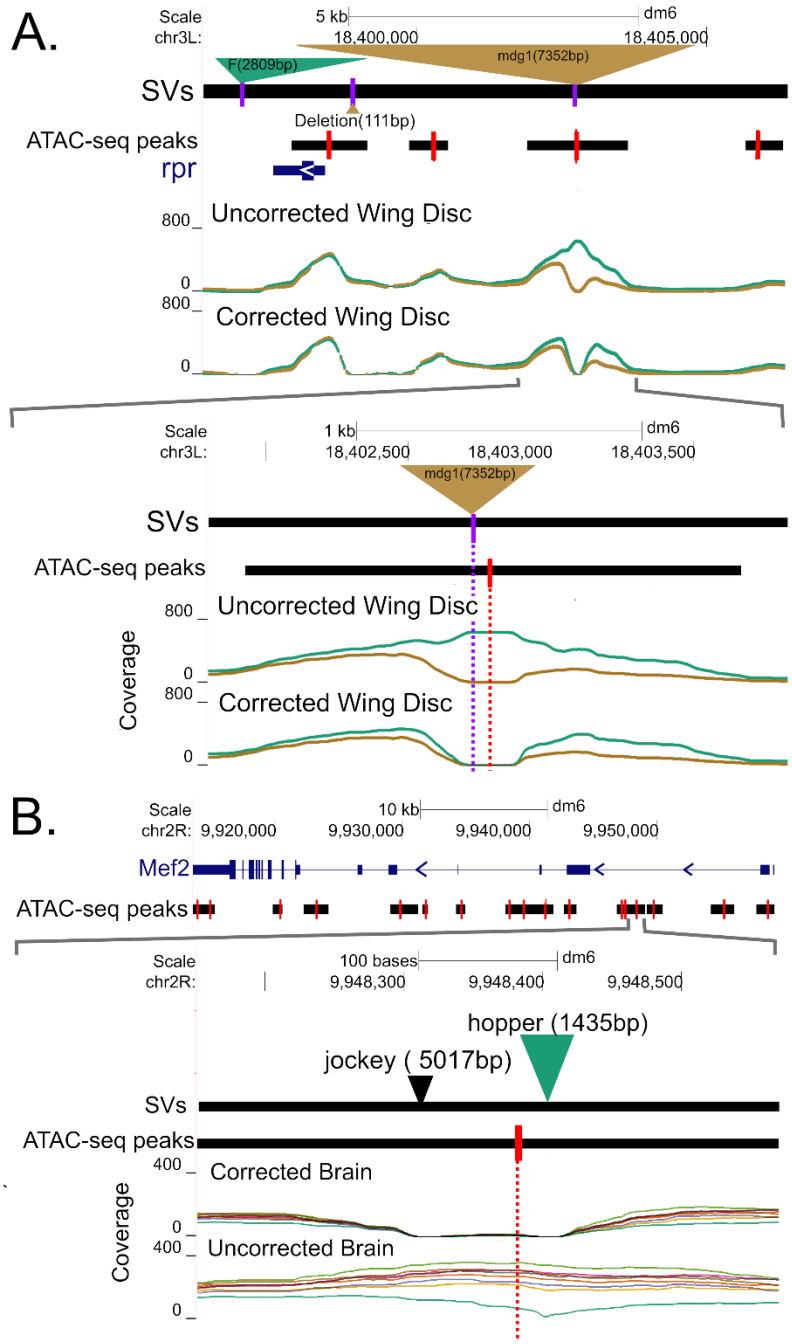


Fig 1.4. Examples illustrating the effects of SV correction on coverage.
(A) After correcting for a large insertion of a mdg1 transposable elements upstream of rpr in strain B6 (brown) the apparent difference in coverage between strains B6 and A4 (green) is largely eliminated. (B) Correcting for the effect of a hopper TE in an intron of Mef2 in the A4 genotype largely eliminates an apparent difference in chromatin configuration. Tracks are structural variant, ATAC-seq peaks, gene, and coverage tracks.

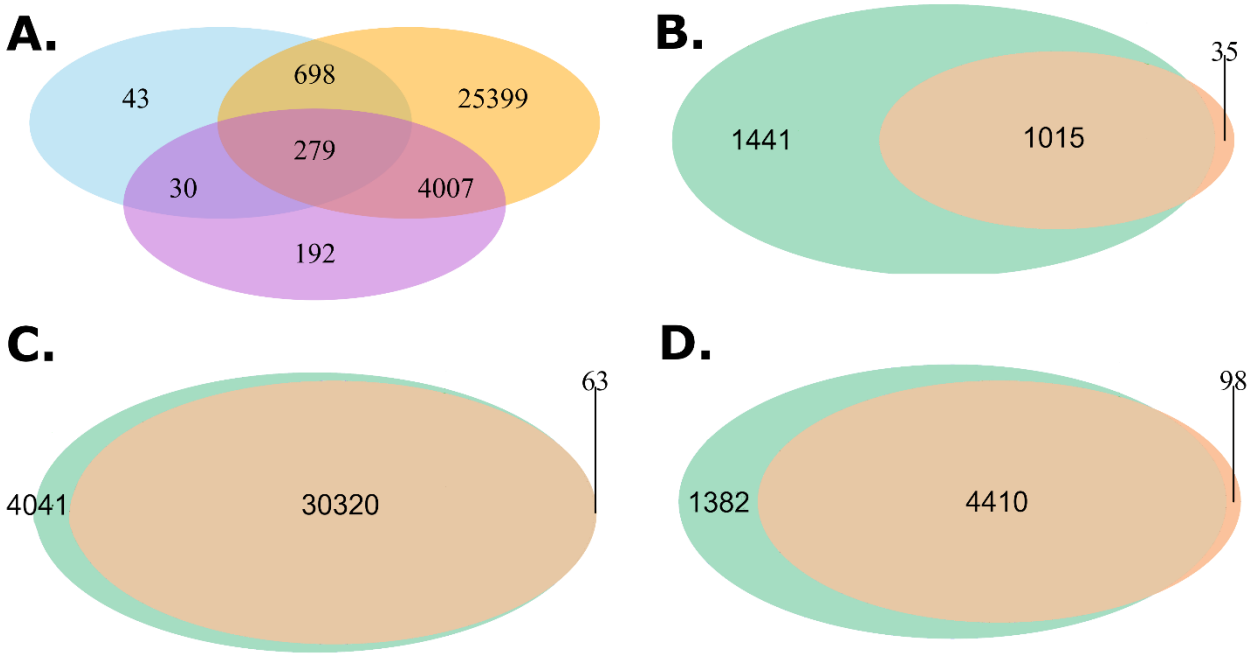


Fig 1.5. Venn diagrams showing overlapping peaks by ANOVA categories and SV correction status.
(A) Venn diagram showing overlap of regions significant (FDR < 0.5%) for Genotype (blue), Tissue (orange), or G:T (orchid) for the SV-corrected data. (B-D) Venn diagrams showing the number of peaks significant G,T, or a G:T interaction, respectively. Green are tests carried out without correction for known SVs, and brown after SV-correction.

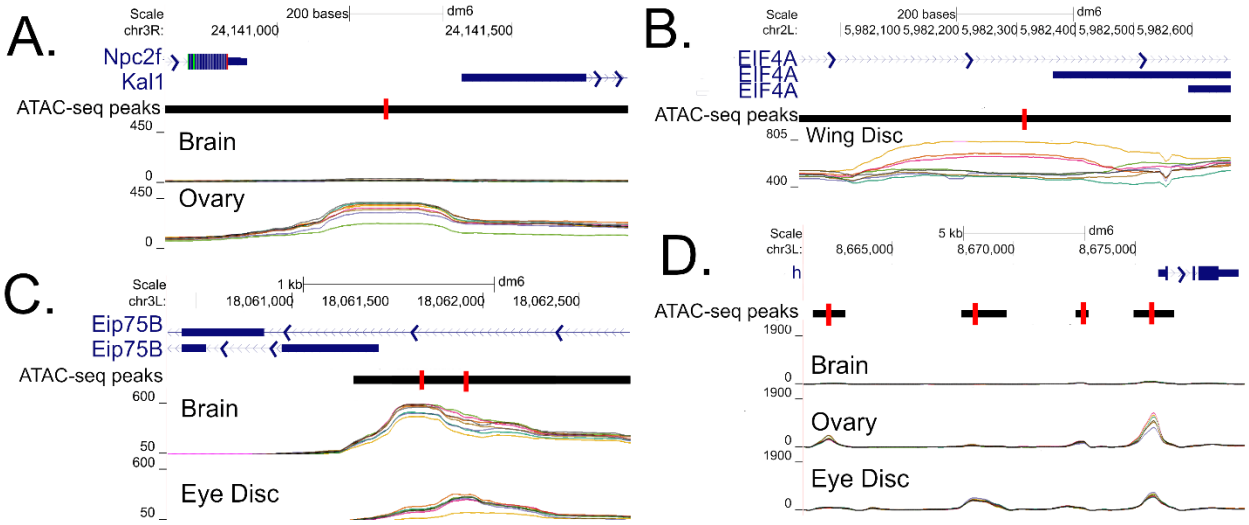


Fig 1.6. Illustrative examples of polymorphic chromatin configuration.
The images depict regions upstream of the TSS of Kal1 (A), upstream of the TSS of a Eip75B isoform (B), upstream of the TSS of a Eip75B isoform (C), and a large region known to harbor cis-regulatory element upstream of hairy (D). SV-corrected coverage is given for a subset of interesting tissue. Tracks are gene, ATAC-seq peaks, and coverage tracks.

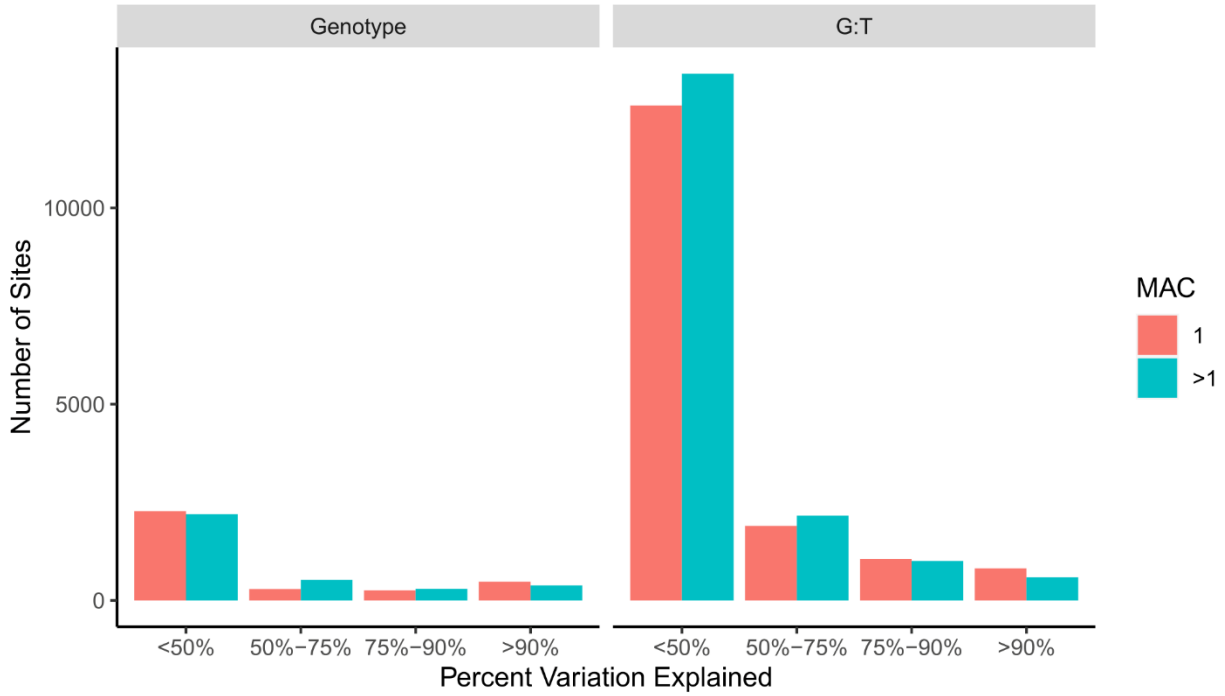


Fig 1.7. ATAC-seq peak coverage variation explained by nearby polymorphisms. *Peaks significant for genotype or genotype by tissue are on the left and right respectively. The number of sites are grouped by Minor Allele Count (MAC).*

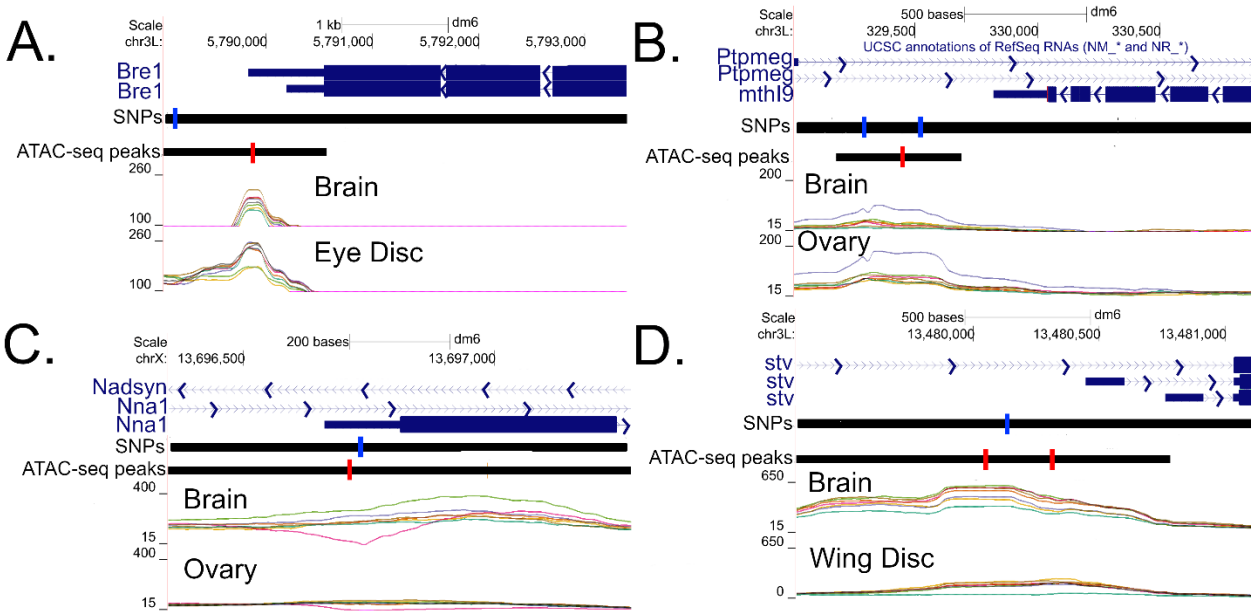
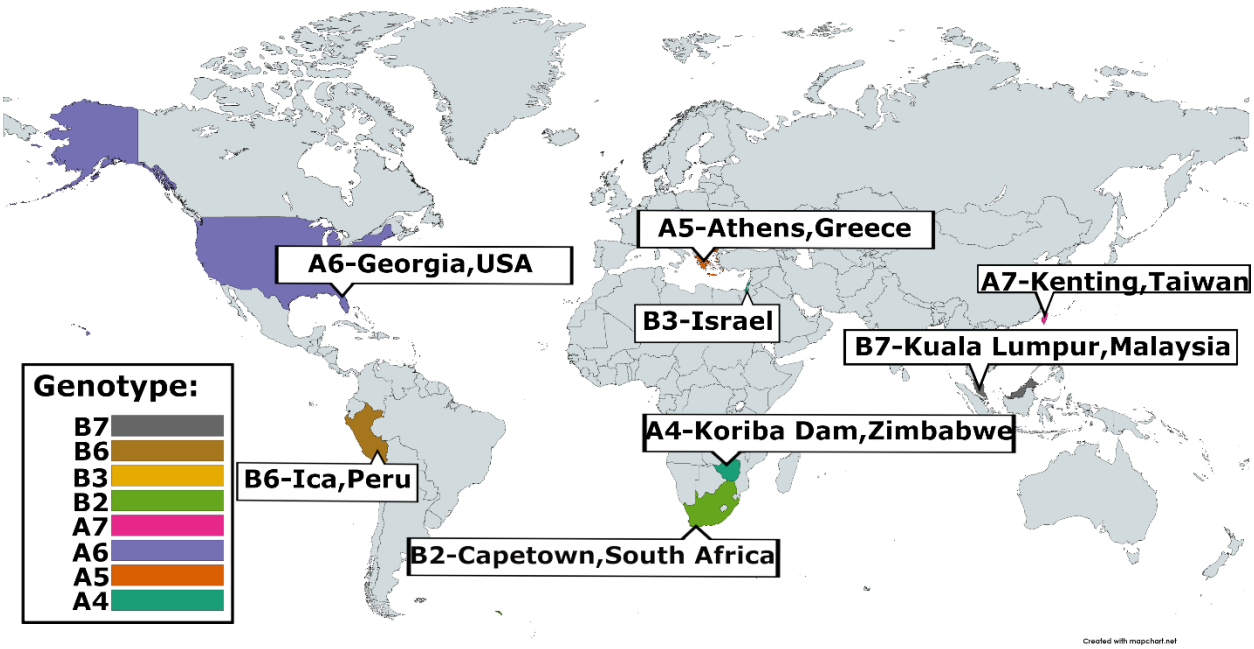
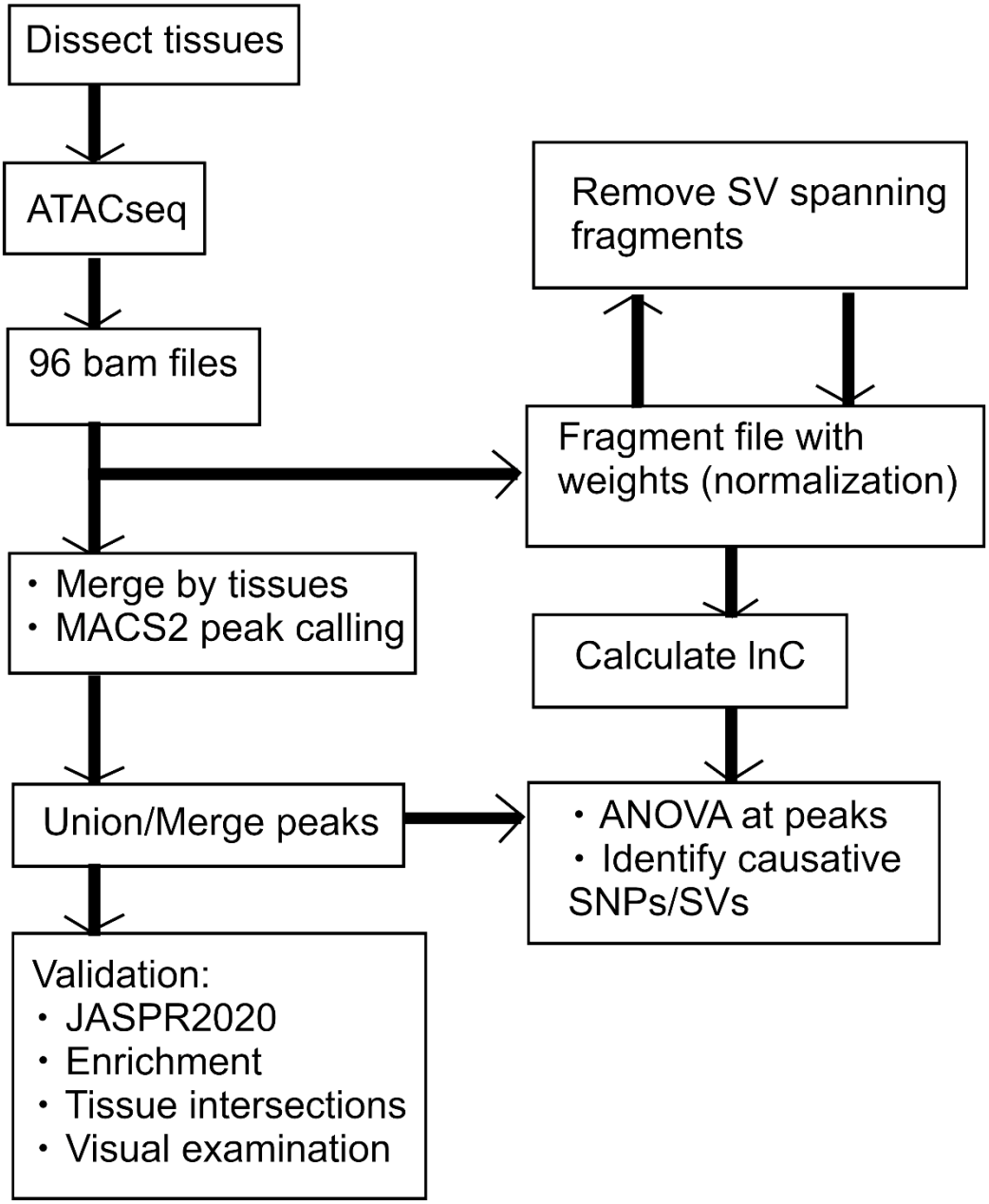


Fig 1.8. Illustrative examples of putatively causal SNPs:
Regions are depicted downstream of the TTS of Bre1 (A) , downstream of the TTS of mthl9 (B), the 5'UTR of a Nna1 isoform (C), and upstream of the TSSs of two stv isoforms (D). Only shows SNPs explaining > 80% of variation in Genotype of a G:T interaction (blue) are depicted. Tracks are gene, SNP location, ATAC-seq peaks, and coverage tracks.

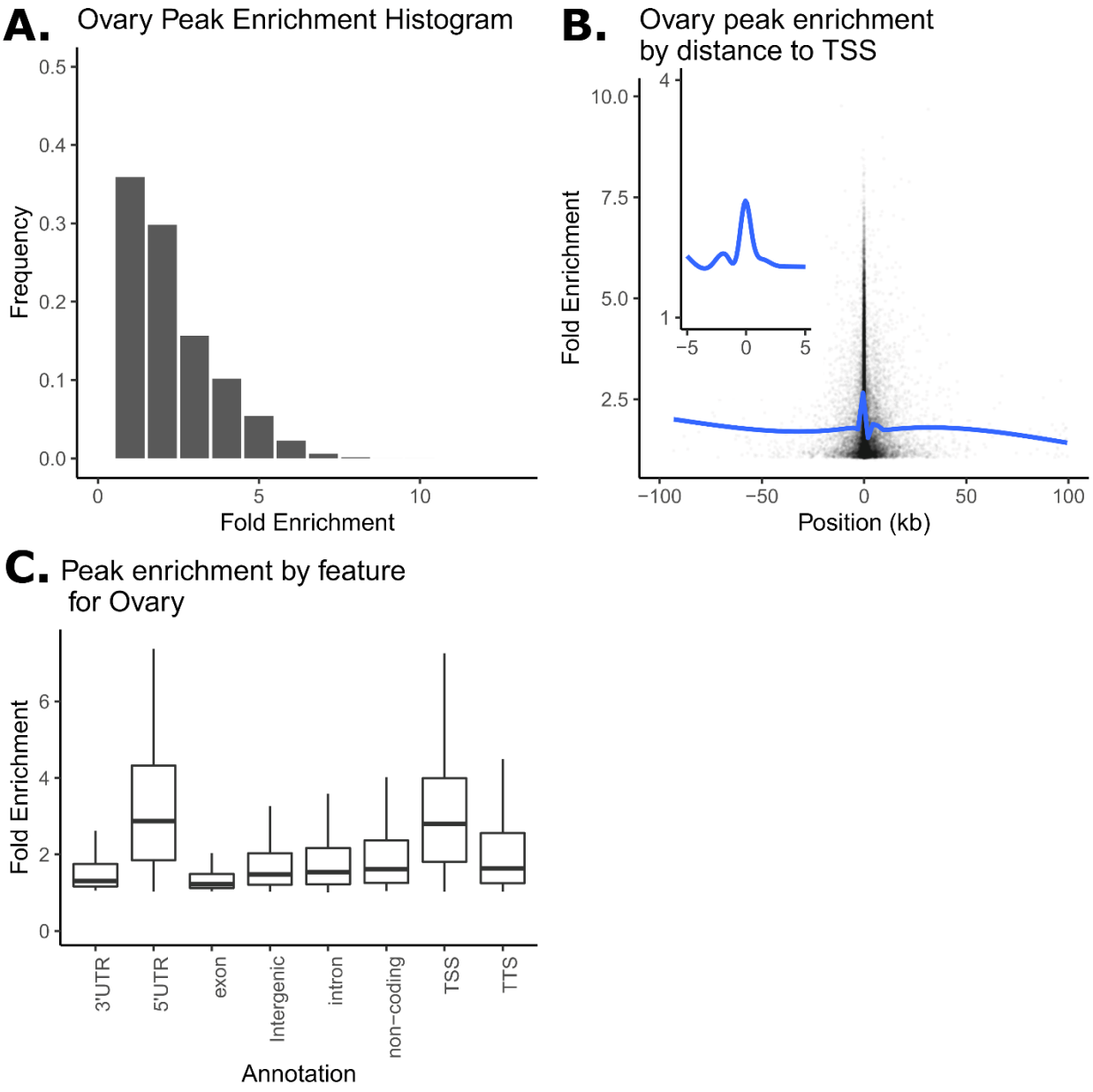


S1.1 Fig. World map showing the collection locations and color legend for all genotypes.

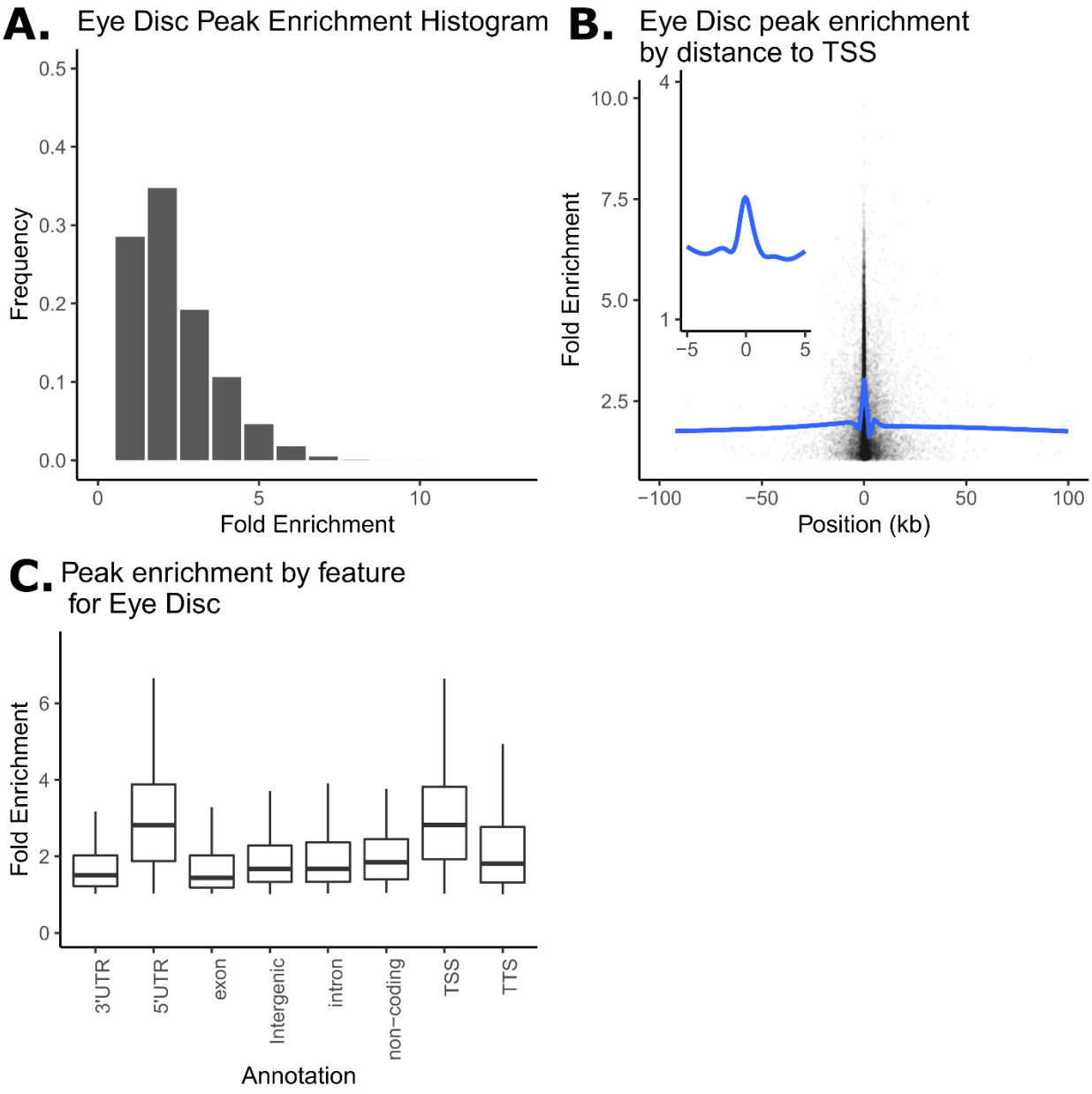
The color legend for genotypes is also kept constant throughout the paper. This map was created using mapchart.net, licensed under This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).



