INVESTIGATING TRANSCRIPTIONAL AND TRANSLATIONAL REGULATION OF GENE EXPRESSION WITHIN AND BETWEEN DROSOPHILA SPECIES

By

Fei Zhang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Animal Science — Doctor of Philosophy

2024

ABSTRACT

Gene expression regulation involves intricate genetic and environmental interactions that shape phenotypic diversity. This dissertation explores both intraspecific and interspecific regulatory variation using F1 hybrid models in Drosophila. The first part of the study focuses on *Drosophila melanogaster*, characterizing *cis*- and *trans*-regulatory effects within the species. By combining allele-specific mapping and RNA sequencing, the analysis reveals that *cis*-regulatory effects are predominant, driving much of the variation in gene expression. *Trans*-regulatory effects, while less frequent, may indicate complex interactions between regulatory elements.

The second part extends the investigation to interspecific gene regulation between *Drosophila melanogaster* and *Drosophila simulans*. By integrating ribosome profiling with RNA sequencing, the study evaluates both transcriptional and translational divergence in hybrid offspring. This analysis identifies significant differences in both *cis*- and *trans*-regulatory effects, with trans effects contributing prominently to translation efficiency variation. Temperature-dependent changes further modulate both transcription and translation, highlighting the role of environmental factors in regulatory divergence between species.

Overall, the findings demonstrate that gene expression regulation is shaped by both local genetic elements and broader regulatory networks, modulated by environmental conditions. The comparative analysis of intraspecific and interspecific hybrids provides insights into the evolutionary dynamics of gene regulation, emphasizing the buffering role of translational regulation and the adaptive potential of *trans*-regulatory elements. This comprehensive framework enhances our understanding of gene expression regulation and its implications for evolutionary biology

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Dr. Wen Huang, for his exceptional mentorship, unwavering support, and insightful guidance throughout my PhD journey. His expertise, dedication, and encouragement have profoundly shaped my research and academic growth. I am truly honored to have had the opportunity to learn from and work under his supervision.

I would also like to extend my heartfelt thanks to my beloved wife, Li Feng, for her constant love, patience, and encouragement. Her unwavering support has been my greatest source of strength, especially during the most challenging times.

To my parents and family, your love, sacrifices, and faith in me have always been my driving force. I am grateful for your endless encouragement and support, which have been instrumental in reaching this milestone.

I am sincerely thankful to my committee members: Dr. Catherine W. Ernst, Dr. David N. Arnosti, Dr. Yuehua Cui and Dr. Juan P. Steibel for their invaluable feedback, guidance, and support throughout my research. Your insights and commitment have been critical in shaping this dissertation.

To our lab members, thank you for fostering a collaborative and motivating environment. Your camaraderie and shared passion for science have made my research experience fulfilling and enjoyable.

Finally, I am grateful to my friends, whose encouragement, support, and shared moments have brought balance and joy to this journey.

Thank you all for making this achievement possible.

TABLE OF CONTENTS

INTRODUCTION

Gene expression regulation is a central mechanism underlying phenotypic diversity, evolution, and adaptation. It involves complex layers of control from transcription to translation, each influenced by genetic variations and environmental factors. Understanding how these regulatory mechanisms contribute to gene expression variation within and between species is fundamental for elucidating the genetic basis of evolution. This dissertation investigates the *cis*- and *trans*-regulatory effects shaping gene expression in *Drosophila*, with a focus on both intraspecific variation within *Drosophila melanogaster* and interspecific divergence between *Drosophila melanogaster* and *Drosophila simulans*.

Cis-regulatory elements are typically located near the gene they regulate and include promoters, enhancers, and untranslated regions, which affect transcriptional or post-transcriptional processes. In contrast, *trans*-regulatory elements include diffusible factors like transcription factors, RNA-binding proteins, and microRNAs that act across loci to modify gene expression. Previous studies suggest that while *trans*-regulatory variation often dominates within species, *cis*-regulatory variation becomes more significant in the context of interspecies divergence (Vande Zande 2022). This interplay between cis and trans effects not only drives gene expression variation but also impacts hybrid fitness and adaptive traits.

While much research has focused on transcriptional regulation, less is known about how post-transcriptional mechanisms, such as translation efficiency, contribute to regulatory variation. Translational regulation adds another layer of complexity, as it can either buffer or amplify transcriptional differences. Recent advances in high-throughput sequencing, including RNA sequencing (RNA-Seq) and ribosome profiling (Ribo-seq), have enabled simultaneous measurement of mRNA abundance and translation efficiency. These tools provide a powerful means to dissect allele-specific regulatory effects in hybrids, offering a unique opportunity to understand the genetic architecture of gene expression across different biological levels.

This dissertation is organized into two main studies. The first study focuses on intraspecific regulatory variation within *Drosophila melanogaster*. By using allele-specific RNA-Seq in F1 hybrids derived from genetically distinct inbred lines, this study quantifies the contributions of *cis*- and *trans*-regulatory elements to transcriptional variation. The second study extends this approach to interspecific hybrids between *Drosophila melanogaster* and *Drosophila simulans*. Integrating RNA-Seq with ribosome profiling, it explores how regulatory divergence at both transcriptional and translational levels contributes to phenotypic variation between these species. Additionally, temperature-controlled experiments are used in both studies to examine how environmental factors modulate regulatory interactions.

Through these complementary analyses, this dissertation provides a comprehensive view of how *cis*- and *trans*-regulatory elements shape gene expression within and between species. The findings offer new insights into the evolutionary dynamics of gene regulation, emphasizing the roles of genetic architecture and environmental factors in phenotypic diversity.

Vande Zande, P., Hill, M. S., & Wittkopp, P. J. (2022). Pleiotropic effects of trans-regulatory mutations on fitness and gene expression. Science, 377(6601), 105-109.

CHAPTER 1: LITERATURE REVIEW

TRANSCRIPTIONAL, POSTTRANSCRIPTIONAL, AND TRANSLATIONAL REGULATION OF GENE EXPRESSION: MECHANISMS, METHODS, AND RECENT ADVANCES IN DROSOPHILA

Overview of the gene expression process in eukaryotes

Gene expression in eukaryotes is a complex, multi-step process that transforms genetic information encoded in DNA into functional proteins (**Figure 1.1**), which in turn determine phenotypic traits across organisms. This process is tightly regulated at each stage, maintaining phenotypic consistency across individuals while enabling adaptive variability in response to environmental changes. The initial step, transcription, involves the synthesis of messenger RNA (mRNA) from DNA by RNA polymerase, guided by a network of transcription factors, chromatin accessibility, and epigenetic modifications. Transcriptional regulation in eukaryotic cells is a complex process, where RNA polymerases play a central role in synthesizing RNA from DNA templates, a well-established understanding of the gene expression process. Regulatory sequences such as promoters, enhancers, and silencers coordinate with DNA-binding proteins to modulate transcription initiation, elongation, and termination (Cramer, 2019).

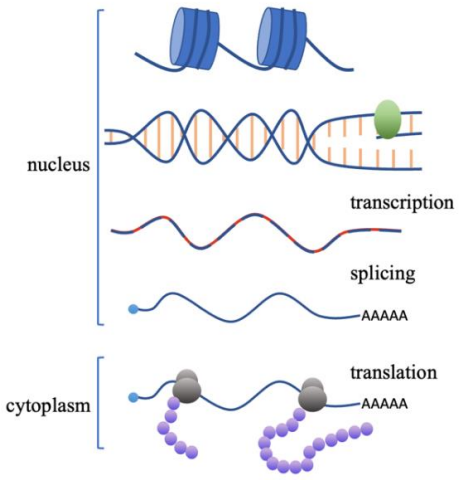


Figure 1.1 Overview of the gene expression process in eukaryotes.

Epigenetic modifications regulate gene expression by altering chromatin structure and accessibility. DNA methylation at CpG islands often represses transcription, while histone modifications, like acetylation and methylation, shift chromatin states between heterochromatin (repressive) and euchromatin (permissive). Histone acetyltransferases (HATs) generally enhance transcription by loosening chromatin, whereas histone methylation can activate or repress transcription depending on the residues affected. Chromatin remodeling complexes, such as SWI/SNF, reposition nucleosomes to regulate DNA accessibility (Jones, 2012; Allis & Jenuwein, 2016; Clapier & Cairns, 2009).

Concomitant with transcription, the nascent mRNA undergoes several posttranscriptional modifications, including splicing, 5' capping, RNA editing, and 3' polyadenylation. These modifications are crucial for mRNA stability, nuclear export, and translation efficiency(Nilsen & Graveley, 2010; Darnell, 2013). Alternative splicing allows a single gene to produce multiple protein isoforms, increasing proteome diversity and enabling tissue-specific functions (Kornblihtt et al., 2013; Baralle & Giudice, 2017). In recent years, RNA modifications such as N6-methyladenosine (m6A) have emerged as important regulators of RNA metabolism, affecting splicing, stability, and translation (Zhao, Roundtree, & He, 2017; Yang, Hsu, Chen, & Yang, 2018; Dominissini et al., 2012)

Once processed, mature mRNAs are exported to the cytoplasm, where they serve as templates for translation. This step is influenced by factors such as ribosome binding, mRNA localization, codon usage bias, mRNA secondary structure, and mRNA stability. Additionally, the availability of tRNAs, regulatory proteins (e.g., initiation and elongation factors), and microRNAs can modulate translation efficiency (Hinnebusch & Lorsch, 2012; Gingold & Pilpel, 2011)

After translation, newly synthesized proteins often undergo posttranslational modifications like phosphorylation, acetylation, or ubiquitination, which regulate their activity, stability, localization, or interaction with other proteins (Mann & Jensen, 2003). This regulatory

layer allows cells to respond rapidly to environmental changes and maintain precise cellular functions. The regulation of gene expression at transcriptional, posttranscriptional, translational, and posttranslational levels ensures that genetic information is dynamically expressed in a way that supports organismal adaptation and phenotypic variation (Albert & Kruglyak, 2015).

Regulation of the transcription process

Transcription is a critical regulatory step in the gene expression pathway, serving as a key control point for many genes in eukaryotes. This process, which occurs in the cell nucleus, is highly regulated and consists of several distinct stages: promoter recognition, promoter opening, transcription initiation, and elongation. The initial phase involves the binding of RNA polymerase to promoter regions, which are specific DNA sequences that signal the start of a gene.

This binding is influenced by *cis*-regulatory elements, such as promoters, enhancers, and silencers, which serve as docking sites for *trans*-regulatory factors, including transcription factors, co-activators, and repressors. Promoters, located near the transcription start site, interact directly with the transcription machinery, including TFIID and RNA polymerase II, to initiate transcription. Core promoter elements, such as the TATA box and Inr, facilitate this binding. Enhancers, often located far from the target gene, influence transcription by looping the DNA to bring activators closer to promoters. Silencers recruit repressors and co-repressors, which condense chromatin or block activator access, thereby repressing transcription. Transcription factors modulate gene expression by either enhancing or inhibiting transcription through direct interactions with these cis-elements and chromatin remodeling enzymes. Their combined action shapes gene expression patterns across different cells, tissues, and responses. (Levine & Tjian, 2003; Spitz & Furlong, 2012). These interactions establish transcriptional specificity and are further modulated by epigenetic factors such as DNA methylation, histone modifications, and chromatin remodeling, which alter chromatin accessibility and gene activity (Allis & Jenuwein, 2016).

The outcomes of transcriptional regulation are diverse and influence development, cell differentiation, and responses to environmental signals. For example, HOX genes, which determine body plan development in animals, are tightly regulated at the transcriptional level through a combination of enhancer sequences, chromatin modifications, and long non-coding RNAs (lncRNAs) (Alexander, Nolte, & Krumlauf, 2009; Rinn et al., 2007). Another example is p53, a tumor suppressor gene whose transcriptional activity is modulated by a network of enhancers, promoters, and epigenetic marks. The regulation of p53 expression is crucial for cellular responses to stress, DNA damage, and oncogenic signals, demonstrating the impact of transcriptional control on tumor suppression and cell cycle regulation (Vousden & Prives, 2009; Sullivan, Galbraith, Andrysik, & Espinosa, 2018).

Epigenetic regulation, such as DNA methylation, also has clear implications for disease. For instance, hypermethylation of promoter regions in tumor suppressor genes is a common hallmark in many cancers, leading to transcriptional silencing and tumor progression (Jones & Baylin, 2002; Esteller, 2008). In neurological disorders, aberrant histone modifications have been linked to diseases like Rett syndrome, where mutations in the MECP2 gene result in improper regulation of transcription, altering neuronal gene expression and leading to developmental defects (Chahrour & Zoghbi, 2007; Lyst & Bird, 2015).

Advancements in sequencing technologies have significantly enhanced our understanding of transcriptional regulation, with each method offering unique insights into various aspects of the transcription process. RNA sequencing (RNA-Seq), one of the most widely used techniques, detects the results of transcriptional regulation by measuring gene expression levels across different conditions, tissues, or cell types. By comparing RNA-Seq data from treated versus untreated samples, researchers can identify differential gene expression and infer regulatory changes caused by transcription factors, enhancers, or other regulatory elements.

Additionally, RNA-Seq can be combined with other techniques to further dissect the mechanisms of transcriptional control. For example, Chromatin immunoprecipitation followed

by sequencing (ChIP-Seq) is instrumental in identifying genome-wide binding sites of transcription factors and mapping histone modifications, revealing the regulatory landscapes that influence transcription (Johnson, Mortazavi, Myers, & Wold, 2007; Farnham, 2009). When combined with RNA-Seq, ChIP-Seq can link transcription factor binding to downstream gene expression changes, providing a more complete picture of transcriptional regulation. ATAC-Seq (Assay for Transposase-Accessible Chromatin) offers insights into chromatin accessibility, helping researchers' study how open or closed chromatin states correlate with transcriptional activation or repression (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013; Corces et al., 2017). It provides information about the physical state of chromatin, identifying regions of open chromatin that are often associated with active promoters, enhancers, or insulators. Integrating ATAC-Seq with RNA-Seq allows researchers to correlate chromatin accessibility changes with gene expression outcomes, showing how dynamic chromatin states influence transcription. RNA Polymerase II ChIP-Seq and PRO-Seq (Precision Run-On sequencing) have enabled fine mapping of transcription initiation and elongation, offering a detailed view of active transcription sites and elongation rates across the genome (Core, Waterfall, & Lis, 2008; Kwak, Fuda, Core, & Lis, 2013). Through these techniques, researchers can monitor how RNA polymerase activity correlates with gene expression patterns, revealing mechanisms such as promoter-proximal pausing or transcriptional elongation regulation.

Emerging technologies like Hi-C and CUT&RUN sequencing have further illuminated the spatial organization of chromatin and its role in transcriptional regulation. Hi-C reveals the three-dimensional architecture of chromatin, showing how enhancers can physically interact with distant promoters to regulate gene expression (Lieberman-Aiden et al., 2009; Rao et al., 2014)

Finally, single-cell technologies like single-cell RNA-Seq (scRNA-Seq), single-cell ATAC-Seq and spatio-temporal single-cell sequencings integrates spatial and temporal dimensions with gene expression dynamics, allowing researchers to map transcriptional changes across time and space within tissues. For example, spatio-temporal scRNA-Seq has been applied to study cell lineage trajectories, tumor evolution, and neuronal differentiation, revealing how

transcriptional regulation drives developmental processes and disease progression at single-cell resolution (Farrell et al., 2018; Sathe et al., 2020).

In summary, transcriptional regulation in eukaryotes involves a complex network of *cis*-elements, *trans*-factors, epigenetic modifications, and chromatin dynamics. These mechanisms ensure that gene expression is precisely controlled in response to developmental cues, environmental stimuli, and cellular signals. The regulatory outcomes of these processes influence cell identity, adaptation, and disease progression, highlighting the importance of transcriptional control in shaping phenotypic diversity and maintaining cellular homeostasis (Spitz & Furlong, 2012).

Posttranscriptional regulation of gene expression

Posttranscriptional regulation represents a crucial layer of gene expression control in eukaryotes, adding complexity to the transcriptome by generating diverse mRNA isoforms and influencing mRNA stability, localization, and translation efficiency. The most prominent mechanism in this regulatory layer is alternative splicing, where pre-mRNA can be spliced in multiple ways, producing different transcript isoforms from a single gene. This variability significantly expands the diversity of both the transcriptome and proteome, enabling different protein products with distinct functions to be produced in a cell-type or developmental stage-specific manner. The regulation of alternative splicing is mediated by a complex interplay between RNA-binding proteins (RBPs), RNA structure, and chromatin state: RBPs such as SR proteins (Serine/Arginine-rich proteins) and hnRNPs (heterogeneous nuclear ribonucleoproteins) bind to cis-regulatory sequences, either promoting or inhibiting spliceosome assembly. The secondary structure of pre-mRNA can influence splicing outcomes by sequestering or exposing cis-elements, affecting the binding of splicing factors. Alternative splicing is often co-regulated with transcription, where the rate of RNA polymerase II elongation can affect splice site choice. Slow elongation rates allow more time for upstream weak splice sites to be recognized, while fast rates favor downstream strong splice sites (Kornblihtt et al., 2013; Baralle, Singh, & Stamm,

2019). Other mechanisms, including alternative transcription initiation, RNA editing, and alternative polyadenylation, further contribute to transcriptome diversity, and each can be regulated by cis-regulatory elements, trans-acting factors, or environmental signals (Zhao et al., 2017; Tian & Manley, 2017).