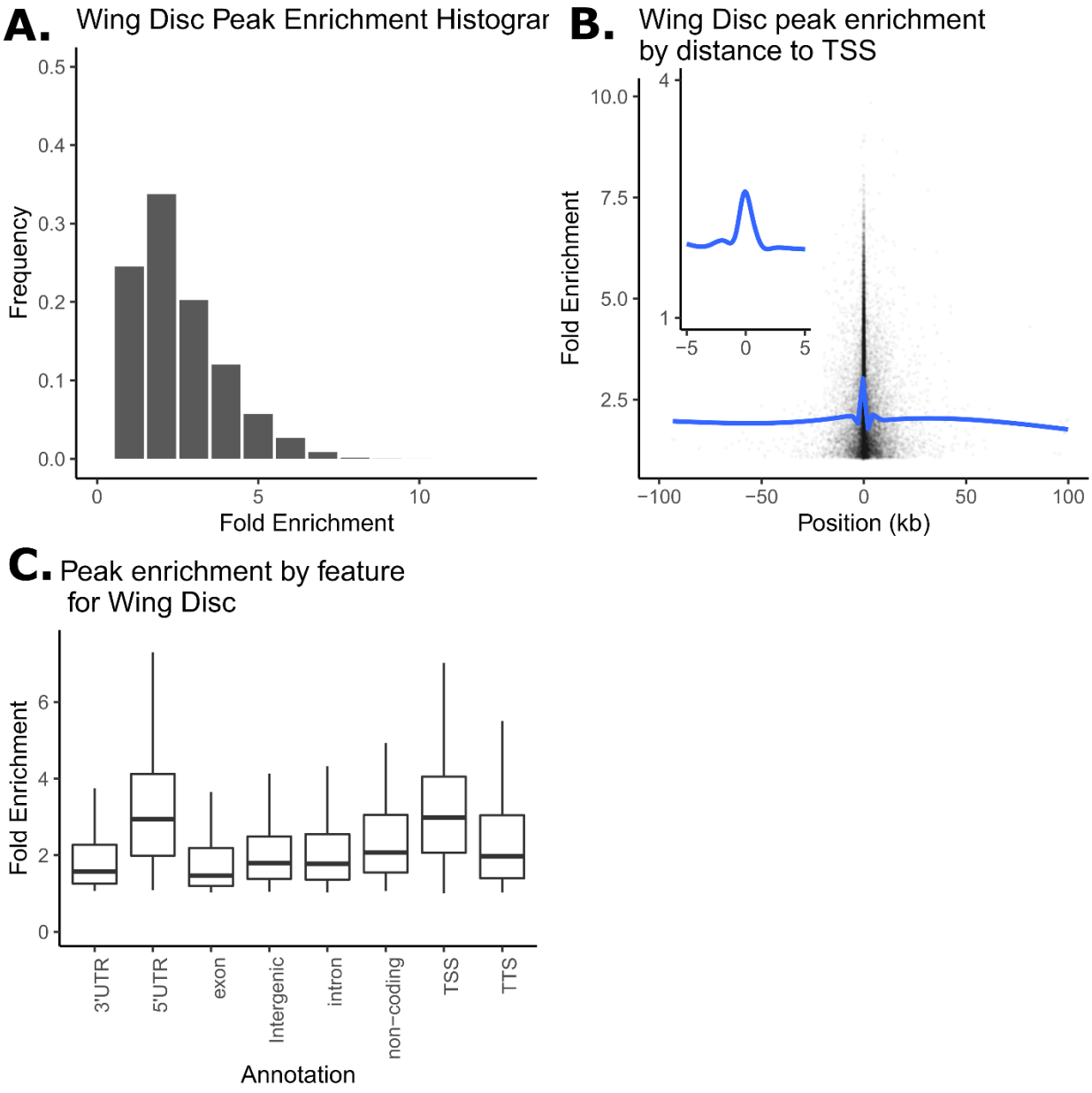
S1.2 Fig. Workflow for ATAC-seq study.



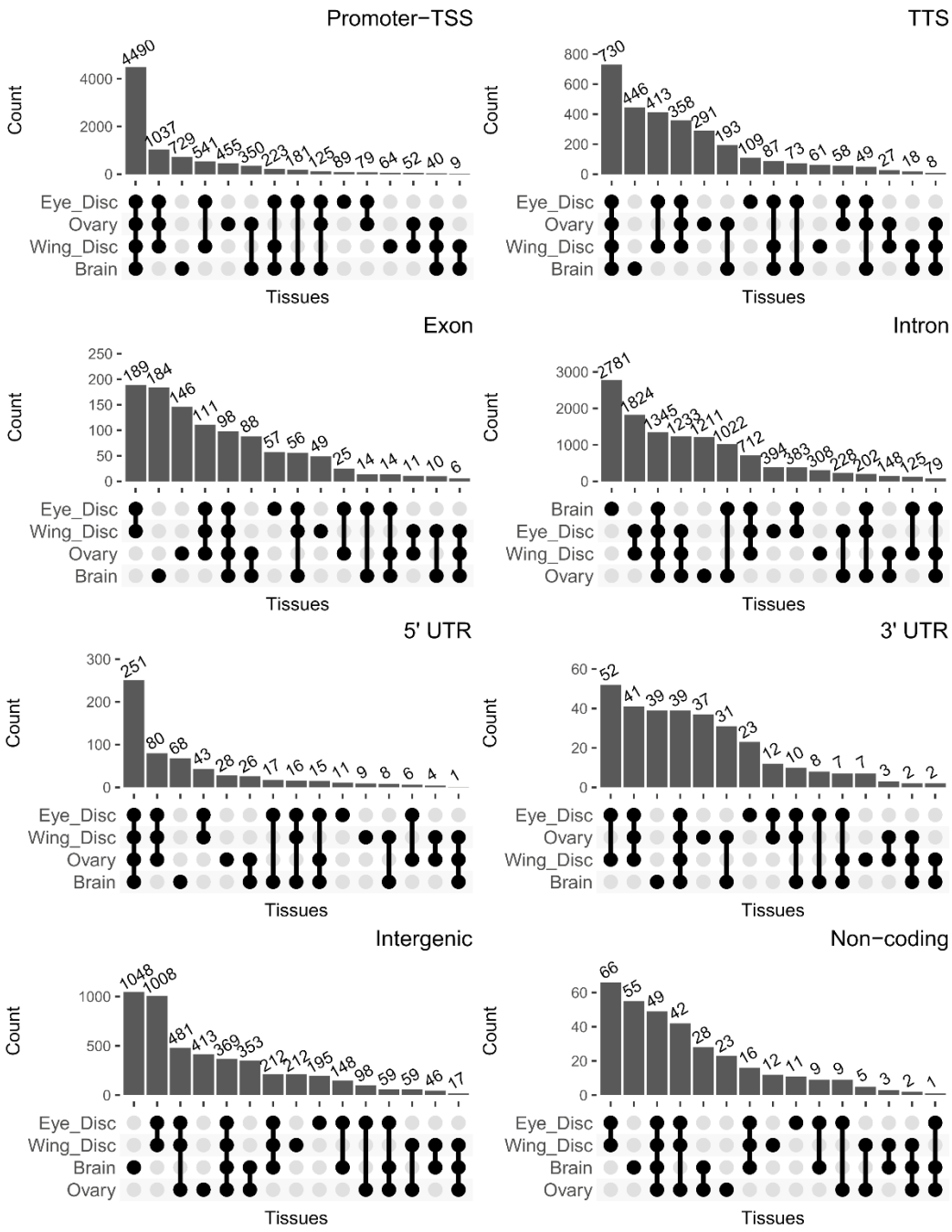
S1.3 Fig. Summary statistics for peaks called for ovary samples.
(A) Distribution of peak enrichment scores for the ovary samples. (B): Peak enrichment scores as a function of distance to the nearest transcription start site with a smoothing line for the ovary samples. Insert focuses on peaks within 10kb of the TSS and showing only the smoothing line. (C): Peak enrichment distribution as a function of genomic feature for the ovary samples.



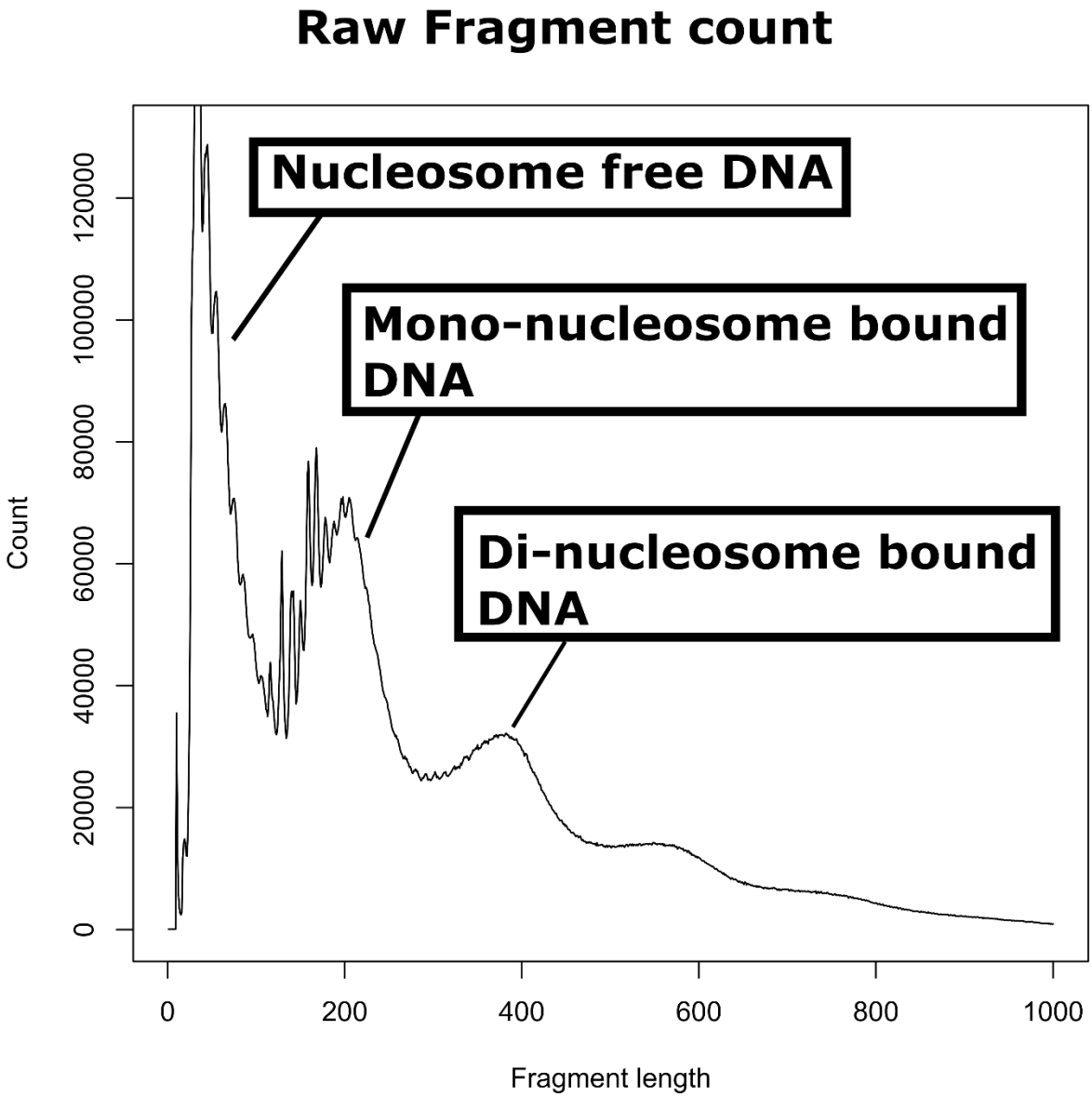
S1.4 Fig. Summary statistics for peaks called for eye disc samples.
(A) Distribution of peak enrichment scores for the eye disc samples. (B): Peak enrichment scores as a function of distance to the nearest transcription start site with a smoothing line for the eye disc samples. Insert focuses on peaks within 10kb of the TSS and showing only the smoothing line. (C): Peak enrichment distribution as a function of genomic feature for the ovary samples for the eye disc samples.



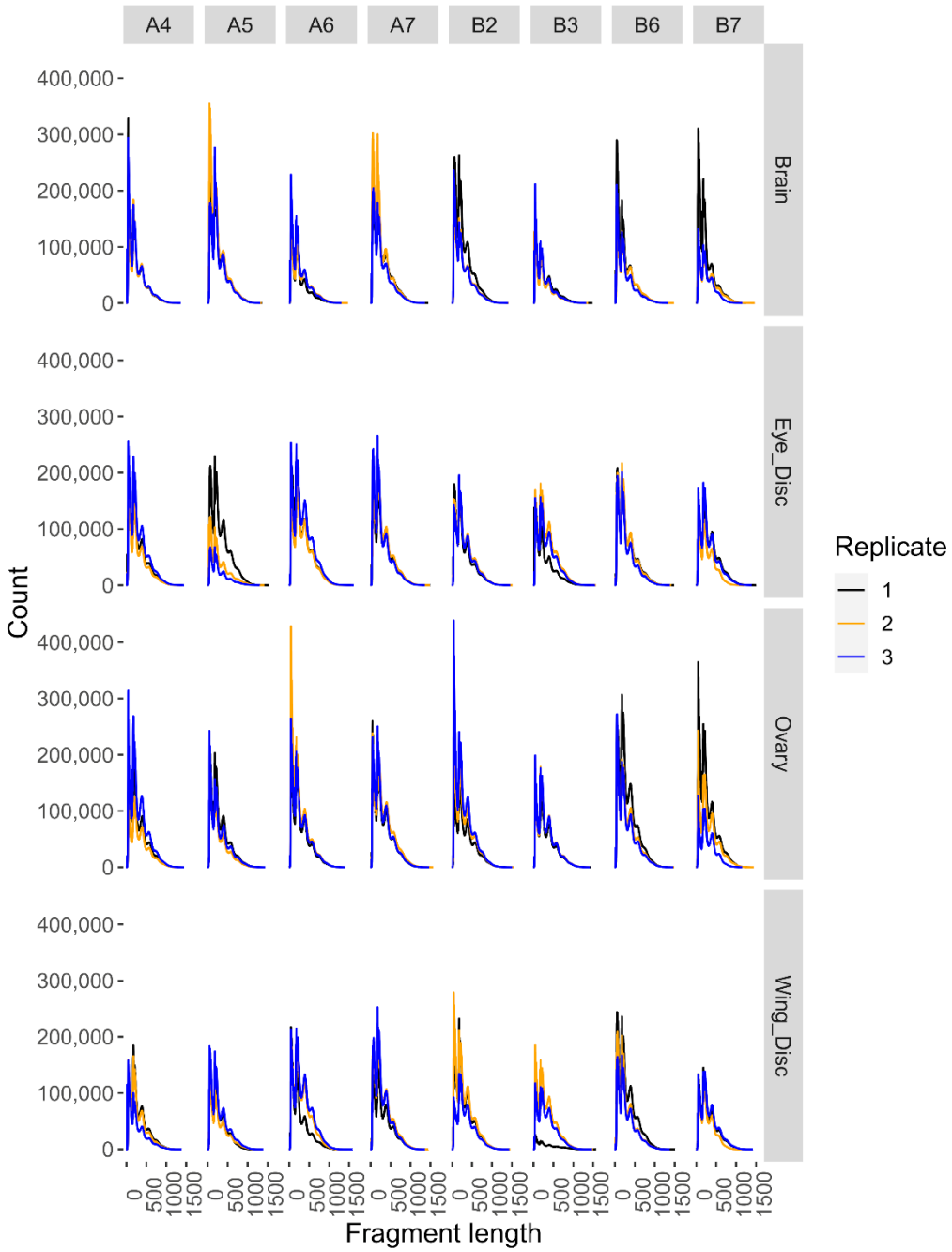
S1.5 Fig. Summary statistics for peaks called for the wing disc samples.
(A) Distribution of peak enrichment scores for the wing disc samples. (B): Peak enrichment scores as a function of distance to the nearest transcription start site with a smoothing line for the wing disc samples. Insert focuses on peaks within 10kb of the TSS and showing only the smoothing line. (C): Peak enrichment distribution as a function of genomic feature for the wing disc samples.



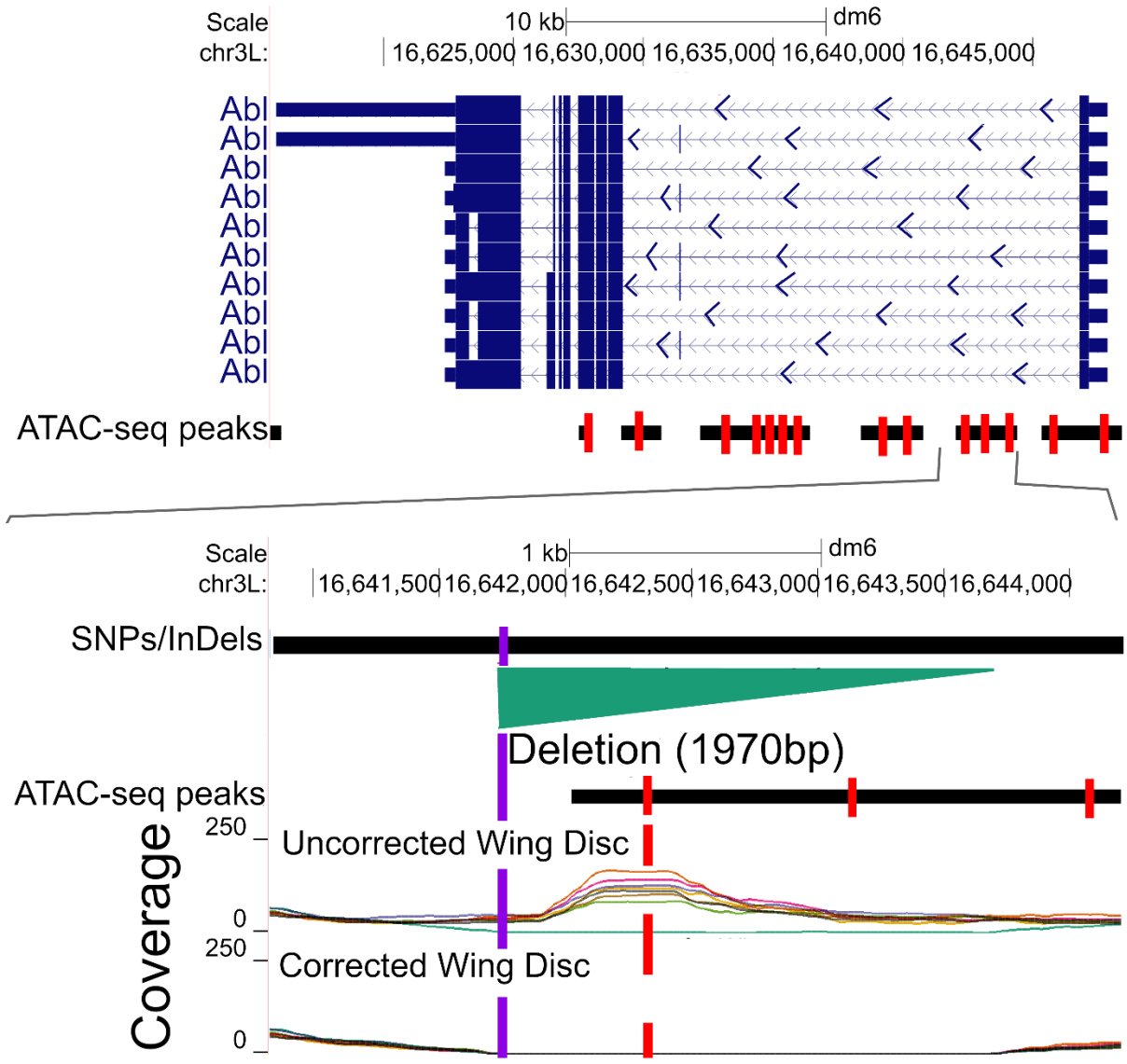
S1.6 Fig. Peak sharing among tissues as a function of feature type.



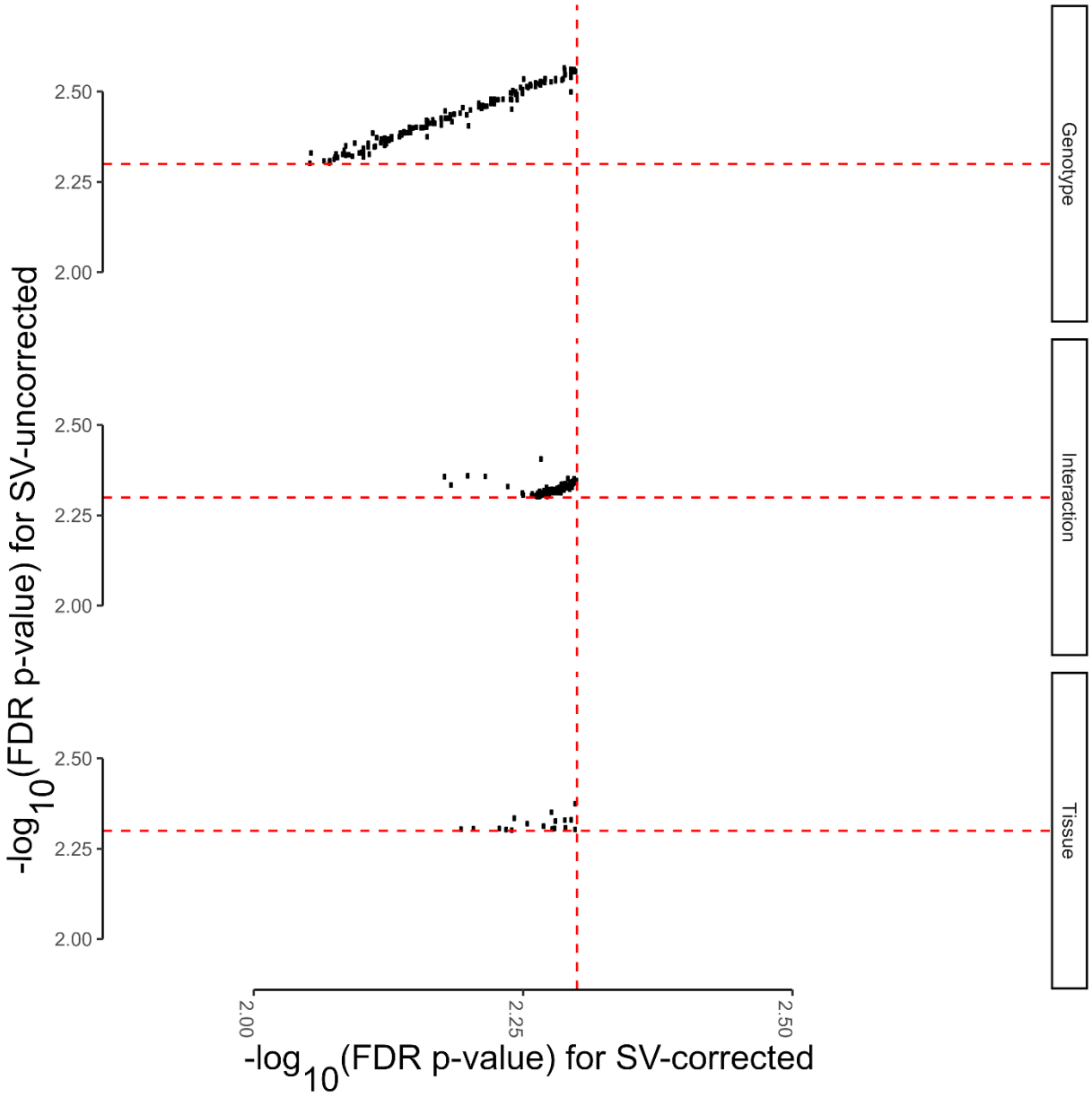
S1.7 Fig. Fragment length distribution and the nucleosome binding configuration depicted by the fragment length.



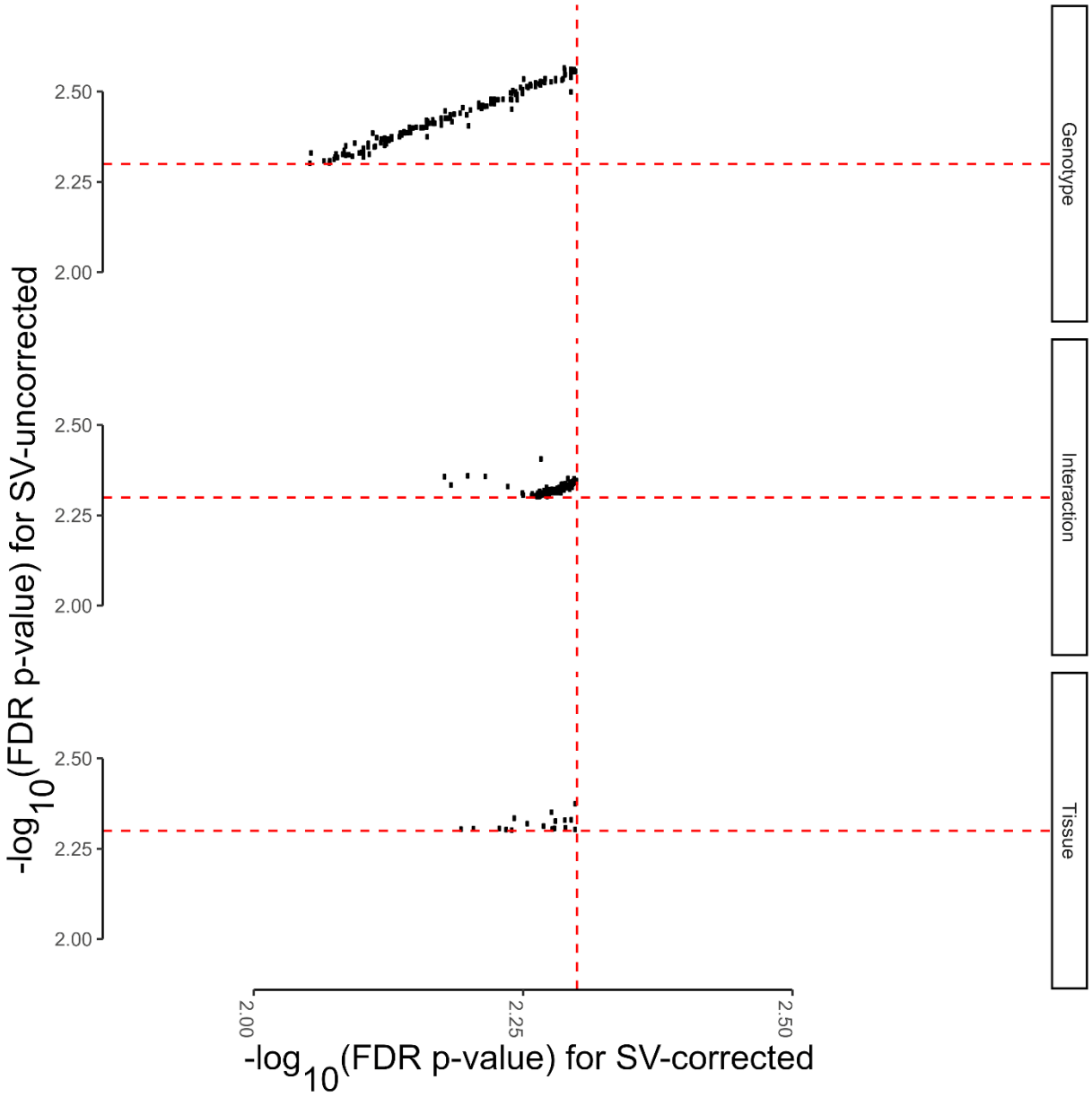
S1.8 Fig. Fragment length distribution for all replicates, tissues, and genotypes.



S1.9 Fig. A polymorphic deletion relative to the reference leads to the incorrect inference of close chromatin in strain A4.
Tracks are gene, ATAC-seq peaks (A), and SNP/InDel location, ATAC-seq peaks, and coverage tracks (B).

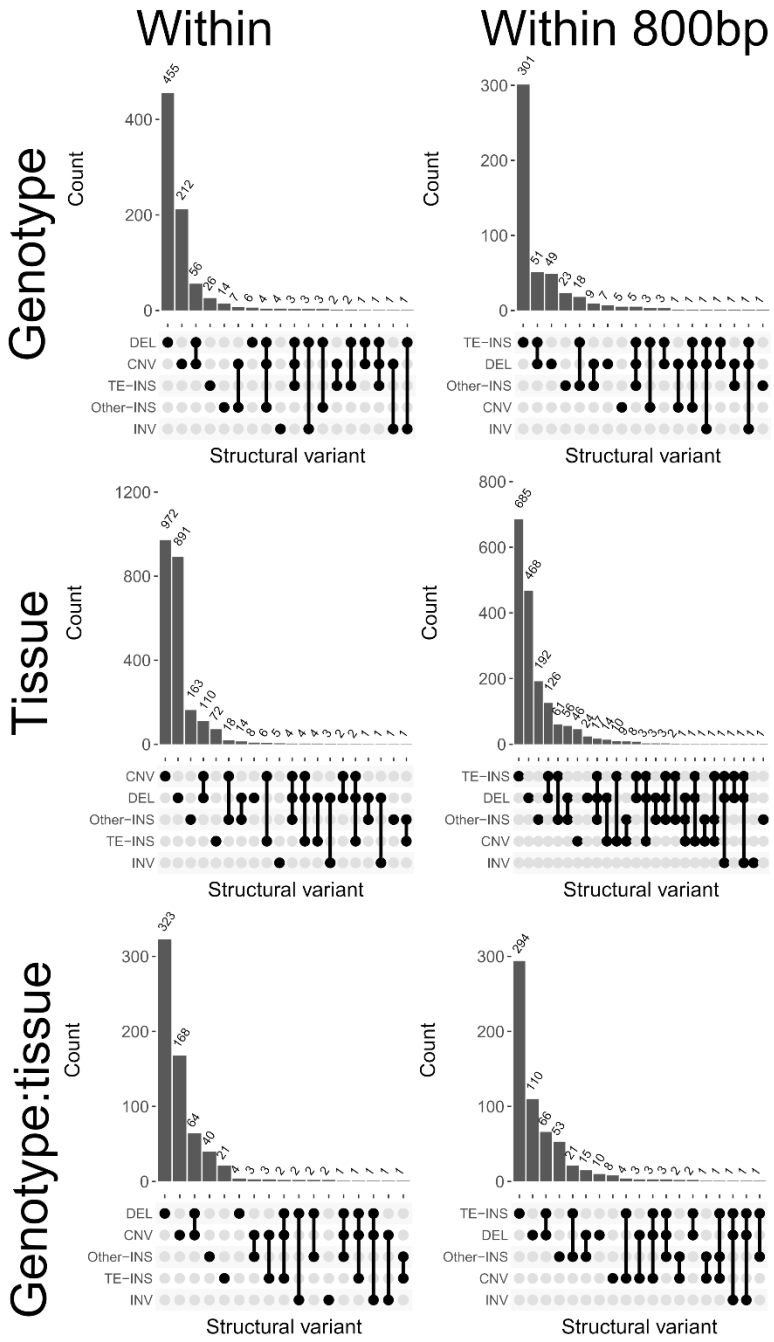


S1.10 Fig. Manhattan plots showing that significant (FDR < 0.005) peaks do not show strong evidence for spatial clustering throughout the genome.
(A): Manhattan plot for significant peaks by genotype. (B): Manhattan plot for significant peaks by tissue. (C): Manhattan plot for significant peaks by genotype and tissue interaction.