Alternative splicing not only drives developmental processes but is also involved in generating phenotypic variation and disease outcomes. For instance, in *Drosophila*, the Dscam1 gene undergoes extensive alternative splicing to potentially produce over 38,000 distinct mRNA isoforms, enabling specific synaptic connections crucial for neural wiring (Schmucker et al., 2000). In humans, splicing of the PTBP1 gene results in isoforms that either promote neuronal differentiation or maintain pluripotency in stem cells, illustrating the impact of splicing on cellular identity (Linares et al., 2015). Aberrant splicing patterns contribute to diseases such as spinal muscular atrophy (SMA), where mutations affect splicing of the SMN2 gene. Therapeutic approaches like Spinraza enhance the production of functional SMN protein and improve patient outcomes, demonstrating how regulating splicing can treat genetic diseases (Singh & Singh, 2018). In cancer, the oncogene Bcl-x undergoes alternative splicing to produce both pro-apoptotic (Bcl-xS) and anti-apoptotic (Bcl-xL) isoforms, influencing cell survival and tumor progression (Mercatante, Bortner, Cidlowski, & Kole, 2001).

mRNA stability also plays a key role in posttranscriptional regulation. The untranslated regions (UTRs) of mRNA contain cis-acting elements that regulate translation efficiency, localization, and degradation rates. Cis-elements, like AU-rich elements (AREs), and RNA-binding proteins (RBPs) (e.g., HuR and tristetraprolin (TTP)), play key roles in stabilizing or destabilizing mRNA, with their activity often controlled by post-translational modifications. mRNA decay impacts cellular processes like stress response, immune signaling, and homeostasis, and it adjusts rapidly in response to stimuli, the cell cycle, or intracellular signals. AU-rich elements (AREs) in the 3' UTRs of mRNAs contribute to ARE-mediated decay, which is crucial in immune response regulation by modulating the stability of cytokine mRNAs like TNF-α and IL-6. These mRNAs are usually rapidly degraded to prevent excessive inflammation

but are stabilized under stress or infection (Kontoyiannis, Pasparakis, Pizarro, Cominelli, & Kollias, 1999). Dysregulation of this process can lead to chronic inflammation and cancer due to the prolonged stability of mRNAs that drive persistent cytokine expression and disease progression (Stumpo, Lai, & Blackshear, 2010). In diseases like rheumatoid arthritis or systemic lupus erythematosus (SLE), impaired mRNA decay results in excessive cytokine levels, highlighting the therapeutic potential of targeting mRNA stability to resolve inflammation (Uchida, Chiba, Kurimoto, & Asahara, 2019).

Nonsense-mediated decay (NMD) functions as a quality control pathway while modulating the levels of specific mRNAs. It plays a role in regulating the expression of genes involved in metabolism and immune responses, such as MHC class I proteins, which are crucial for antigen presentation in immune cells (Lejeune, 2022; Lindeboom, Supek, & Lehner, 2016). Dysregulation of NMD can contribute to diseases like cancer, where the accumulation of aberrant proteins contributes to tumorigenesis (Popp & Maquat, 2016).

Advancements in sequencing technologies have significantly advanced our understanding of posttranscriptional regulation. Traditional short-read RNA-seq, though instrumental for detecting alternative splicing events, has limitations in identifying full-length isoforms and understanding splicing complexity. Third-generation sequencing technologies, such as PacBio Iso-Seq and Oxford Nanopore sequencing, allow for full-length cDNA sequencing, providing a more comprehensive view of the transcriptome. These technologies have improved the detection of novel splicing variants, RNA editing events, and alternative polyadenylation, complementing short-read data and offering a detailed picture of posttranscriptional modifications (Tilgner, Grubert, Sharon, & Snyder, 2014; Wyman & Mortazavi, 2019). PacBio Iso-Seq uses circular consensus sequencing to generate high-accuracy, full-length cDNA sequences, revealing alternative splicing events, RNA editing, and alternative polyadenylation with greater precision. Oxford Nanopore sequencing is similar but offers real-time sequencing capabilities. It detects RNA modifications, such as N6-methyladenosine (m6A), directly from RNA molecules, allowing researchers to study RNA modifications and their impact on mRNA stability and

translation. The field of epitranscriptomics has gained attention for adding another layer of posttranscriptional control, involving RNA modifications such as N6-methyladenosine (m6A). These modifications can regulate splicing, translation, and RNA stability. For example, m6A modifications impact the export of mRNA from the nucleus, modulate alternative splicing by recruiting specific RNA-binding proteins, and regulate mRNA degradation (Zhao et al., 2017; Roundtree, Evans, Pan, & He, 2017).

Ribosome Profiling (Ribo-Seq) is an invaluable tool for studying post-transcriptional regulation, as it provides a genome-wide view of translation dynamics by identifying ribosome-protected mRNA fragments (Ingolia, Ghaemmaghami, Newman, & Weissman, 2009; Ingolia, 2016). This technique reveals how gene expression is regulated at the translational level by tracking ribosome positions along mRNAs, offering insights into translation initiation, elongation, and termination (Brar & Weissman, 2015). It has been particularly effective in analyzing elongation dynamics, where variations in speed arise from factors like codon usage, tRNA availability, and mRNA structure, which can influence co-translational events such as protein folding and targeting (Radhakrishnan & Green, 2016; Huch & Nissan, 2014).

Ribo-Seq also aids in identifying translation initiation sites, including non-canonical starts like upstream open reading frames (uORFs), which serve as key regulatory elements in stress responses and other cellular processes (Calviello & Ohler, 2017). Combining Ribo-Seq with complementary techniques such as mRNA-Seq enables researchers to distinguish between changes in mRNA abundance and translational regulation, providing a clearer understanding of how gene expression is controlled at both transcriptional and translational levels. Moreover, Ribo-Seq reveals novel translated regions, including those derived from long non-coding RNAs (lncRNAs), highlighting previously unknown protein-coding potential and expanding our understanding of the functional translatome (van Heesch et al., 2019). This technique has proven particularly valuable in studying the translational response to stress, where cells rapidly reprogram translation to adapt to changes like hypoxia, nutrient deprivation, or viral infection (Zhou et al., 2018). Complementary approaches like polysome profiling, which examines

ribosome density across mRNAs, and RNC-Seq, which focuses on ribosome-nascent chain complexes, provide additional insights into translation rates, efficiency, and co-translational folding (T. Wang et al., 2013).

To conclude, the complex nature of post-transcriptional regulation is pivotal to gene expression, driven by mechanisms like alternative splicing, mRNA stability, and translation dynamics. Advanced sequencing techniques, including Ribo-Seq, PacBio Iso-Seq, and Oxford Nanopore, have greatly expanded our understanding of these regulatory processes, revealing their roles in cellular function, development, and disease. These advancements offer a comprehensive view of how cells fine-tune gene expression beyond transcription, showcasing the intricate layers of regulation that maintain cellular diversity and adaptability.

Genetic and environmental effects on gene expression regulation

Gene expression is shaped not only by genetic factors but also by environmental influences, affecting every stage of the gene expression pathway. Genetic variants—including single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variations—contribute to gene expression variation by altering regulatory elements and impacting transcription, splicing, translation, and posttranslational modifications. For instance, *cis*-acting expression quantitative trait loci (*cis*-eQTLs) near promoters or enhancers can modify transcription factor binding sites, influencing transcription initiation rates, while *trans*-eQTLs can alter the expression of distal genes by affecting transcription factors or signaling molecules (Albert & Kruglyak, 2015; Consortium, 2020).

Environmental factors—including internal changes like developmental stages, aging, and hormonal fluctuations, as well as external stimuli such as temperature, diet, stress, and pathogen exposure—play significant roles in gene expression regulation. Developmental cues drastically alter transcriptional programs, leading to stage-specific gene expression patterns essential for differentiation and tissue specification. During the transition from embryonic to fetal development, distinct changes in transcriptional and translational profiles occur, driven by

growth factors and hormones acting as internal environmental signals (Nord et al., 2013). Aging is also associated with changes in chromatin structure and gene expression, often resulting in increased variability in gene expression among individuals, particularly in genes related to inflammation and metabolic processes (Stegeman & Weake, 2017; Benayoun et al., 2019).

External environmental factors influence gene expression at multiple regulatory levels. For instance, heat shock proteins (HSPs) are rapidly upregulated in response to elevated temperatures, demonstrating a transcriptional response that helps cells cope with stress by maintaining protein folding and function (Lindquist, 1986; Akerfelt, Morimoto, & Sistonen, 2010). Exposure to UV radiation can activate DNA repair genes, while nutrient availability affects the translation of mRNAs involved in metabolism via pathways like mTOR, which integrates signals from amino acids and glucose to regulate global protein synthesis (Hay & Sonenberg, 2004; Saxton & Sabatini, 2017).

Environmental factors also play a role in posttranscriptional regulation. For instance, hypoxia (low oxygen conditions) alters splicing patterns of genes involved in angiogenesis, allowing for adaptive responses that promote blood vessel growth (Mazzone et al., 2009). During cellular stress, mRNA stability is often modulated by stress-induced RNA-binding proteins (RBPs) or microRNAs (miRNAs), affecting the half-life of stress-responsive mRNAs like ATF4 or HIF1A (Kilberg, Shan, & Su, 2009). Translational regulation can also be affected by environmental changes; for instance, nutrient deprivation triggers phosphorylation of eIF2α, reducing global translation but selectively enhancing translation of mRNAs with upstream open reading frames (uORFs), such as those encoding stress-response proteins (Hinnebusch, 2011; Pakos-Zebrucka et al., 2016).

Research into gene-environment interactions has been facilitated by recent advances in sequencing technologies. RNA sequencing (RNA-seq) combined with genotyping has been employed to identify eQTLs that vary under different environmental conditions, revealing context-specific gene expression changes (Montgomery & Dermitzakis, 2011). ATAC-seq and

ChIP-seq have provided insights into how environmental factors influence chromatin states and transcriptional regulation, while ribosome profiling (Ribo-seq) has enabled the study of translation dynamics under different stress conditions, such as nutrient deprivation or hypoxia (Ingolia et al., 2009; Brar & Weissman, 2015). Single-cell RNA-seq (scRNA-seq) and multi-omics approaches have further facilitated the exploration of gene expression changes during development, aging, and environmental responses, providing a comprehensive view of gene-environment interactions across life stages (Tang et al., 2009; Lopez-Otin, Blasco, Partridge, Serrano, & Kroemer, 2013; Ziegenhain et al., 2017).

In summary, gene expression is dynamically shaped by both genetic and environmental factors. Developmental stages, aging, and external stimuli all contribute to phenotypic diversity and adaptation. Understanding how these factors interact is essential for elucidating complex traits, evolutionary adaptations, and disease mechanisms, underscoring the importance of gene-environment interactions in shaping cellular functions and organismal phenotypes.

Cis- and *trans*-regulatory effects on gene expression

Gene expression is regulated by both *cis*- and *trans*-regulatory elements across all stages of the gene expression pathway, from transcription to posttranslational modification. *Cis*-regulatory elements include promoters, enhancers, and untranslated regions (UTRs), which act locally by interacting with proteins or non-coding RNAs, thereby influencing transcription initiation, splicing, mRNA stability, and translation efficiency (Levine & Tjian, 2003). In contrast, *trans*-regulatory factors include diffusible proteins such as transcription factors, RNA-binding proteins (RBPs), and microRNAs (miRNAs), which can regulate multiple genes across the genome. Together, these regulatory effects drive phenotypic diversity within species and evolutionary divergence between species (Wittkopp, Haerum, & Clark, 2008; Signor & Nuzhdin, 2018).

Mapping cis- and trans-regulatory effects is fundamental for understanding gene expression variation, with two primary approaches being F1 hybrid mapping and expression

quantitative trait loci (eQTL) mapping. In F1 hybrids, which inherit alleles from two different parental strains or species, differences in allele-specific expression (ASE) can be used to identify cis- and trans-regulatory effects. Cis effects are detected when one parental allele consistently shows higher or lower expression than the other within the same cellular environment, indicating sequence differences in local regulatory elements. Conversely, trans effects are inferred when both alleles have similar expression in the F1 hybrid but differ between the parental strains, suggesting that variations in trans-acting factors drive the expression differences (Wittkopp, Haerum, & Clark, 2004). Technologies such as RNA-seq have been widely used to measure ASE at the transcriptional level, while other sequencing approaches extend this analysis to additional regulatory layers. For instance, ribosome profiling (Ribo-seq) has been applied to F1 hybrids to detect allele-specific translation, revealing how cis-regulatory differences not only affect mRNA abundance but also translation efficiency (Ingolia et al., 2009; Artieri & Fraser, 2014). Similarly, ATAC-seq has been used to assess allele-specific chromatin accessibility, identifying cis effects on chromatin states, while ChIP-seq reveals how cis-regulatory variants influence transcription factor binding (Buenrostro et al., 2013; Core et al., 2014).

eQTL mapping identifies genetic loci associated with gene expression variation across a population, distinguishing between *cis*- and *trans*-eQTLs. *Cis*-eQTLs are typically located near the gene they regulate, often within promoters or enhancers, leading to direct changes in transcription or splicing. In contrast, *trans*-eQTLs are located farther away and influence gene networks by altering the expression or function of *trans*-acting factors such as transcription factors or RBPs (Albert & Kruglyak, 2015; Consortium, 2020). eQTL mapping has traditionally relied on RNA-seq or microarrays to measure gene expression levels across individuals, combined with genotyping to identify genetic variants linked to expression changes. Recent advances such as single-cell RNA-seq (scRNA-seq) provide insights into eQTLs at the single-cell level, enabling fine-scale mapping of gene expression variation in complex tissues (Tang et al., 2009). Additionally, multi-omics approaches—integrating RNA-seq with ATAC-seq or ChIP-seq—have refined eQTL mapping by linking genetic variants to chromatin accessibility or

transcription factor binding (Battle et al., 2014; Furey, 2012). For example, in human populations, eQTL mapping has uncovered both *cis*- and *trans*-eQTLs associated with complex traits such as asthma and diabetes, highlighting how genetic variation shapes gene expression (Gamazon et al., 2018). In crops like maize, eQTL mapping has identified both *cis*- and *trans*-eQTLs that influence stress response genes, shedding light on gene-environment interactions in plant adaptation (Kremling et al., 2018).

Several other approaches complement F1 hybrid mapping and eQTL mapping in distinguishing cis- and trans-regulatory effects. The common reference design involves using a pooled reference RNA sample to normalize expression differences between two biological samples, facilitating the detection of both cis- and trans-effects. While it lacks the precision of allele-specific methods, it is useful for broad comparisons of gene expression across individuals or species (Landry et al., 2005). Allele-specific chromatin accessibility (ASCA), employing techniques such as ATAC-seq or DNase-seq in heterozygotes or F1 hybrids, can reveal how cis-regulatory variants affect chromatin states, while the absence of allele-specific differences suggests trans-regulatory influences (Degner et al., 2012; Bryois et al., 2018). CRISPR/Cas9 editing has also been applied to introduce specific mutations into cis-elements or trans-acting genes, providing direct evidence of their regulatory roles. For example, editing a promoter sequence can confirm a cis-effect, while altering a transcription factor gene can uncover trans-effects (Gasperini et al., 2019). Massively parallel reporter assays (MPRAs) allow thousands of cis-regulatory sequences to be tested simultaneously for their effects on gene expression, primarily revealing cis-effects but also providing insights into trans-effects when tested in different cellular backgrounds (Tewhey et al., 2016).

In summary, cis- and trans-regulatory effects influence gene expression across all stages, driving phenotypic diversity and evolutionary changes within and between species. Techniques such as F1 hybrid mapping, eQTL mapping, and complementary methods like ASCA, CRISPR editing, and MPRAs provide comprehensive insights into these regulatory mechanisms. By leveraging advanced sequencing technologies, researchers can better understand the complex

interplay of cis and trans-regulation, revealing how genetic and environmental factors shape gene expression dynamics.

Recent advances in flies

Drosophila has long been a model organism in genetic and molecular biology research, recognized for its many advantages. Its short life cycle, high reproductive rate, and genetic tractability make it ideal for studying gene expression regulation. The species' well-annotated genome, coupled with resources like the *Drosophila* Genetic Reference Panel (DGRP), comprising over 200 inbred lines from a natural population of *D. melanogaster*, enhances its utility (Mackay et al., 2012). Additionally, projects like "The *Drosophila* Genome Nexus" and "101 Drosophila Genomes" have expanded genomic resources, providing insights into global genetic diversity and regulatory variation (Lack et al., 2015; Kim et al., 2021) (Lack et al., 2015; Kim et al., 2021). These advancements have been pivotal in identifying both cis- and trans-regulatory variants, enabling detailed mapping of gene expression traits across developmental, physiological, and evolutionary contexts. Combined with advancements in sequencing and gene-editing technologies, these resources make *Drosophila* an unparalleled model for understanding gene expression regulation.

Recent studies have revealed detailed maps of cis-regulatory elements (CREs) using high-throughput techniques like ChIP-seq and ATAC-seq, which have refined our understanding of enhancer dynamics across developmental stages and tissues. For instance, a genome-wide enhancer activity mapping method called STARR-seq was applied to *D. melanogaster* to enable direct, quantitative identification of enhancers, revealing complex transcriptional regulation with multiple independent enhancers per gene (Arnold et al., 2013). Furthermore, single-cell RNA sequencing (scRNA-seq) has provided unprecedented resolution of transcriptional states, helping to identify lineage-specific transcription factors and their regulatory roles in cell differentiation (Li, 2021).

Post-transcriptional regulation in *Drosophila* has been increasingly dissected through techniques like CLIP-seq, which maps RNA-protein interactions, revealing how RNA-binding proteins influence splicing, stability, and localization of mRNAs (Hansen et al., 2015). Recent advances in long-read sequencing technologies, such as PacBio and Nanopore, have enabled the detection of full-length isoforms, uncovering complex alternative splicing patterns that affect developmental transitions (Zhang, Bae, Cuddleston, & Miura, 2023). Recent developments in spatial transcriptomics provides high-resolution 3D transcriptomic maps of *Drosophila* embryos and larvae. It captures spatial gene expression patterns, enabling detailed analyses of gene regulatory networks, tissue-specific dynamics, and cell state changes during development. (M. Wang et al., 2022).

REFERENCES

Cramer, P. (2019). Organization and regulation of gene transcription. *Nature, 573*(7772), 45-54. doi:10.1038/s41586-019-1517-4

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet, 13*(7), 484-492. doi:10.1038/nrg3230

Allis, C. D., & Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nat Rev Genet, 17*(8), 487-500. doi:10.1038/nrg.2016.59

Clapier, C. R., & Cairns, B. R. (2009). The biology of chromatin remodeling complexes. *Annu Rev Biochem, 78*, 273-304. doi:10.1146/annurev.biochem.77.062706.153223

Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature, 463*(7280), 457-463. doi:10.1038/nature08909

Darnell, J. E., Jr. (2013). Reflections on the history of pre-mRNA processing and highlights of current knowledge: a unified picture. *RNA, 19*(4), 443-460. doi:10.1261/rna.038596.113

Kornblihtt, A. R., Schor, I. E., Allo, M., Dujardin, G., Petrillo, E., & Munoz, M. J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol, 14*(3), 153-165. doi:10.1038/nrm3525

Baralle, F. E., & Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol, 18*(7), 437-451. doi:10.1038/nrm.2017.27

Zhao, B. S., Roundtree, I. A., & He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. *Nat Rev Mol Cell Biol, 18*(1), 31-42. doi:10.1038/nrm.2016.132

Yang, Y., Hsu, P. J., Chen, Y. S., & Yang, Y. G. (2018). Dynamic transcriptomic m(6)A decoration: writers, erasers, readers and functions in RNA metabolism. *Cell Res, 28*(6), 616-624. doi:10.1038/s41422-018-0040-8

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., . . . Rechavi, G. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature, 485*(7397), 201-206. doi:10.1038/nature11112

Hinnebusch, A. G., & Lorsch, J. R. (2012). The mechanism of eukaryotic translation initiation: new insights and challenges. *Cold Spring Harb Perspect Biol, 4*(10). doi:10.1101/cshperspect.a011544

Gingold, H., & Pilpel, Y. (2011). Determinants of translation efficiency and accuracy. *Mol Syst Biol, 7*, 481. doi:10.1038/msb.2011.14

Mann, M., & Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nat Biotechnol, 21*(3), 255-261. doi:10.1038/nbt0303-255

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat Rev Genet, 16*(4), 197-212. doi:10.1038/nrg3891

Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature, 424*(6945), 147-151. doi:10.1038/nature01763

Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet, 13*(9), 613-626. doi:10.1038/nrg3207

Alexander, T., Nolte, C., & Krumlauf, R. (2009). Hox genes and segmentation of the hindbrain and axial skeleton. *Annu Rev Cell Dev Biol, 25*, 431-456. doi:10.1146/annurev.cellbio.042308.113423

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., . . . Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell, 129*(7), 1311-1323. doi:10.1016/j.cell.2007.05.022

Vousden, K. H., & Prives, C. (2009). Blinded by the Light: The Growing Complexity of p53. *Cell, 137*(3), 413-431. doi:10.1016/j.cell.2009.04.037

Sullivan, K. D., Galbraith, M. D., Andrysik, Z., & Espinosa, J. M. (2018). Mechanisms of transcriptional regulation by p53. *Cell Death Differ, 25*(1), 133-143. doi:10.1038/cdd.2017.174

Jones, P. A., & Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nat Rev Genet, 3*(6), 415-428. doi:10.1038/nrg816

Esteller, M. (2008). Epigenetics in cancer. *N Engl J Med, 358*(11), 1148-1159. doi:10.1056/NEJMra072067

Chahrour, M., & Zoghbi, H. Y. (2007). The story of Rett syndrome: from clinic to neurobiology. *Neuron, 56*(3), 422-437. doi:10.1016/j.neuron.2007.10.001

Lyst, M. J., & Bird, A. (2015). Rett syndrome: a complex disorder with simple roots. *Nat Rev Genet, 16*(5), 261-275. doi:10.1038/nrg3897

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science, 316*(5830), 1497-1502. doi:10.1126/science.1141319

Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. *Nat Rev Genet, 10*(9), 605-616. doi:10.1038/nrg2636

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods, 10*(12), 1213-1218. doi:10.1038/nmeth.2688

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., . . . Chang, H. Y. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods, 14*(10), 959-962. doi:10.1038/nmeth.4396

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science, 322*(5909), 1845-1848. doi:10.1126/science.1162228

Kwak, H., Fuda, N. J., Core, L. J., & Lis, J. T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science, 339*(6122), 950-953. doi:10.1126/science.1229386

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science, 326*(5950), 289-293. doi:10.1126/science.1181369

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., . . . Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell, 159*(7), 1665-1680. doi:10.1016/j.cell.2014.11.021

Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A., & Schier, A. F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science, 360*(6392). doi:10.1126/science.aar3131

Sathe, A., Grimes, S. M., Lau, B. T., Chen, J., Suarez, C., Huang, R. J., . . . Ji, H. P. (2020). Single-Cell Genomic Characterization Reveals the Cellular Reprogramming of the Gastric Tumor Microenvironment. *Clin Cancer Res, 26*(11), 2640-2653. doi:10.1158/1078-0432.CCR-19-3231

Baralle, F. E., Singh, R. N., & Stamm, S. (2019). RNA structure and splicing regulation. *Biochim Biophys Acta Gene Regul Mech, 1862*(11-12), 194448. doi:10.1016/j.bbagrm.2019.194448

Tian, B., & Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol, 18*(1), 18-30. doi:10.1038/nrm.2016.116

Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., . . . Zipursky, S. L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell, 101*(6), 671-684. doi:10.1016/s0092-8674(00)80878-8

Linares, A. J., Lin, C. H., Damianov, A., Adams, K. L., Novitch, B. G., & Black, D. L. (2015). The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. *Elife, 4*, e09268. doi:10.7554/eLife.09268

Singh, R. N., & Singh, N. N. (2018). Mechanism of Splicing Regulation of Spinal Muscular Atrophy Genes. *Adv Neurobiol, 20*, 31-61. doi:10.1007/978-3-319-89689-2_2

Mercatante, D. R., Bortner, C. D., Cidlowski, J. A., & Kole, R. (2001). Modification of alternative splicing of Bcl-x pre-mRNA in prostate and breast cancer cells. analysis of apoptosis and cell death. *J Biol Chem, 276*(19), 16411-16417. doi:10.1074/jbc.M009256200

Kontoyiannis, D., Pasparakis, M., Pizarro, T. T., Cominelli, F., & Kollias, G. (1999). Impaired on/off regulation of TNF biosynthesis in mice lacking TNF AU-rich elements: implications for joint and gut-associated immunopathologies. *Immunity, 10*(3), 387-398. doi:10.1016/s1074-7613(00)80038-2

Stumpo, D. J., Lai, W. S., & Blackshear, P. J. (2010). Inflammation: cytokines and RNA-based regulation. *Wiley Interdiscip Rev RNA, 1*(1), 60-80. doi:10.1002/wrna.1

Uchida, Y., Chiba, T., Kurimoto, R., & Asahara, H. (2019). Post-transcriptional regulation of inflammation by RNA-binding proteins via cis-elements of mRNAs. *J Biochem, 166*(5), 375-382. doi:10.1093/jb/mvz067

Lejeune, F. (2022). Nonsense-Mediated mRNA Decay, a Finely Regulated Mechanism. *Biomedicines, 10*(1). doi:10.3390/biomedicines10010141

Lindeboom, R. G., Supek, F., & Lehner, B. (2016). The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat Genet, 48*(10), 1112-1118. doi:10.1038/ng.3664

Popp, M. W., & Maquat, L. E. (2016). Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine. *Cell, 165*(6), 1319-1322. doi:10.1016/j.cell.2016.05.053

Tilgner, H., Grubert, F., Sharon, D., & Snyder, M. P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc Natl Acad Sci U S A, 111*(27), 9869-9874. doi:10.1073/pnas.1400447111

Wyman, D., & Mortazavi, A. (2019). TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics, 35*(2), 340-342. doi:10.1093/bioinformatics/bty483

Roundtree, I. A., Evans, M. E., Pan, T., & He, C. (2017). Dynamic RNA Modifications in Gene Expression Regulation. *Cell, 169*(7), 1187-1200. doi:10.1016/j.cell.2017.05.045

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science, 324*(5924), 218-223. doi:10.1126/science.1168978

Ingolia, N. T. (2016). Ribosome Footprint Profiling of Translation throughout the Genome. *Cell, 165*(1), 22-33. doi:10.1016/j.cell.2016.02.066

Brar, G. A., & Weissman, J. S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol, 16*(11), 651-664. doi:10.1038/nrm4069

Radhakrishnan, A., & Green, R. (2016). Connections Underlying Translation and mRNA Stability. *J Mol Biol, 428*(18), 3558-3564. doi:10.1016/j.jmb.2016.05.025

Huch, S., & Nissan, T. (2014). Interrelations between translation and general mRNA degradation in yeast. *Wiley Interdiscip Rev RNA, 5*(6), 747-763. doi:10.1002/wrna.1244

Calviello, L., & Ohler, U. (2017). Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. *Trends Genet, 33*(10), 728-744. doi:10.1016/j.tig.2017.08.003

van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J. F., Adami, E., Faber, A. B., . . . Hubner, N. (2019). The Translational Landscape of the Human Heart. *Cell, 178*(1), 242-260 e229. doi:10.1016/j.cell.2019.05.010

Zhou, J., Wan, J., Shu, X. E., Mao, Y., Liu, X. M., Yuan, X., . . . Qian, S. B. (2018). N(6)-Methyladenosine Guides mRNA Alternative Translation during Integrated Stress Response. *Mol Cell, 69*(4), 636-647 e637. doi:10.1016/j.molcel.2018.01.019

Wang, T., Cui, Y., Jin, J., Guo, J., Wang, G., Yin, X., . . . Zhang, G. (2013). Translating mRNAs strongly correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. *Nucleic Acids Res, 41*(9), 4743-4754. doi:10.1093/nar/gkt178

Consortium, G. T. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science, 369*(6509), 1318-1330. doi:10.1126/science.aaz1776

Nord, A. S., Blow, M. J., Attanasio, C., Akiyama, J. A., Holt, A., Hosseini, R., . . . Visel, A. (2013). Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell, 155*(7), 1521-1531. doi:10.1016/j.cell.2013.11.033

Stegeman, R., & Weake, V. M. (2017). Transcriptional Signatures of Aging. *J Mol Biol, 429*(16), 2427-2437. doi:10.1016/j.jmb.2017.06.019

Benayoun, B. A., Pollina, E. A., Singh, P. P., Mahmoudi, S., Harel, I., Casey, K. M., . . . Brunet, A. (2019). Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses. *Genome Res, 29*(4), 697-709. doi:10.1101/gr.240093.118

Lindquist, S. (1986). The heat-shock response. *Annu Rev Biochem, 55*, 1151-1191. doi:10.1146/annurev.bi.55.070186.005443

Akerfelt, M., Morimoto, R. I., & Sistonen, L. (2010). Heat shock factors: integrators of cell stress, development and lifespan. *Nat Rev Mol Cell Biol, 11*(8), 545-555. doi:10.1038/nrm2938

Hay, N., & Sonenberg, N. (2004). Upstream and downstream of mTOR. *Genes Dev, 18*(16), 1926-1945. doi:10.1101/gad.1212704

Saxton, R. A., & Sabatini, D. M. (2017). mTOR Signaling in Growth, Metabolism, and Disease. *Cell, 168*(6), 960-976. doi:10.1016/j.cell.2017.02.004

Mazzone, M., Dettori, D., de Oliveira, R. L., Loges, S., Schmidt, T., Jonckx, B., . . . Carmeliet, P. (2009). Heterozygous deficiency of PHD2 restores tumor oxygenation and inhibits metastasis via endothelial normalization. *Cell, 136*(5), 839-851. doi:10.1016/j.cell.2009.01.020

Kilberg, M. S., Shan, J., & Su, N. (2009). ATF4-dependent transcription mediates signaling of amino acid limitation. *Trends Endocrinol Metab, 20*(9), 436-443. doi:10.1016/j.tem.2009.05.008

Hinnebusch, A. G. (2011). Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiol Mol Biol Rev, 75*(3), 434-467, first page of table of contents. doi:10.1128/MMBR.00008-11

Pakos-Zebrucka, K., Koryga, I., Mnich, K., Ljujic, M., Samali, A., & Gorman, A. M. (2016). The integrated stress response. *EMBO Rep, 17*(10), 1374-1395. doi:10.15252/embr.201642195

Montgomery, S. B., & Dermitzakis, E. T. (2011). From expression QTLs to personalized transcriptomics. *Nat Rev Genet, 12*(4), 277-282. doi:10.1038/nrg2969

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., . . . Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods, 6*(5), 377-382. doi:10.1038/nmeth.1315

Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2013). The hallmarks of aging. *Cell, 153*(6), 1194-1217. doi:10.1016/j.cell.2013.05.039

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., . . . Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell, 65*(4), 631-643 e634. doi:10.1016/j.molcel.2017.01.023

Wittkopp, P. J., Haerum, B. K., & Clark, A. G. (2008). Regulatory changes underlying expression differences within and between Drosophila species. *Nat Genet, 40*(3), 346-350. doi:10.1038/ng.77

Signor, S. A., & Nuzhdin, S. V. (2018). The Evolution of Gene Expression in cis and trans. *Trends Genet, 34*(7), 532-544. doi:10.1016/j.tig.2018.03.007

Wittkopp, P. J., Haerum, B. K., & Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature, 430*(6995), 85-88. doi:10.1038/nature02698

Artieri, C. G., & Fraser, H. B. (2014). Evolution at two levels of gene expression in yeast. *Genome Res, 24*(3), 411-421. doi:10.1101/gr.165522.113

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet, 46*(12), 1311-1320. doi:10.1038/ng.3142

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., . . . Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res, 24*(1), 14-24. doi:10.1101/gr.155192.113

Furey, T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet, 13*(12), 840-852. doi:10.1038/nrg3306

Gamazon, E. R., Segre, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., . . . Ardlie, K. G. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat Genet, 50*(7), 956-967. doi:10.1038/s41588-018-0154-4

Kremling, K. A. G., Chen, S. Y., Su, M. H., Lepak, N. K., Romay, M. C., Swarts, K. L., . . . Buckler, E. S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature, 555*(7697), 520-523. doi:10.1038/nature25966

Landry, C. R., Wittkopp, P. J., Taubes, C. H., Ranz, J. M., Clark, A. G., & Hartl, D. L. (2005). Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. *Genetics, 171*(4), 1813-1822. doi:10.1534/genetics.105.047449

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J. B., Gaffney, D. J., Pickrell, J. K., . . . Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature, 482*(7385), 390-394. doi:10.1038/nature10808

Bryois, J., Garrett, M. E., Song, L., Safi, A., Giusti-Rodriguez, P., Johnson, G. D., . . . Crawford, G. E. (2018). Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. *Nat Commun, 9*(1), 3121. doi:10.1038/s41467-018-05379-y

Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., . . . Shendure, J. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell, 176*(1-2), 377-390 e319. doi:10.1016/j.cell.2018.11.029