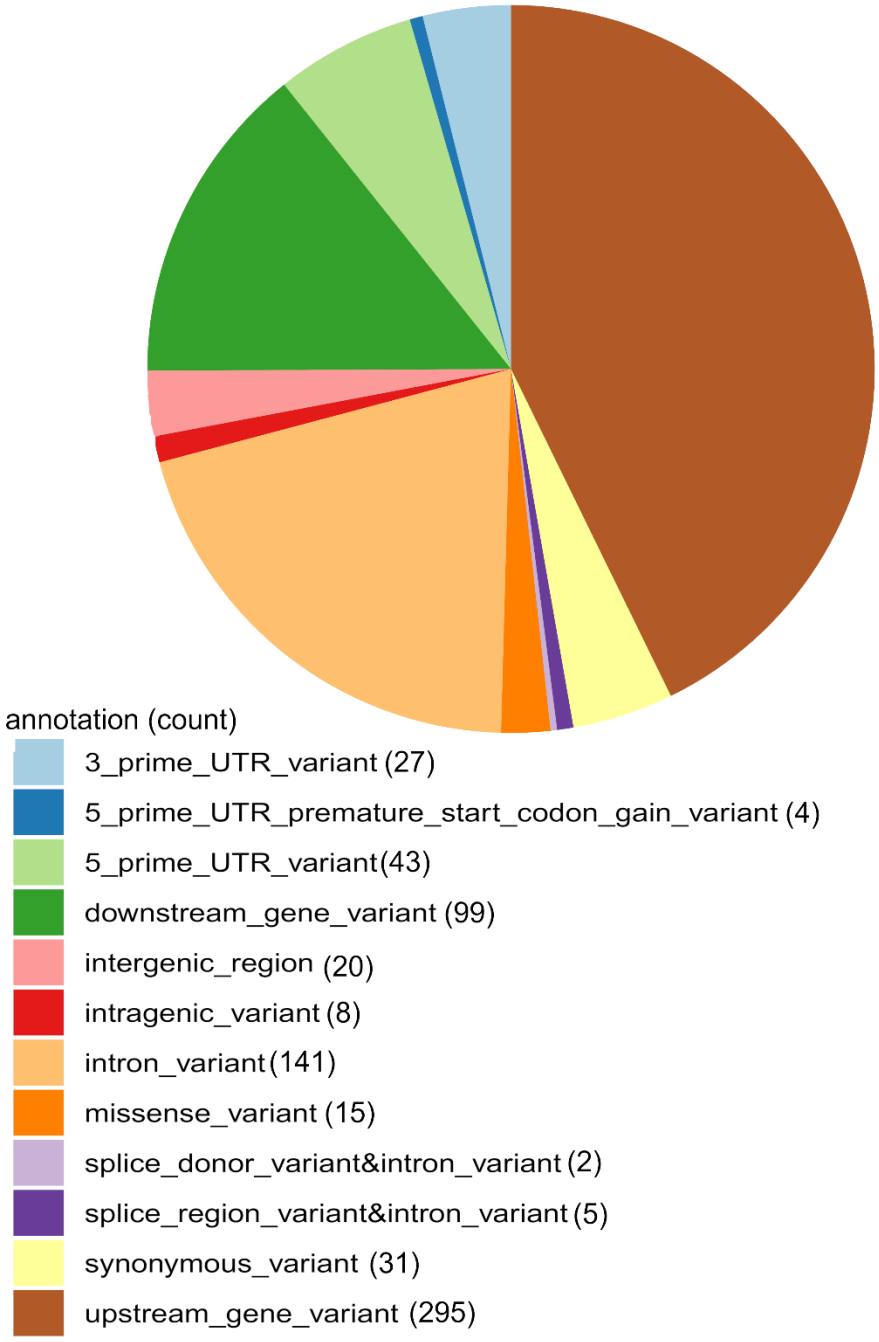


S1.11 Fig. -log(FDR adjusted p-value) scatterplot comparison between SV-corrected data and SV-uncorrected data for false positive peaks that falls outside of SV affected regions.

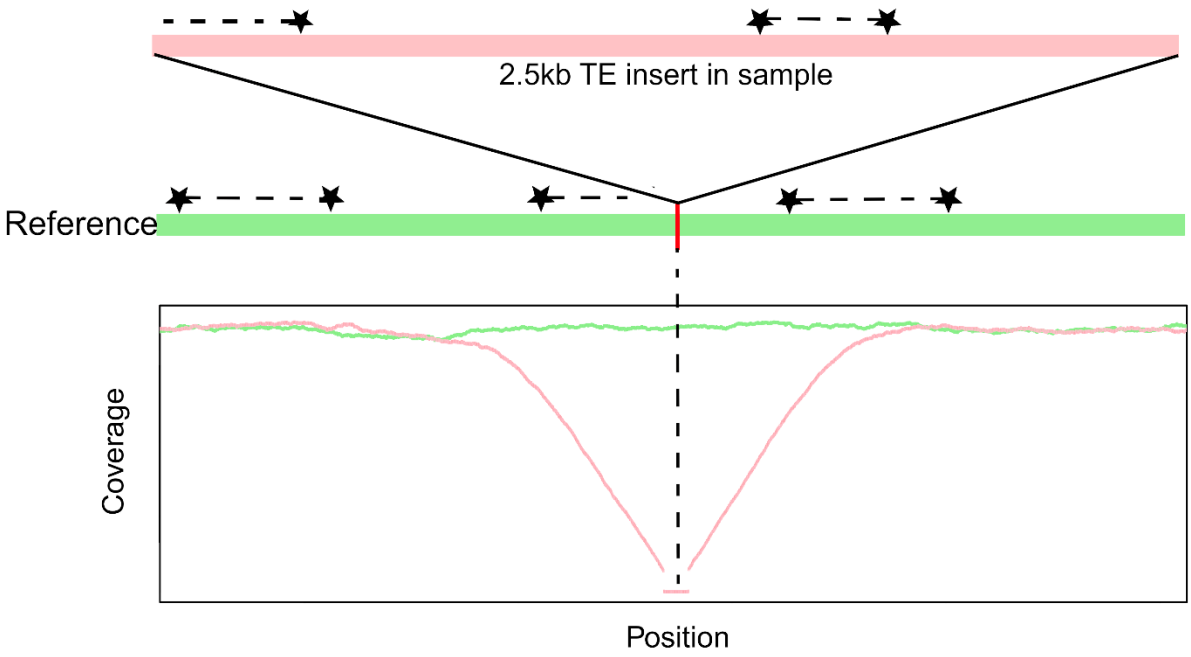
Red dashed lines showing -log(p-value) = 2.3 for SV-uncorrected (horizontal) and SV-corrected (vertical). Sampling variation likely drives the observed differences, as hits tend to be just beyond the significance level in the uncorrected dataset.



S1.12 Fig. Number of false positive significant peaks within structural variants (left) and within 800bp of the structural variants (right) by statistical test carried out (Genotype, Tissue, or Genotype by Tissue).
Variant types are deletion relative to reference (DEL), insertion due to TE (TE-INS), non-TE insertions (Other-INS), inversion (INV), or copy number variant (CNV). The categories are non-exclusive as multiple SV events could be close to an ATAC-seq peak, especially since we integrate over all strains.



S1.13 Fig. SnpEff annotation for causative SNPs that explain 100% variation.



S1.14 Fig. Example showing the effect of structural variant on inferred fragment coverage.
The green track depicts the reference 2kb sequence with the pink track depicting a non-reference 2.5kb transposon insertion (not drawn to scale) at the location of the red dash. The bottom plot depicts the coverage of reference (green) or non-reference sample aligned to the reference genome. The stars depict short read (forward and reverse reads) pairs and dashed lines the fragments created by read pairs. Chimeric fragments with one read in the TE insertion are mis-mapped, resulting in strong dips in read coverage at locations close to the TE insertion site.

1.8 TABLES

Table 1.1: Number of peaks showing significant variation at an FDR of 0.5%

Statistical Test	SV-corrected	SV-uncorrected
Genotype	1050	2456
Tissue	30383	34361
Genotype:Tissue	4508	5792
Total	31769	36059

Table 1.2: Number of peaks that are only significant in SV uncorrected data as a function of statistical test and distance from nearest SV.

Statistical Test	**Number of Peaks**			
	within	±800bp	> ±800 bp	Total
Genotype	801	481	159	1441
Tissue	2282	1736	23	4041
Genotype:Tissue	639	599	144	1382

Table 1.3: Number of SNPs within 250 bp and explaining >=80% of the variation in coverage for peaks significantly varying by Genotype or G:T.

	Peaks that vary by:	
Tissue	**Genotype**	**Genotype:Tissue**
Genotype	1253	NA
G:T	NA	2735
Brain	1620	6485
Ovary	1299	5441
Eye Disc	1401	6780
Wing Disc	1465	7385
Total Tests	6707	33570

S1.1 Table: Details of the eight strains examined in this study.
All eight strains are P-element and Wolbachia free, were brother sister mated for up to 18 generations, and are highly isogenic [48]*. Each strain, bar B7, is associated with a reference quality de novo genome assembly* [43]*. The Stock Number is the Bloomington ('b') or Tucson/San Diego ('t') Drosophila Stock Center code, although these strains are no longer available from these centers. The stock Full Name, if any, is also given.*

Name	Stock Number	Full Name	Collection details
A4	b.3852	KSA 2	Koriba Dam, Zimbabwe, 1963
A5	b.3875	VAG 1	Athens, Greece, 1965
A6	b.3886	Wild 5B	Red Top Mountain, Georgia, USA, 1966
A7	t.14021-0231.7	-	Ken-ting,Taiwan, 1968
B2	b. 3846	CA 1	Cape town, South Africa, 1954
B3	b.3864	QI 2	Israel, 1954
B6	t.14021-0231.1	-	Ica, Peru, 1956
B7	t.14021-0231.4	-	Kuala Lumpur, Malaysia, 1962

S1.2 Table: Euchromatin boundaries employed in this work (dm6 coordinates).

Chromosome	Euchromatin Start	Euchromatin End
2L	82455	22011009
2R	5398184	24684540
3L	158639	22962476
3R	4552934	31845060
X	277911	22628490

S1.3 Table: Raw euchromatin peak count by tissue for each feature type. The Genome column is the percent of each feature type in the genome.

Feature	Brain	Ovary	Eye Disc	Wing Disc	Genome
Total count	25464	18111	18496	17413	na
TSS	27.90%	39%	37.50%	38.20%	16.20%
TTS	8.60%	10.50%	10.30%	10.20%	11.30%
Exon	2.90%	3.70%	3.10%	3.30%	12.90%
5' UTR	1.60%	1.70%	2.10%	2.20%	1.30%
3' UTR	0.80%	1.00%	1.00%	1.00%	2.00%
Intron	39.50%	30.90%	30.50%	30.00%	36.70%
Intergenic	17.70%	12.30%	14.40%	14.10%	18.80%
non-coding RNA	1.00%	0.80%	1.10%	1.10%	0.90%

S1.4 Table: Mapping statistics

Genotype	tissue	replicate	N map read pairs	N Q30 mapped	Pass QC	N euchromatin Q30	N euchromatin Q30 SV correct
A4	BR	1	71771158	52274422	pass	44045024	39906836
A4	BR	2	62368290	48936438	pass	39888632	35951940
A4	BR	3	75783419	53540250	pass	46957102	42648788
A5	BR	1	93960376	66682740	pass	60104462	54641732
A5	BR	2	104946756	78893144	pass	69588914	63294036
A5	BR	3	74716529	61709316	pass	54206298	49143044
A6	BR	1	60036748	35474110	pass	30280784	27372406
A6	BR	2	50750255	39838160	pass	34505018	31120250
A6	BR	3	69539697	47686952	pass	41529350	37706314
A7	BR	1	81170471	65444566	pass	56952234	51674972
A7	BR	2	95112122	76447960	pass	64960418	59009104
A7	BR	3	81603713	56964342	pass	51014078	46486386
B2	BR	1	98211903	81500160	pass	71810664	65186390
B2	BR	2	59244102	50617958	pass	44335324	40220006
B2	BR	3	61223391	52463710	pass	45736186	41560158
B3	BR	1	48140599	34596332	pass	30134474	27262448
B3	BR	2	47302786	30924424	pass	26149100	23611046
B3	BR	3	58958036	36076368	pass	31083868	28169286
B6	BR	1	75492051	58128380	pass	51409164	46875760
B6	BR	2	53692598	45076220	pass	39648610	35912142
B6	BR	3	66214673	43063540	pass	36645936	33227368
B7	BR	1	100137834	64327562	pass	55669962	49994944
B7	BR	2	43326751	34532064	pass	30298136	27062024
B7	BR	3	40448122	31455402	pass	27907736	25190998
A4	ED	1	80656000	59815788	pass	52992504	48309038
A4	ED	2	61685334	49750814	pass	44107418	40239934
A4	ED	3	91787435	73777002	pass	64862128	59050680
A5	ED	1	96264676	77976884	pass	68514152	62077086
A5	ED	2	56723122	32797120	pass	28439586	25709724
A5	ED	3	22048156	18520246	pass	16160534	14711476
A6	ED	1	98103076	79547840	pass	69447044	62939648
A6	ED	2	80190328	68546420	pass	60569824	54744804
A6	ED	3	112075924	92510282	pass	81750116	73876196
A7	ED	1	76871348	63396382	pass	55405028	50246964
A7	ED	2	87796280	70520590	pass	61553424	55931196
A7	ED	3	99631145	70239416	pass	61033526	55545294
B2	ED	1	75581524	55412894	pass	47953644	43423182
B2	ED	2	69879775	60423182	pass	52980234	47985102
B2	ED	3	67310535	57811538	pass	50663066	45951628
B3	ED	1	50705365	39819642	pass	34722654	31549212

B3	ED	2	83872346	69014394	pass	60078854	54296438
B3	ED	3	70051686	58709896	pass	51575088	46676514
B6	ED	1	80196358	66764068	pass	58238558	52980476
B6	ED	2	81189648	66477162	pass	57788830	52440592
B6	ED	3	69560875	56079856	pass	49220370	44807844
B7	ED	1	70660653	59552736	pass	51445154	45967382
B7	ED	2	46142279	39290060	pass	34717954	31347758
B7	ED	3	76766961	59126162	pass	51841356	46522222
A4	OV	1	81532698	62735178	pass	52623390	47337012
A4	OV	2	58997754	45029450	pass	37189040	33327282
A4	OV	3	103428127	81046954	pass	68790262	61811300
A5	OV	1	77113758	63529554	pass	54558530	49299252
A5	OV	2	59234650	46582400	pass	38048650	34327612
A5	OV	3	76530838	55396006	pass	44949670	40410692
A6	OV	1	75719086	62022436	pass	52350264	47158736
A6	OV	2	109355567	81718154	pass	68233836	61612174
A6	OV	3	89618699	68649814	pass	57097412	51406606
A7	OV	1	89598938	71171406	pass	60046320	54176252
A7	OV	2	105624957	80165870	pass	66653750	59823442
A7	OV	3	91350105	76028296	pass	63641408	57227690
B2	OV	1	66225646	54281112	pass	45318216	40545374
B2	OV	2	87852534	72597616	pass	61582210	55398620
B2	OV	3	110184396	89180084	pass	75773800	68275370
B3	OV	1	68449030	50788994	pass	41882250	37674094
B3	OV	2	71617172	56847672	pass	47878182	42987852
B3	OV	3	71205534	56702676	pass	48197708	43278522
B6	OV	1	121770614	99565362	pass	86695036	78192806
B6	OV	2	85845549	71108898	pass	60337072	54305822
B6	OV	3	91793419	69257402	pass	57245346	51556746
B7	OV	1	147722618	89146168	pass	73717892	65795130
B7	OV	2	79761516	60599588	pass	50038230	44409656
B7	OV	3	42962580	33720300	pass	27556906	24587532
A4	WD	1	60515151	49738892	pass	44243336	40315938
A4	WD	2	56078307	45685146	pass	40602590	37064638
A4	WD	3	51314785	32872384	pass	28625726	26134694
A5	WD	1	50354595	43154900	pass	38044562	34770630
A5	WD	2	45469759	35994894	pass	31439700	28660300
A5	WD	3	65988568	52040818	pass	45442736	41354866
A6	WD	1	64728252	47195092	pass	41612180	37933684
A6	WD	2	82221797	70455610	pass	62819410	57102528
A6	WD	3	93572161	78823382	pass	69735446	63243370
A7	WD	1	61092104	52066686	pass	45519698	41418234
A7	WD	2	84757693	71119418	pass	62719728	57210128
A7	WD	3	86537558	70314768	pass	61882494	56504754
B2	WD	1	83582025	66927970	pass	58428114	52995220
B2	WD	2	85742323	73722046	pass	65226462	59299268
B2	WD	3	52279250	45647482	pass	39921550	36149674
B3	WD	1	20539457	7216490	fail	NA	NA
B3	WD	2	68864938	58342154	pass	51721966	47097194
B3	WD	3	53466204	44496162	pass	38890938	35325288
B6	WD	1	93000699	77663020	pass	68424846	62396890

B6	WD	2	71531414	57303772	pass	50990114	46560980
B6	WD	3	59763212	50287802	pass	44357814	40478002
B7	WD	1	54458465	46577628	pass	40953970	36921480
B7	WD	2	42220164	36215608	pass	31970698	28937010
B7	WD	3	57232749	49048528	pass	43086702	38768168

Chapter 2

Cis and *trans* nature of genetic variation in chromatin state in Drosophila melanogaster

2.1 ABSTRACT

We use ATAC-seq to examine chromatin accessibility in *Drosophila melanogaster* ovaries. The tissues are collected from two isogenic strains (A4,B6), which have a reference quality genome assembly, and their F1 hybrid offspring. We utilize our developed quantile normalization of ATAC-seq data,SV-correction, and ANOVA-based statistical analysis on 44099 ATAC-seq peaks identified in Huynh et al. 2022 [1].We also performed read phasing for our F1 hybrid samples to separate ATAC-seq reads in F1 hybrid samples into A4 genome or B6 genomes using SNPs. We identified 3006 ATAC-seq peaks that are significantly different between parental genotypes. Out of those ATAC-seq peaks, 106 and 45 peaks are identified to be *cis* and *trans* regulatory respectively using cis-*trans* value.

2.2 INTRODUCTION

Historically, complex trait community has been using Genome-Wide Association Studies (GWAS) (GWAS study 14000 cases of seven common diseases [2]), and QTL mapping (QTL mapping studies in yeast [3,4], mouse [5–7], and *Drosophila* [8]) to identify causal loci linked to a particular trait. Despite the fact that both methods are well developed and powerful, the exact causative variants underlying risk remain hidden

[2,9], and an appreciable fraction of heritable variation remains unexplained [10]. New emerging evidence, however, gave rise to the idea that variations found in complex traits are due to regulatory variants [11–15]. Thus,a new strategy in studying complex traits is to identify polymorphic non-coding regions with regulatory function. This is done utilizing DNase-I HS (DNase-I hypersensitive site) sequencing [16] and/or the more recent and experimentally straightforward ATAC-seq (Assay for Transposase Accessible Chromatin) approach [17]. Among the two methods, ATAC-seq is especially beneficial due to its low requirement of sample concentration. It employs Tn5 transposon (Nextera) sequencing chemistry to make an Illumina paired end library using nucleosome bound DNA as template for the transposition reaction. As a result, only open chromatin regions, which likely function as regulatory features, are cut by the Tn5 and result in high sequence coverage. Both easy to use and straightforward natures of ATAC-seq allows for regulatory elements characterization of large panels of genotypes [18,19].

Given the importance of nucleosome free DNA regions, great effort has been spent to elucidate the mechanism of nucleosome eviction. DNA sequences are known to bend differently depending on their nucleotide sequences [20–22]. Thus, nucleosome stability is dependent on histone affinity to specific DNA sequences [23,24]. As a result, the less affinity there is between histone and DNA sequence, the less stable a nucleosome is. In this case, the resulting openness of chromatin due to nucleosome eviction can be in *cis* due to substantial DNA sequence reference. This can be observed for the example of A4 genotype in supplementary figure 2.1. In addition, it has also been shown that nucleosome positions can be regulated in *trans* by ATP-dependent

nucleosome remodeling complexes [25,26]. In this case, the nucleosomes are being evicted by these complexes resulting in the observed chromatin accessibility for A4 genotype in supplementary figure 2.1.

In order to elucidate the *cis* and *trans* nature of open chromatin regions, creating a F1 cross between two isogenic F0 parents is a viable option. This is due to F1 samples carrying both alleles of the parents while being under effect of the same transcription factor. Open chromatin regions that are in *trans* would manifest as an intermediate expression for both alleles in F1 hybrid regardless of the level of expression of the same gene from parental genotypes [27]. This can be observed in the example in supplementary figure 2.1 top. In this example, the nucleosome modification complex from the A4 genotype exerts the same effect on the B6 allele resulting in the same chromatin accessibility for both hybrid alleles. The observed ratio of ATAC-seq coverage (chromatin accessibility) between the two hybrid alleles would be different compared to the ratio of ATAC-seq peak coverage between the two parental genotypes. In contrast, open chromatin regions that are in *cis* would manifest as an imbalanced coverage between the two alleles in F1 hybrid with each allele having the same coverage as their parents of origin (S2.1 bottom). This is due to allele specific DNA property instead of diffusible nucleosome modification complexes.

However, as most ATAC-seq studies don't perform haplotype phasing for heterozygous diploid samples, it is difficult to verify the *cis* or *trans* nature of any identified variations in regulatory elements since it isn't possible to observe which of the two parent chromosomes a variant allele is acting on [28]. Furthermore, haplotype phasing doesn't appear to be a standard for the ATAC-seq studies in complex trait

community since a quick search on pubmed.gov prior to March 30,2023 with three keywords "ATAC-seq","phase", and "haplotype" yielded a single paper discussing the diploid nature of species and sequencing method for haplotype phasing utilizing human data [29]. As a result, while characterizing polymorphic transcription factor binding sites (TFBS) with ATAC-seq is helpful in elucidating the variations in regulation for complex traits, the *cis* and *trans* nature of those open chromatin regions remain poorly understood.

Another concern raised by the ATAC-seq studies mentioned above is the lack of controls for structural variant impact on alignment failure. None of the *Drosophila* ATAC-seq studies mentioned above have used a strain containing the *Drosophila* reference genome, against which their ATAC-seq data are aligned. In fact, they all use different mutant strains and wildtype strains that lack a reference quality genome *de novo* assemblies. This presents an issue as any pair of Drosophila strains have been shown to contain a plethora of differential structural variants, short insertion/deletion, and SNPs [30]. These events can have artificial effects on the alignment of data as demonstrated in RNAseq data [31].

Here we carry out a biologically replicated ATAC-seq experiment in two of highly-characterized isogenic genotypes of *D. melanogaster* [30], and their F1 hybrid offsprings. Then, statistical tests are done using a set of open chromatin peaks identified in Huynh et al. 2022 [1] to identify peaks that are polymorphic in open chromatin configuration between two parents. As both isogenic parental strains in use are highly-characterized with reference quality *de novo* assemblies [30], we can also correct for SV impacts on alignment artifacts which can produce a false positive rate

higher than 50% [1]. Furthermore, we also have a complete list of SNP differences in both F0 (parental) strains compared to the dm6 reference genome. Thus, we finally perform haplotype phasing on a F1 (child) offspring from the two parental strains A4 and B6 for statistical analysis to distinguish between *cis* and *trans* natures of identified variants.

2.3 RESULT

Workflow and quality control: We dissected ovaries from adult A4 and B6 strains from *Drosophila* Synthetic Population Resource [3], whose founder strain genomes have been extremely well characterized [49], and the F1 offspring of those two strains. For each genotype (strain), we obtained seven biological replicates. Both A4 and B6 are highly inbred strains and are collected from two different geographic locations shown in Supplementary Table 1. Dissected samples were immediately processed to make indexed ATAC-seq libraries [19], and sequenced to obtain 206-493 million Illumina PE reads per sample (mean=393M, sd=70M). Reads were aligned to the dm6 *Drosophila* reference genome.

After alignment, ATACseqQC is used to generate mononucleosome and nucleosome read density against distance to TSS as a means to quantify tagmentation. Figure 2.1 shows an example of an expected pattern (Fig 2.1 top) and an over-tagmented pattern (Fig 2.1 bottom) using two replicates from B6 genotype samples. Thus, we only kept samples that have the pattern as similarly to the expected pattern as possible. Supplementary figures 2.1,2.2,2.3 show the mononucleosome density against nucleosome-free patterns for all selected samples from A4,B6,and Hybrid genotypes

since the pattern matched with the expected pattern. These samples are kept for downstream analysis with replicate numbers being renumbered from 1 to 5.

Samples were then normalized to obtain weighted coverage at each ATAC-seq peak identified in Huynh et al. 2022 [1]. We finally corrected read coverage statistics using the identified polymorphic structural variants (SVs) for A4 and B6 genotypes, and carried out statistical tests at peaks to identify chromatin structures that varied among the two parental genotypes (A4 and B6). Supplementary figure 2.4 and 2.5 shows the fragment length distribution for each sample before and after normalization respectively. For each peak that is significantly different between two parental strains, their parental coverage ratios are then compared to the phased coverage ratios from the Hybrid data. Our general workflow is depicted in Supplementary Figure 2.6.

ATAC-seq identifies polymorphic open chromatin regions between A4 and B6 parental genotypes: Out of 44099 peaks identified in Huynh et al. 2022 [1], only 36863 peaks that have coverage higher than the cut-off of 50 for both A4 and B6 genotypes. Among these, only 3006 peaks are significantly different between A4 and B6 samples with the FDR adjusted p-value < 0.1 ($-\log_{10}$(FDR p-value) > 1). Figure 2.2A shows the histogram of the FDR adjusted p-value (FDR p-value) for all peaks. Figure 2.2B shows the distribution of FDR adjusted p-value by chromosome position with the significant peaks highlighted in red. Since both A4 and B6 are only different strains of the same species, it is as expected that 91.85% of the ATAC-seq peaks are not significantly different between the two genotypes. However, the 3006 significant peaks are of great interest as they represent the potential differences in gene expression regulation

between the two genotypes due to the differences in open chromatin regions. Interestingly, all of the significant peaks appear to be randomly distributed (Fig 2.2B).

Example of significantly different open chromatin regions between A4 and B6 parental genotypes: Figure 2.3A depicts a peak at chr3R:31444502 on the intron 3, and 2 of isoforms A, and B respectively of the gene *Gprk2* (G-protein-coupled receptor kinase) which encodes a family of serine/threonine kinases. This gene is interesting as it has been shown to regulate female fertility [51], egg shape [52], and egg cAMP level [53]. *Gprk2* is also involved in regulation of Hedgehog signaling [54] , and in regulation of rhythmic olfactory response [55]. However, our data are only collected from ovary samples, no implication can be made regarding the *Gprk2* functions in Hedgehog signaling and olfactory response regulation. Regardless, the importance of *Gprk2* in regulation of female fertility and egg related aspects would still be interesting for the developmental biological community. This peak is specific by genotype with a FDR p-value of 0.0020. This can be observed as B6 parent (brown) appears to be more open compared to A4 parent (green). The result suggests that this region is being regulated in a genotype dependent manner. Thus, given the functions of *Gprk2* gene, this peak should be an interesting future study target.

The second example is the ATAC-seq peak at chrX:2299228 located on intron 2 of *Raf* gene which is involved with proliferation of stem cells [56], and cell differentiation [57] (Fig 2.3B). This ATAC-seq peak has a FDR p-value of 0.098 suggesting that there is significant difference in the openness of chromatin at this region between the two parental genotypes A4, and B6. This can be observed as the B6 (brown) genotype has

a higher coverage compared to the A4 (green) genotype. This result implies that the *Raf* gene is being regulated in a genotype dependent manner in this region. Given the location of the peak inside the intron 2 of *Raf* gene, this differential in regulation might be due to an alternative splicing process since alternative splicing has been shown to be regulated by transcriptional factor [58]. In addition, intronic enhancers have also been shown to be involved in fine-tuning developmental specific gene expression in plants [59], and in humans [60]. Thus, this enhancer would be an interesting target for future functional study related to stem cell proliferation or differentiation since our experiment is not sufficient to test its function.

Figure 2.3C depicts another example of a polymorphic open chromatin region at chr2L:2959038, which is associated with the gene *Rbp9,* with a FDR p-value of 0.027. Rbp9 gene encodes a putative RNA binding protein which has been shown to be involved in establishment of blood brain barrier [61],regulation of germ cell proliferation [62] by inducing apoptosis in egg chambers [63], and maintenance of germline sexuality [64]. Furthermore, *Rbp9* gene, which is also homologous to human Hu gene, has been shown to be involved in causing ovarian cancer in flies if it is mutated [65]. Since the FDR p-value of this peak is 0.027, this peak is highly specific by genotype with A4 parent (green) having more open configuration compared to the B6 parent (brown) suggesting that the gene *Rbp9* is being regulated in a genotype dependent manner at this location. Due to *Rbp9*'s important functions, this peak would be an interesting study target for future studies. However, as our samples are collected solely from adult ovary, the implication from our data can only be associated with the function of Rbp9 gene on germ cell regulation. Furthermore, this peak is located 109 bp upstream from TSS of

isoforms C,F,K, located on intron 2 of the isoforms H,I, and located on intron 1 of isoform B,E,J,G. This suggests that the genotype specificity at this peak potentially affects gene expression through intron [66] or through transcription starting site which is a more canonical regulator of gene expression [67,68].