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., . . . Sabeti, P. C. (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell, 165*(6), 1519-1529. doi:10.1016/j.cell.2016.04.027

Mackay, T. F., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., . . . Gibbs, R. A. (2012). The *Drosophila melanogaster* Genetic Reference Panel. *Nature, 482*(7384), 173-178. doi:10.1038/nature10811

Lack, J. B., Cardeno, C. M., Crepeau, M. W., Taylor, W., Corbett-Detig, R. B., Stevens, K. A., . . . Pool, J. E. (2015). The Drosophila genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics, 199*(4), 1229-1241. doi:10.1534/genetics.115.174664

Kim, B. Y., Wang, J. R., Miller, D. E., Barmina, O., Delaney, E., Thompson, A., . . . Petrov, D. A. (2021). Highly contiguous assemblies of 101 drosophilid genomes. *Elife, 10*. doi:10.7554/eLife.66405

Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science, 339*(6123), 1074-1077. doi:10.1126/science.1232542

Li, H. (2021). Single-cell RNA sequencing in Drosophila: Technologies and applications. *Wiley Interdiscip Rev Dev Biol, 10*(5), e396. doi:10.1002/wdev.396

Hansen, H. T., Rasmussen, S. H., Adolph, S. K., Plass, M., Krogh, A., Sanford, J., . . . Christiansen, J. (2015). Drosophila Imp iCLIP identifies an RNA assemblage coordinating F-actin formation. *Genome Biol, 16*(1), 123. doi:10.1186/s13059-015-0687-0

Zhang, Z., Bae, B., Cuddleston, W. H., & Miura, P. (2023). Coordination of alternative splicing and alternative polyadenylation revealed by targeted long read sequencing. *Nat Commun, 14*(1), 5506. doi:10.1038/s41467-023-41207-8

Wang, M., Hu, Q., Lv, T., Wang, Y., Lan, Q., Xiang, R., . . . Liu, L. (2022). High-resolution 3D spatiotemporal transcriptomic maps of developing Drosophila embryos and larvae. *Dev Cell, 57*(10), 1271-1283 e1274. doi:10.1016/j.devcel.2022.04.006

CHAPTER 2: INTRASPECIFIC HYBRID ALLELE-SPECIFIC MAPPING EFFICIENTLY DETERMINES MODE OF REGULATORY VARIATION IN *DROSOPHILA MELANOGASTER*

ABSTRACT

Gene expression regulation is modulated by both genetic and environmental factors, leading to variation in regulatory effects across individuals in a population. This study employs an F1 hybrid allele-specific mapping approach to partition regulatory variation into cis- and trans-regulatory effects in intraspecific reciprocal crosses of *Drosophila melanogaster* inbred lines. Whole body RNA sequencing was performed for females from both parental lines and their F1 hybrids, enabling the identification of allele-specific expression patterns. Results showed a predominance of cis regulatory effects over trans regulatory effects across all crosses, corroborating previous findings within *Drosophila* populations. Additionally, the concordance between hybrid mapping and expression quantitative trait loci (eQTL) mapping was notably strong for cis effects but weaker for trans effects. The study demonstrates that allele-specific mapping not only effectively characterizes the regulatory architecture but also provides a cost-efficient alternative to population-scale eQTL mapping, offering potential insights into gene-environment interactions and rare variant effects. These findings highlight the utility of hybrid designs for elucidating complex gene regulatory mechanisms.

INTRODUCTION

Among phenotypically divergent individuals between and within species there is substantial regulatory variation of gene expression that involves changes in abundance and spatiotemporal distribution without structural alterations (King and Wilson 1975; Lemos et al. 2008; Wittkopp et al. 2008). Regulatory variation can arise due to genetic and environmental perturbations to the regulatory programs of gene expression. Environmental perturbations may broadly include stage of development and differentiation, physiological and disease status, and external stimuli. For example, expression of heat shock protein 70 (hsp70) in Drosophila cells is rapidly induced upon exposure to high temperature (Spradling et al. 1975). Genetic perturbations, on the other hand, are ultimately due to mutations in DNA sequences that are heritable. For example, sequence variation in the *yellow* gene in different Drosophila species leads to gene expression divergence and ultimately pigmentation variation (Wittkopp et al. 2002).

According to their underlying molecular mechanisms, regulatory genetic variation can be classified into two modes, cis and trans regulatory variation. Cis regulatory variation alters DNA sequences that have regulatory potential for genes in physical proximity. For example, mutations in the *yellow* gene in Drosophila lead to the gain of multiple transcription factor binding sites, thus enhancing *yellow* transcription and leading to male specific wing spot pigmentation (Gompel et al. 2005). Trans regulatory variation affects diffusible factors such as transcription factors, microRNAs, and RNA binding proteins that can act on distal genes. For example, mutations in a single D-MEF2 transcription factor can cause myogenesis deficiency in larval and adult Drosophila (Ranganayakulu et al. 1995). Previous studies have invariably shown that cis regulatory variation accounts for the majority of intra- and inter-specific gene expression differences (Wittkopp et al. 2004; Tirosh et al. 2009), whereas contribution of trans regulatory variation appears to be larger between species than within species (Lemos et al. 2008; Osada et al. 2017).

There are two major strategies to genetically characterize regulatory variation. In diploid organisms, allele specific expression (ASE) is a widespread phenomenon where one allele is preferentially expressed over the other (Ge et al. 2009; Crowley et al. 2015). In heterozygous individuals ASE is due to cis regulatory effects because both alleles share the same trans-acting factors and cellular environment whereas in inbred parents differential expression is due to both cis and trans regulatory effects. Therefore, comparison of allelic abundance of RNA transcripts in F1 hybrids and in parents can distinguish cis and trans regulatory effects. This hybrid allele specific mapping approach can also be used to identify allele-specific epigenetic effects such as parent of origin effects (Takada et al. 2017; Floc'hlay et al. 2021).

Expression quantitative trait loci (eQTL) mapping is another strategy to identify regulatory effects. This strategy relies on large populations with genetically divergent individuals in the same species. In humans, large scale eQTL mapping has identified eQTLs that are shared between or specific to different tissues (GTEx Consortium et al. 2017). In Drosophila, eQTL mapping studies have also found environmentally plastic regulatory variation (Cannavò et al. 2017; Huang et al. 2020). These eQTL mapping studies either focused on cis-eQTLs only or found cis-eQTLs to be more prevalent, suggesting that cis-regulatory variation was more prominent within species. While the F1 hybrid design is able to identify genes under cis- and/or trans regulation, eQTL mapping can provide information on sequence variation associated with the regulatory effects.

In this study, we apply the F1 hybrid allele specific mapping design to characterize the regulatory landscape of gene expression in reciprocal crosses of inbred lines from the Drosophila melanogaster Genetic Reference Panel (DGRP), where eQTLs have also been mapped previously. Although many previous studies have applied the hybrid allele specific mapping approach to characterize regulatory landscape in flies, this study is the first to do so in the presence of eQTL mapping information in the same population. This allows us to directly compare these two approaches in their abilities to characterize regulatory variation.

RESULTS

Counting allele-specific RNA-Seq Reads

To identify genes under the control of cis versus trans regulation, we reciprocally crossed four pairs of DGRP lines that were chosen to be maximally divergent in exons. Within each pair, we sequenced whole-body RNA of 50 pooled 3-5 day old females for both parental lines and both reciprocal F1 crosses, each with two biological replicates. To eliminate mapping bias (Stevenson et al. 2013), we constructed line specific genomes by substituting variant sites in the BDGP6 reference genome with alleles of the DGRP lines. Reference annotations were also modified according to cumulative coordinate changes by indels between the reference genome and the line-specific genomes. For each RNA-Seq sample, we mapped sequence reads to both parental line specific genomes and assigned each read to one of the parents according to the alleles it carried. We sequenced on average 22.6M 150 bp single-end reads, of which 75.4% could be uniquely mapped to either genome. Between 23.95% and 26.07% of the uniquely mapped reads overlapped with informative DNA variant sites, which allowed us to classify them according to their parental origin. Remarkably, we were able to classify the vast majority (>97%) of reads derived from the parental samples correctly to their own genome (**Figure 2.1**), suggesting that the bioinformatic method was effective. In hybrids, except for one cross (Figure 1a, 138x819), approximately 50% of reads were assigned to either parent, consistent with the expectation the parental genomes should on average produce an equal amount of RNA transcripts. We retained three crosses where the hybrids had a ratio of parental reads within the range of 43-57% (**Figure 2.1**b-d).

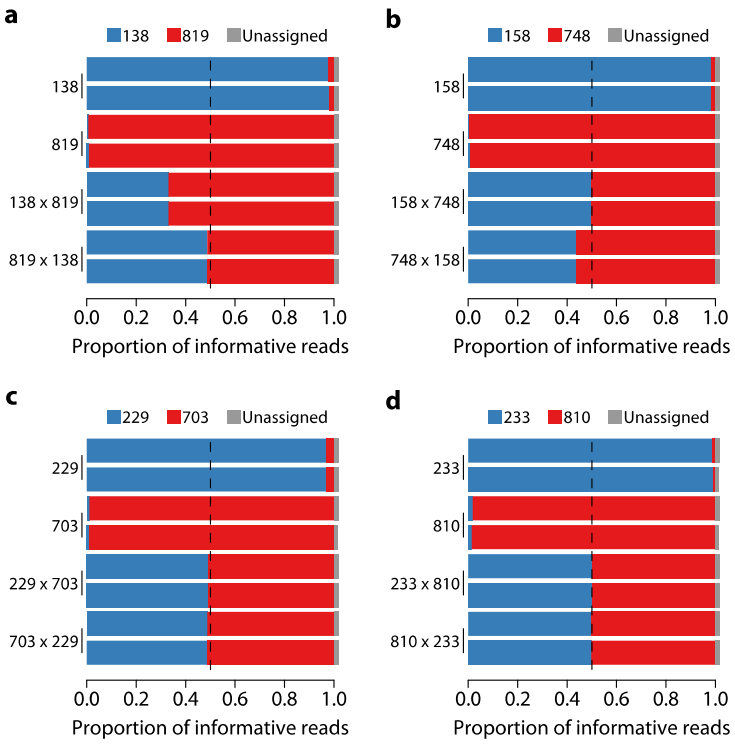


Figure 2.1 Mapping reads to parental genomes. RNA sequence reads were mapped to line-specific genomes. Informative reads that overlap variant sites were assigned to one of the two parental genomes. Proportion of informative reads over total assigned reads was plotted for each sample. Proportion of unassigned reads was plotted as a fraction of total assigned reads. (a) Crosses between 138 and 819; (b) 158 and 748; (c) 229 and 703; (d) 233 and 810.

Identification of genes with cis or trans regulatory effects

To identify genes under the influence of cis and/or trans regulatory effects, we fitted generalized linear models comparing read counts derived from the parental alleles in the inbred parents and the F1 hybrids using edgeR (Robinson et al. 2010), which models read counts distributed as a negative binomial distribution. Because different crosses have different informative variant sites, the models were fitted to all eight biological replicates (two for each of the parents and two for each of the reciprocal crosses), but separately for each of the three pairs of lines. This method takes advantage of edgeR's generalized linear model framework and its flexible parameterization to quickly identify cis and trans effects. First, if the two alleles are differentially expressed in the F1 hybrids, cis regulatory effects are called. Second, if there is a

difference in the allelic effects between parental strains and F1 hybrids, trans regulatory effects are present. It is important to note that trans effects can enhance the allelic effects or compensate the changes. Although there are more complex classifications of regulatory effects, we chose to focus on these two easily interpretable effects.

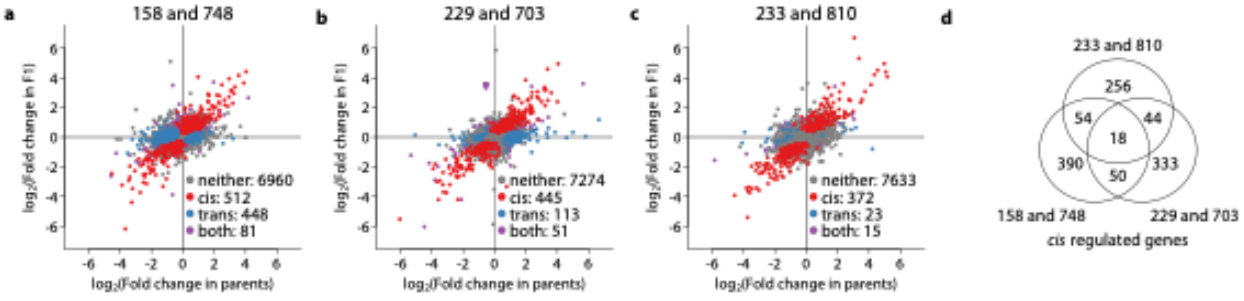


Figure 2.2 Characterizing regulatory architecture by hybrid allele-specific mapping. Cis and trans regulatory effects were called (FDR = 0.05) in each of the three reciprocal crosses (a – c). (d) Overlap between gene identities of cis regulated genes identified in each cross.

We identified between 512, 445, 372 cis regulated genes in crosses between 158 and 748, 229 and 703, and 233 and 810 respectively (**Figure 2.2**a-c) as well as 448, 113, 23 trans regulated genes, and 81, 51, and 15 that were regulated by both mechanisms. In general, there were more cis regulated genes than trans regulated genes and cis effects were larger, consistent with previous studies within and between Drosophila species (Wittkopp et al. 2004; Osada et al. 2017). Overlap between genes identified in different crosses was moderate (**Figure 2.2**d). However, it is difficult to attribute the lack of overlap to genetic reasons because not all pairs of crosses have the same divergence between cis regulatory elements.

No evidence for widespread parent-of-origin effect in gene expression

Our reciprocal design allowed us to identify parent-of-origin effects, if any. We tested for allelic differences (maternal versus paternal) in both of the reciprocal crosses and look for differences that were persistent in both crosses. No genes were significantly differentially expressed in the same parental direction in both crosses by more than two-folds in all three pairs

of crosses (**Figure 2.3**). Consistent with previous work (Lyko et al. 2000), this analysis indicated that parent-of-origin effect in steady state RNA abundance was not a widespread phenomenon.

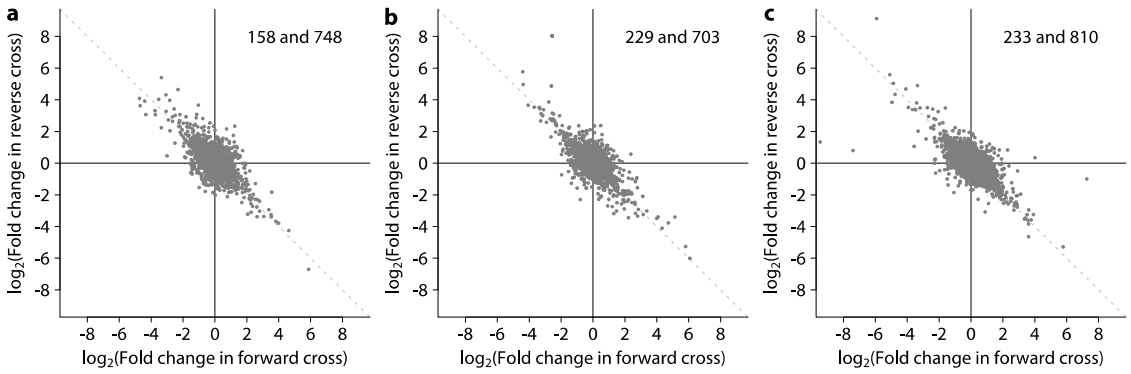


Figure 2.3 Parent-of-origin effects in reciprocal crosses. Log fold changes of maternal versus paternal in forward (male parent is the first of the indicated pair) and reverse (female parent is the first of the indicated pair) crosses are plotted against each other. Significant effects that are of the same sign indicate consistent maternal versus paternal effect or parent-of-origin effect.

Hybrid allele-specific mapping largely recapitulates cis-eQTL mapping effects

We have demonstrated that hybrid allele-specific mapping is able to partition the genome into those that are under *cis* or *trans* regulation in gene expression. To further evaluate its effectiveness and accuracy, we compared hybrid allele-specific mapping to eQTL mapping, which was performed in the DGRP including all lines used in this study (Tan et al. 2024 under review). We predicted gene expression based on the sum of eQTL allelic effects each line carried, which represented the total effects of all mapped eQTLs. Most genes had fewer than five eQTLs. This allowed us to compare fold changes of gene expression between a pair of DGRP lines based on 1) difference between inbred parents; 2) difference between the parental haplotypes within the F1 hybrid; 3) difference between eQTL predicted parental line gene expression.

First, we compared fold changes between eQTL (both *cis* and *trans*) predicted gene expression in the parental lines with observed fold changes. Strong concordance was found (correlation between 0.51 and 0.56, **Figure 2.4**) in all three pairs of lines, suggesting that eQTL

effects were highly reproducible. Interestingly, there were genes that showed no difference based on eQTL effects but large differences in the parental lines. These may be due to rare variants that were not tested in the eQTL mapping.

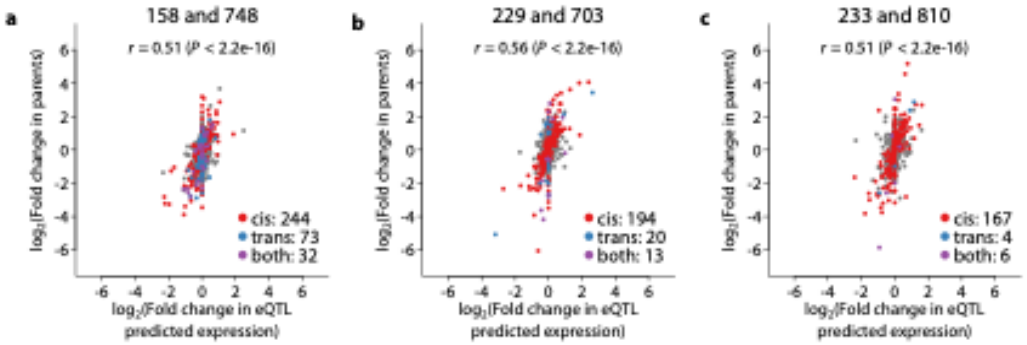


Figure 2.4 Comparison between eQTL predicted fold changes and observed fold changes.

Second, we compared fold changes between *cis*-eQTL predicted gene expression in the parental lines with fold changes between the alleles in the F1 hybrids for genes that showed at least some *cis* differences in the hybrid allele-specific mapping. Remarkably, *cis*-eQTL predicted expression was highly correlated with *cis* effects estimated in the F1 hybrids (correlation between 0.64 and 0.73,**Figure 2.5**). Not surprisingly, when restricted to genes with mapped *cis* effects in F1 hybrids, the correlations were substantially higher than comparing eQTL predicted expression and observed parental expression. This result suggested that hybrid allele-specific mapping largely recapitulates *cis*-eQTL mapping effects.

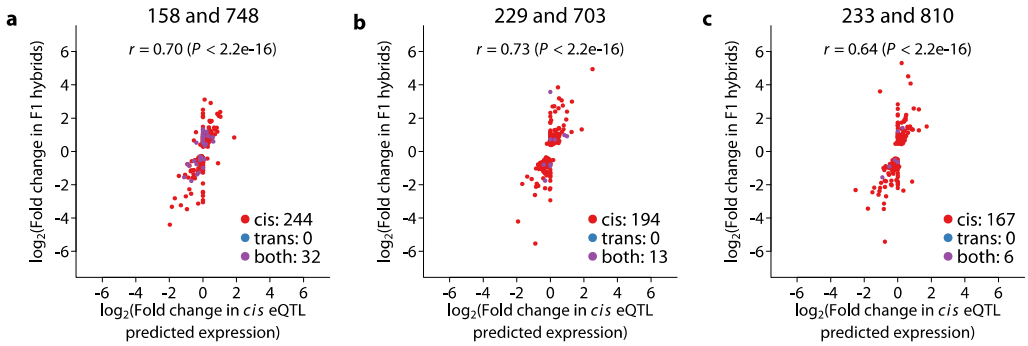


Figure 2.5 Comparison between cis-eQTL predicted fold changes and observed fold changes in F1 hybrids.

Finally, we compared fold changes between *trans*-eQTL predicted gene expression in the parental lines with fold changes between the parental for genes that showed at least some *trans* differences in the hybrid allele-specific mapping. The correlations, for much fewer genes, were much weaker (**Figure 2.6**), suggesting that *trans* effects were less conserved across lines.

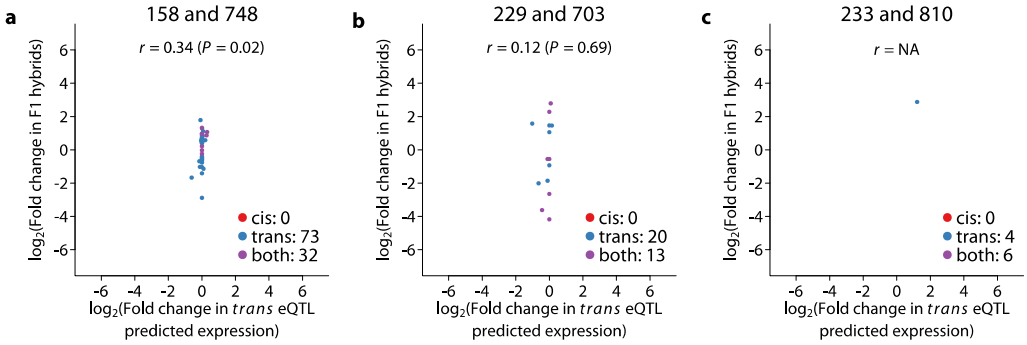


Figure 2.6 Comparison between trans-eQTL predicted fold changes and observed fold changes in parental lines.

Taken together, these results suggested that hybrid allele-specific mapping is able to largely recapitulate *cis* eQTL mapping performed at a population scale.

DISCUSSION

We performed one of the few intra-specific hybrid allele-specific mapping studies in *Drosophila melanogaster*. Our results indicated the approach was highly effective, partitioning the Drosophila genome into cis and trans regulatory effects in specific pairs of inbred strains. More importantly, the allelic effects estimated in hybrid allele-specific mapping largely agreed with those estimated from a population scale eQTL mapping study, especially for cis eQTLs. This is important, because it allows characterization of regulatory variation to be performed in a much more cost-effective and efficient way.

Our method of estimating allele specific expression using read counts as opposed to SNP allele counts (van de Geijn et al. 2015) was novel and reduced reference bias. Hybrids produced roughly 50% of reads originating from either parents. The read count based approach is able to leverage existing RNA-Seq analytical framework and avoids integrating results across SNPs, which can be challenging.

There are at least two advantages for such an effective strategy. First, application of this approach across treatments and conditions can be used to characterize interactions between regulatory variation and environmental factors as opposed to applying environmental treatments to many inbred strains, which may be cost prohibitive (Chapter 3 of this thesis applies this strategy). Moreover, eQTL mapping requires high frequency of mutations to be properly tested. On the other hand, hybrid allele-specific mapping can work with rare and even private mutations, a significant advantage of this approach. Indeed, there were many genes whose eQTLs cannot be mapped but significant cis effects were observed in hybrids (**Figure 2.5**).

METHODS

DGRP lines cross and RNA-Seq

Four pairs of DGRP lines were reciprocally crossed: line 138 and line 819, line 158 and line 748, line 229 and line 703, line 233 and line 810. RNA-Seq was performed as previously described (Huang et al. 2020). Briefly, female 3-5 day mated adults of the parental lines and F1 hybrids were collected and whole body RNA was extracted from pool of 50 flies. The total RNA was subjected to Illumina stranded mRNA sequencing by 125bp single-end. An average of 30.66 million reads and 60.02 million reads were sequenced for inbred lines and hybrid lines respectively.

RNA-Seq mapping

Line specific reference genomes of involved DGRP lines were constructed by modifying the *Drosophila melanogaster* reference genome (BDGP6) using variants called in the Drosophila melanogaster Genetic Reference Panel (Huang et al. 2014). For each cross, sequenced reads of parents and F1 were mapped to both line specific reference genomes. Reads overlapping variants between the parental strains were retained and assigned to either genome based on matches. We required that the reads must map better to one of the parental genomes than the other based on numbers of mismatches and otherwise considered the reads ambiguous. We were able to uniquely assign 23.95% to 26.07% of reads (**Table 2.1**), indicating high diversity in the DGRP to apply the hybrid design. Importantly, among reads that overlapped informative SNPs, the proportions of ambiguous reads were typically less than 2%, indicating strong performance of the read assignment procedure. Ratios of reads from two lines in most of hybrid lines were approximately 1:1. Deviation of the ratio from 1:1 indicates possible contamination and one of our four pairs were removed due to this reason.

Cis and *trans* effect identification

Genome-wide read counts were analyzed using edgeR (Robinson et al. 2010). Read counts were normalized and fitted to generalized linear models implemented in edgeR that

models read counts as negative binomial distributions. We test for differences between alleles of genes using the model counts ~ allele + cross + cross:allele, where counts is the response variable for read counts, allele represents origin of the alleles from which read counts are derived, cross represents where the counts are derived from F1 hybrid or parents and the interaction between the two variables. Significant allele effect when the cross is F1 indicates significant *cis* effect, while significant interaction between cross and allele indicates *trans* effect. This allows us to partition the genome into genes that have 1) *cis* effect only, 2) *trans* effect only, 3) both, and 4) neither.

To test for parent-of-origin effect, differential expression was called within either direction of the reciprocal cross and significant differential expression in both directions of crosses but of the same sign was considered parent-of-origin effect.

Finally, we used eQTL mapping results obtained from another study where all DGRP lines were profiled for gene expression using the same RNA-Seq procedure and eQTLs were mapped for all common variants (MAF > 0.05). eQTLs were partitioned into *cis* (< 5kb from transcription start site) and *trans* and used to predict gene expression in the parental lines using estimated eQTL effects. The predicted gene expression was compared with observed expression in the parental lines as well as estimated *cis* effects in the F1 hybrids.

REFERENCES

Cannavò E, Koelling N, Harnett D, Garfield D, Casale FP, Ciglar L, Gustafson HE, Viales RR, Marco-Ferreres R, Degner JF, et al. 2017. Genetic variants regulating expression levels and isoform diversity during embryogenesis. Nature. 541(7637):402–406. doi:10.1038/nature20802.

Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, Kim Y, Wang JR, Morgan AP, Calaway JD, Aylor DL, et al. 2015. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat Genet. 47(4):353–360. doi:10.1038/ng.3222.

Floc'hlay S, Wong ES, Zhao B, Viales RR, Thomas-Chollier M, Thieffry D, Garfield DA, Furlong EEM. 2021. Cis-acting variation is common across regulatory layers but is often buffered during embryonic development. Genome Res. 31(2):211–224. doi:10.1101/gr.266338.120.

Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KCL, Gagné V, et al. 2009. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nat Genet. 41(11):1216–1222. doi:10.1038/ng.473.

van de Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 12(11):1061–1063. doi:10.1038/nmeth.3582.

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB. 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature. 433(7025):481–487. doi:10.1038/nature03235.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. 2017. Genetic effects on gene expression across human tissues. Nature. 550(7675):204–213. doi:10.1038/nature24277.

Huang W, Carbone MA, Lyman RF, Anholt RRH, Mackay TFC. 2020. Genotype by environment interaction for gene expression in Drosophila melanogaster. Nat Commun. 11(1):5451. doi:10.1038/s41467-020-19131-y.

Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF, et al. 2014. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. 24(7):1193–1208. doi:10.1101/gr.171546.113.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science. 188(4184):107–116. doi:10.1126/science.1090005.

Lemos B, Araripe LO, Fontanillas P, Hartl DL. 2008. Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. Proc Natl Acad Sci U S A. 105(38):14471–14476. doi:10.1073/pnas.0805160105.

Osada N, Miyagi R, Takahashi A. 2017. Cis- and Trans-regulatory Effects on Gene Expression in a Natural Population of Drosophila melanogaster. Genetics. 206(4):2139–2148. doi:10.1534/genetics.117.201459.

Ranganayakulu G, Zhao B, Dokidis A, Molkentin JD, Olson EN, Schulz RA. 1995. A series of mutations in the D-MEF2 transcription factor reveal multiple functions in larval and adult myogenesis in Drosophila. Dev Biol. 171(1):169–181. doi:10.1006/dbio.1995.1269.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26(1):139–140. doi:10.1093/bioinformatics/btp616.

Spradling A, Penman S, Pardue ML. 1975. Analysis of drosophila mRNA by in situ hybridization: sequences transcribed in normal and heat shocked cultured cells. Cell. 4(4):395–404. doi:10.1016/0092-8674(75)90160-9.

Stevenson KR, Coolon JD, Wittkopp PJ. 2013. Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. BMC Genomics. 14:536. doi:10.1186/1471-2164-14-536.

Takada Y, Miyagi R, Takahashi A, Endo T, Osada N. 2017. A Generalized Linear Model for Decomposing Cis-regulatory, Parent-of-Origin, and Maternal Effects on Allele-Specific Gene Expression. G3 (Bethesda). 7(7):2227–2234. doi:10.1534/g3.117.042895.

Tirosh I, Reikhav S, Levy AA, Barkai N. 2009. A yeast hybrid provides insight into the evolution of gene expression regulation. Science. 324(5927):659–662. doi:10.1126/science.1169766.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. Nature. 430(6995):85–88. doi:10.1038/nature02698.

Wittkopp PJ, Haerum BK, Clark AG. 2008. Regulatory changes underlying expression differences within and between Drosophila species. Nat Genet. 40(3):346–350. doi:10.1038/ng.77.

Wittkopp PJ, Vaccaro K, Carroll SB. 2002. Evolution of *yellow* gene regulation and pigmentation in Drosophila. Curr Biol. 12(18):1547–1556. doi:10.1016/s0960-9822(02)01113-2.

Lyko, F., Ramsahoye, B. H., & Jaenisch, R. (2000). DNA methylation in *Drosophila melanogaster*. *Nature*, 408(6812), 538-540.

APPENDIX

samples	Mapped reads	Assigned reads
138_1	11017922	2796900(25.39%)
138_2	11249223	2885285(25.65%)
138x819_1	20114075	5096831(25.34%)
138x819_2	20940982	5290737(25.26%)
158_1	12219168	3166694(25.92%)
158_2	9251009	2404577(25.99%)
158x748_1	23435876	6003262(25.62%)
158x748_2	24252451	6222797(25.66%)
229_1	14482971	3527671(24.36%)
229_2	13311243	3255936(24.46%)
229x703_1	28876456	6916797(23.95%)
229x703_2	29868185	7162258(23.98%)
233_1	10442643	2672663(25.59%)
233_2	12396063	3221797(25.99%)
233x810_1	19962393	5027879(25.19%)
233x810_2	20660782	5208263(25.21%)
703_1	10474606	2571082(24.55%)
703_2	9476565	2254403(23.79%)
703x229_1	22693993	5461213(24.06%)
703x229_2	23497586	5649571(24.04%)
748_1	10114496	2615169(25.86%)
748_2	10206187	2637279(25.84%)
748x158_1	19630915	5037129(25.66%)
748x158_2	20316873	5214217(25.66%)
810_1	11411180	2906068(25.47%)
810_2	11916936	3025834(25.39%)
810x233_1	23355609	5935253(25.41%)
810x233_2	24232357	6141953(25.35%)
819_1	9973306	2558508(25.65%)
819_2	11591755	3022120(26.07%)
819x138_1	23608656	6026538(25.53%)
819x138_2	24518927	6253184(25.50%)

Table 2.1 Percentage of assigned reads in mapped reads for all sequenced samples

CHAPTER 3: HYBRID ALLELE-SPECIFIC MAPPING OF TRANSLATIONAL REGULATORY VARIATION IN DROSOPHILA

ABSTRACT

Gene expression regulation involves intricate mechanisms at multiple levels, including transcription of RNAs and translation of mRNAs to proteins. Variation in gene expression within and between species contribute significantly to phenotypic diversity. While regulatory variation in steady state mRNA level has been well characterized in many populations and species, the genetic variation and evolution of translational control remains largely unexplored. In this study, by combining the hybrid allele-specific mapping design and ribosome profiling (Ribo-Seq), we investigated the regulatory architecture divergence of two Drosophila species. We first developed high quality genome assemblies for *Drosophila melanogaster* and *Drosophila simulans* that had N50 approaching full chromosomes. We then profiled RNA abundance, ribosome occupancy, and translation efficiency in both species and their F1 hybrids at 3rd instar larval stage, and raised under two different temperatures. This allowed us to test for allelic effects that contributed to regulatory variation at multiple levels and partition the genome into those under cis and trans regulation. We found that the majority of allelic difference between the two species can be attributed to cis effects at all of RNA, ribosome, and translation efficiency levels.

INTRODUCTION

Regulatory genetic variation contributes to species divergence. The contribution of regulatory variation in addition to coding variation to species divergence is well established (King and Wilson 1975). Regulatory genetic variation can be broadly divided into those that change cis (proximal) regulatory sequence elements and variants altering trans (distal) acting factors. Although the relative contribution of cis and trans regulatory variation is highly variable across studies, it is generally accepted that both modes of regulatory variation lead to the expression divergence among species and cis regulatory effects tend to be individually larger (Wittkopp et al. 2004). Although gene expression regulation involves complex layers of control, from transcription to translation, the contribution of regulatory variation to species divergence beyond steady state RNA abundance remains largely unknown. Recent studies suggested that buffering between the regulatory layers led to lower divergence in translation than transcription (Wang et al. 2020).