Figure 2.3D shows an open chromatin region at chr3L:3428317 which is located downstream of the *eIF5B* gene known to enable translational factor ability through interaction with *Vasa* during development [69,70]. This peak is located 341 bp downstream of TTS of isoform D, E, and 348 bp downstream of TTS of isoform B,C,F. Since the region is located directly downstream of TTS of all isoforms, it is implied that this open chromatin region is regulating gene expression of *eIF5B* through interaction with TTS. Thus, the polymorphism observed in open chromatin configuration at this location with FDR p-value of 0.092 suggests that the region is being regulated in a genotype dependent manner with B6 parent (brown) appearing to be more open with higher coverage compared to A4 parent (green)

Phasing quality control: SNP count after each step of filtering is provided in table 1. After identifying peaks that are significantly different between parental A4 and B6 genotypes, the next step is to classify those peaks into *cis*- or *trans*- regulatory elements categories to further elucidate their functions in gene regulation. However, since this task involves comparing differences in coverage between parental genotypes and phased hybrid genotypes, it is essential to ensure that all the phasing is done correctly, and to have a cut-off value for phase percentage. Since both A4 and B6 genotypes are isogenic genotypes from the Drosophila Synthetic Population Resource

DSPR [3], we expected that most of the fragments from A4 genotype,or B6 genotype should be phased into genome 2 (A4), or genome1 (B6) respectively. Thus, we have used the data from these two genotypes to control for the quality of our custom phasing script. Figure 2.4A depicts the fragment counts for the fragments that are phased into genome A4 versus B6 for each peak using samples from A4 and B6 genotypes. Each dot in figure 2.4A represents a peak with the peak being phased correctly (if it is sample A4/B6, most fragments should be phased into A4/B6 respectively) colored as red. Due to our rigorous selection of SNPs, all of the peaks are phased correctly as expected (Fig 2.4A). In addition to quality control of custom phasing script using SNP phase percentage, we also select a cut-off value for phase percentage for the phased coverage from Hybrid samples. We have decided to pick a cut-off at 33% phased for phase percentage. This is due to the 36863 peaks used in parental ANOVA having the 50th percentile phase percentage as 24.5% (Fig 2.4B).

Cis-trans value quality control: Out of 3006 peaks that have been identified to be polymorphic by genotype between two A4 and B6 parental genotypes, only 1398 peaks can pass our FDR p-value cut-off at 0.1 and our phase percentage cut-off at 33% (supplementary data 1). However, Loess smoothing curves for the plots of cis-trans value as a function of mean phase percentage (Fig 2.5A), parent FDR p-value (Fig 2.5B), and mean parent coverage (Fig 2.5C) reveal potential association between them and the cis-trans value. This is a potential problem as the cis-trans value could be under the effect of technical reasons rather than biological reason. For example, the increase in the phase percentage appears to drive the cis-trans classification toward *cis* whereas

the increase in parental FDR p-value and parental coverage appears to drive the cis-trans classification toward *trans*. Therefore, this reinforces the need of an additional stringent quality control -as described in method section- by comparing alignments to three different genomes in order to completely remove potential artifacts due to unknown SVs, or SNPs effect on alignment rate (coverage).

Surprisingly, 4 peak regions have multiple duplications when they are matched to A4 and B6 genomes using Blastn [71–75]. Thus, these four peaks are removed since it is impossible to make a comparison for alignment rate. Then, the first criteria that we looked at is whether the peaks are near any known SVs at all. Supplementary figure 2.7 depicts the average percentage error in alignment to A4 or B6 genome compared to alignment to dm6 genome across replicates for A4 samples and B6 samples. As expected, 22.17% of the 1394 peaks have 1 or more SVs located within 800bp of the peak. While our procedure includes correction for SVs using dm6 genome coordinates, we decide that it is better to remove these peaks. This is due to the difficulty in obtaining the correct A4 and B6 genome coordinates for all of our known SVs. As a result, only 1085 peaks remain. Supplementary figure 2.8 shows that by removing these peaks, the majority of the peaks -which are in the two extreme parental coverage difference quantile bins (bottom 10% (S2.8 Fig left) and top 10% (S2.8 Fig right)) of parental difference -have an absolute value of percentage error in alignment at peak being less than 5% with only a few outliers.

The next criteria that we look at is the average percentage difference at peak. As stated in the material and method section, we decide on the cut-off being 5%. This means that the difference in fragment counts have to be less than 5% for both

alignment to dm6 versus A4 and alignment to dm6 versus B6. Supplementary figure 2.9 shows the cis-trans value for the remaining 965 peaks after applying this filter for all three quantile bins of parental coverage difference: bottom 10% [-3.7,0.88], middle 80% (-0.88,0.753], and top 10% (0.753,3.34]. Most peaks have a percentage error less than 3% regardless of difference between parents. Thus, we are confident that the cis-trans values for these remaining 965 peaks are truly representative of biological factors and not of technical factors due to stringent cut-off for SNP phase percentage, SVs locations, and percentage error in alignment rate between 3 genomes.

Allele ratio comparison: In addition to the mentioned criteria, we also fit a linear regression line to each quantile bins of parental coverage difference for all 3006 original peaks. Figure 2.6A depicts the three linear regression lines (black line shows the slopes for these lines) and loess smoothing curve (gray) by parental coverage difference quantile bins. Surprisingly, the slope of the linear regression line for the (-0.88,0.753] quantile bin has a slope, and R^2 of only 0.52 and 0.5 respectively. The R^2 indicates that the linear regression line wasn't a good fit and that the peaks are more scattered. Furthermore, the slope of only 0.52 suggests that there isn't a strong relationship between the parental coverage ratio and the phased hybrid coverage ratio. Thus, it is very difficult to discern whether these open chromatin regions are in *cis* or in *trans* because *cis* and *trans* open chromatin regions should be clustered on diagonal line, and horizontal y=0 line respectively. In contrast, for the two extreme quantile bins of [-3.7,-0.88], and (0.753, 3.34], the R^2 and the slope values are 0.901, 1.01, and 0.86, 1.06 respectively. Both R^2 suggest that all the peaks in these two quantile bins fit really well

on the linear regression models, and have a strong *cis*- pattern. Furthermore, both slope values suggest that there is a strong relationship between the parental coverage ratio and the phased hybrid coverage ratio Similar results can be observed in the 965 peaks that pass all of our quality control steps (Fig 2.6B) implying that the observed trends are due to biological reasons. Thus, we would have the best ability to detect coverage ratio change between parents and phased hybrid genotypes. Therefore, we only classify *cis* and *trans* for peaks with $\log_2$(A4/B6) > 0.753 and < -0.88.

Figure 2.6C depicts the ratio in coverage between parental genotypes and the phased hybrid genotypes. As expected, *cis* nature open chromatin regions are distributed along the diagonal line whereas *trans* nature open chromatin regions are distributed along the horizontal zero line. This trend is in line to the result published in McManus et al., 2010 [58]. Out of 965 peaks that pass our quality control.10.98% of peaks, and 4.66% of peaks are assigned as *cis* and *trans* respectively (Table 2). The count for *cis* and *trans* classification for peaks before the alignment comparison quality control step is also provided in table 2.

Illustrative examples of *cis*- and *trans*- open chromatin regions: Figure 2.7A depicts one example of a cis open chromatin region at chr2L:3346050 located on intron 2 of the *E23* gene which has been shown to be involved in ATPase-coupled transmembrane transporter activity [77] and to be capable of repressing ecdysone-mediated gene activation [77]. Since ecdysone is a major steroid hormone which is known for its role in coordinating developmental processes, such as metamorphosis [78], *E23* gene ,thus, is implied to be a newly elucidated regulator for hormone

signaling. Given its important role, This ATAC-seq peak is an interesting future study target for the developmental community. The chr2L:3346050 ATAC-seq peak can be observed to be significant by genotype with A4 genotype (green) has a lower coverage compared to B6 genotype (brown). This suggests that the B6 genotype is more open than A4 genotype in this region. This is also supported by the FDR p-value of 0.00196. Beside being regulated in a genotype dependent manner, the region also appears to be regulated in a *cis*-manner. This can be observed as the coverage for H_A4 (light green) and A4 (green) is similar to each other. Similarly, the coverage for H_B6 (light brown) and B6 (brown) is also relatively similar to each other. Thus, this suggests that the ratio of H_A4_coverage/H_B6_coverage is the same as A4_coverage/B6_coverage. This result is also supported by a *cis*-tran value of 1.44. Due to *E23* function and intron open chromatin region's ability to regulate gene expression [59], or regulate alternative splicing [79], these results imply that ATPase-coupled transmembrane transporter activity, or even ecdysone-mediated gene activation is being regulated in a *cis*- acting and genotype dependent manner.

Figure 2.7B depicts another ATAC-seq peak example at chr2R:6110252 located on intron 2 of *EcR* isoform B,G, intron 1 of *EcR* isoform C, intron 3 of *EcR* isoform A, and intron 4 of *EcR* isoform E,D. *EcR* encodes a receptor for ecdysone which has been shown to involve in regulation of normal oogenesis [62], metamorphosis [63,64], sleep [60],and early germline differentiation [65]. Given *EcR* important function, it is interesting to see polymorphism in open chromatin at this chr2R:6110252 loci. This ATAC-seq peak chr2R:6110252 has FDR p-value and *cis*-tran value at 0.0167 and 0.24 respectively. The p-value suggests that there is a genotype dependent polymorphism in

open chromatin configuration between the two parental A4 and B6 genotypes. This can be observed as B6 (brown) appears to be more open than A4 (green) with higher coverage. This region also appears to be regulated in a *trans* manner with a *cis*-trans value at 0.24. This is an indication that the H_A4 (light green) and H_B6 (light brown) coverage ratio from hybrid samples is different compared to A4 (green) and B6 (brown) parental genotypes. These ratios can be observed in figure 2.7B. Thus, the results imply that the *EcR* gene is being regulated in a *trans* manner in this region. Due to *EcR* functions, our results imply that early germline differentiation, metamorphosis,sleep, and oogenesis are being regulated in a genotype and *cis* manner at this open chromatin region. However, figure 2.7B has shown a potential issue. Instead of being a *trans* peak as supported by our analysis, this peak can be classified as *cis* if it is moved 100 bp upstream.

2.4 DISCUSSION

Previous ATAC-seq studies only aligned their data to either dm3 or dm6 reference genome without any haplotype phasing despite the heterozygous diploid nature of the data [22–37]. Furthermore, haplotype phasing doesn't appear to be a standard for the ATAC-seq studies in complex trait community since a quick search on pubmed.gov with three keywords "ATAC-seq","phase", and "haplotype" yielded a single paper discussing the diploid nature of species and sequencing method for haplotype phasing utilizing human data [42]. This presents an issue as gene expression has been shown to be regulated in both *cis*- and *trans*- manner [38–40]. Here, we carried out ATAC-seq experiment on adult ovary samples due to its ease of collection and its

important function as a model for study on the maintenance and regrowth of new organs by stem cell units, germline, and somatic follicle stem cells [84]. Furthermore, our data consists of ATAC-seq on two founder genotypes A4, B6, and their F1 offsprings from the established Drosophila Synthetic Population Resource [10]. This allows us to perform haplotype phasing using the known SNPs from both parental genotypes. As a result, we can classify the polymorphic open chromatin regions between parental genotypes -identified through statistical test- as *cis*-acting or *trans*-acting by comparing their coverage ratio with the phased coverage ratio from the hybrid sample. We expect that our data would be of great utility for developmental biology and complex trait communities.

In this work we performed ANOVA statistical test to characterize polymorphic chromatin profiles using obtained nuclei from two founder genotypes and their F1 heterozygous offsprings. Both founder genotypes are collected from Drosophila Synthetic Population Resource [10], are highly isogenic, and have *de novo* assemblies to reference quality [49]. Thus, this allows us to correct for alignment artifacts caused by structural variants which can cause an upward of 50% false positive rate to the statistical test [1]. After correcting for SVs and ANOVA tests, 3006 ATAC-seq peaks have been identified to be significantly different between two parental genotypes.

After identifying polymorphic open chromatin regions, we also perform haplotype phasing on hybrid samples using known SNPs from the two founder genotypes. Furthermore, we also performed quality control tests and only kept SNPs that could phase 90% of read spanning the loci correctly. This removes 76.40% of SNPs that can be considered bad SNPs ensuring a robust and accurate haplotype phasing. Next, we

attempted to classify polymorphic open chromatin regions as *cis*-acting or *trans*-acting using our simple cis-trans value calculated using the parental (A4,B6) genotype coverage and the phased haplotype coverage (H_A4,H_B6) from hybrid samples. Out of 3006 polymorphic peaks, only 1398 peaks that have FDR p-value less than 0.1 and have more than 33% fragment spanning the peaks phased correctly.

However, Loess smoothing curves for the plots of cis-trans value as a function of mean phase percentage (Fig 2.5A), parent FDR p-value (Fig 2.5B), and mean parent coverage (Fig 2.5C) reveal potential association between them and the cis-trans value. This is a potential problem as the cis-trans value could be under the effect of technical reasons rather than biological reason. Thus, we performed additional quality control rounds using A4 and B6 samples which are aligned to dm6, A4, and B6 genomes in order to calculate the percentage of correct alignment by comparing all three genome alignments. Only peaks with more than 95% correct alignment between all three genomes are kept. However, since it is near impossible to lift-over the dm6 coordinate of known SVs to the coordinate of A4 genome or B6 genome, we also decide to remove all peaks that have any known SVs within 800bp to ensure much more restricted but robust results. As a result, only 965 peaks remain that pass all quality controls steps , and are significantly different between two parental genotypes.

We also attempt to classify ATAC-peak by *cis* and *trans* nature using cis-trans value. However, three separate linear regression lines applied to each parental coverage difference quantile bin suggests that we would have the best ability to detect coverage ratio change between parents and phased hybrid genotypes, and to classify *cis* or *trans* for only peaks with $\log_2(A4/B6) > 0.753$ and < -0.88. The *cis* and *trans*

regulatory classification distribution is 10.98% and 4.66% respectively out of the 965 peaks. These cis-trans classified ATAC-seq peaks would no doubt be interesting targets for future functional studies using the Crispr/Cas9/allele swapping methods [86–90] to further understand gene regulations-especially for important genes associated with these *cis* and *trans* polymorphic open chromatin regions. However, there is still a potential quality control step or a hidden error being missed. This can be seen in the example shown in figure 2.7B. The ATAC-seq peak in figure 2.7B can be classified differently if it is moved 100 bp upstream. This indicates that there are still hidden issues that need to be resolved. Once this final hurdle is solved, the chapter will be ready for publication, and will provide interesting targets for future functional studies on complex traits and variants in regulatory elements.

2.5 MATERIALS AND METHODS

Strains: All the strains used are founder strains A4 and B6 from DSPR, and their F1 offspring (S1 Table).

Tissue dissection and ATAC-seq library prep: Our procedure for tissue dissection and ATAC-seq prep is the same as described in Huynh et al. 2022 [1]. The 2 inbred parent strains (A4, and B6), and F1 offspring were raised and maintained in regular narrow fly vials on a standard cornmeal-yeast-molasses media in an incubator set to 25°C, 50% relative humidity, and a 12 hour Light : 12 hour Dark cycle. Ovaries were dissected and pooled from five 2 day old females per replicate. All dissections were carried out 1-9 hours after lights on, and following dissection all samples were

immediately subjected to nuclei isolation. The full protocol for ATAC-seq library construction can be found in Huynh et al. 2022 [1].

Read processing and normalization: Adapters were trimmed from the raw reads using Trimgalore-0.4.5 [90,91], and were aligned to dm6 *D. melanogaster* reference genome with bwa-0.7.17 [92]. Only primary reads and their mates that were properly mapped were kept using samtools 1.15.1 [93]. All duplicated reads were removed using picard 2.18.7 [93,94] via MarkDuplicates. Only reads aligning to the five major chromosome arms - X, 2R, 2L, 3R, and 3L - were retained for analysis. Density plots by distance to TSS of nucleosome-free fragments and mononucleosome fragments produced by ATACseqQC 1.18.1 [95] were used for quality control for each sample.

After processing sequencing data, we also performed normalization by assigning "weight" to each fragment length by samples. This method is inspired by “quantile normalization” methods used in the field of gene expression [96]. Detailed read processing and normalization are described in our previous paper [1]. After normalization, we calculated the weighted Coverage (C) at peak loci , which is the set of peaks identified in our previous paper [1], by summing all the weights at said loci. Result files contain coverage for all five replicates from A4 genotype, B6 genotype, and Hybrid genotypes. Additional standard quantile normalization is also applied to data prior to ANOVA step.

Accounting for hidden structural variants: Typical ATAC-seq experiment reads are aligned to reference genomes. However, the results are then normalized without

correcting for any potential artifacts in alignment errors due to structural variants. Thus, as both A4 and B6 strains in use for this study have complete de novo reference quality genome assemblies [49], we are able to correct the structural variant -found in both strains- effect on read alignments. Both the reasons for such correction and the correction process are described in full in our previous paper [1].

Statistical testing: We carry out ANOVA statistical test for the two parental strains (A4,B6) at peaks to identify loci with polymorphic weighted log transformed Coverage (lnC = ln(C+5)) after peaks with a weighted average coverage < 50 were dropped as:

lnC ~ geno

A False Discovery Rate (FDR) was calculated using the p.adjust function in R (Benjamini and Hochberg 1995; Yekutieli and Benjamini 1999) [97,98]. Tests with FDR adjusted p-values < 0.1 (or $-\log_{10}$(FDR p-value) > 1) are considered significant.

Peak annotation for significant peaks: For peaks identified to be significantly different between A4 and B6 genotypes, peak positions are converted to a pseudo bed file with columns as: chromosome, peak position -1 , peak position +1, p-value, "+". HOMER v 4.11.1 [99] is then used to annotate each peak as belonging to eight feature type groups based on the peak locations. The eight feature types are 3' UTR, 5' UTR, promoter-TSS (TSS), transcription termination site (TTS), intron, exon, intergenic, and non-coding. The

detailed description of each feature type is shown on HOMER main website or in our previous paper Huynh et al. 2022 [1]. The peak count for each feature type is shown in supplementary table 2.

Haplotype phasing: Since all of our SNPs are provided by Drosophila Synthetic Population Resource, we can remove any SNPs that have heterozygous genotypes (1/0 or 0/1 instead of 0/0 or 1/1) for any of the eight isogenic strains. Furthermore, only SNPs that are heterozygous between parental strains A4 and B6 genotype (1/1 and 0/0 for A4, B6 or vice versa) will be kept. Then, haplotype phasing is performed using bam files and a custom script utilizing pysam [2] which is a wrapper around htslib [100] and samtools [93,101]. The custom pysam script takes in heterozygous SNPs between the two parental strains A4 (genome 2) and B6 (genome 1) to separate reads into genome 1 or genome 2 by comparing the DNA bases at SNP positions to the A4 and B6 specific bases found at said SNPs. If reads do not span any SNPs, or have mates phased into incorrect genomes, they will be assigned as unassigned. Afterward, SNPs are further filtered by using the SNP phase percentage correctly (the percentage of reads that are phased into the correct genome out of the total reads spanning said SNP). Only SNPs that can phase correctly for all five replicates in both genotypes A4 and B6 will be kept. The cut-off SNP phase percentage correctly used for this step is set at higher than 90% since we will lose close to or higher than 50% of total SNPs with higher percentage. Hybrid data that are phased into the A4 genome and into the B6 genome are labeled as H_A4 and H_B6 respectively. Then, weight from the same Hybrid samples are transferred to the phased data from the same sample using fragment

length. Phased coverage at peak loci was calculated similarly to the procedure described above. However, the phased coverage is also up-scaled by multiplying to a factor 2/(phase percentage at peak) (phase percentage at peak is (H_A4_coverage + H_B6_coverage)/H_coverage) to account for the haploid nature of phased data. Any mention of phased coverage for H_A4 and H_B6 for the rest of this paper will be the up-scaled coverage if not mentioned otherwise.

Haplotype phasing quality control: As a means to control the quality of our phasing scripts, we have used the data from A4 and B6 genotypes as controls. Since they are both isogenic strains from DSPR, we expect that the fragments belonging to A4 or B6 genotypes will be phased into the same genotypes. Thus, we have calculated the actual raw count of fragments that span each peak that has coverage > 50. If the samples belong to A4 genotypes, most fragments should be phased into genotype A4 with a negligible count of fragments phased into genotype B6. The same can be expected of samples belonging to B6 genotype with most fragments phased into genotype B6. In addition to ensuring the phasing to be correct, we also select a cut-off phase percentage -which is simply 100*(un-scaled H_A4 coverage + unscaled H_B6 coverage) / (H_coverage) , and is not the same as the SNP phase percentage mentioned above-. Only peaks that have phase percentage for all Hybrid samples higher than this cut-off will remain for downstream analysis.

Allele ratio comparisons: As stated in the introduction, we expect that the openness of these regions would manifest similarly to the gene expression observed in the

comparison study between the F1 hybrid and the F0 isogenic parents [44]. Therefore, if the open chromatin regions are regulated in a *trans*-manner, H_A4_coverage/ H_B6_coverage would be different from A4_coverage/B6_coverage. In contrast, if the open chromatin regions are regulated in *cis*-manner, H_A4_coverage/ H_B6_coverage would be the same as A4_coverage/B6_coverage**.** Therefore, in order to assign *cis*- or *trans*- to identified open chromatin regions, we would attempt to calculate cis-trans value. For each peak that was significant between parents, we extracted all peak loci that had fragments phased into two haplotypes (genome 1 (B6), genome 2 (A4)). Then, we calculated a cis-tran value which is simply:

cis-trans = log2(mean_H_A4_coverage/ mean_H_B6_coverage) – log2(mean_A4_coverage/ mean_B6_coverage)

We assigned any peaks- that pass all additional filters mentioned in the cis-trans value quality control section below with cis-trans value- within $(-\infty,-0.5]$ or $[0.5,\infty)$ as cis as this indicates the difference in coverage ratio between hybrid and parental samples. Any peak with *cis*-trans value within (-0.5,0.5) is assigned as trans since this indicates that there is no difference in coverage ratio between hybrid and parental samples. Then, for each peak, we plotted $\log_2$(mean_H_A4_coverage/ mean_H_B6_coverage) of the child samples against $\log_2$(mean_A4_coverage/ mean_B6_coverage) from the two parental pairs. This plot is inspired by the paper Wang *et al.*, 2019 [102].

Cis-trans value quality control: Since the cis-trans value is highly dependent on the difference between A4_coverage and B6_coverage, or between H_A4_coverage and H_B6_coverage, we feel that it is necessary to apply more stringent quality control tests to ensure accurate cis-trans value representation. Therefore, we look at the cis-trans value as a function of mean phase percentage from the Hybrid samples, as a function of parental FDR p-value, and as a function of the mean parental coverage. These functions will highlight any obvious bias in cis-trans value calculation due to those three factors.

Furthermore, we also separate peaks by difference between parental coverage into three quantile bins at [0%,10%],(10%,90%],and (90%,100%] (which are [-3.7,0.88], (-0.88,0.753], and (0.753,3.34] respectively). Next, we extracted 100bp, 2kb, and 20kb nucleotide sequences centered on each peak. Blastn is then used to match those regions from the dm6 genome to A4 and B6 genomes to get the A4 and B6 genomes coordinates for the same regions. Only regions with the greatest number of nucleotides being matched will be kept. Afterward, we align A4 and B6 samples to both A4 and B6 genomes. Fragments are then extracted from each genome alignment using appropriate coordinates. As a result, we are able to compare the actual fragment count spanning 100 bp regions (peak loci ±50bp) centered on each peak between alignments to three reference genomes (dm6, A4, and B6). This allows us to remove any peaks that have the absolute percentage error (in alignment) at peak > 5% between alignment to dm6 reference genome and alignment to both A4, B6 genomes. The discrepancy at peak is simply calculated as follows:

Percentage error at peak =100% $(N_X - N_{dm6})/N_{dm6}$

N_X is the total fragment count mapped to the X genome being A4 or B6 for the 100bp peak region. N_{dm6} is the total fragment count mapped to dm6 for the 100bp peak region. Furthermore, if any peak has structural variant(s) within 800bp upstream and downstream of the peak, it will be removed from downstream analysis. This is because it is extremely difficult to acquire the correct A4 or B6 genome coordinates for any of the known structural variants which are identified using dm6 coordinates. Last but not least, we also fitted three different linear regression models to each of the three parental coverage difference quantile bins [-3.7,-0.88], (-0.88,0.753], and (0.753,3.34] to select a cut-off for parental coverage difference to avoid misclassification due to subtle difference in parental coverage. However, despite best effort, figure 2.7B has indicated that there is a missing quality control step or a hidden error in my method. This can be seen with the completely different possible nature of *cis*- and *trans*- of this ATAC-seq peak if its location is moved 100 bp upstream. Once this final hurdle is cleared, the second chapter would be ready for publication.

2.6 FIGURES

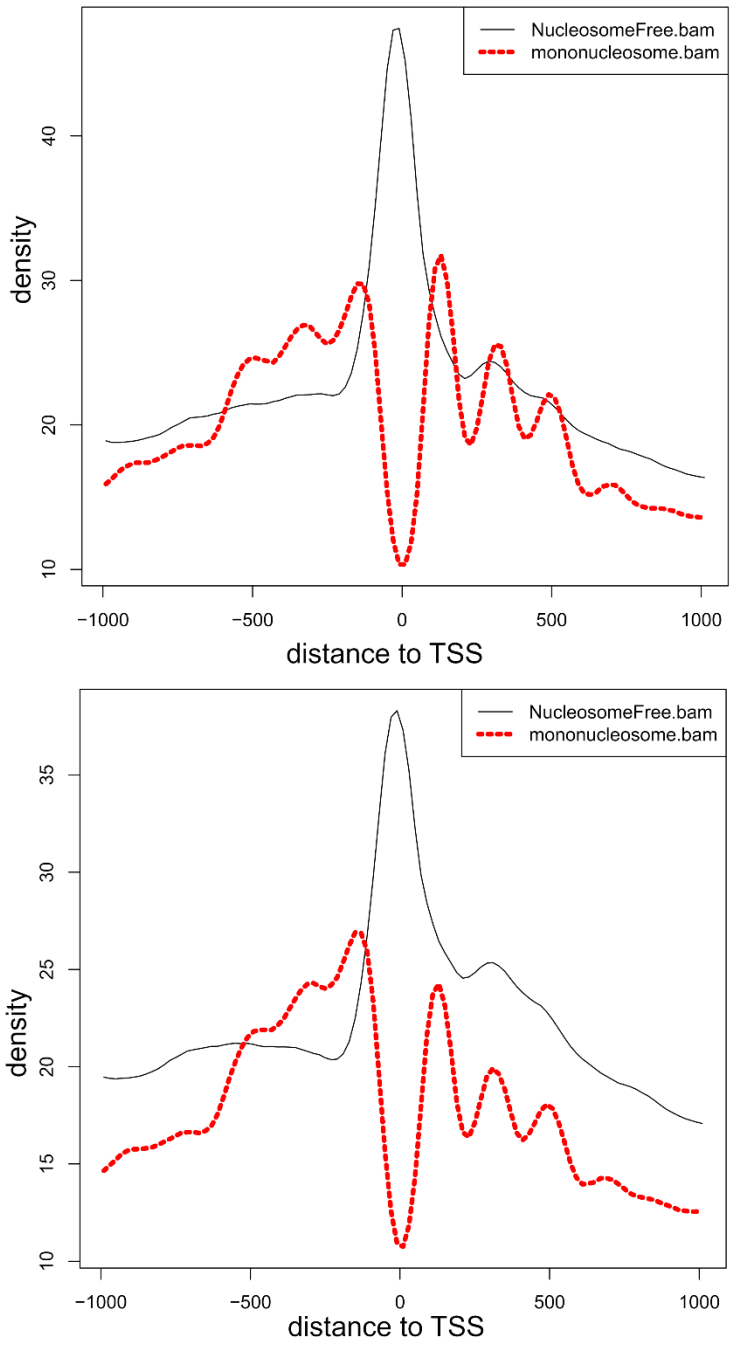


Fig 2.1. ATACseqQC mononucleosome and nucleosomes-free read density against distance to TSS.
(A) Plot showing expected pattern for correct tagmentation. (B) Plot showing density pattern for over-tagmentation

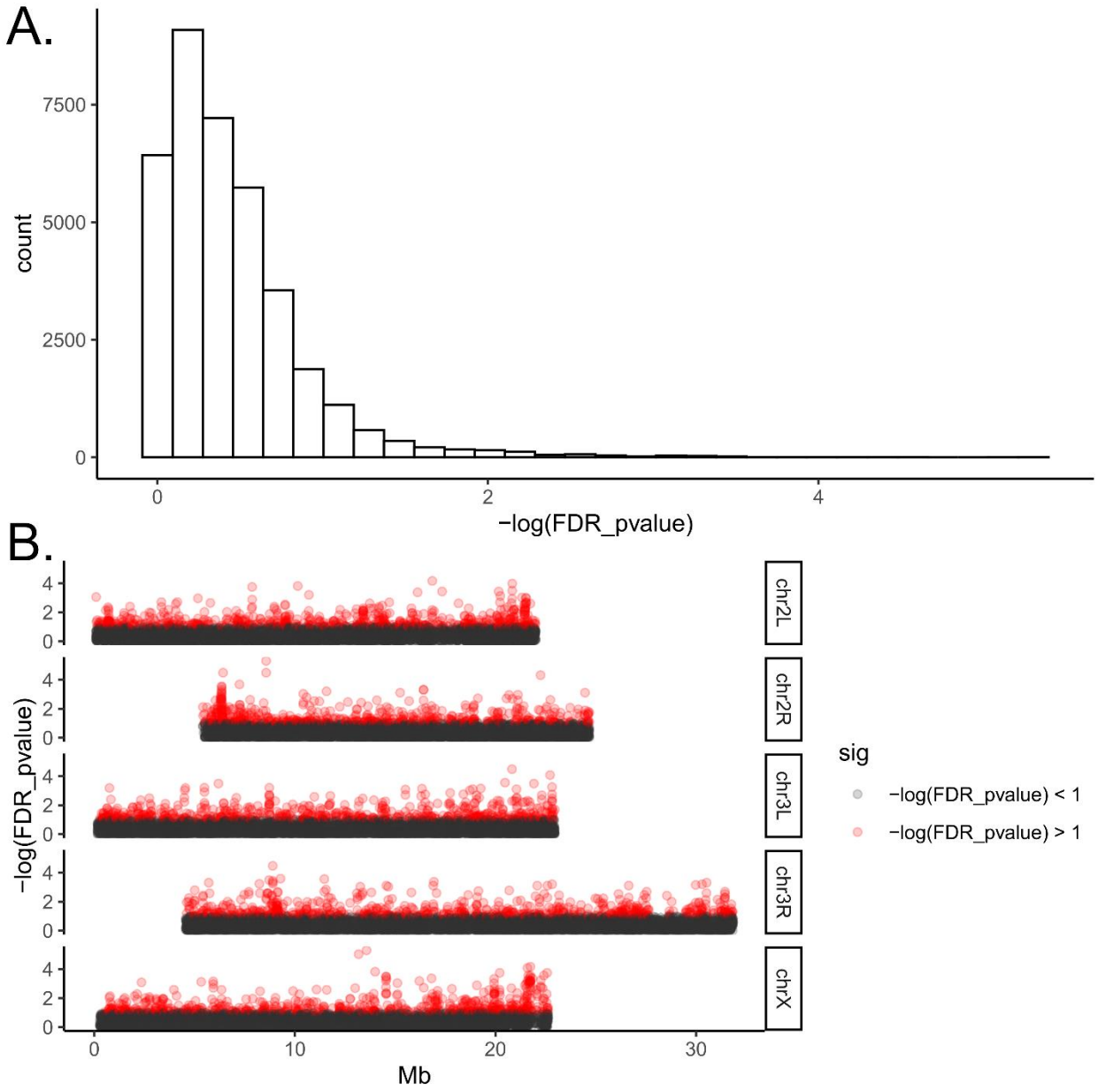


Fig 2.2. Distribution of FDR p-value for ANOVA statistical test of parental coverage.