Hybrid allele-specific mapping is a powerful approach for dissecting the contributions of cis- and trans-regulatory effects to gene expression differences. In hybrids, each gene has two alleles—one inherited from each parent—and their expression can be compared directly within the same cellular environment. By comparing allele-specific expression levels in the parental lines and F1 hybrids, we can infer whether observed differences are due to cis-regulatory elements, which are linked to the gene itself, or trans-regulatory elements, which involve diffusible factors that affect multiple genes. This approach allowed us to identify genes that are differentially regulated by cis- or trans-factors, providing a comprehensive understanding of the genetic basis of regulatory divergence. This approach has been applied to study transcriptional regulatory variation between and within Drosophila species (Wittkopp et al. 2004; Wittkopp et al. 2008; Graze et al. 2009; Coolon et al. 2014; Osada et al. 2017) and others as well (Emerson et al. 2010; Osada et al. 2017). The advent of high throughput sequencing made this approach more readily applicable due to the ability of sequence reads to interrogate allele identity and abundance simultaneously.

Post-transcriptional gene expression regulation is another crucial aspect of gene expression. While transcriptional regulation determines the abundance of mRNA, post-transcriptional mechanisms, such as translation efficiency, play a significant role in determining protein levels. Translation efficiency, which refers to the rate at which mRNA is translated into protein, can vary between alleles and is influenced by both cis- and trans-regulatory factors.

Ribosome profiling (Ribo-Seq) is sequencing based approach that turns ribosome occupancy on mRNAs into digital counts by sequencing ribosome protected fragments after nuclease digestion of mRNAs (Ingolia et al. 2009; Dunn et al. 2013). Similar to RNA-Seq, Ribo-Seq also provides information on both allelic identity and abundance, making it amenable to hybrid allele-specific mapping. For example, Ribo-Seq has been applied to hybrids in rice (Zhu et al. 2023) and mice (Hou et al. 2015) to characterize the regulatory landscape of translation.

Cis- and trans-regulatory elements play critical roles in modulating gene expression, contributing to phenotypic variation between species and within hybrids. The study of cis- and trans-effects is crucial for understanding how genetic variation shape gene expression in hybrid individuals. Here, we used hybrids of two closely related species, *Drosophila melanogaster* and *Drosophila simulans*, to investigate regulatory differences in gene expression at transcriptional and translational levels. Our experimental approach utilized RNA-Seq and Ribo-Seq to distinguish between transcriptional and translational regulation, supported by highly contiguous T2T genome assemblies for precise allele-specific analyses.

The availability of high quality genome assemblies for the used strains of *Drosophila melanogaster* and *Drosophila simulans* was a critical aspect of our study. The assemblies provide complete, high-quality genome sequences, enabling accurate mapping of reads to specific alleles and reducing ambiguity in allele-specific analyses. This level of precision is essential for distinguishing between cis- and trans-regulatory effects in hybrids, as it allows us to assign expression differences to either parental allele with confidence. The trio-binning approach used

for assembly ensured that each parental genome was assembled independently, providing a robust foundation for downstream analyses of hybrid gene expression.

High-quality genome assembly and annotation of *Drosophila simulans* and *Drosophila melanogaster*

To assist the hybrid allele specific mapping process, we generated high-quality genome assemblies and annotations for *Drosophila simulans* and *Drosophila melanogaster* using the trio-binning strategy (Koren et al., 2018). The sequencing depth included approximately 51× coverage of PacBio HiFi reads with 99.8% of reads above Q20, 55× coverage of Oxford Nanopore Technologies (ONT) ultra-long reads (maximum read length of 661 kb and N50 of 83.6 kb) in the hybrid (**Figure 3.1**a, b), and 83× and 71× coverage of Illumina 150 PE reads for parental D. *simulans* and *D. melanogaster*, respectively. Importantly, there are two prominent peaks representing the one-copy and two-copy peaks for the F1 hybrid, which provided a strong basis for the trio-binning approach to work (**Figure 3.1**c).

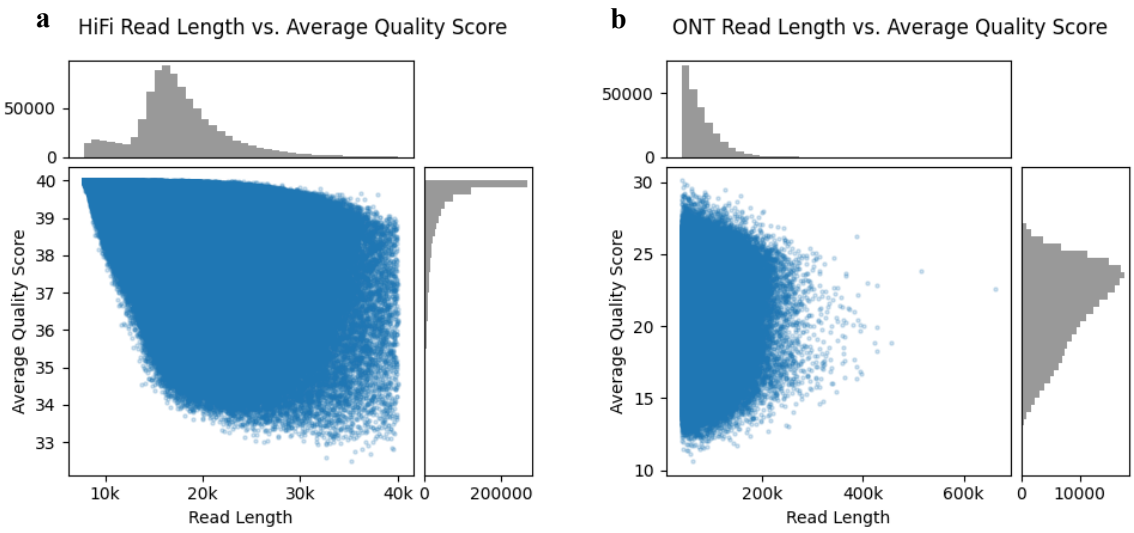
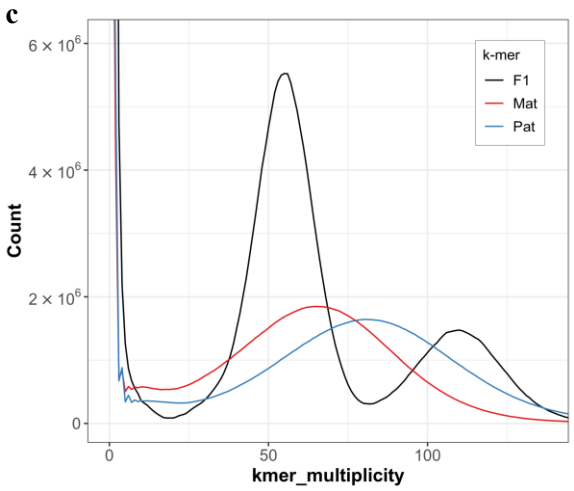


Figure 3.1 Sequence data for genome assembly. (a) Scatter plot of PacBio HiFi reads by length and average read quality. (b) Scatter plot of ONT ultra-long reads by length and average read quality. (c) K-mer frequency plot of paternal (*D. simulans*) DNA-Seq, maternal (*D. melanogaster*) DNA-Seq reads and F1 hybrid HiFi reads.

Figure 3.1 (cont'd)



Assemblies were generated using the trio module implemented in Hifiasm (Cheng et al. 2021) and subsequently polished with NextPolish2 (Hu et al., 2024) to correct for potential errors involving single nucleotide variants and small insertions/deletions (InDels). The assemblies reached chromosome-level resolution without detectable haplotype switch errors (**Figure 3.2**). We evaluated assembly quality using metrics implemented in four software packages: Merqury, BUSCO, QUAST, and Nanopore Pomoxis. The assemblies achieved high quality, as evidenced by a Merqury quality value (QV) exceeding 60, >99.3% coverage of the *Drosophila* gene set, and >92% alignment to published genome references (**Table 3.1**).

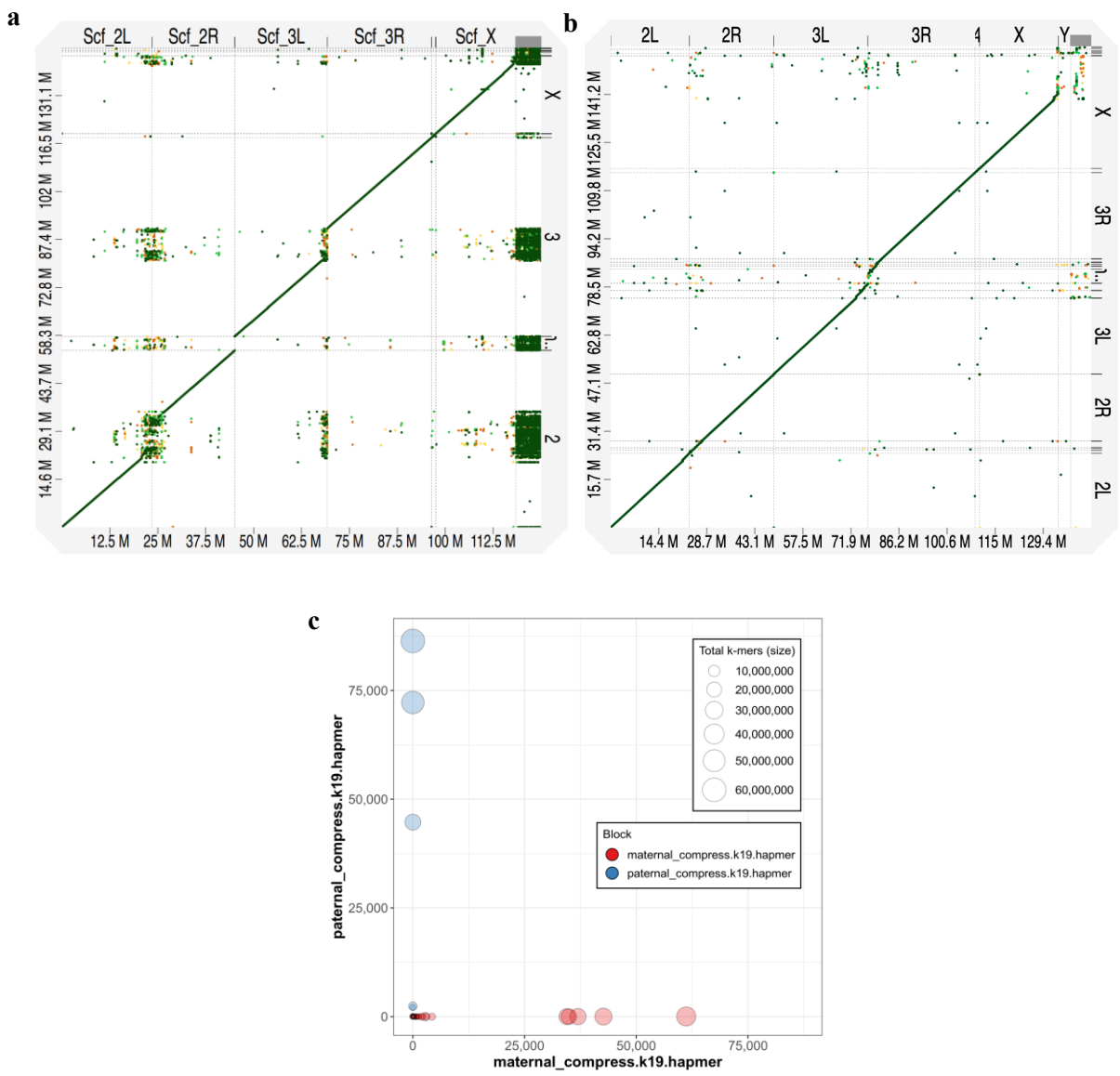


Figure 3.2 Genome assembly quality evaluation. (**a**) *D. simulans* assembly aligns with an existing genome assembly. (**b**) *D. melanogaster* assembly aligns with the existing genome reference (BDGP6). (**c**) Haplotype switch error test for paternal and maternal assemblies.

	D.simulans	*D.melanogaster*
Number of Contigs	18	46
total_bps	145,656,428	156,917,366
longest	60,327,594	36,759,586
median	199,098	159,901
N50	53,690,975	24,690,071
Merqury QV	63.7	61.4
BUSCO	99.3%	99.5%
Genome fraction	94.1%	92.5%

Table 3.1 Summary of genome assemblies and quality evaluation

Decontamination of the assemblies was performed using an integrated approach involving BUSCO (Manni et al., 2021), BLASTp, and FCS (Astashyn et al., 2024). Repetitive regions were predicted and masked using RepeatMasker, while rRNA regions were identified using RNAmmer. Gene annotations from the novel assemblies were aligned with published annotations (Flybase) using Liftoff (Shumate and Salzberg, 2021). To evaluate structural consistency, D-Genies was employed to align the newly assembled genomes to reference genomes using Minimap2, generating a collinearity plot that demonstrated high concordance in scaffolding between the newly assembled genomes and existing references (**Figure 3.2**a, b).

To ensure that the trio-binning approach did not introduce haplotype switch errors, high-quality unique k-mers from the parental genomes were aligned to the assemblies using Merqury. We found undetectable levels of haplotype switch errors, demonstrating the high phasing accuracy of the assemblies (**Figure 3.2**c). The resulting assemblies and annotations were used for downstream data analysis, including the identification of orthologous gene pairs between the two species using Syngap (Wu et al., 2024), which combines collinearity and reciprocal BLAST results.

Sequencing the transcriptome and translatome of *D.melanogaster* and *D.simulans* and their F1 hybrids

To characterize the regulatory architecture of gene expression and the influence of environment, we crossed *D.melanogaster* females (virgin) to *D.simulans* males and collected third instar larvae from both the parental strains and the F1 hybrids. The flies were subjected to two temperature treatment, including the standard 25 °C and a low temperature 20 °C which slows down development. We aimed for four biological replicates for each strain (parental or F1) in each thermal environment. Total RNA was extracted from each sample and sequenced by stranded mRNA-Seq. Matched cytoplasmic lysate was extracted and subjected to Ribo-Seq.

To filter out rRNA, tRNA, snRNA or snoRNA contamination in Ribosome profiling reads, sequenced reads were initially mapped to a combined set of rRNA, tRNA, snRNA, and snoRNA sequences from the Rfam database (Kalvari et al., 2021) using Bowtie (Langmead et al., 2009). Unmapped reads of length 25-34 nt were retained as clean reads for subsequent analyses

RNA-Seq reads and Ribo-Seq ribosome-protected fragments (RPFs) were mapped to a combined *D. simulans* and *D. melanogaster* genome using HISAT2 (Kim et al., 2019). Uniquely mapped reads were phased between the species and used for allele-specific expression analysis. The phasing rates for RNA-Seq reads ranged from 37-41%, while Ribo-Seq reads had phasing rates between 24-31% (**Table 3.2**), confirming the feasibility of using ribosome profiling data for allele-specific mapping studies.

sample type	average reads	informative reads
D.mel RNA	27,095,977	11,107,955(41%)
D.sim RNA	26,531,779	10,785,880(40.7%)
Hybrid RNA	26,963,103	9,881,899(36.6%)
D.mel Ribo	17,110,481	4,017,502(23.5%)
D.sim Ribo	18,660,037	5,705,419(30.6%)
Hybrid Ribo	20,425,331	5,595,723(27.4%)

Table 3.2 Read count summary for RNA-Seq and Ribosome profiling

All uniquely mapped and phased reads were counted for all coding sequences (CDS regions) for RPFs and RNA-Seq reads in both species. For samples derived from the parental species, the mapped reads exhibited nearly 100% species-specificity for both RNA-Seq and Ribo-Seq while hybrid samples displayed a roughly 50:50 distribution (Figure 3.4). Allele-specific translation efficiency was calculated as the TPM of RPFs in CDS regions divided by the TPM of genes.

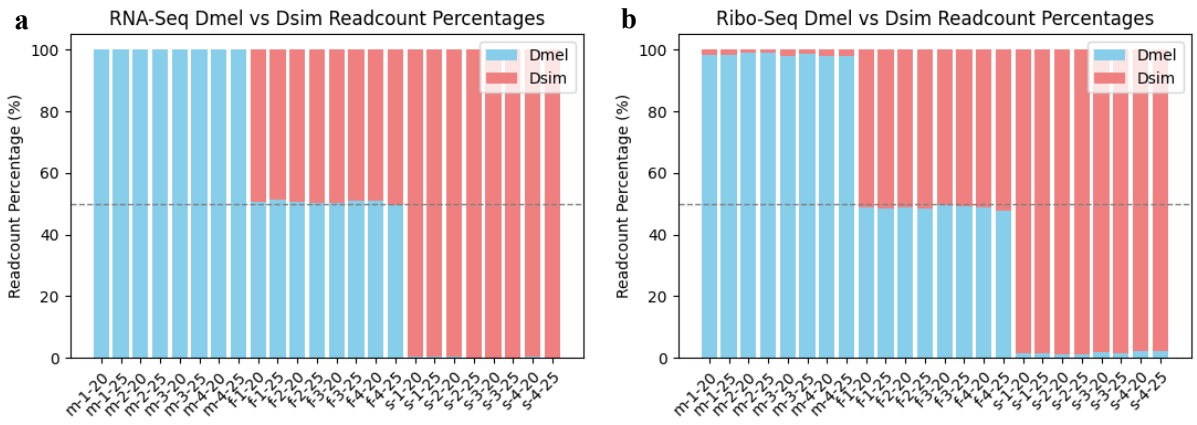


Figure 3.3 Allele-specific read phasing in RNA-Seq and Ribo-Seq. Barplots showing proportions of reads mapped to either genome in RNA-Seq (a) and Ribo-Seq (b). The sample labels follow the format of species-replicate-temperature. m=*melanogaster*, s=*simulans*.