(A) Histogram of FDR p-value. (B) Distribution of FDR p-value by chromosome locations with red color being FDR p-value < 0.1

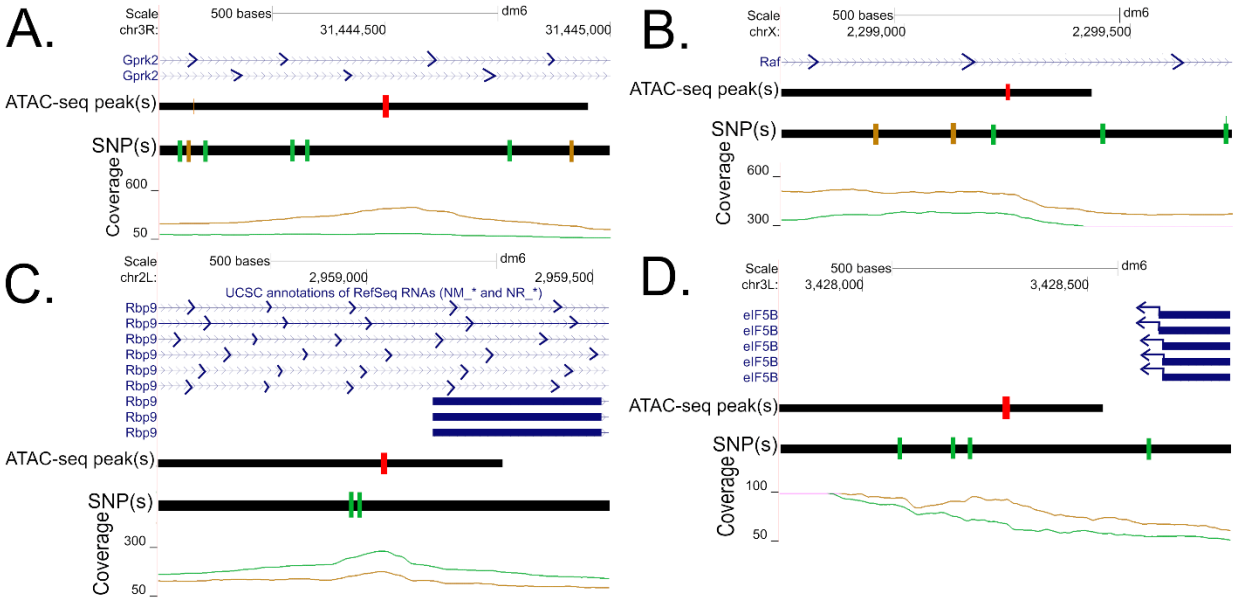


Fig 2.3. Illustrative examples of polymorphic chromatin configurations.
The images depict a region on intron of Gprk2 (A), on intron of Raf (B), on intron and upstream of the TSS of a Rbp9 isoform (C), and downstream of TTS of eIF5B (D). Tracks are gene, ATAC-seq peaks, and coverage tracks.

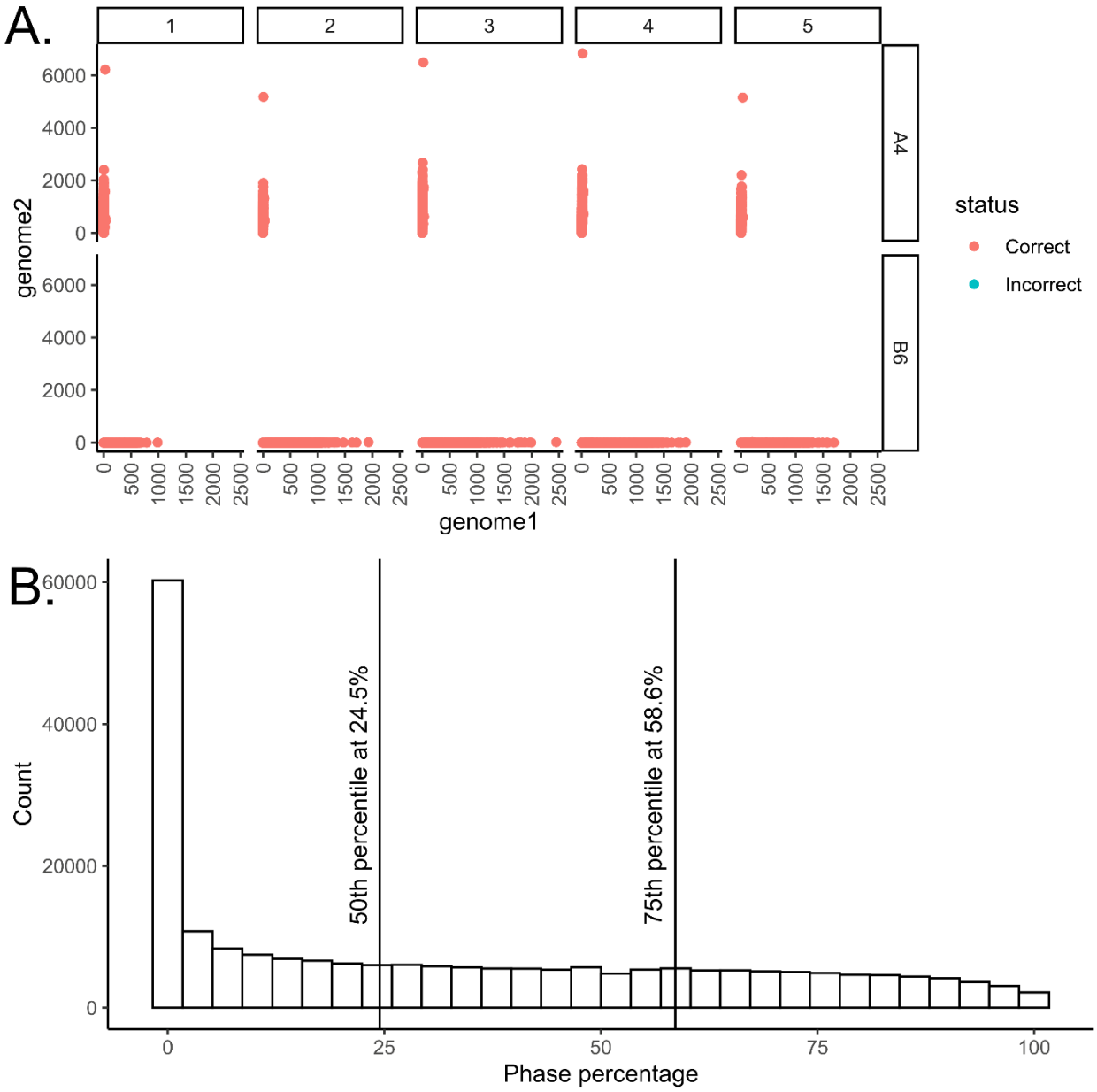


Fig 2.4. Quality control plots for phasing.

(A) Count of fragments phased into genome 1 and genome 2 for A4, B6 samples. (B) distribution of phase percentage from Hybrid sample.

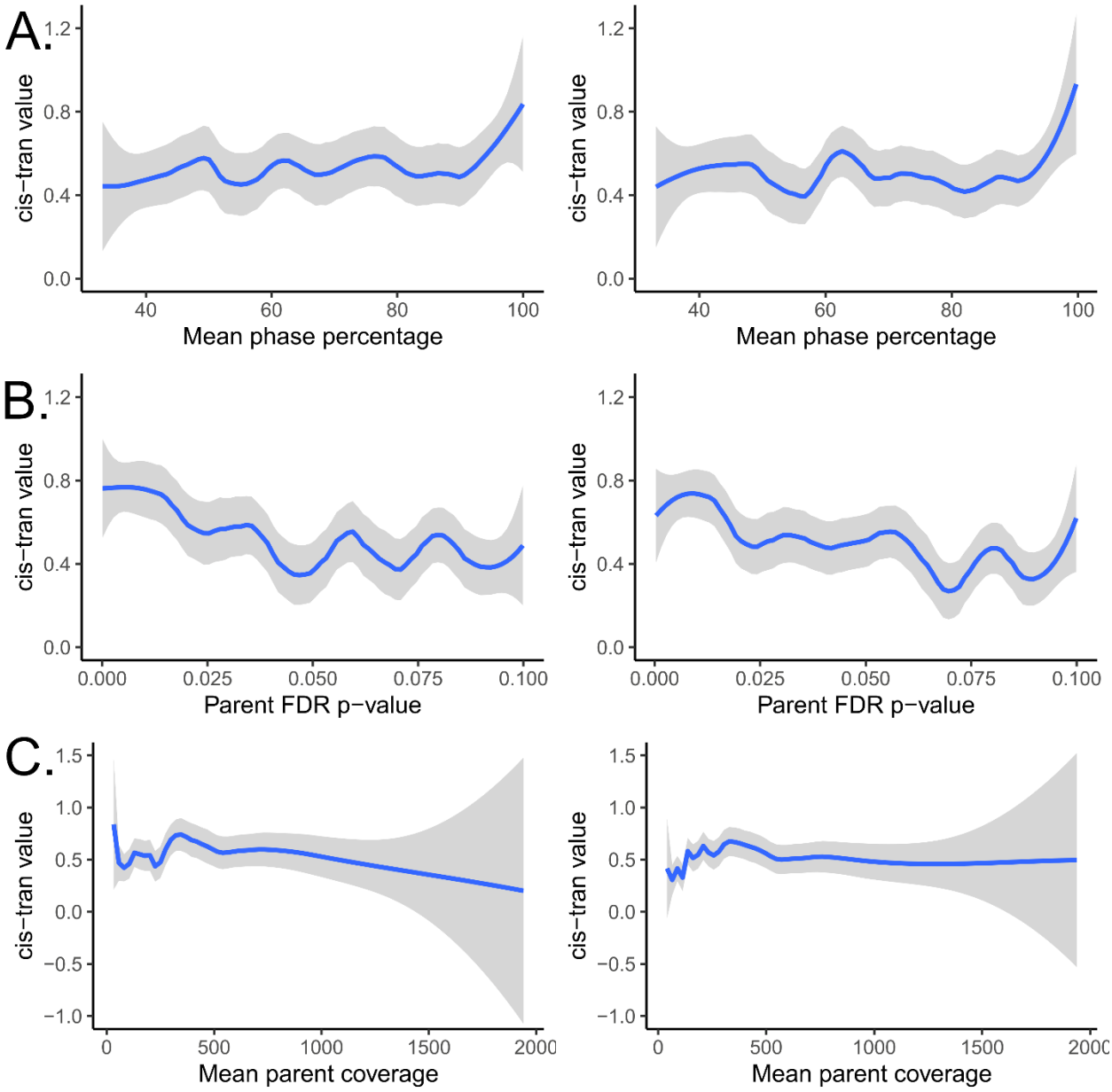


Fig 2.5. Cis-trans value quality control.
(A): cis-trans value against mean phase percentage before (left) and after (right) alignment QC. (B): cis-trans value against parent FDR p-value before (left) and after (right) alignment QC. (C): cis-trans value against mean parent coverage before (left) and after (right) alignment QC.

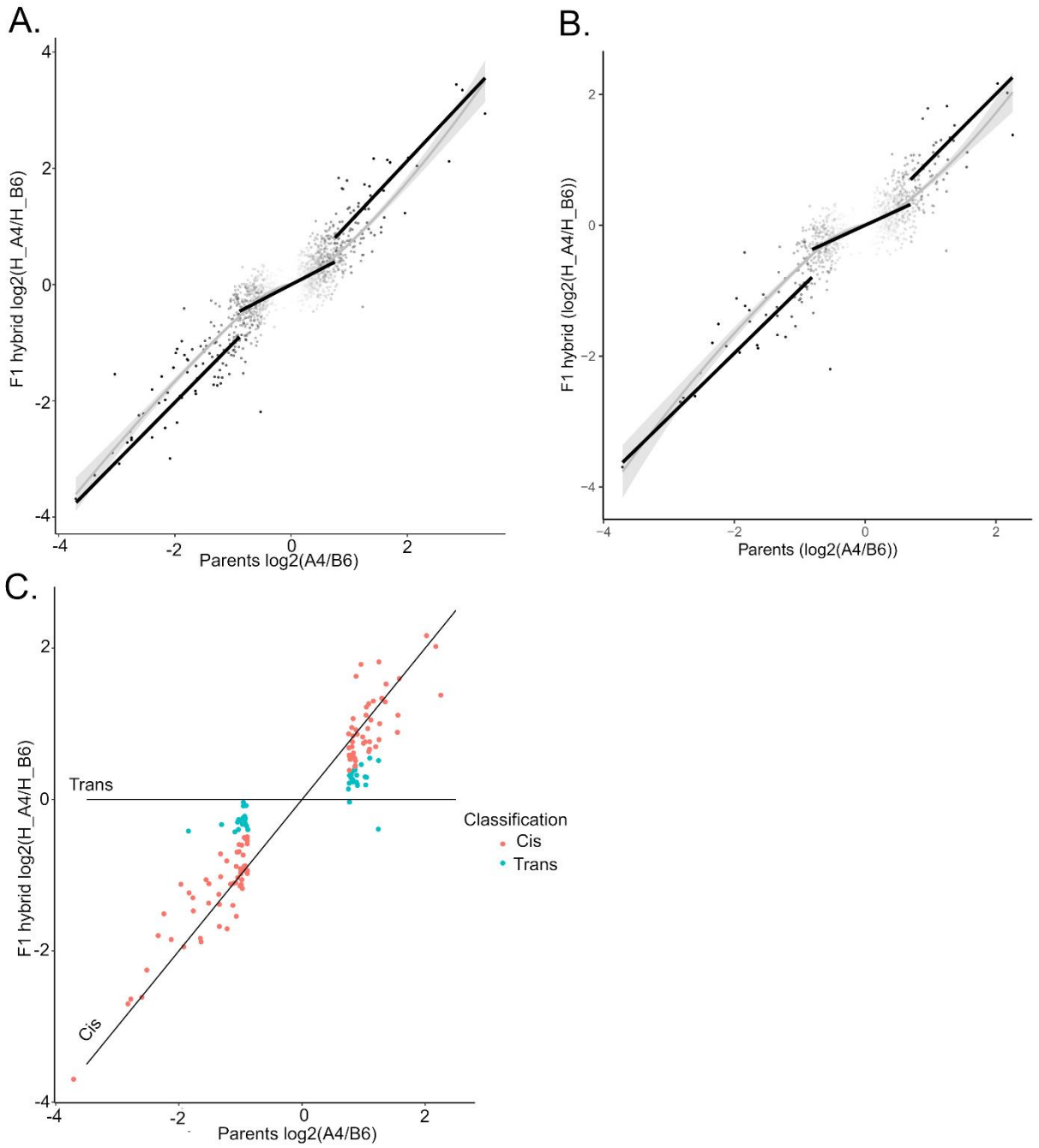


Fig 2.6. Parents log2(A4/B6) vs F1 log2(H_A4/H_B6).
(A): All 3006 ATAC-seq peaks that are significantly different between parental genotypes. The lines are the slopes of three fitted linear regression. (B): All classified ATAC-seq peaks colored by classification.

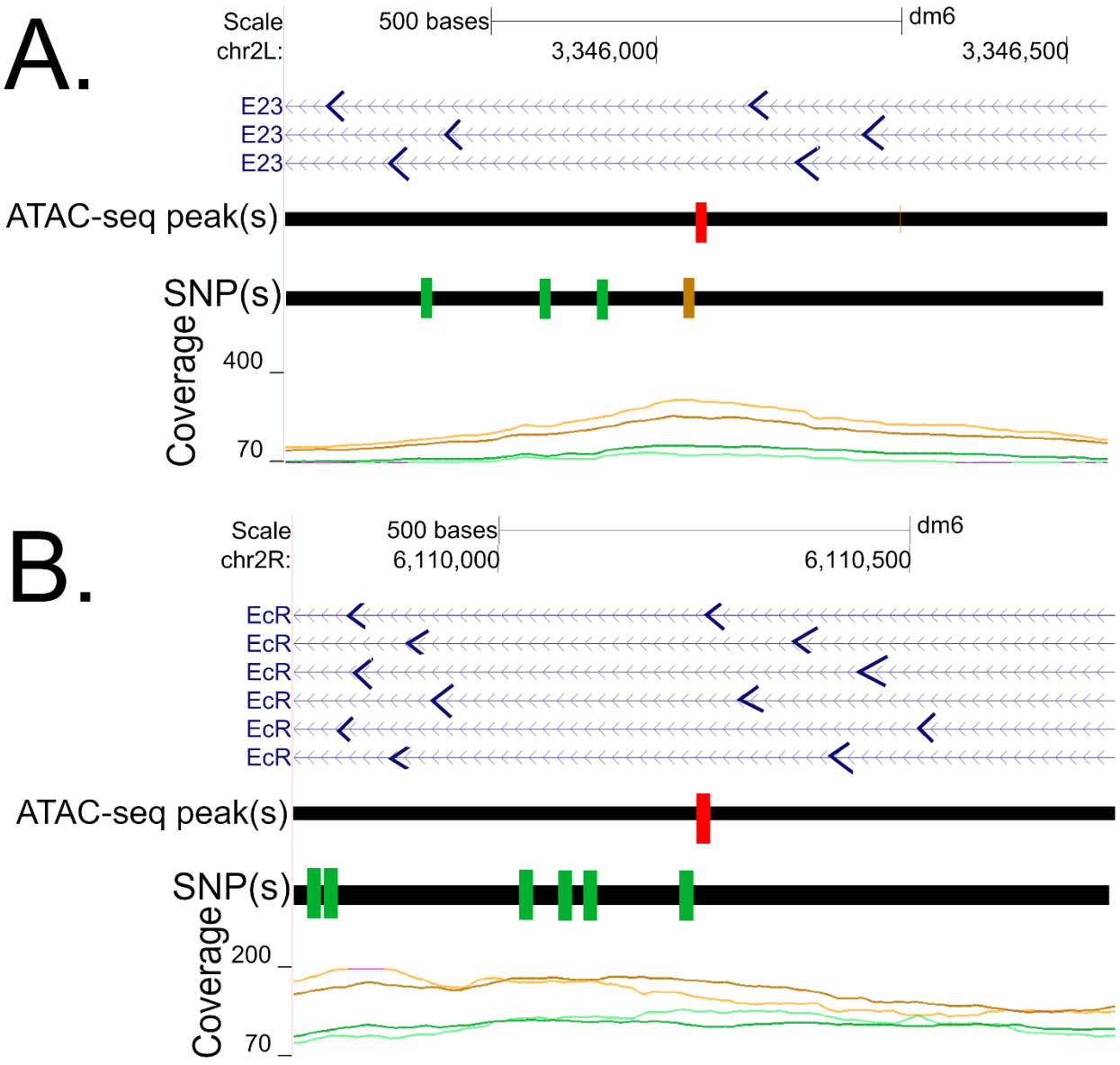
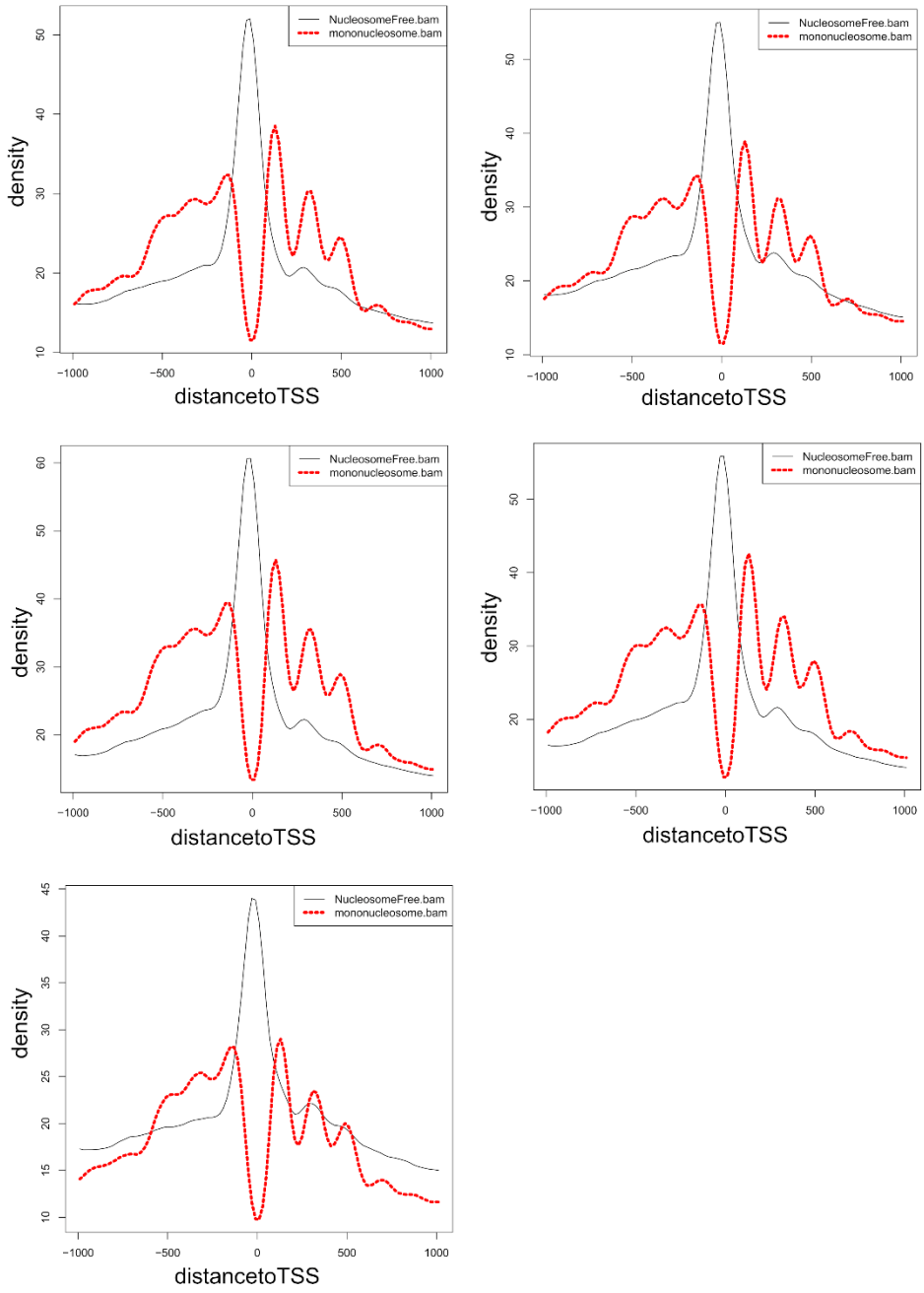
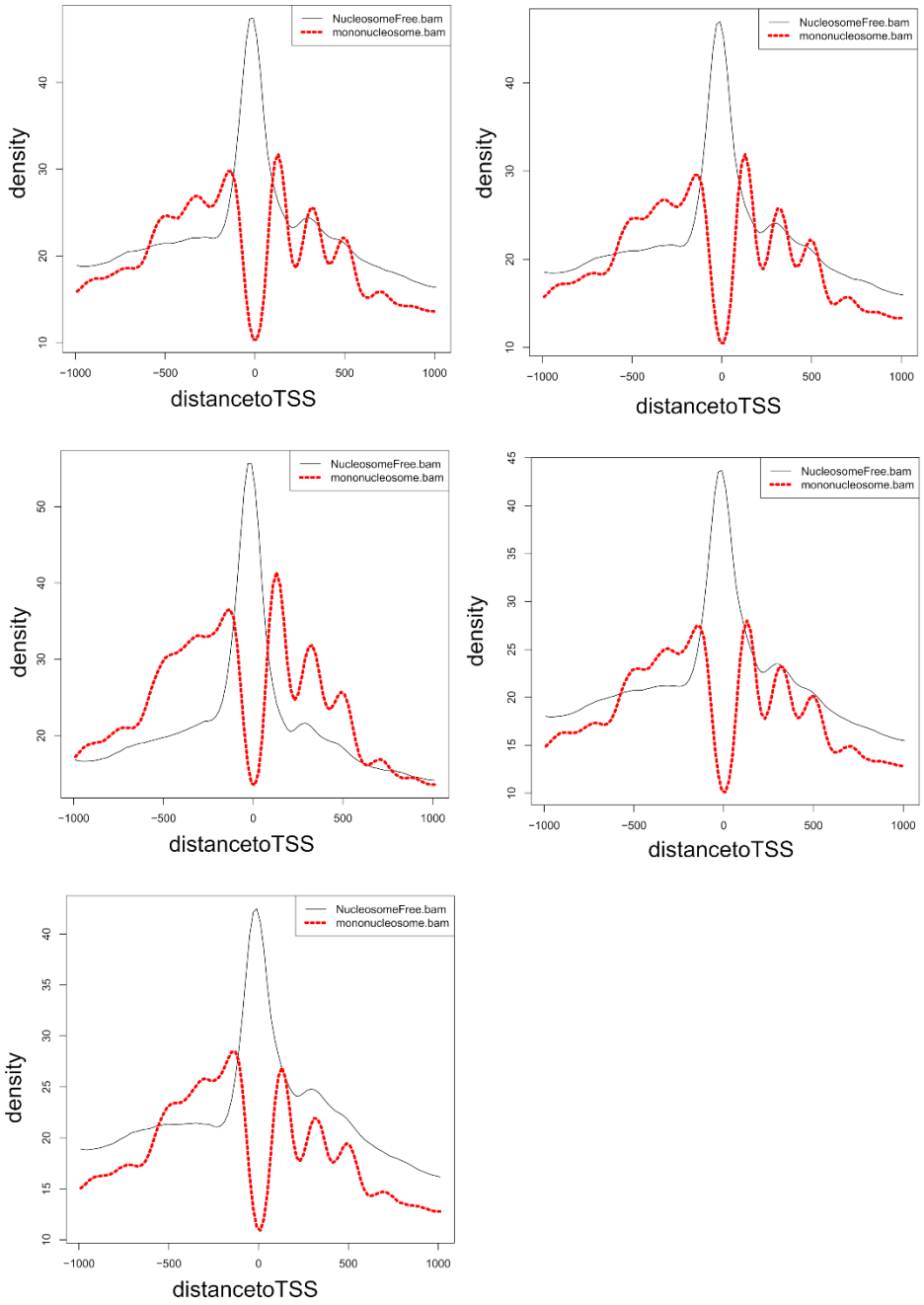


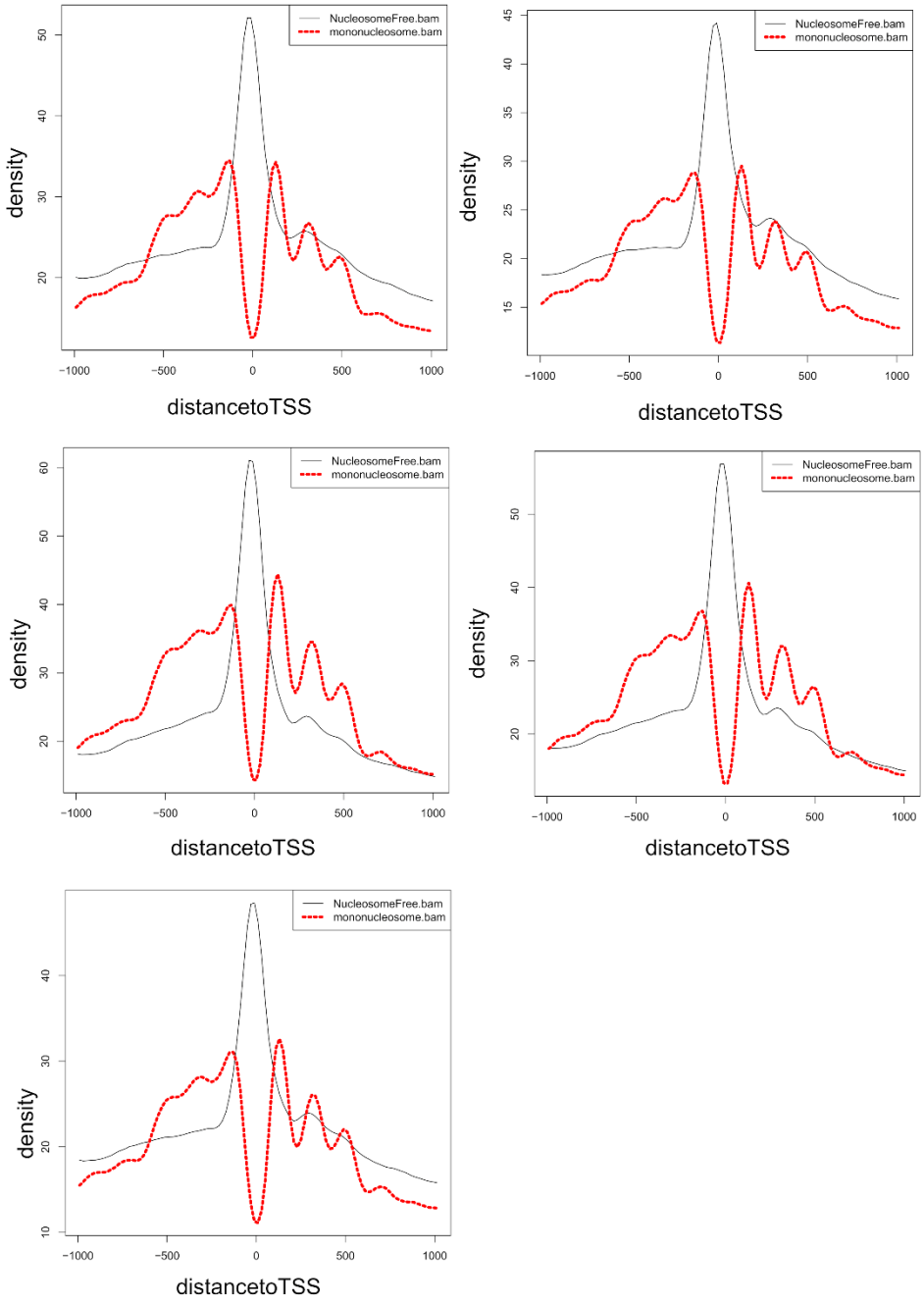
Fig 2.7. Illustrative examples of cis- and trans- chromatin configurations.
The images depict a region on intron of E23 (A), on intron of EcR (B). SNPs are colored by the genotypes that have alternate base at SNPs. Brown, light brown are B6, H_B6 respectively. Green, light green are A4, H_A4 respectively. Tracks are gene, ATAC-seq peaks, SNP location, and coverage tracks.