When plotted along the chromosomes, there were roughly equal coverage for the *melanogaster* and *simulans* alleles (**Figure 3.4**) and coverage from RFPs were similar to mRNA, suggesting that the primary determinants of RFP coverage was still transcription. However, there were also differences in RFP and mRNA that may represent regulation at the post-transcriptional level.

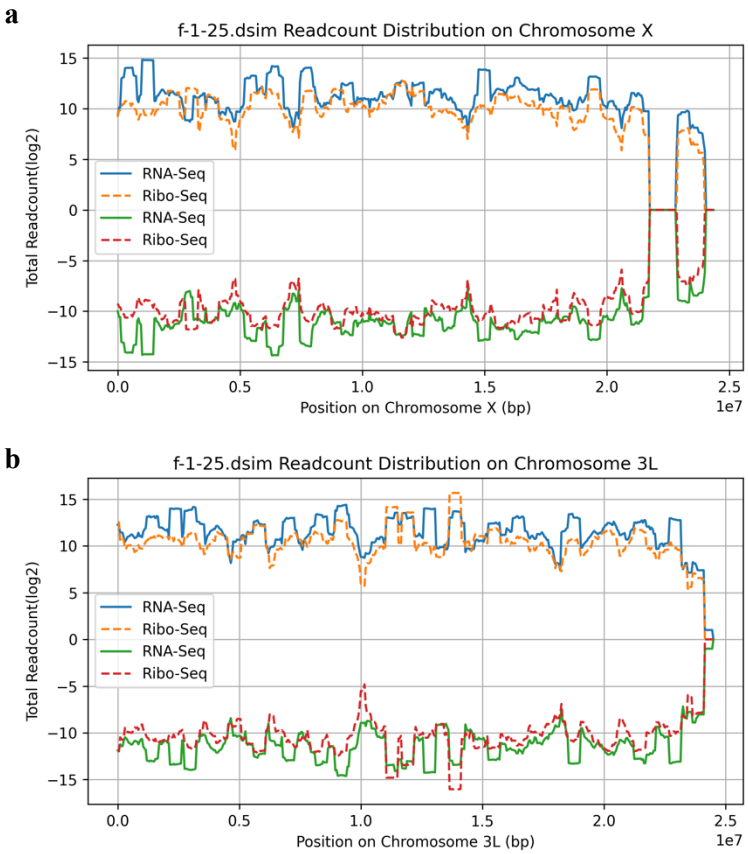


Figure 3.4 Read count distribution on chromosome of one representative hybrid sample F-1-25. Distribution of phased RNA-Seq reads and Ribo-Seq reads of hybrid on Chromosome X (a) and Chromosome 3L(b). The lines above and below zero represent the *melanogaster* and *simulans* genome derived reads respectively.

To evaluate the quality and reproducibility of using mRNA-Seq and Ribo-Seq to estimate allele specific RNA abundance and ribosome occupancy, we obtained RNA expression and RFP expressed as transcripts per million reads (TPMs) and compared them between different biological replicates, species, and temperature treatments. Remarkably, the samples fell into distinct clusters for RNA expression, primarily based on parental strains versus F1 hybrids and secondarily based on origin of species within hybrids (**Error! Reference source not found.**a). T here was no clustering between samples treated with the same temperature environment suggesting that at a global level, temperature does not change RNA abundance dramatically, at least less so than species divergence. Furthermore, biological replicates were highly correlated with each other (**Error! Reference source not found.**b), suggesting high reproducibility.

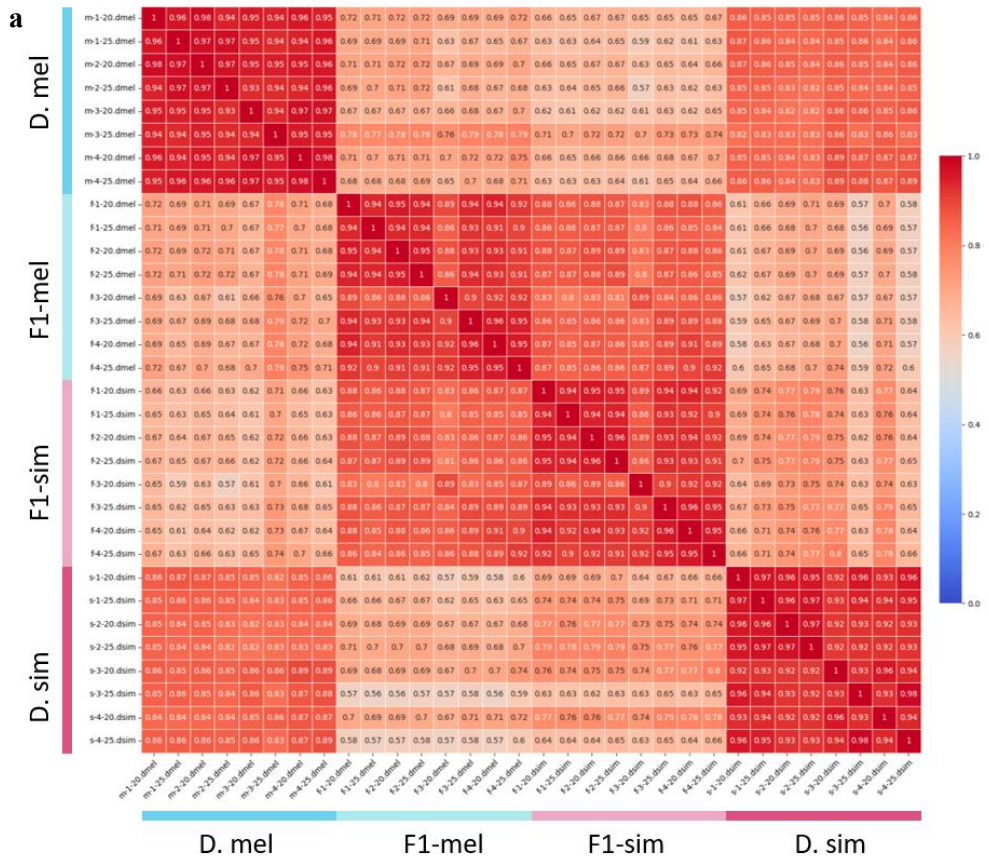
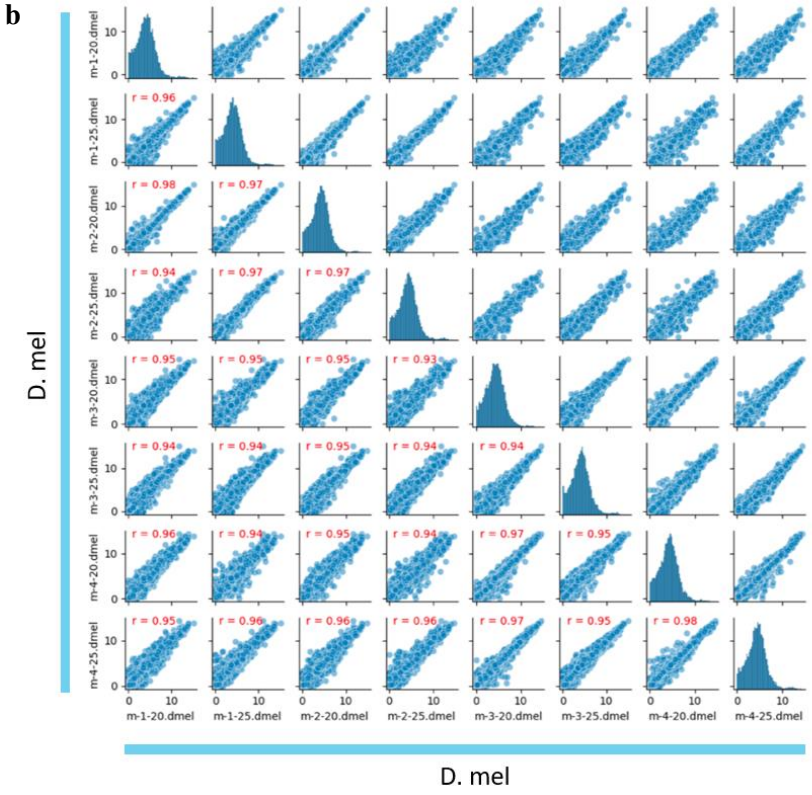


Figure 3.5 Quality and reproducibility of RNA abundance estimation. (a) Heatmap showing the correlation between RNA-abundance in parental lines and hybrids. (b) Correlation scatterplot of all *D.melanogaster* samples as a representative to show reproducibility.

Figure 3.5 (cont'd)



Ribo-Seq was more complex than mRNA-Seq due to its shorter fragment length and contamination of ribosomal RNA reads. After filtering out RFPs that mapped to ribosomal and small structural RNAs, we found a strong enrichment of RFPs at 30 nt long (**Figure 3.6**a), consistent with previous work that RFPs were around 30 nt (Ingolia et al. 2009; Dunn et al. 2013). Furthermore, over 90% of of RPFs mapped to coding sequence (CDS), followed by mapping to 5' UTRs and 3' UTRs (**Figure 3.6**b) suggesting that the RFPs are true RFPs and are associated with bona fide translation. The correlation between samples were generally lower for TPMs of RFPs than of mRNAs and there was no strong clustering. Nevertheless, correlation between biological replicates remained high, which suggested high reproducibility.

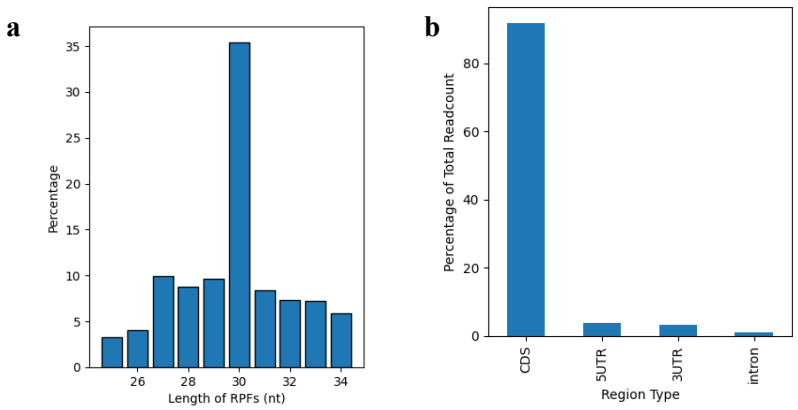


Figure 3.6 Ribosome profiling reads quality control. (a) Distribution of ribosome protected fragments (RPFs) length. (b) Percentages of RPFs mapped to genomic features.

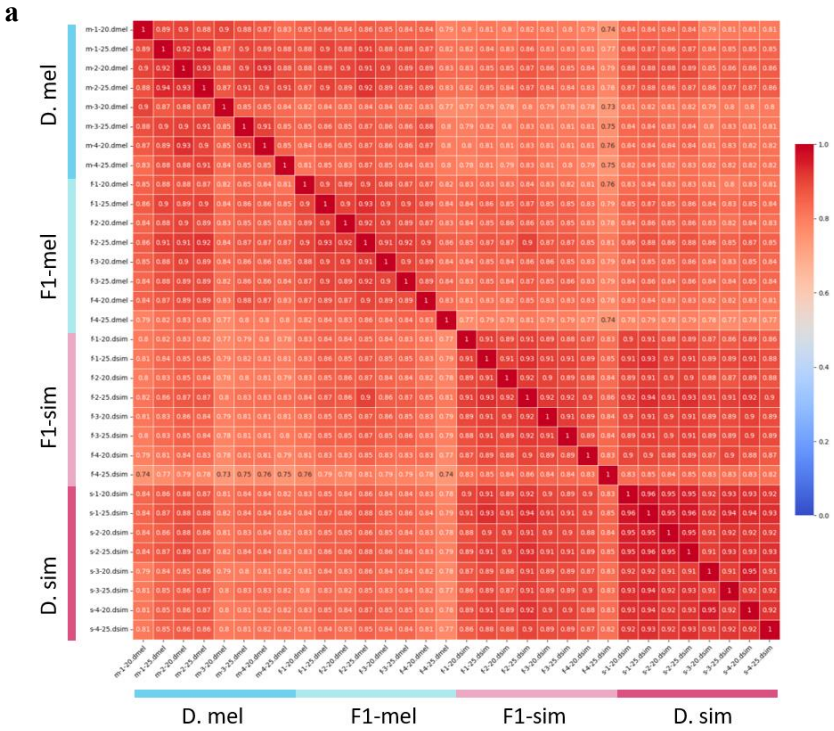
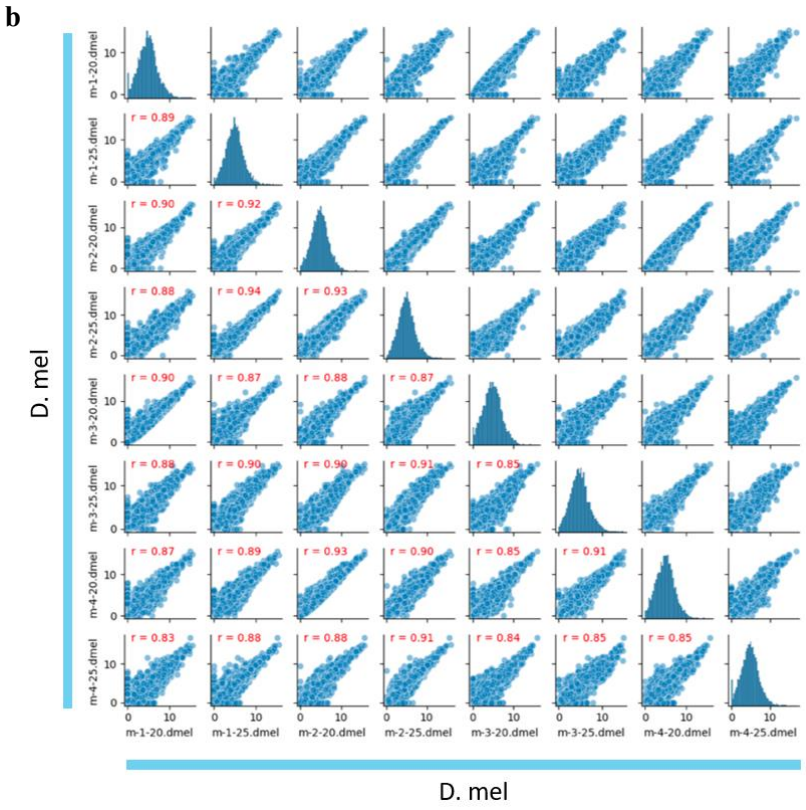


Figure 3.7 Quality and reproducibility of RPF abundance estimation. (a) Heatmap showing the correlation between RPF-abundance in parental lines and hybrids. (b) Correlation scatterplot of all *D.melanogaster* samples as a representative to show reproducibility.

Figure 3.7 (cont'd)



Extensive gene expression and translation differences between *D. simulans* and *D. melanogaster* alleles

The applications of both mRNA-Seq and Ribo-Seq to the hybrid allele specific mapping design allow us to characterize the regulatory architecture at both the transcriptional and post-transcriptional levels.

In both parental strains, the correlation between RNA abundance and RFP abundance (ribosome occupancy) was high (Figure 3.7, 0.74 in *D.mel* and 0.73 in D.sim) but substantially different from unity. This suggested that RFP abundance was dependent on transcription. The more mRNA there are, the more they are translated. However, when normalizing RFP abundance by mRNA abundance to obtain translation efficiency, the correlation was much weaker (Figure 3.7), suggesting that regulation of translation is not completely coupled with regulation of transcription.