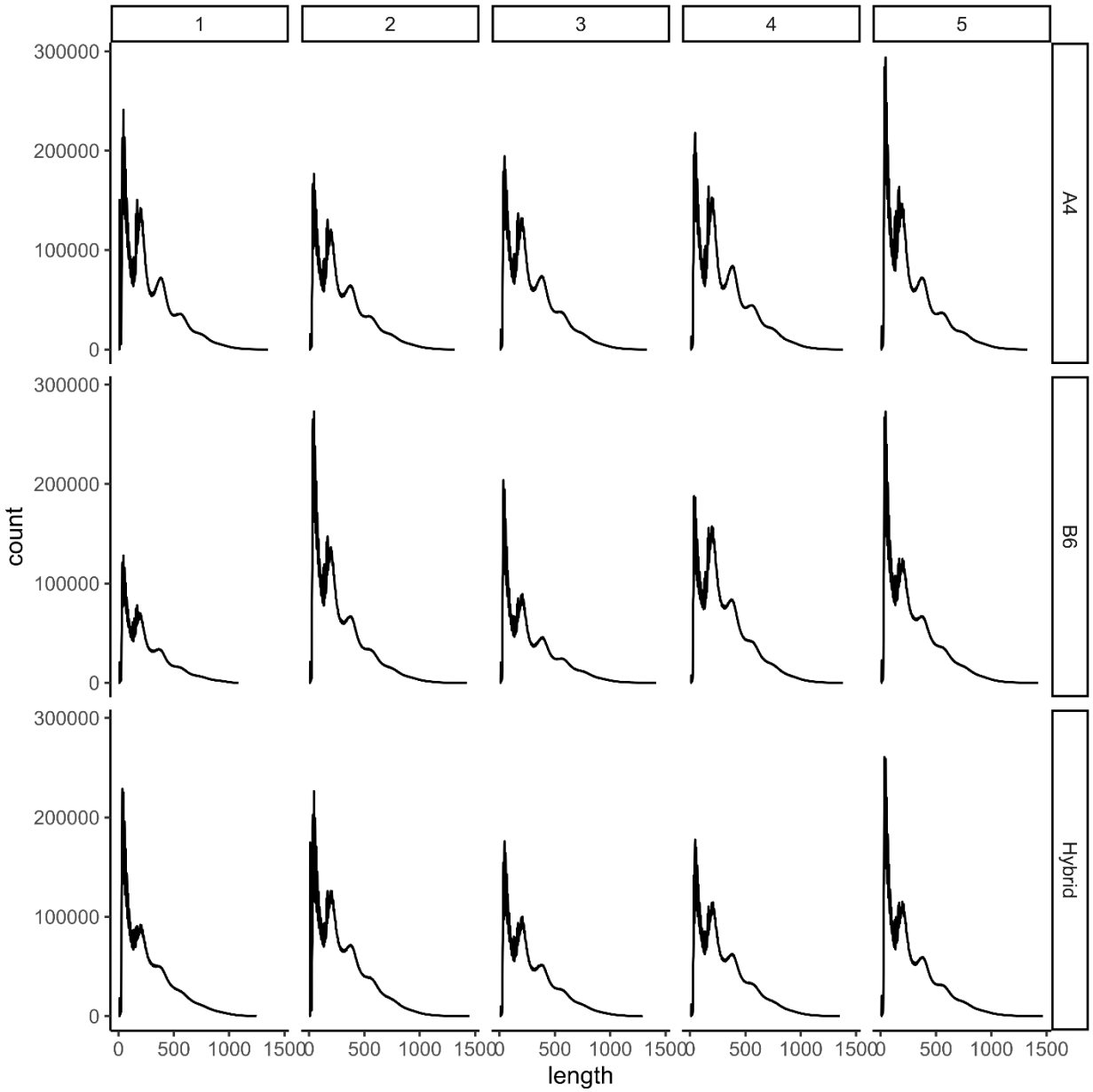
S2.1 Fig. ATACseqQC mononucleosome and nucleosomes-free read density against distance to TSS from A4 genotypes.



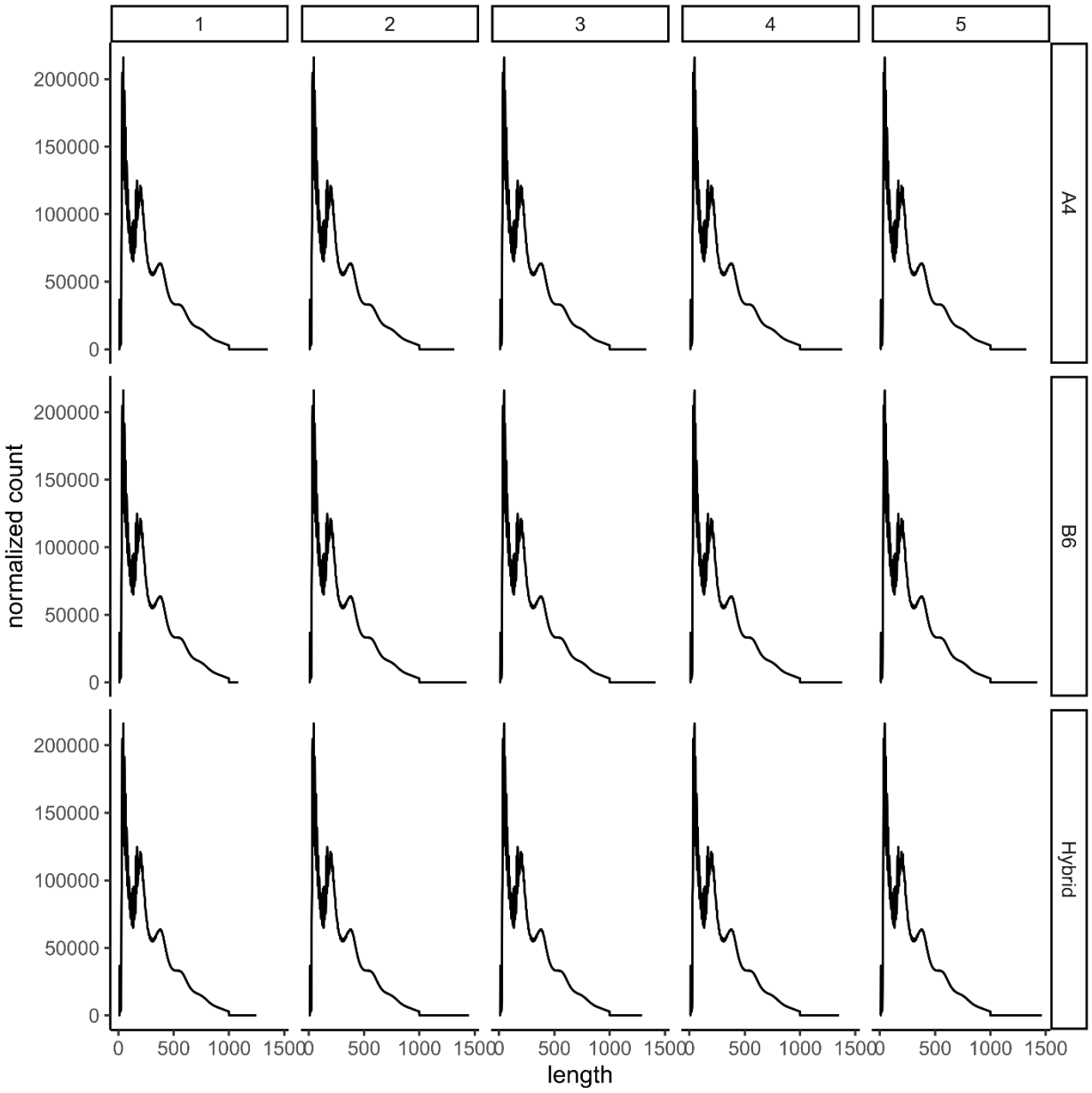
S2.2 Fig. ATACseqQC mononucleosome and nucleosomes-free read density against distance to TSS from B6 genotypes.



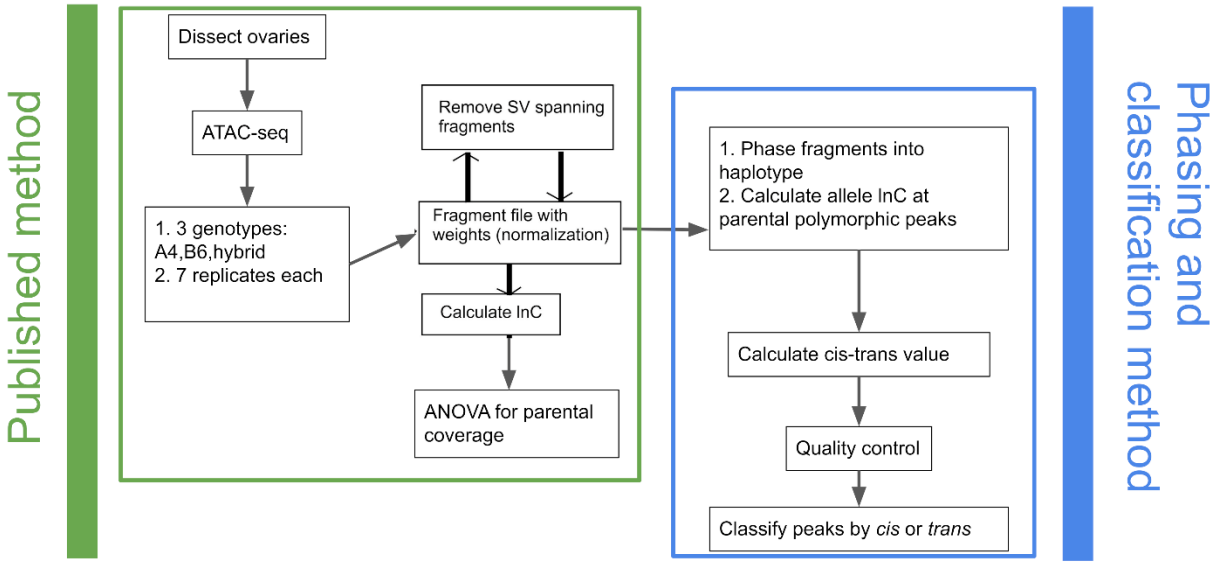
S2.3 Fig. ATACseqQC mononucleosome and nucleosomes-free read density against distance to TSS from Hybrid genotypes.



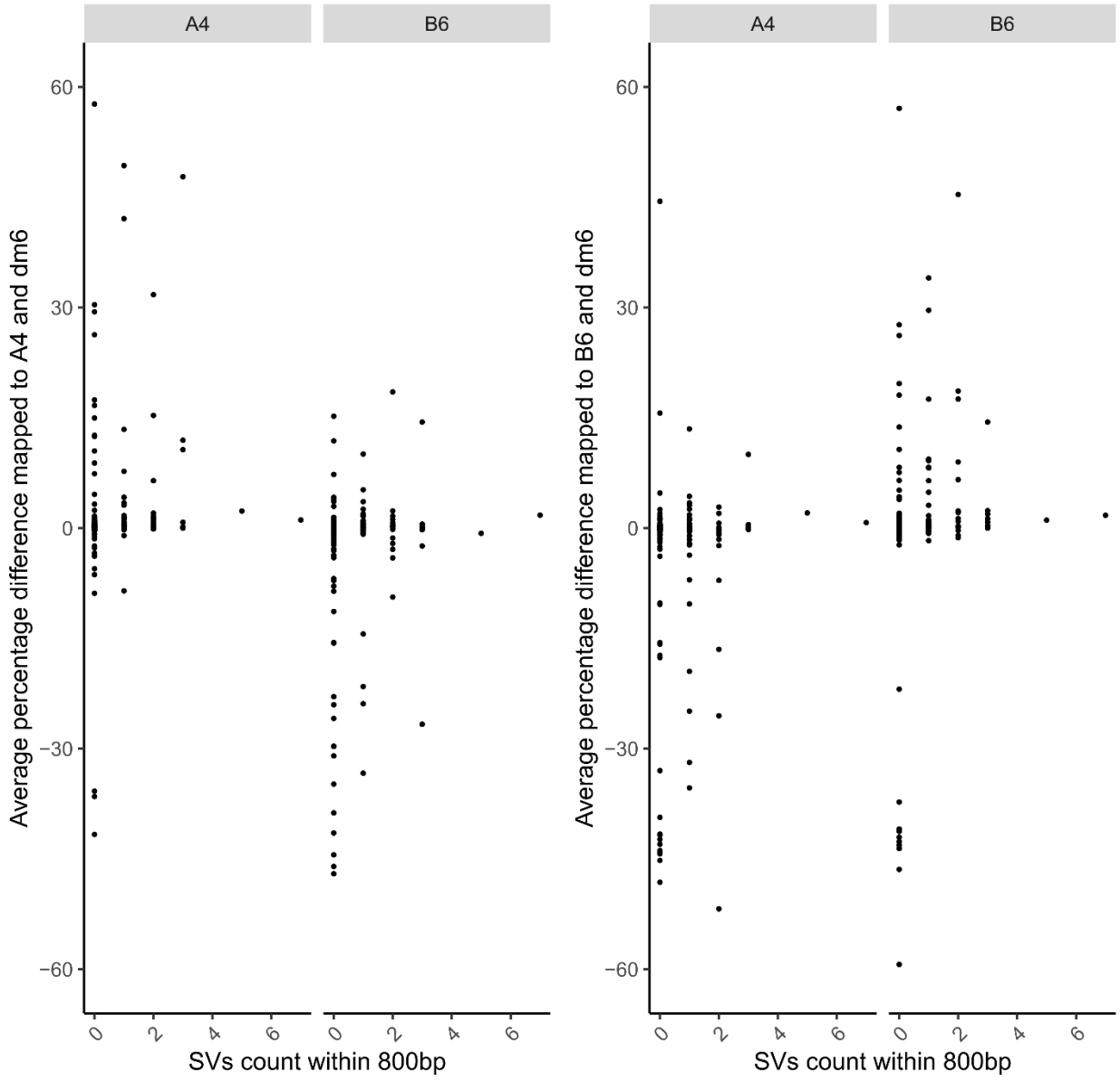
S2.4 Fig. Raw fragment distribution of ATAC-seq samples.
Replicates are shown by columns. Samples are from A4,B6, and Hybrid genotypes from top to bottom.



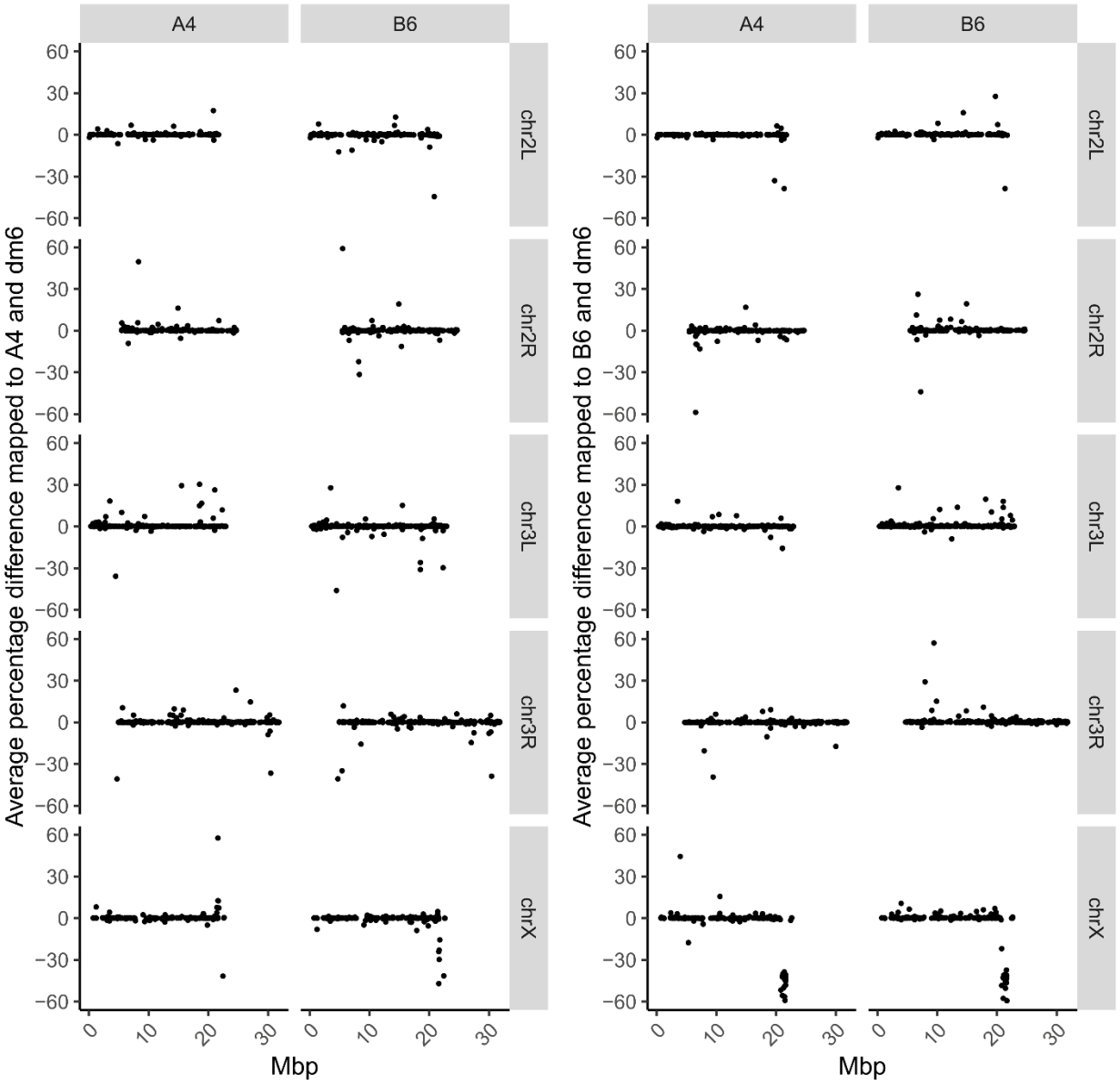
S2.5 Fig. Normalized fragment distribution of ATAC-seq samples.
Replicates are shown by columns. Samples are from A4,B6, and Hybrid genotypes from top to bottom.



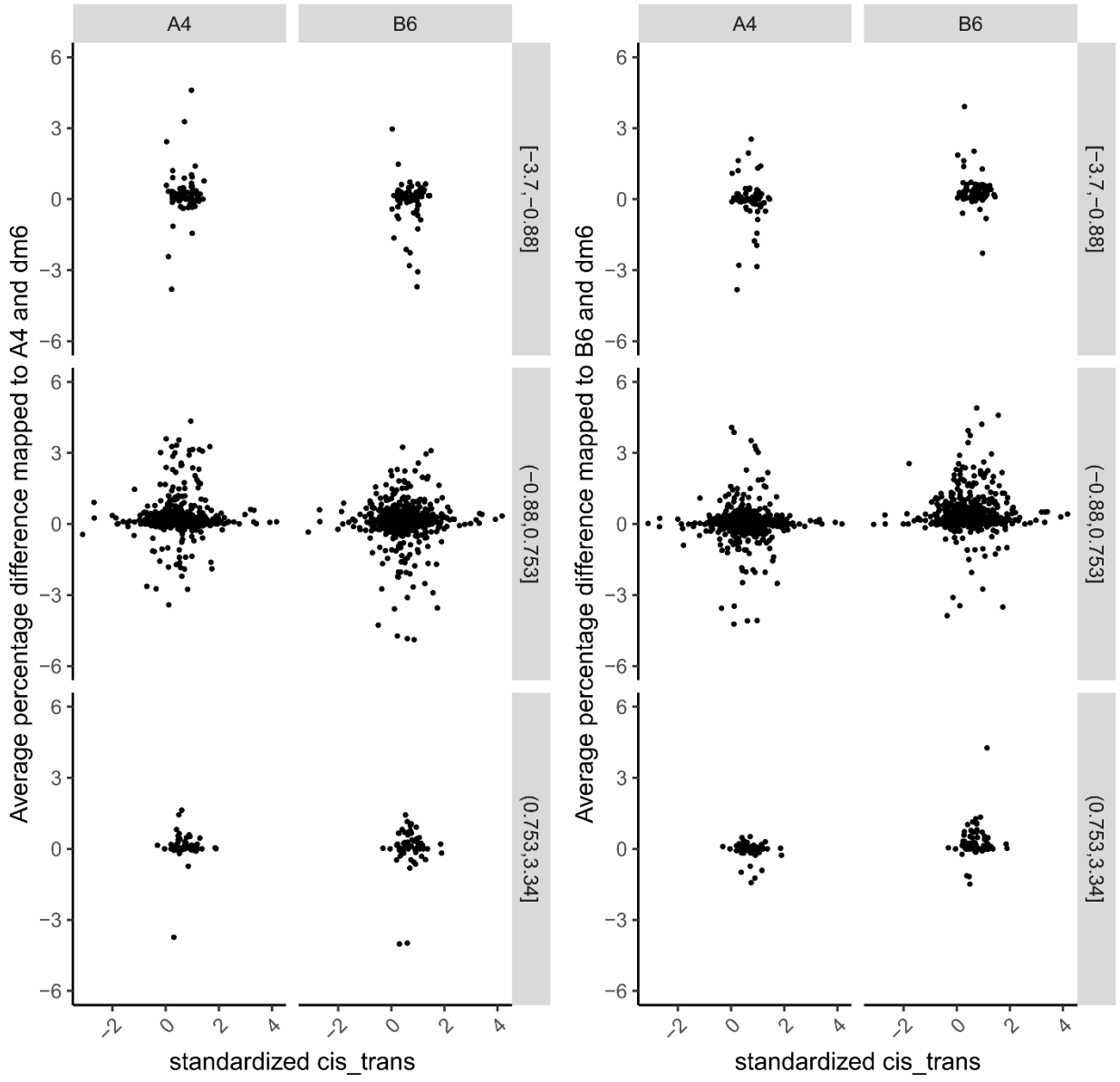
S2.6 Fig. ATAC-seq and phasing workflow



S2.7 Fig. Count of SVs distance to ATAC-seq peaks by average percentage difference in fragment count mapped to A4/B6 vs dm6 genomes.
(Left): comparison between alignment mapped to A4 and dm6. (Right): comparison between alignment mapped to B6 and dm6



S2.8 Fig. Genome distribution of average percentage difference in fragment count mapped to A4/B6 compared to dm6 genomes.
(Left): comparison between alignment mapped to A4 and dm6. (Right): comparison between alignment mapped to B6 and dm6



S2.9 Fig. Cis-trans value by average percentage difference in fragment count mapped to A4/B6 vs dm6 genomes.
(Left): comparison between alignment mapped to A4 and dm6. (Right): comparison between alignment mapped to B6 and dm6

2.7 TABLES

Table 2.1: SNP counts after filtering steps.

SNP genotypes that are not 0/1 or 1/0	Heterozygous between A4 and B6	Biallelic SNP with SNP phase percentage > 90%	Euchromatin biallelic SNP with SNP phase percentage > 90%
1565413	438470	372844	369353

Table 2.2: Peak count by cis-trans classification before and after quality control using three genome alignment comparison.

QC status	Cis	Trans	Total
Before	212 (15.16%)	70 (5.00%)	1398
After	106 (10.98%)	45 (4.66%)	965

S2.1 Table: Details of all strains examined in this study.
The A4,B6 parental strains are P-element and Wolbachia free, were brother and sister mated for at least 18 generations, and are highly isogenic ([49]). Each parental strain is associated with a reference quality de novo genome assembly ([46]). The Hybrid strain is the heterozygous offspring of the A4 and B6 strains

Name	Stock Number[1]	Full Name[2]	Collection details
A4	b.3852	KSA 2	Koriba Dam, Zimbabwe, 1963
B6	t.14021-0231.1	-	Ica, Peru, 1956
Hybrid	-	-	Heterozygous offspring of A4 and B6

S2 Table: Feature type annotation for polymorphic peaks.

Feature type	Polymorphic peak count (percentage of total count)
Total	3006 (100%)
TSS	764 (25.41%)
TTS	296 (9.85%)
3' UTR	28 (0.93%)
5' UTR	46 (1.53%)
intron	1188 (39.52%)
exon	94 (3.13%)
intergenic	558 (18.56%)
non-coding	32 (1.06%)

Chapter 3

Dissertation conclusion

Historically, complex traits were considered to be monogenic phenotypes contributed by large effect causal genes despite the biometricians' argument that such contribution couldn't explain the continuous variation observed in many phenotypic traits [1]. This argument was only resolved with the discovery of the "infinitesimal model" published by Fisher in 1918 [2]. However, the number of causal genes per traits, and their effect size remains unclear until now [1]. As a result, great effort from the complex trait community has been made to identify these statistics with GWAS, and QTL-mapping. Furthermore, causal genes have been found to directly contribute to complex traits variation through changing the proteins [3], non-coding genetic variants can also participate in driving the variations of these traits as regulatory elements [4–6]. Thus, DNase-I HS [7] and ATAC-seq [8] have been used to great success in identifying these regulatory elements by characterizing open chromatin state in large panels of genotypes [9,10].

However, there are three shortcomings in the field of ATAC-seq as discussed in this work. The first shortcoming is the usage of primarily embryo samples. This would limit the utility of any identified genome-wide chromatin state landscape since they would not be applicable to adult tissues. The second shortcoming is the lack of reference quality genome sequence for the genotypes used in ATAC-seq studies. This shortcoming would limit the utility of those ATAC-seq studies because we can't address

the possible mis-alignment caused by the effect of SVs on read coverage which has been well-documented leading to incorrect inference of coverage [11], and significant contributions to complex trait variation by hidden variants, such as SNPs or SVs [12,13]. The last shortcoming is the lack of haplotype phasing. Without haplotype phasing, it is impossible to associate identified variants in open chromatin states to alleles resulting in extreme difficulty in characterizing *cis*-acting or *trans*-acting nature of open chromatin regions. Therefore, the first chapter of this work aims at addressing first two shortcomings by characterizing genome-wide chromatin state landscape using ATAC-seq samples from eight different genotypes and four different tissues (adult brain, adult ovary, and embryo imaginary wing disc, and eye disc). Then, the second chapter aims at addressing the third shortcoming by performing haplotype phasing using the identified ATAC-seq peaks in chapter one, and ATAC-seq samples from two different parental genotypes (A4, and B6), and their F1 offspring (hybrid).

In chapter one, I have performed ATAC-seq on eight genotypes from DSPR which all have complete *de novo* reference quality genome. Thus, I can address the lack of adult tissue studies and the lack of reference quality genome sequences in the ATAC-seq field. With complete *de novo* reference quality genome, correction for SV effect on coverage can be corrected. The SV effect can be seen as B6 genotype (brown) coverage is significantly affected in figure 1.4. After correcting for SVs effect on coverage, 9.5%, 50.8%, and 34.3% peaks are identified to be false positive in inference of differences in coverage between tissue, genotype, and genotype:tissue interaction respectively. This further reinforces the need of reference quality genome sequences for all genotypes used for ATAC-seq studies. Without such sequences, it is near impossible

to correct for SVs which lead to incorrect inference of polymorphism in any identified open chromatin regions.

As the goal of the first chapter is to characterize a genome-wide chromatin state landscape, we have identified a total of 44099 ATAC-seq peaks using mostly standard ATAC-seq pipeline. Out of these, 30383, 1050, and 4508 peaks are identified to be polymorphic for coverage by tissue, by genotype, and by genotype:tissue respectively following rigorous statistical testing. The polymorphic peaks by tissue, and by genotype will absolutely be of use for complex trait communities. Furthermore, the polymorphic peaks by genotype:tissue interaction are even of greater interest since they are regions that are regulated in a genotype dependent manner for some tissues. These likely represent QTL that are tissue or complex trait specific with less pleiotropy. Following the identification of polymorphic ATACseq peaks in coverage, we also test all SNPs and SVs within 250 bp or within 800 bp of each significant peak by genotype and by genotype:tissue interaction. 597 SNPs and 55 SVs were identified to explain 100% of the variation. However, as these SNPs/SVs are collected from only 8 genotypes, a much larger set of genotypes would be necessary for association studies to avoid over-fitting. In conclusion, the identified ATAC-seq peaks in chapter 1 would have great utility for the complex trait community since we have performed SV correction to eliminate false positive inference of polymorphic open chromatin states, and have included data from multiple genotypes and tissues.

After the first chapter, my next goal is to address the lack of haplotype phasing in the field by attempting to perform haplotype phasing and to classify the identified polymorphic ATAC-seq peak in first chapter by their *cis*- or *trans*- acting natures. In

chapter two, I carried out an ATAC-seq experiment on adult ovary samples collected from two DSPR founder genotypes A4, B6, and their F1 offsprings [14]. The first step in chapter two is to identify ATAC-seq peaks that are different between the two parental genotypes (A4,B6) among the 44099 peaks from chapter one.

Since both A4 and B6 have known SNP lists, I was able to perform haplotype phasing. However, the first step was to filter SNPs to only keep the best SNPs that are heterozygous between the two parental genotypes. After correcting for SVs and ANOVA tests, I identified 3006 ATAC-seq peaks that are significantly different between two parental genotypes with FDR p-value < 0.1. Similar to the values of identified polymorphic ATAC-seq peaks in chapter one, these peaks would be interesting targets for future complex trait studies, albeit with less utility since only two genotypes and one tissue were used.

Then, we performed haplotype phasing in order to address the third shortcoming of the field. The first step is to filter SNPs for A4 and B6 genotypes in order to select only the best possible SNPs that are heterozygous for the two parental genotypes. Only 369353 SNPs remain to ensure robust and accurate haplotype phasing. After SNPs selection, haplotype phasing was carried out. Phasing percentage was also calculated. Only 1398 peaks remain out of 3006 peaks due to FDR p-value cutoff of < 0.1, and phasing percentage cutoff of > 33%. Then, we calculated cis-trans values for *cis*- and *trans*- classification. However, Loess smoothing curves for the plots of cis-trans value as a function of mean phase percentage (Figure 2.5A), parent FDR p-value (Figure 2.5B), and mean parent coverage (Figure 2.5C) reveal potential association between them and the cis-trans value. Thus, we applied an additional filter step to ensure that

only peaks correctly aligned are chosen. After aligning our A4, and B6 data to dm6, A4, and B6 genomes, only 965 peaks remain after removing those with SVs within 800 bp and percentage of correct alignment between all three genomes being > 95%. Furthermore, three separate linear regression lines applied to each parental coverage difference quantile bin suggests that we would have the best ability to detect coverage ratio change between parents and phased hybrid genotypes, and to classify *cis*- or *trans*- for only peaks with $\log_2(A4/B6) > 0.753$ and < -0.88. We have identified 106, and 45 ATAC-seq to be *cis*- and *trans*- acting respectively. However, despite best effort, figure 2.7B has indicated that there is a missing quality control step or a hidden error in my method. This can be seen with the completely different possible nature of cis- and trans of this ATAC-seq peak if its location is moved 100 bp upstream. Once this final hurdle is cleared, these cis-trans classified ATAC-seq peaks would no doubt be interesting targets for future functional studies using the Crispr/Cas9/allele swapping methods [15–19] to further understand gene regulations-especially for important genes associated with these *cis*- and *trans*-acting polymorphic open chromatin regions.

Reference

I. Abstract and Introduction reference:

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753.

2. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16: 197–212.

3. Floc'hlay S, Wong ES, Zhao B, Viales RR, Thomas-Chollier M, Thieffry D, et al. Cis-acting variation is common across regulatory layers but is often buffered during embryonic development. Genome Res. 2021;31: 211.

4. Hill MS, Vande Zande P, Wittkopp PJ. Molecular and evolutionary processes generating variation in gene expression. Nat Rev Genet. 2020;22: 203–215.

5. Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco-Ferreres R, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature. 2018;555: 538–542.

6. Davie K, Janssens J, Koldere D, De Waegeneer M, Pech U, Kreft Ł, et al. A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. Cell. 2018;174: 982.

7. Falconer DS, Mackay TFC. Introduction to quantitative genetics, Longman. Essex, England. 1996.

8. Lynch M, Walsh B, Others. Genetics and analysis of quantitative traits. 1998. Available: http://www.invemar.org.co/redcostera1/invemar/docs/RinconLiterario/2011/febrero/AG_8.pdf

9. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33: D514–7.

10. Genetic Alliance, District of Columbia Department of Health. Diagnosis of a Genetic Disease. Genetic Alliance; 2010.

11. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42: D1001–6.

12. Abbate R, Sticchi E, Fatini C. Genetics of cardiovascular disease. Clin Cases Miner Bone Metab. 2008;5: 63.

13. Gelernter J. Genetics of complex traits in psychiatry. Biol Psychiatry. 2015;77: 36.

14. Zaitlen N, Kraft P, Patterson N, Pasaniuc B, Bhatia G, Pollack S, et al. Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. PLoS Genet. 2013;9. doi:10.1371/journal.pgen.1003520

15. Kadri NK, Guldbrandtsen B, Lund MS, Sahana G. Genetic dissection of milk yield traits and mastitis resistance quantitative trait loci on chromosome 20 in dairy cattle. J Dairy Sci. 2015;98: 9015–9025.

16. Su J, Xu K, Li Z, Hu Y, Hu Z, Zheng X, et al. Genome-wide association study and Mendelian randomization analysis provide insights for improving rice yield potential. Sci Rep. 2021;11: 6894.

17. Shadan S. Genomics: The long and the short of it. Nature. 2010;467: 539.

18. Goddard ME, Kemper KE, MacLeod IM, Chamberlain AJ, Hayes BJ. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. Proceedings of the Royal Society B: Biological Sciences. 2016 [cited 31 Mar 2022]. doi:10.1098/rspb.2016.0569

19. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169: 1177–1186.

20. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh. 1919. pp. 399–433. doi:10.1017/s0080456800012163

21. Barton NH, Etheridge AM, Véber A. The infinitesimal model. bioRxiv. 2016. p. 039768. doi:10.1101/039768

22. Barton NH. What role does natural selection play in speciation? Philos Trans R Soc Lond B Biol Sci. 2010;365: 1825–1840.

23. Weissman DB, Barton NH. Limits to the rate of adaptive substitution in sexual populations. PLoS Genet. 2012;8: e1002740.

24. Barton NH, Etheridge AM. The relation between reproductive value and genetic contribution. Genetics. 2011;188: 953–973.

25. Mackay TFC. Quantitative trait loci in Drosophila. Nat Rev Genet. 2001;2: 11–20.

26. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20: 467–484.

27. Long AD, Grote MN, Langley CH. Genetic analysis of complex diseases. Science. 1997. p. 1328; author reply 1329–30.

28. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273: 1516–1517.

29. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet. 2007;39: 857–864.

30. Hakonarson H, Grant SFA, Bradfield JP, Marchand L, Kim CE, Glessner JT, et al. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature. 2007;448: 591–594.

31. Leahy JL. A genome-wide association study identifies novel risk loci for type 2 diabetes. Yearbook of Endocrinology. 2008. pp. 36–37. doi:10.1016/s0084-3741(08)79221-7

32. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science. 2007;316: 1336–1341.

33. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science. 2007;316: 1341–1345.

34. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PIW, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. 2007;316: 1331–1336.

35. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, et al. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet. 2007;39: 770–775.

36. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet. 2007;39: 631–637.

37. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet. 2007;39: 645–649.

38. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. 2007;447: 1087–1093.

39. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet. 2007;39: 870–874.

40. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447: 661–678.

41. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187: 367–383.

42. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42: 565–569.

43. Shi H, Kichaev G, Pasaniuc B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am J Hum Genet. 2016;99: 139–153.

44. Bloom JS, Kotenko I, Sadhu MJ, Treusch S, Albert FW, Kruglyak L. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nature Communications. 2015. doi:10.1038/ncomms9712

45. Bloom JS, Boocock J, Treusch S, Sadhu MJ, Day L, Oates-Barker H, et al. Rare variants contribute disproportionately to quantitative trait variation in yeast. Elife. 2019;8. doi:10.7554/eLife.49212

46. Chesler EJ. Out of the bottleneck: the Diversity Outcross and Collaborative Cross mouse populations in behavioral genetics research. Mamm Genome. 2014;25: 3–11.

47. Hook M, Roy S, Williams EG, Bou Sleiman M, Mozhui K, Nelson JF, et al. Genetic cartography of longevity in humans and mice: Current landscape and horizons. Biochim Biophys Acta Mol Basis Dis. 2018;1864: 2718–2732.

48. Saul MC, Philip VM, Reinholdt LG, Center for Systems Neurogenetics of Addiction, Chesler EJ. High-Diversity Mouse Populations for Complex Traits. Trends Genet. 2019;35: 501–514.

49. Long AD, Macdonald SJ, King EG. Dissecting complex traits using the Drosophila Synthetic Population Resource. Trends Genet. 2014;30: 488–495.

50. Mackay T. Trudy Mackay. Current Biology. 2006. pp. R659–R661. doi:10.1016/j.cub.2006.08.016

51. Nelson RM, Pettersson ME, Carlborg Ö. A century after Fisher: time for a new paradigm in quantitative genetics. Trends Genet. 2013;29: 669–676.

52. Rönnegård L, Valdar W. Detecting major genetic loci controlling phenotypic variability in experimental crosses. Genetics. 2011;188: 435–447.

53. Forsberg SKG, Andreatta ME, Huang X-Y, Danku J, Salt DE, Carlborg Ö. The Multi-allelic Genetic Architecture of a Variance-Heterogeneity Locus for Molybdenum Concentration in Leaves Acts as a Source of Unexplained Additive Genetic Variance. PLoS Genet. 2015;11: e1005648.

54. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9: 29.

55. Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature. 1988;335: 721–726.

56. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. 1989;121: 185–199.

57. Chitre AS, Polesskaya O, Munro D, Cheng R, Mohammadi P, Holl K, et al. Exponential increase in QTL detection with increased sample size. Genetics. 2023. doi:10.1093/genetics/iyad054

58. Price AH. Believe it or not, QTLs are accurate! Trends Plant Sci. 2006;11: 213–216.

59. Smith R, Sheppard K, DiPetrillo K, Churchill G. Quantitative trait locus analysis using J/qtl. Methods Mol Biol. 2009;573: 175–188.

60. Grisel JE, Crabbe JC. Quantitative Trait Loci Mapping. Alcohol Health Res World. 1995;19: 220–227.

61. Broman KW. The genomes of recombinant inbred lines. Genetics. 2005;169: 1133–1146.

62. King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S, Broman KW, et al. Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome Res. 2012;22: 1558–1566.

63. King EG, Macdonald SJ, Long AD. Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. Genetics. 2012;191: 935–949.

64. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33 Suppl: 228–237.

65. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014;94: 559–573.

66. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016;352: 600–604.

67. Torres JM, Gamazon ER, Parra EJ, Below JE, Valladares-Salgado A, Wacher N, et al. Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. Am J Hum Genet. 2014;95: 521–534.

68. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518: 337–343.

69. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6: e1000888.

70. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337: 1190–1195.

71. Pai AA, Pritchard JK, Gilad Y. The Genetic and Mechanistic Basis for Variation in Gene Regulation. PLoS Genet. 2015;11. doi:10.1371/journal.pgen.1004857

72. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10: 1213–1218.

73. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet. 2016;48: 206–213.

74. Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nature Genetics. 2018. pp. 1140–1150. doi:10.1038/s41588-018-0156-2

75. Richmond TJ, Davey CA. The structure of DNA in the nucleosome core. Nature. 2003;423: 145–150.

76. Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, et al. A genomic code for nucleosome positioning. Nature. 2006;442: 772–778.

77. Jin J, Bai L, Johnson DS, Fulbright RM, Kireeva ML, Kashlev M, et al. Synergistic action of RNA polymerases in overcoming the nucleosomal barrier. Nat Struct Mol Biol. 2010;17: 745–752.

78. Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat Genet. 2012;44: 743–750.

79. Teves SS, Weber CM, Henikoff S. Transcribing through the nucleosome. Trends Biochem Sci. 2014;39: 577–586.

80. Hartzog GA. Transcription elongation by RNA polymerase II. Curr Opin Genet Dev. 2003;13: 119–126.

81. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489: 75–82.

82. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012;489: 83–90.

83. Jenuwein T, Allis CD. Translating the histone code. Science. 2001;293: 1074–1080.

84. Satchwell SC, Drew HR, Travers AA. Sequence periodicities in chicken nucleosome core DNA. J Mol Biol. 1986;191: 659–675.

85. Widom J. Role of DNA sequence in nucleosome stability and dynamics. Q Rev Biophys. 2001;34: 269–324.

86. Trifonov EN. Sequence-dependent deformational anisotropy of chromatin DNA. Nucleic Acids Res. 1980;8: 4041–4053.

87. Sekinger EA, Moqtaderi Z, Struhl K. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. Mol Cell. 2005;18: 735–748.

88. Anderson JD, Widom J. Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. Mol Cell Biol. 2001;21: 3830–3839.

89. Thåström A, Lowary PT, Widlund HR, Cao H, Kubista M, Widom J. Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. J Mol Biol. 1999;288: 213–229.

90. Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, et al. Global analysis of protein expression in yeast. Nature. 2003;425: 737–741.

91. Cairns BR. Chromatin remodeling complexes: strength in diversity, precision through specialization. Curr Opin Genet Dev. 2005;15: 185–190.

92. Bravo González-Blas C, Quan X-J, Duran-Romaña R, Taskiran II, Koldere D, Davie K, et al. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. Mol Syst Biol. 2020;16: e9438.

93. Jacobs J, Atkins M, Davie K, Imrichova H, Romanelli L, Christiaens V, et al. The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. Nat Genet. 2018;50: 1011–1020.

94. Davie K, Jacobs J, Atkins M, Potier D, Christiaens V, Halder G, et al. Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. PLoS Genet. 2015;11: e1004994.

95. Salces-Ortiz J, Vargas-Chavez C, Guio L, Rech GE, González J. Transposable elements contribute to the genomic response to insecticides in Drosophila melanogaster. Philos Trans R Soc Lond B Biol Sci. 2020;375: 20190341.

96. Witt E, Svetec N, Benjamin S, Zhao L. Transcription Factors Drive Opposite Relationships between Gene Age and Tissue Specificity in Male and Female Drosophila Gonads. Mol Biol Evol. 2021;38: 2104–2115.

97. Ramalingam V, Natarajan M, Johnston J, Zeitlinger J. TATA and paused promoters active in differentiated tissues have distinct expression characteristics. Mol Syst Biol. 2021;17: e9866.

98. Zeitlinger J, Stark A, Kellis M, Hong J-W, Nechaev S, Adelman K, et al. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat Genet. 2007;39: 1512–1516.

99. Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, Grissom SF, et al. RNA polymerase is poised for activation across the genome. Nat Genet. 2007;39: 1507–1511.

100. Engström PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res. 2007;17: 1898–1908.

101. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet. 2012;13: 233–245.

102. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, et al. A promoter-level mammalian expression atlas. Nature. 2014;507: 462–470.

103. FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C. Comparative genomics of Drosophila and human core promoters. Genome Biol. 2006;7: R53.

104. Day DS, Zhang B, Stevens SM, Ferrari F, Larschan EN, Park PJ, et al. Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. Genome Biol. 2016;17: 120.

105. Gaertner B, Johnston J, Chen K, Wallaschek N, Paulson A, Garruss AS, et al. Poised RNA polymerase II changes over developmental time and prepares genes for future expression. Cell Rep. 2012;2: 1670–1683.

106. Boettiger AN, Levine M. Synchronous and stochastic patterns of gene activation in the Drosophila embryo. Science. 2009;325: 471–473.

107. Lagha M, Bothma JP, Esposito E, Ng S, Stefanik L, Tsui C, et al. Paused Pol II coordinates tissue morphogenesis in the Drosophila embryo. Cell. 2013;153: 976–987.

108. Lifton RP, Goldberg ML, Karp RW, Hogness DS. The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. Cold Spring Harb Symp Quant Biol. 1978;42 Pt 2: 1047–1051.

109. Raser JM, O'Shea EK. Control of stochasticity in eukaryotic gene expression. Science. 2004;304: 1811–1814.

110. Blake WJ, Balázsi G, Kohanski MA, Isaacs FJ, Murphy KF, Kuang Y, et al. Phenotypic consequences of promoter-mediated transcriptional noise. Mol Cell. 2006;24: 853–865.

111. Tirosh I, Weinberger A, Carmi M, Barkai N. A genetic signature of interspecies variations in gene expression. Nat Genet. 2006;38: 830–834.

112. Schug J, Schuller W-P, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ Jr. Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol. 2005;6: R33.

113. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 2006;38: 626–635.

114. Dong C, Simonett SP, Shin S, Stapleton DS, Schueler KL, Churchill GA, et al. INFIMA leverages multi-omics model organism data to identify effector genes of human GWAS variants. Genome Biol. 2021;22: 241.

115. Porcu E, Sadler MC, Lepik K, Auwerx C, Wood AR, Weihs A, et al. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. Nat Commun. 2021;12: 5647.

116. Bozek M, Cortini R, Storti AE, Unnerstall U, Gaul U, Gompel N. ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the Drosophila blastoderm. Genome Res. 2019;29: 771–783.

117. Koromila T, Gao F, Iwasaki Y, He P, Pachter L, Gergen JP, et al. Odd-paired is a pioneer-like factor that coordinates with Zelda to control gene expression in embryos. eLife. 2020. doi:10.7554/eLife.59610

118. Soluri IV, Zumerling LM, Payan Parra OA, Clark EG, Blythe SA. Zygotic pioneer factor activity of Odd-paired/Zic is necessary for late function of the Drosophila segmentation network. eLife. 2020. doi:10.7554/elife.53916

119. Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. Nat Commun. 2019;10: 4872.

120. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20: 246.

121. Munger SC, Raghupathy N, Choi K, Simons AK, Gatti DM, Hinerfeld DA, et al. RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. Genetics. 2014;198: 59–73.

122. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11: 446–450.

123. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009;10: 241–251.

124. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature. 2015;527. doi:10.1038/nature15518

125. Wong ES, Schmitt BM, Kazachenka A, Thybert D, Redmond A, Connor F, et al. Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. Nat Commun. 2017;8. doi:10.1038/s41467-017-01037-x

126. Tirosh I, Reikhav S, Levy AA, Barkai N. A yeast hybrid provides insight into the evolution of gene expression regulation. Science. 2009;324: 659–662.

127. Wang D, Sung H-M, Wang T-Y, Huang C-J, Yang P, Chang T, et al. Expression evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors. Genome Res. 2007;17: 1161–1169.

128. Springer NM, Stupar RM. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. Plant Cell. 2007;19: 2391–2402.

129. Lemos B, Araripe LO, Fontanillas P, Hartl DL. Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. Proc Natl Acad Sci U S A. 2008;105: 14471–14476.

130. Wittkopp PJ, Haerum BK, Clark AG. Regulatory changes underlying expression differences within and between Drosophila species. Nat Genet. 2008;40: 346–350.

131. Connelly CF, Wakefield J, Akey JM. Evolution and genetic architecture of chromatin accessibility and function in yeast. PLoS Genet. 2014;10: e1004427.

II. Chapter 1 reference:

1. Zaitlen N, Kraft P, Patterson N, Pasaniuc B, Bhatia G, Pollack S, et al. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet. 2013;9: e1003520.

2. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447: 661–678.

3. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187: 367–383.

4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753.

5. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42: 565–569.

6. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169: 1177–1186.

7. Shi H, Kichaev G, Pasaniuc B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am J Hum Genet. 2016;99: 139–153.

8. Bloom JS, Kotenko I, Sadhu MJ, Treusch S, Albert FW, Kruglyak L. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nature Communications. 2015. doi:10.1038/ncomms9712

9. Bloom JS, Boocock J, Treusch S, Sadhu MJ, Day L, Oates-Barker H, et al. Rare variants contribute disproportionately to quantitative trait variation in yeast. Elife. 2019;8. doi:10.7554/eLife.49212

10. Chesler EJ. Out of the bottleneck: the Diversity Outcross and Collaborative Cross mouse populations in behavioral genetics research. Mamm Genome. 2014;25: 3–11.

11. Hook M, Roy S, Williams EG, Bou Sleiman M, Mozhui K, Nelson JF, et al. Genetic cartography of longevity in humans and mice: Current landscape and horizons. Biochim Biophys Acta Mol Basis Dis. 2018;1864: 2718–2732.

12. Saul MC, Philip VM, Reinholdt LG, Center for Systems Neurogenetics of Addiction, Chesler EJ. High-Diversity Mouse Populations for Complex Traits. Trends Genet. 2019;35: 501–514.

13. Long AD, Macdonald SJ, King EG. Dissecting complex traits using the Drosophila Synthetic Population Resource. Trends Genet. 2014;30: 488–495.

14. Mackay T. Trudy Mackay. Current Biology. 2006. pp. R659–R661. doi:10.1016/j.cub.2006.08.016

15. Torres JM, Gamazon ER, Parra EJ, Below JE, Valladares-Salgado A, Wacher N, et al. Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. Am J Hum Genet. 2014;95: 521–534.

16. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518: 337–343.

17. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6: e1000888.

18. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337: 1190–1195.

19. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. Am J Hum Genet. 2018;102: 717–730.

20. Hoekstra HE, Coyne JA. The locus of evolution: evo devo and the genetics of adaptation. Evolution. 2007;61: 995–1016.

21. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132: 311–322.

22. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10: 1213–1218.

23. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet. 2016;48: 206–213.

24. Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nature Genetics. 2018. pp. 1140–1150. doi:10.1038/s41588-018-0156-2

25. Gorkin DU, Barozzi I, Zhao Y, Zhang Y, Huang H, Lee AY, et al. An atlas of dynamic chromatin landscapes in mouse fetal development. Nature. 2020;583: 744–751.

26. Bozek M, Cortini R, Storti AE, Unnerstall U, Gaul U, Gompel N. ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the Drosophila blastoderm. Genome Res. 2019;29: 771–783.

27. Calderon D, Nguyen MLT, Mezger A, Kathiria A, Müller F, Nguyen V, et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat Genet. 2019;51: 1494–1505.

28. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, et al. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature. 2011;471: 480–485.

29. Davie K, Jacobs J, Atkins M, Potier D, Christiaens V, Halder G, et al. Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. PLoS Genet. 2015;11: e1004994.

30. Koenecke N, Johnston J, Gaertner B, Natarajan M, Zeitlinger J. Genome-wide identification of Drosophila dorso-ventral enhancers by differential histone acetylation analysis. Genome Biol. 2016;17: 196.

31. Hannon CE, Blythe SA, Wieschaus EF. Concentration dependent chromatin states induced by the bicoid morphogen gradient. Elife. 2017;6. doi:10.7554/eLife.28275

32. Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco-Ferreres R, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature. 2018;555: 538–542.

33. Jacobs J, Atkins M, Davie K, Imrichova H, Romanelli L, Christiaens V, et al. The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. Nat Genet. 2018;50: 1011–1020.

34. Haines JE, Eisen MB. Patterns of chromatin accessibility along the anterior-posterior axis in the early Drosophila embryo. PLoS Genet. 2018;14: e1007367.

35. Soluri IV, Zumerling LM, Payan Parra OA, Clark EG, Blythe SA. Zygotic pioneer factor activity of Odd-paired/Zic is necessary for late function of the Drosophila segmentation network. eLife. 2020. doi:10.7554/elife.53916

36. Koromila T, Gao F, Iwasaki Y, He P, Pachter L, Gergen JP, et al. Odd-paired is a pioneer-like factor that coordinates with Zelda to control gene expression in embryos. eLife. 2020. doi:10.7554/eLife.59610

37. Reddington JP, Garfield DA, Sigalova OM, Karabacak Calviello A, Marco-Ferreres R, Girardot C, et al. Lineage-Resolved Enhancer and Promoter Usage during a Time Course of Embryogenesis. Dev Cell. 2020;55: 648-664.e9.

38. Bravo González-Blas C, Quan X-J, Duran-Romaña R, Taskiran II, Koldere D, Davie K, et al. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. Mol Syst Biol. 2020;16: e9438.

39. Salces-Ortiz J, Vargas-Chavez C, Guio L, Rech GE, González J. Transposable elements contribute to the genomic response to insecticides in Drosophila melanogaster. Philos Trans R Soc Lond B Biol Sci. 2020;375: 20190341.

40. Witt E, Svetec N, Benjamin S, Zhao L. Transcription Factors Drive Opposite Relationships between Gene Age and Tissue Specificity in Male and Female Drosophila Gonads. Mol Biol Evol. 2021;38: 2104–2115.

41. Ramalingam V, Natarajan M, Johnston J, Zeitlinger J. TATA and paused promoters active in differentiated tissues have distinct expression characteristics. Mol Syst Biol. 2021;17: e9866.

42. Ruiz JL, Ranford-Cartwright LC, Gómez-Díaz E. The regulatory genome of the malaria vector Anopheles gambiae: integrating chromatin accessibility and gene expression.

Cold Spring Harbor Laboratory. 2020. p. 2020.06.22.164228. doi:10.1101/2020.06.22.164228

43. Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. Nat Commun. 2019;10: 4872.

44. Munger SC, Raghupathy N, Choi K, Simons AK, Gatti DM, Hinerfeld DA, et al. RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. Genetics. 2014;198: 59–73.

45. Mackay TFC, Huang W. Charting the genotype-phenotype map: lessons from the Drosophila melanogaster Genetic Reference Panel. Wiley Interdiscip Rev Dev Biol. 2018;7: e289.

46. Qiu X, Wu H, Hu R. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. BMC Bioinformatics. 2013;14: 124.

47. King EG, Macdonald SJ, Long AD. Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. Genetics. 2012;191: 935–949.

48. King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S, Broman KW, et al. Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome Res. 2012;22: 1558–1566.

49. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9: R137.

50. Koohy H, Down TA, Spivakov M, Hubbard T. A comparison of peak callers used for DNase-Seq data. PLoS One. 2014;9: e96303.

51. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014;94: 559–573.

52. Lu RJ-H, Liu Y-T, Huang CW, Yen M-R, Lin C-Y, Chen P-Y. ATACgraph: Profiling Genome-Wide Chromatin Accessibility From ATAC-seq. Front Genet. 2020;11: 618478.

53. Jenull S, Tscherner M, Mair T, Kuchler K. ATAC-Seq Identifies Chromatin Landscapes Linked to the Regulation of Oxidative Stress in the Human Fungal Pathogen Candida albicans. J Fungi (Basel). 2020;6. doi:10.3390/jof6030182

54. Bysani M, Agren R, Davegårdh C, Volkov P, Rönn T, Unneberg P, et al. ATAC-seq reveals alterations in open chromatin in pancreatic islets from subjects with type 2 diabetes. Sci Rep. 2019;9: 7785.

55. Orchard P, Kyono Y, Hensley J, Kitzman JO, Parker SCJ. Quantification, Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. Cell Syst. 2020;10: 298-306.e4.

56. Meers MP, Adelman K, Duronio RJ, Strahl BD, McKay DJ, Matera AG. Transcription start site profiling uncovers divergent transcription and enhancer-associated RNAs in Drosophila melanogaster. BMC Genomics. 2018;19: 157.

57. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19: 185–193.

58. Bour BA, O'Brien MA, Lockwood WL, Goldstein ES, Bodmer R, Taghert PH, et al. Drosophila MEF2, a transcription factor that is essential for myogenesis. Genes & Development. 1995. pp. 730–741. doi:10.1101/gad.9.6.730

59. Huang X, Warren JT, Buchanan J, Gilbert LI, Scott MP. Drosophila Niemann-Pick type C-2 genes control sterol homeostasis and steroid biosynthesis: a model of human neurodegenerative disease. Development. 2007;134. doi:10.1242/dev.004572

60. Andrenacci D, Grimaldi MR, Panetta V, Riano E, Rugarli EI, Graziani F. Functional dissection of the Drosophila Kallmann's syndrome protein DmKal-1. BMC Genet. 2006;7: 47.

61. Li Q, Imataka H, Morino S, Rogers GW, Jr, Richter-Cook NJ, et al. Eukaryotic Translation Initiation Factor 4AIII (eIF4AIII) Is Functionally Distinct from eIF4AI and eIF4AII. Mol Cell Biol. 1999;19: 7336.

62. Cáceres L, Necakov AS, Schwartz C, Kimber S, Roberts IJH, Krause HM. Nitric oxide coordinates metabolism, growth, and development via the nuclear receptor E75. Genes Dev. 2011;25: 1476–1485.

63. Schleinitz D, Böttcher Y, Blüher M, Kovacs P. The genetics of fat distribution. Diabetologia. 2014;57: 1276–1286.

64. Eleanor R. Grimm NIS. Genetics of Eating Behavior: Established and Emerging Concepts. Nutr Rev. 2011;69: 52.

65. Brown NL, Sattler CA, Markey DR, Carroll SB. hairy gene function in the Drosophila eye: normal expression is dispensable but ectopic expression alters cell fates. Development. 1991;113: 1245–1256.

66. Carroll SB, Laughon A, Thalley BS. Expression, function, and regulation of the hairy segmentation protein in the Drosophila embryo. Genes Dev. 1988;2: 883–890.

67. Carroll SB and Whyte JS. The role of the hairy gene during Drosophila morphogenesis: stripes in imaginal discs. Genes Dev. 1989;3: 905–916.

68. Robin C, Lyman RF, Long AD, Langley CH, Mackay TFC. hairy: A quantitative trait locus for drosophila sensory bristle number. Genetics. 2002;162: 155–164.

69. Macdonald SJ, Long AD. A potential regulatory polymorphism upstream of hairy is not associated with bristle number variation in wild-caught Drosophila. Genetics. 2004;167: 2127–2131.

70. Long AD, Mullaney SL, Reid LA, Fry JD, Langley CH, Mackay TF. High resolution mapping of genetic factors affecting abdominal bristle number in Drosophila melanogaster. Genetics. 1995;139: 1273–1291.

71. Riddihough G, Ish-Horowicz D. Individual stripe regulatory elements in the Drosophila hairy promoter respond to maternal, gap, and pair-rule genes. Genes Dev. 1991;5: 840–854.

72. Small S, Arnosti DN. Transcriptional Enhancers in Drosophila. Genetics. 2020;216: 1–26.

73. Bray S, Musisi H, Bienz M. Bre1 is required for Notch signaling and histone modification. Dev Cell. 2005;8: 279–286.

74. Urbanek K, Lesiak M, Krakowian D, Koryciak-Komarska H, Likus W, Czekaj P, et al. Notch signaling pathway and gene expression profiles during early in vitro differentiation of liver-derived mesenchymal stromal cells to osteoblasts. Lab Invest. 2017;97: 1225–1234.

75. Yu X, Zou J, Ye Z, Hammond H, Chen G, Tokunaga A, et al. Notch signaling activation in human embryonic stem cells is required for embryonic but not trophoblastic lineage commitment. Cell Stem Cell. 2008;2: 461.

76. Whited JL, Robichaux MB, Yang JC, Garrity PA. Ptpmeg is required for the proper establishment and maintenance of axon projections in the central brain of Drosophila. Development. 2007;134: 43–53.

77. Li M-Y, Lai P-L, Chou Y-T, Chi A-P, Mi Y-Z, Khoo K-H, et al. Protein tyrosine phosphatase PTPN3 inhibits lung cancer cell proliferation and migration by promoting EGFR endocytic degradation. Oncogene. 2015;34: 3791–3803.

78. Lin YJ, Seroude L, Benzer S. Extended life-span and stress resistance in the Drosophila mutant methuselah. Science. 1998;282: 943–946.

79. Arndt V, Dick N, Tawo R, Dreiseidler M, Wenzel D, Hesse M, et al. Chaperone-assisted selective autophagy is essential for muscle maintenance. Curr Biol. 2010;20. doi:10.1016/j.cub.2009.11.022

80. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The Drosophila melanogaster Genetic Reference Panel. Nature. 2012;482: 173–178.

81. Gao B, Huang Q, Baudis M. segment_liftover : a Python tool to convert segments between genome assemblies. F1000Res. 2018;7: 319.

82. Vinkhuyzen AAE, Pedersen NL, Yang J, Lee SH, Magnusson PKE, Iacono WG, et al. Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. Transl Psychiatry. 2012;2: e102.

83. Caballero A, Tenesa A, Keightley PD. The Nature of Genetic Variation for Complex Traits Revealed by GWAS and Regional Heritability Mapping Analyses. Genetics. 2015;201: 1601–1613.

84. O'Connor LJ, Schoech AP, Hormozdiari F, Gazal S, Patterson N, Price AL. Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. Am J Hum Genet. 2019;105: 456–476.

85. Fournier T, Abou Saada O, Hou J, Peter J, Caudal E, Schacherer J. Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. Elife. 2019;8. doi:10.7554/eLife.49258

86. Lamb AM, Walker EA, Wittkopp PJ. Tools and strategies for scarless allele replacement in Drosophila using CRISPR/Cas9. Fly . 2017;11: 53–64.

87. Port F, Chen H-M, Lee T, Bullock SL. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. Proc Natl Acad Sci U S A. 2014;111: E2967-76.

88. Gratz SJ, Ukken FP, Rubinstein CD, Thiede G, Donohue LK, Cummings AM, et al. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. Genetics. 2014;196: 961–971.

89. Gratz SJ, Cummings AM, Nguyen JN, Hamm DC, Donohue LK, Harrison MM, et al. Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. Genetics. 2013;194: 1029–1035.

90. Ren X, Sun J, Housden BE, Hu Y, Roesel C, Lin S, et al. Optimized gene editing technology for Drosophila melanogaster using germ line-specific Cas9. Proc Natl Acad Sci U S A. 2013;110: 19012–19017.

91. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17: 10–12.

92. Krueger F. TrimGalore. A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. TrimGalore (accessed on 27 August 2019). 2016.

93. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6 reference sequence of the Drosophila melanogaster genome. Genome Res. 2015;25: 445–458.

94. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009. pp. 1754–1760. doi:10.1093/bioinformatics/btp324

95. Li H, Bob H, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map (SAM) Format and. 2009 [cited 12 Jan 2021]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.325.1516

96. Broad Institute. Picard version 2.18.27. In: Broad Institute: Picard [Internet]. [cited 2019]. Available: http://broadinstitute.github.io/picard/

97. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.

98. Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, et al. Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res. 2011;21: 147–163.

99. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38: 576–589.

100. Zhou C, Yuan Z, Ma X, Yang H, Wang P, Zheng L, et al. Accessible chromatin regions and their functional interrelations with gene transcription and epigenetic modifications in sorghum genome. Plant Commun. 2021;2: 100140.

101. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12: 996–1006.

102. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32: D493-6.

103. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010;26: 2204–2207.

104. Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics. 2014;30: 1003–1005.

105. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc. 1995;57: 289–300.

106. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Stat Plan Inference. 1999;82: 171–196.

107. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . 2012;6: 80–92.

III. Chapter 2 reference:

1. Huynh K, Smith BR, Macdonald SJ, Long AD. Genetic Variation in Chromatin State Across Multiple Tissues in Drosophila melanogaster. bioRxiv. 2022. p. 2022.09.26.509449. doi:10.1101/2022.09.26.509449

2. pysam: Pysam is a Python module for reading and manipulating SAM/BAM/VCF/BCF files. It’s a lightweight wrapper of the htslib C-API, the same one that powers samtools, bcftools, and tabix. Github; Available: https://github.com/pysam-developers/pysam

3. King EG, Macdonald SJ, Long AD. Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. Genetics. 2012;191: 935–949.

4. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447: 661–678.

5. Bloom JS, Kotenko I, Sadhu MJ, Treusch S, Albert FW, Kruglyak L. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nature Communications. 2015. doi:10.1038/ncomms9712

6. Bloom JS, Boocock J, Treusch S, Sadhu MJ, Day L, Oates-Barker H, et al. Rare variants contribute disproportionately to quantitative trait variation in yeast. Elife. 2019;8. doi:10.7554/eLife.49212

7. Chesler EJ. Out of the bottleneck: the Diversity Outcross and Collaborative Cross mouse populations in behavioral genetics research. Mamm Genome. 2014;25: 3–11.

8. Hook M, Roy S, Williams EG, Bou Sleiman M, Mozhui K, Nelson JF, et al. Genetic cartography of longevity in humans and mice: Current landscape and horizons. Biochim Biophys Acta Mol Basis Dis. 2018;1864: 2718–2732.

9. Saul MC, Philip VM, Reinholdt LG, Center for Systems Neurogenetics of Addiction, Chesler EJ. High-Diversity Mouse Populations for Complex Traits. Trends Genet. 2019;35: 501–514.

10. Long AD, Macdonald SJ, King EG. Dissecting complex traits using the Drosophila Synthetic Population Resource. Trends Genet. 2014;30: 488–495.

11. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187: 367–383.

12. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753.

13. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169: 1177–1186.

14. Torres JM, Gamazon ER, Parra EJ, Below JE, Valladares-Salgado A, Wacher N, et al. Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. Am J Hum Genet. 2014;95: 521–534.

15. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518: 337–343.

16. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6: e1000888.

17. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337: 1190–1195.

18. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132: 311–322.

19. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10: 1213–1218.

20. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet. 2016;48: 206–213.

21. Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nature Genetics. 2018. pp. 1140–1150. doi:10.1038/s41588-018-0156-2

22. Koenecke N, Johnston J, Gaertner B, Natarajan M, Zeitlinger J. Genome-wide identification of Drosophila dorso-ventral enhancers by differential histone acetylation analysis. Genome Biol. 2016;17: 196.

23. Hannon CE, Blythe SA, Wieschaus EF. Concentration dependent chromatin states induced by the bicoid morphogen gradient. Elife. 2017;6. doi:10.7554/eLife.28275

24. Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco-Ferreres R, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature. 2018;555: 538–542.

25. Haines JE, Eisen MB. Patterns of chromatin accessibility along the anterior-posterior axis in the early Drosophila embryo. PLoS Genet. 2018;14: e1007367.

26. Bozek M, Cortini R, Storti AE, Unnerstall U, Gaul U, Gompel N. ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the Drosophila blastoderm. Genome Res. 2019;29: 771–783.

27. Soluri IV, Zumerling LM, Payan Parra OA, Clark EG, Blythe SA. Zygotic pioneer factor activity of Odd-paired/Zic is necessary for late function of the Drosophila segmentation network. eLife. 2020. doi:10.7554/elife.53916

28. Koromila T, Gao F, Iwasaki Y, He P, Pachter L, Gergen JP, et al. Odd-paired is a pioneer-like factor that coordinates with Zelda to control gene expression in embryos. eLife. 2020. doi:10.7554/eLife.59610

29. Reddington JP, Garfield DA, Sigalova OM, Karabacak Calviello A, Marco-Ferreres R, Girardot C, et al. Lineage-Resolved Enhancer and Promoter Usage during a Time Course of Embryogenesis. Dev Cell. 2020;55: 648–664.e9.

30. Ramalingam V, Natarajan M, Johnston J, Zeitlinger J. TATA and paused promoters active in differentiated tissues have distinct expression characteristics. Mol Syst Biol. 2021;17: e9866.

31. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, et al. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature. 2011;471: 480–485.

32. Merrill CB, Montgomery AB, Pabon MA, Shabalin AA, Rodan AR, Rothenfluh A. Harnessing changes in open chromatin determined by ATAC-seq to generate insulin-responsive reporter constructs. BMC Genomics. 2022;23: 399.

33. Bravo González-Blas C, Quan X-J, Duran-Romaña R, Taskiran II, Koldere D, Davie K, et al. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. Mol Syst Biol. 2020;16: e9438.

34. Jacobs J, Atkins M, Davie K, Imrichova H, Romanelli L, Christiaens V, et al. The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. Nat Genet. 2018;50: 1011–1020.

35. Davie K, Jacobs J, Atkins M, Potier D, Christiaens V, Halder G, et al. Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. PLoS Genet. 2015;11: e1004994.

36. Salces-Ortiz J, Vargas-Chavez C, Guio L, Rech GE, González J. Transposable elements contribute to the genomic response to insecticides in Drosophila melanogaster. Philos Trans R Soc Lond B Biol Sci. 2020;375: 20190341.

37. Witt E, Svetec N, Benjamin S, Zhao L. Transcription Factors Drive Opposite Relationships between Gene Age and Tissue Specificity in Male and Female Drosophila Gonads. Mol Biol Evol. 2021;38: 2104–2115.

38. Arnold M, Ellwanger DC, Hartsperger ML, Pfeufer A, Stümpflen V. Cis-acting polymorphisms affect complex traits through modifications of microRNA regulation pathways. PLoS One. 2012;7: e36694.

39. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell. 2019;177: 1022–1034.e6.

40. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53: 1300–1310.

41. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011;12: 703–714.

42. Christiansen L, Amini S, Zhang F, Ronaghi M, Gunderson KL, Steemers FJ. Contiguity-Preserving Transposition Sequencing (CPT-Seq) for Genome-Wide Haplotyping, Assembly, and Single-Cell ATAC-Seq. In: Tiemann-Boege I, Betancourt A, editors. Haplotyping: Methods and Protocols. New York, NY: Springer New York; 2017. pp. 207–221.

43. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature. 2015;527. doi:10.1038/nature15518

44. Shen SQ, Turro E, Corbo JC. Hybrid mice reveal parent-of-origin and Cis- and trans-regulatory effects in the retina. PLoS One. 2014;9: e109382.

45. Russell ND, Chow CY. The dynamic effect of genetic variation on the in vivo ER stress transcriptional response in different tissues. G3 . 2022;12. doi:10.1093/g3journal/jkac104

46. Zhao N, Ding X, Lian T, Wang M, Tong Y, Liang D, et al. The Effects of Gene Duplication Modes on the Evolution of Regulatory Divergence in Wild and Cultivated Soybean. Front Genet. 2020;11: 601003.

47. Suvorov A, Nolte V, Pandey RV, Franssen SU, Futschik A, Schlötterer C. Intra-specific regulatory variation in Drosophila pseudoobscura. PLoS One. 2013;8: e83547.

48. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol. 2015;16: 144–154.

49. Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. Nat Commun. 2019;10: 4872.

50. Munger SC, Raghupathy N, Choi K, Simons AK, Gatti DM, Hinerfeld DA, et al. RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. Genetics. 2014;198: 59–73.

51. Fan S, Schneider LE. The role of maternal and zygotic Gprk2 expression in Drosophila development. Biochem Biophys Res Commun. 2003;301: 127–135.

52. Schneider LE, Spradling AC. The Drosophila G-protein-coupled receptor kinase homologue Gprk2 is required for egg morphogenesis. Development. 1997;124: 2591–2602.

53. Lannutti BJ, Schneider LE. Gprk2 controls cAMP levels in Drosophila development. Dev Biol. 2001;233: 174–185.

54. Chen Y, Li S, Tong C, Zhao Y, Wang B, Liu Y, et al. G protein-coupled receptor kinase 2 promotes high-level Hedgehog signaling by regulating the active state of Smo through kinase-dependent and kinase-independent mechanisms in Drosophila. Genes Dev. 2010;24: 2054–2067.

55. Tanoue S, Krishnan P, Chatterjee A, Hardin PE. G protein-coupled receptor kinase 2 is required for rhythmic olfactory responses in Drosophila. Curr Biol. 2008;18: 787–794.

56. Nishida Y, Hata M, Ayaki T, Ryo H, Yamagata M, Shimizu K, et al. Proliferation of both somatic and germ cells is affected in the Drosophila mutants of raf proto-oncogene. EMBO J. 1988;7: 775–781.

57. Luo H, Rose PE, Roberts TM, Dearolf CR. The Hopscotch Jak kinase requires the Raf pathway to promote blood cell activation and differentiation in Drosophila. Mol Genet Genomics. 2002;267: 57–63.

58. Kadener S, Fededa JP, Rosbash M, Kornblihtt AR. Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. Proc Natl Acad Sci U S A. 2002;99: 8185–8190.

59. Meng F, Zhao H, Zhu B, Zhang T, Yang M, Li Y, et al. Genomic editing of intronic enhancers unveils their role in fine-tuning tissue-specific gene expression in Arabidopsis thaliana. Plant Cell. 2021;33: 1997–2014.

60. Cheng F, Zheng W, Liu C, Barbuti PA, Yu-Taeger L, Casadei N, et al. Intronic enhancers of the human *SNCA* gene predominantly regulate its expression in brain in vivo. Sci Adv. 2022;8: eabq6324.

61. Kim J, Kim Y-J, Kim-Ha J. Blood-brain barrier defects associated with Rbp9 mutation. Mol Cells. 2010;29: 93–98.

62. Jeong K, Kim-Ha J. Precocious expression of Drosophila Rbp9 inhibits ovarian germ cell proliferation. Mol Cells. 2004;18: 230–236.

63. Jeong K, Kim-Ha J. Expression of Rbp9 during mid-oogenesis induces apoptosis in egg chambers. Mol Cells. 2003;16: 392–396.

64. Lee SH, Kim Y, Kim-Ha J. Requirement of Rbp9 in the maintenance of Drosophila germline sexual identity. FEBS Lett. 2000;465: 165–168.

65. Kim J, Kim-Ha J. Ovarian tumors in Rbp9 mutants of Drosophila induce an immune response. Mol Cells. 2006;22: 228–232.

66. Shaul O. How introns enhance gene expression. Int J Biochem Cell Biol. 2017;91: 145–155.

67. Li H, Bai L, Li H, Li X, Kang Y, Zhang N, et al. Selective translational usage of TSS and core promoters revealed by translatome sequencing. BMC Genomics. 2019;20: 282.

68. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res. 2006;16: 1–10.

69. Johnstone O, Lasko P. Interaction with eIF5B is essential for Vasa function during development. Development. 2004;131: 4167–4178.

70. Carrera P, Johnstone O, Nakamura A, Casanova J, Jäckle H, Lasko P. VASA mediates translation through interaction with a Drosophila yIF2 homolog. Mol Cell. 2000;5: 181–187.

71. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410.

72. Madden TL, Tatusov RL, Zhang J. Applications of network BLAST server. Methods Enzymol. 1996;266: 131–141.

73. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7: 203–214.

74. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421.

75. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008;24: 1757–1764.

76. McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res. 2010;20: 816–825.

77. Hock T, Cottrill T, Keegan J, Garza D. The E23 early gene of Drosophila encodes an ecdysone-inducible ATP-binding cassette transporter capable of repressing ecdysone-mediated gene activation. Proc Natl Acad Sci U S A. 2000;97: 9519–9524.

78. Ishimoto H, Kitamoto T. The steroid molting hormone Ecdysone regulates sleep in adult Drosophila melanogaster. Genetics. 2010;185: 269–281.

79. Ullah F, Hamilton M, Reddy ASN, Ben-Hur A. Exploring the relationship between intron retention and chromatin accessibility in plants. BMC Genomics. 2018;19: 21.

80. Carney GE, Bender M. The Drosophila ecdysone receptor (EcR) gene is required maternally for normal oogenesis. Genetics. 2000;154: 1203–1211.

81. Li T, Bender M. A conditional rescue system reveals essential functions for the ecdysone receptor (EcR) gene during molting and metamorphosis in Drosophila. Development. 2000;127: 2897–2905.

82. D'Avino PP, Thummel CS. The ecdysone regulatory pathway controls wing morphogenesis and integrin expression during Drosophila metamorphosis. Dev Biol. 2000;220: 211–224.

83. König A, Yatsenko AS, Weiss M, Shcherbata HR. Ecdysteroids affect Drosophila ovarian stem cell niche formation and early germline differentiation. EMBO J. 2011;30: 1549–1562.

84. Slaidina M, Banisch TU, Gupta S, Lehmann R. A single-cell atlas of the developing Drosophila ovary identifies follicle stem cell progenitors. Genes Dev. 2020;34: 239–249.

85. Lamb AM, Walker EA, Wittkopp PJ. Tools and strategies for scarless allele replacement in Drosophila using CRISPR/Cas9. Fly . 2017;11: 53–64.

86. Port F, Chen H-M, Lee T, Bullock SL. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. Proc Natl Acad Sci U S A. 2014;111: E2967–76.

87. Gratz SJ, Ukken FP, Rubinstein CD, Thiede G, Donohue LK, Cummings AM, et al. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. Genetics. 2014;196: 961–971.

88. Gratz SJ, Cummings AM, Nguyen JN, Hamm DC, Donohue LK, Harrison MM, et al. Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. Genetics. 2013;194: 1029–1035.

89. Ren X, Sun J, Housden BE, Hu Y, Roesel C, Lin S, et al. Optimized gene editing technology for Drosophila melanogaster using germ line-specific Cas9. Proc Natl Acad Sci U S A. 2013;110: 19012–19017.

90. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17: 10–12.

91. Krueger F. TrimGalore. A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. TrimGalore (accessed on 27 August 2019). 2016.

92. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009. pp. 1754–1760. doi:10.1093/bioinformatics/btp324

93. Li H, Bob H, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map (SAM) Format and. 2009 [cited 12 Jan 2021]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.325.1516

94. Broad Institute. Picard version 2.18.27. In: Broad Institute: Picard [Internet]. [cited 2019]. Available: http://broadinstitute.github.io/picard/

95. Ou J, Liu H, Yu J, Kelliher MA, Castilla LH, Lawson ND, et al. ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. BMC Genomics. 2018;19: 169.

96. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19: 185–193.

97. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol. 1995;57: 289–300.

98. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Stat Plan Inference. 1999;82: 171–196.

99. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38: 576–589.

100. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, et al. HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience. 2021;10. doi:10.1093/gigascience/giab007

101. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10. doi:10.1093/gigascience/giab008

102. Wang Q, Jia Y, Wang Y, Jiang Z, Zhou X, Zhang Z, et al. Evolution of cis- and trans-regulatory divergence in the chicken genome between two contrasting breeds analyzed using three tissue types at one-day-old. BMC Genomics. 2019;20: 933.

IV. Chapter 3 reference:

1. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169: 1177–1186.

2. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh. 1919. pp. 399–433. doi:10.1017/s0080456800012163

3. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33 Suppl: 228–237.

4. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014;94: 559–573.

5. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42: D1001–6.

6. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016;352: 600–604.

7. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132: 311–322.

8. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10: 1213–1218.

9. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet. 2016;48: 206–213.

10. Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nature Genetics. 2018. pp. 1140–1150. doi:10.1038/s41588-018-0156-2

11. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20: 246.

12. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11: 446–450.

13. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009;10: 241–251.

14. Long AD, Macdonald SJ, King EG. Dissecting complex traits using the Drosophila Synthetic Population Resource. Trends Genet. 2014;30: 488–495.

15. Lamb AM, Walker EA, Wittkopp PJ. Tools and strategies for scarless allele replacement in Drosophila using CRISPR/Cas9. Fly . 2017;11: 53–64.

16. Port F, Chen H-M, Lee T, Bullock SL. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. Proc Natl Acad Sci U S A. 2014;111: E2967–76.

17. Gratz SJ, Ukken FP, Rubinstein CD, Thiede G, Donohue LK, Cummings AM, et al. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. Genetics. 2014;196: 961–971.

18. Gratz SJ, Cummings AM, Nguyen JN, Hamm DC, Donohue LK, Harrison MM, et al. Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. Genetics. 2013;194: 1029–1035.

19. Ren X, Sun J, Housden BE, Hu Y, Roesel C, Lin S, et al. Optimized gene editing technology for Drosophila melanogaster using germ line-specific Cas9. Proc Natl Acad Sci U S A. 2013;110: 19012–19017.

APPENDIX A

Chapter 1: supplementary text 1

Brittny Smith
Stuart Macdonald

Nuclei Isolation

1. Dissect tissue of interest in lysis buffer.

   We have successfully started the protocol with:
   - i. Ovaries from 5 adult, mated females,
   - ii. Eye-antennal discs from 5 male third-instar larvae,
   - iii. Wing discs from 5 male third-instar larvae,
   - iv. Brains from 10 adult, mated females.

   Ideally you want to keep everything cold and/or rapidly dissect your tissue out.

   The lysis buffer is:
   10 mM Tris-HCl, pH 7.4
   10 mM NaCl
   3 mM $MgCl_2$
   0.1% IGEPAL CA-630

2. Place into 200-µl lysis buffer on ice in 1.7-ml tube.
3. Manually grind 25 times using a blue plastic pestle (Fisher, K749521-1500).

   A "grind" in this case being loosely defined as a single turn of the pestle with your fingers.

4. Let sample sit on ice for ~1-min.
5. Repeat steps 3 and 4 twice each (a total of 75 grinds).
6. Spin sample at 100-g for 10-min at 4ºC.
7. Remove supernatant. Resuspend sample in 200-µl lysis buffer.
8. Run sample through 30-µm filter cloth.

   The cloth is Nitex Nylon Mesh (30µm) from Genesee Scientific (Cat # 57-105). After it arrives you will want to wash it in water, rinse with ethanol, dry, and cut into ~1-inch squares.

   You'll start this step with a fresh 1.7ml tube in a rack. Remove the pointy end from a standard 1000-µl tip so its got a nice wide bore, and drop into the destination tube. Then on top of the barrel of the tip you can lay a piece of the cloth. Take up your sample from step #7 in a pipette, jam the tip into the cloth (making a little sieve with the barrel of the

tip that's in the destination tube) and fire the resuspended cells through the mesh. In step #9 below you're simply ensuring everything went through by adding more lysis buffer to the "sieve".

9. Wash through cloth with 200-µl lysis buffer.
10. Spin sample at 1000-g for 10-min at 4ºC.
11. Pipette off supernatant.

Tagmentation Reaction

1. To the cell pellet add 25-µl of tagmentation reaction mix:

    12.5µl 2X TD Buffer
    1.25µl Tn5 Transposase
    11.25µl H_2O

The TD Buffer and Tn5 Transposase are from the Illumina Nextera DNA Sample Preparation Kit (Cat # FC-121-1030). Now (mid-2021) you can purchase them separately from Illumina.

2. Pipette to resuspend pellet in the tagmentation mix.
3. Incubate for 30-min at 37ºC.
4. Place sample on ice (if moving forward to purification) or freeze at −20ºC.

Qiagen MinElute Purification

1. Add 125-µl (5 volumes) of PB buffer to the 25-µl (1 volume) tagmentation reaction, and mix.

Use Qiagen MinElute PCR Purification Kit (Cat # 28004). Ensure that the correct volume of 100% ethanol is added to buffer PE concentrate before use.

2. Place MinElute column in a 2-ml collection tube.
3. Add sample to column, and centrifuge at 18,000-g for 1-min at RT.
4. Discard flow-through and place column back in collection tube.
5. Add 750-µl PE buffer to column, and centrifuge at 18,000-g for 1-min at RT.
6. Discard flow-through and place column back in collection tube.
7. Centrifuge at 18,000-g for 1-min at RT.
8. Place column in new 1.7-ml tube.
9. Add 20-µl EB buffer to the center of the column.
10. Let column stand for 1-min, and centrifuge at 18,000-g for 1-min at RT.
11. Place sample on ice (if moving forward to PCR) or freeze at −20ºC.

PCR Amplification

Reaction

1 × 25µl reaction

5-µl	Purified, tagmented DNA
2.5-µl	H_2O
2.5-µl	SJM 7## Indexed primer (@ 12.5-µM)
2.5-µl	SJM 5## Indexed primer (@ 12.5-µM)
12.5-µl	Kapa Master Mix

For ovary samples we used 5-µl of purified tagmented DNA. For discs and brain samples we used 7.5-µl of purified tagmented DNA and eliminated the water from the reaction.

The Kapa Master Mix comes from Cat # KK2612.

The "SJM 7## Indexed primer" (and the 5## version) are custom versions of Illumina Nextera index primers. (You can obviously use those described in the original ATACseq protocol.)

SJM 7## Indexed primer:
xxxxxxx = 7-base i7 index
5'- CAAGCAGAAGACGGCATACGAGATxxxxxxxGTCTCGTGGGCTCGG -3'

SJM 5## Indexed primer:
yyyyyyyy = 8-base i5 index
5'- AATGATACGGCGACCACCGAGATCTACACyyyyyyyyTCGTCGGCAGCGTC -3'

Thermocycling

72ºC 5-min
98ºC 30-sec
12 cycles of:

98ºC	10-sec
63ºC	30-sec
72ºC	1-min

4ºC hold

Bead Cleanup

1. Add 25-µl of beads to sample.

This is a 1X bead cleanup. So provides limited size selection. Beads are Agencourt AMPure XP beads (A63881).

2. Mix and incubate at RT for 5-min.
3. Place sample on magnetic plate for 1-min.
4. Remove and discard supernatant.

5. Wash with 50-µl of 70% ethanol. Repeat the wash.
6. Air dry for 5-min.
7. Resuspend in 20-µl of Qiagen EB buffer.
8. Incubate at RT for 2-min.
9. Place sample on magnetic plate for 1-min.
10. Transfer 20-µl to new 1.7-ml tube.

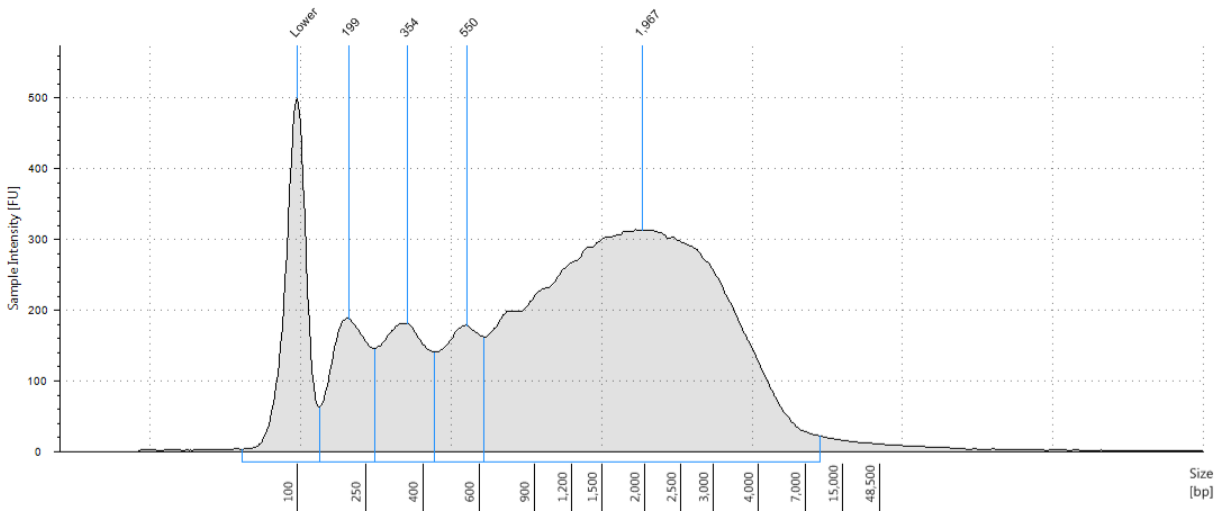
Quality Control

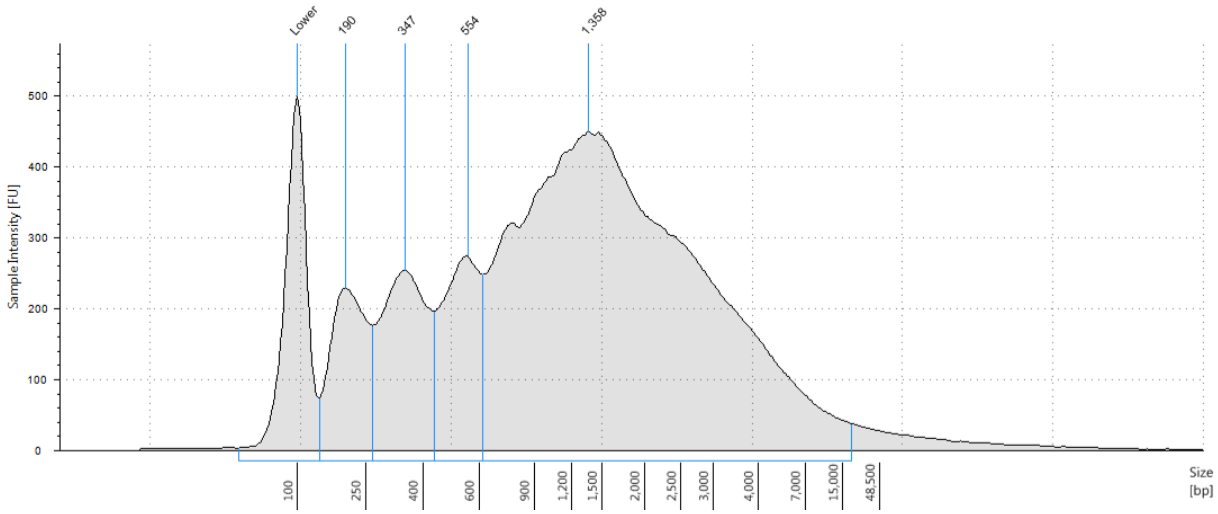
1. Qubit sample using BR (broad range) dsDNA kit (ThermoFisher Q32850).

Not totally clear to what extent the Qubit value informs you about sample quality. We have seen nice ATACseq peaks from samples that had <10 ng/µl values, and similar peak profiles from samples that had >40 ng/µl values.

2. Run on Agilent TapeStation instrument using a Genomic DNA ScreenTape.

The TapeStation is similar to the Agilent BioAnalyzer, but a little lower resolution. Example good pictures (meaning that if you see this, and then sequence that library, the sample has peaks and isn't simply an over-digested mess) are below. You want to see the periodicity at the low molecular weight end of the TapeStation profile.




Lower
190
347
554
1,358
Sample Intensity [FU]
500
400
300
200
100
0
100
250
400
600
900
1,200
1,500
2,000
2,500
3,000
4,000
7,000
15,000
48,500
Size
[bp]