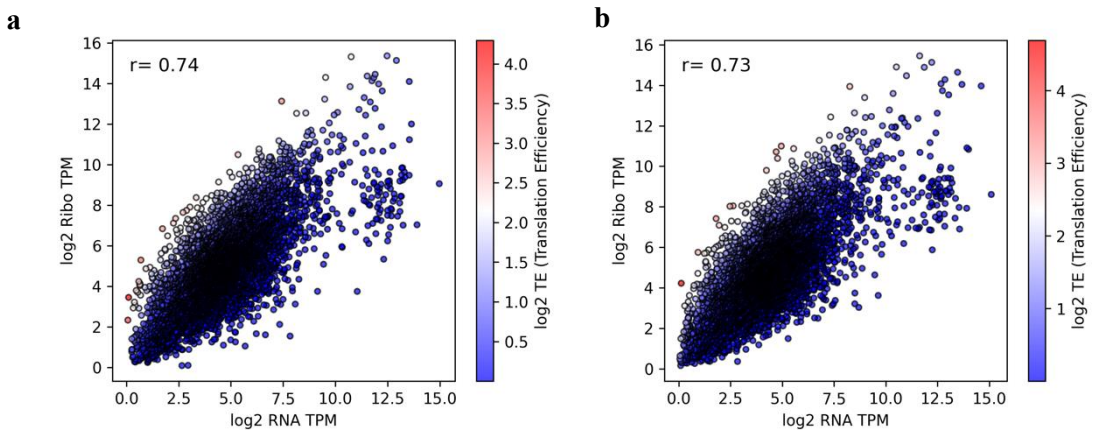


Figure 3.8 Correlation between RNA abundance, RPF abundance and translation efficiency. Scatter plot between log2 RNA abundance in TPM and log2 RPF abundance in TPM in a representative *D.mel* (a) and *D.sim* (b) sample. Gradient color indicates translation efficiency.

To further characterize the effects of the *melanogaster* and *simulans* alleles on transcription and translation, we fitted a full model in edgeR modeling RNA, RFP read counts as well as translation efficiency, in parents and hybrids separately. The model contains effects of the source of the allele (*melanogaster* versus *simulans*), temperature (25 versus 20), and the interaction between the two factors. Interestingly, the allelic effects (log2 fold change) were largely consistent between the two temperatures (**Figure 3.9**). Very few genes were significant (5 in RNA abundance and 1 in RPF abundance at FDR = 0.05) for the source by temperature interaction term. This result suggested that the allelic effects were consistent between the two temperatures in both parental lines and in hybrids.

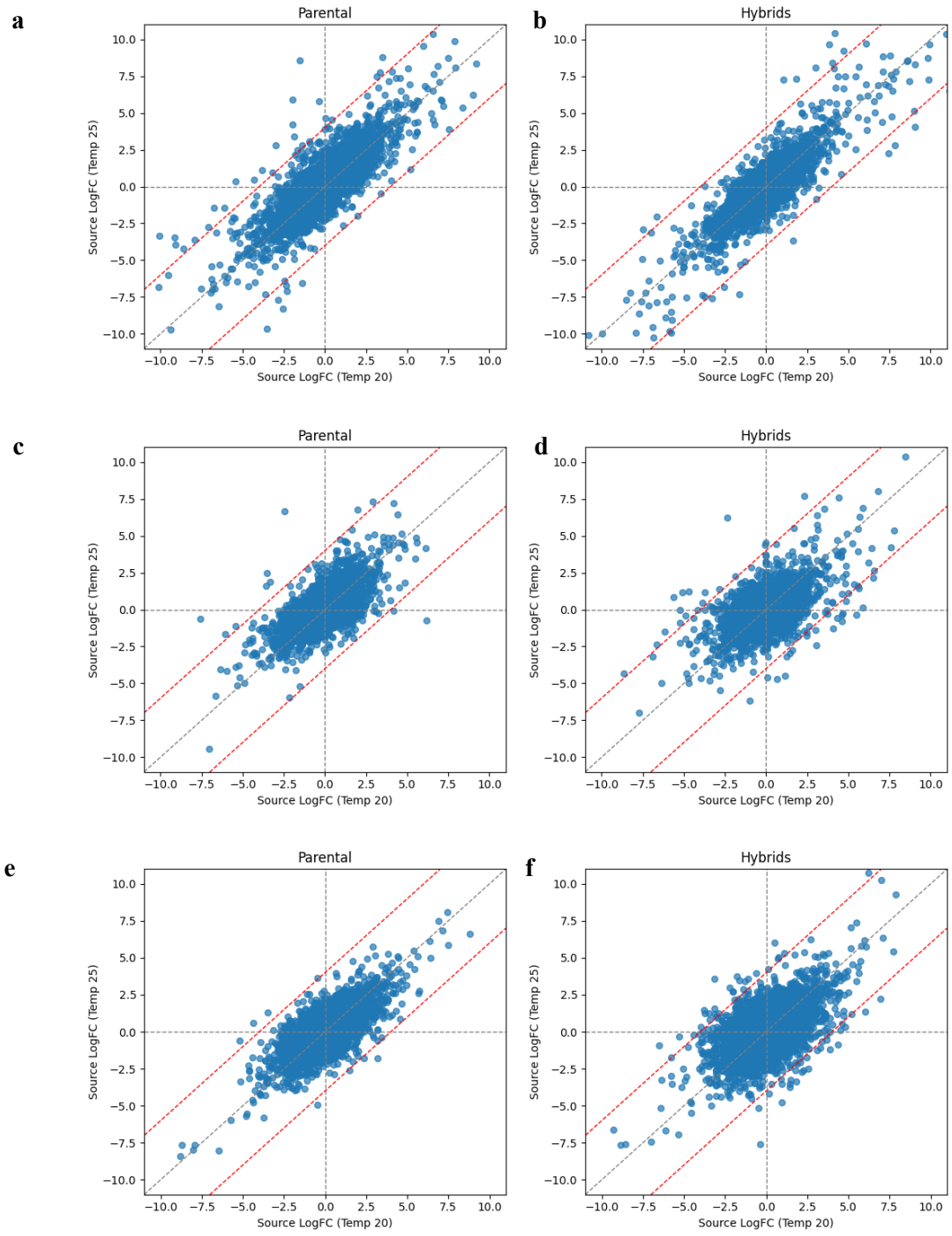


Figure 3.9 Little genotype by temperature interaction in RNA abundance, RPF abundance and translation efficiency. Scatter plot of Log2 allele fold change (*simulans*/*melanogaster*) at 20 ℃ and 25 ℃ in parental lines and F1 hybrids for RNA abundance (a. b) RPF abundance (c, d) and translation efficiency (e, f). Red lines indicate 4-fold difference.

Because the interaction term was not significant, we then tested effects of temperature and source of allele pooling data from both temperatures. This allowed us to identify genes that showed significant differential RNA abundance, RFP abundance, and translation efficiency between the *melanogaster* and *simulans* alleles (source factor) and between the two temperatures (temperature factor). Numbers of showed in **Table 3.3** at threshold of FDR=0.05. The results suggested that allelic effects contributed far more to regulatory variation than environmental effects such as temperatures (**Error! Reference source not found.**).

Regulatory level	Source	Temperature
RNA abundance	3599	80
RPF abundance	563	89
Translation efficiency	2011	1

Table 3.3 cis- and trans- regulatory effects identified in gene expression regulatory levels.

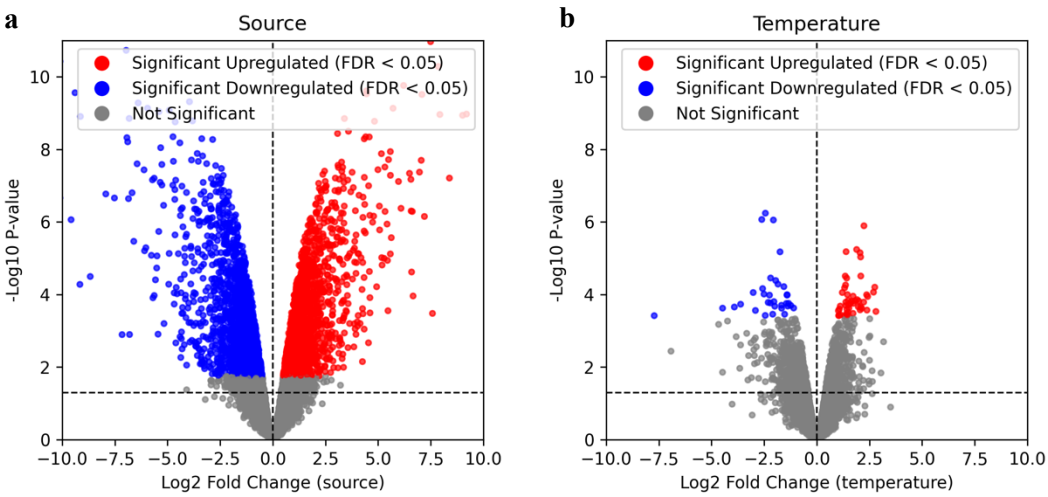
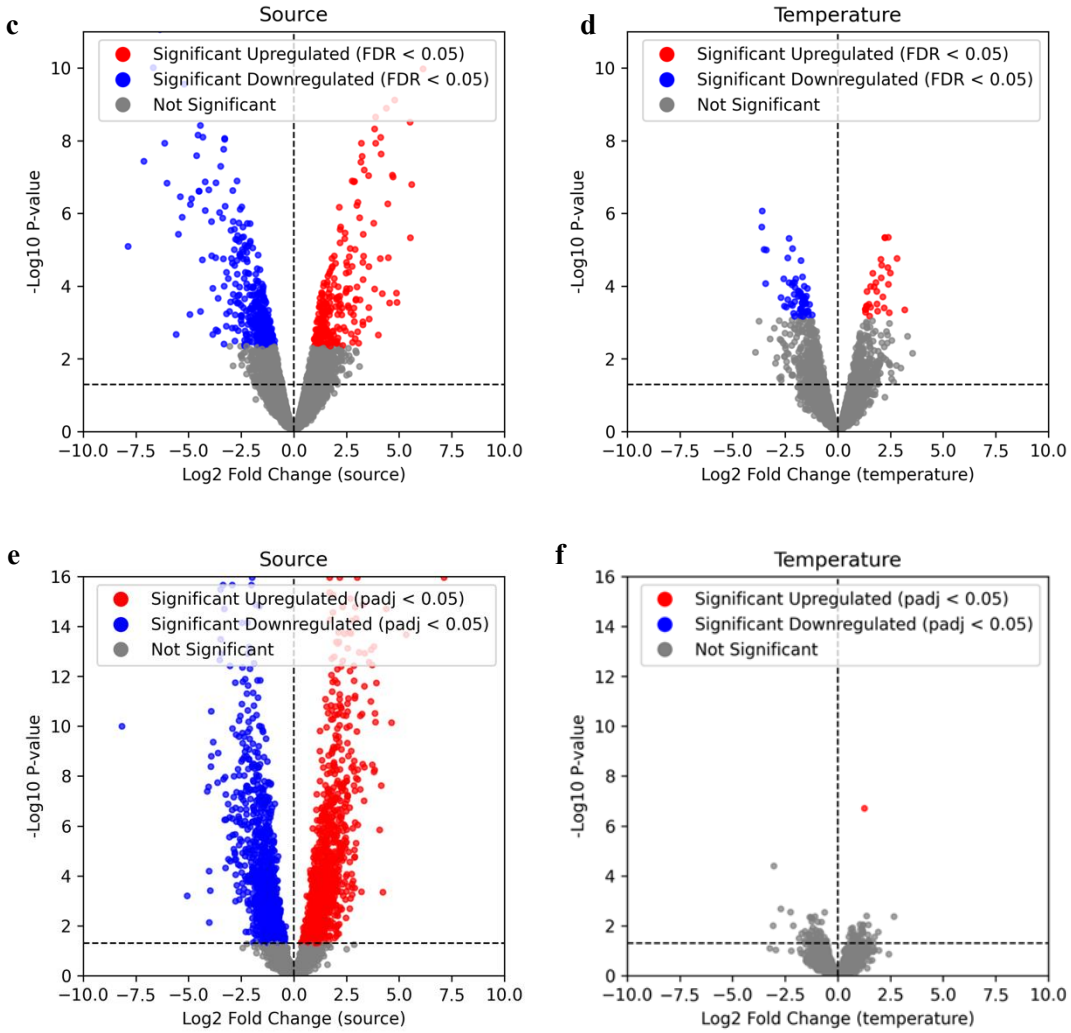


Figure 3.10 Differential RNA abundance, RPF abundance, and translation efficiency due to allelic (source) and temperature effects. Volcano plots showing up and downregulated genes in RNA abundance due to allelic effect (a) and temperature effect (b). (c, d) same plots but for RPF abundance. Same plots but for TE (e, f).

Figure 3.10 (cont'd)



The tests in both parental strains and F1 hybrids allowed us to classify genes into four categories, including genes under *cis* regulation only, *trans* regulation only, both *cis* and *trans*, and neither, for each of RNA abundance, RFP abundance, and translation efficiency. As expected, we found that the majority of regulatory variation in RNA abundance was due to at least some level of *cis* effects (**Figure 3.11**a). A similar pattern was found for RFP abundance (**Figure 3.11**b) but with fewer significant genes. Finally, *cis* effects were also the majority in translation efficiency (**Figure 3.11**c). To test whether the regulation for these three levels was distinct, we asked if the genes identified as *cis* versus *trans* effects overlapped between the three

layers. The sharing between genes under the same mode of regulation was stronger for *cis* effects (**Figure 3.11**d) than for *trans* effects (**Figure 3.11**e).

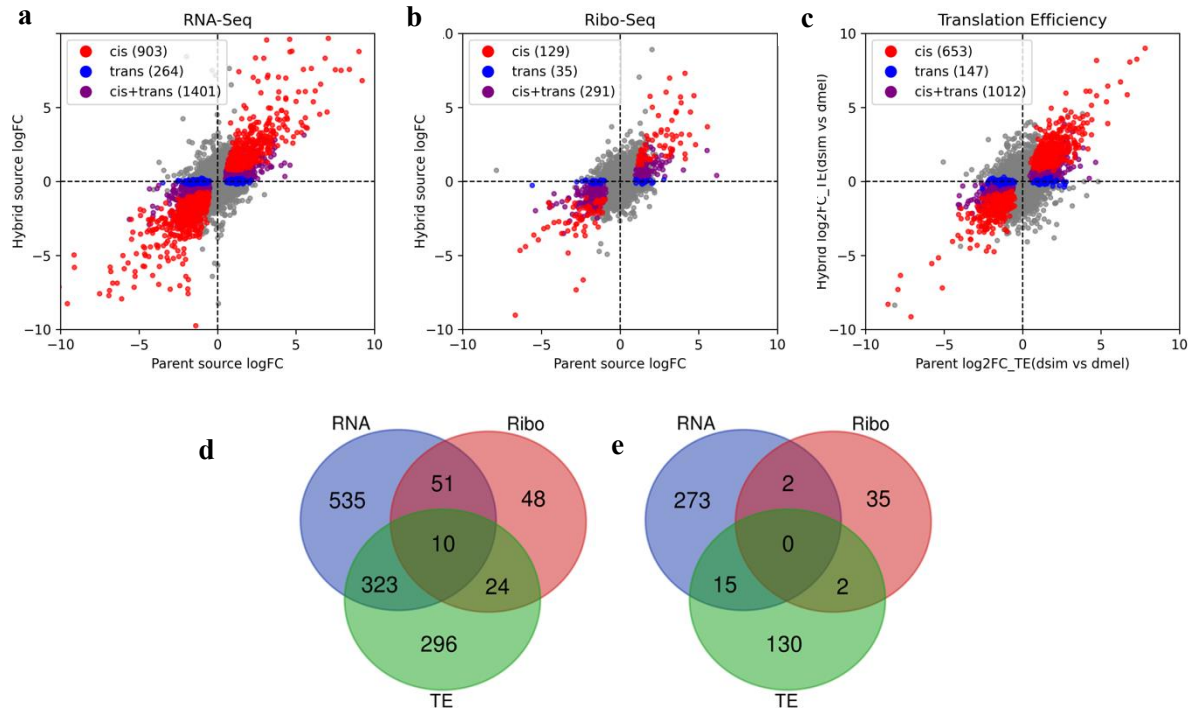


Figure 3.11 Identification of cis- and trans- regulatory effects on RNA abundance RPF abundance and translation efficiency. Scatter plot showing comparison of allelic effect in parents and hybrids for RNA abundance (a) RPF abundance (b) TE (c). Venn diagrams showing gene overlap between three levels of regulation by cis- effect (d) and trans- effect (e).

DISCUSSION

Using a hybrid allele-specific mapping approach combined with RNA-Seq, Ribo-seq, and high quality genome assemblies, we elucidated the contribution of cis- and trans-regulatory effects to gene expression regulation in Drosophila species divergence at multiple levels. Although regulation at multiple levels were distinct and uncoupled, we found cis regulation to be the predominant mode of regulation in both transcriptional and post-transcriptional levels.

This study represents the first in flies that applies the hybrid allele-specific mapping approach to characterize translational control as measured using Ribo-Seq. While it's possible to map cis and trans regulatory effects by population scale mapping, which is possible in flies, it remains a costly approach in many instances. The hybrid mapping design offers a quick and cost-effective way to identify genes that are controlled by cis regulatory elements versus trans acting factors. Chapter 2 of this thesis demonstrated that at the RNA abundance level, hybrid mapping design can largely recapitulated effects estimated in population scale eQTL mapping. In this study, we identified many genes in either category that can be further investigated. For example, it would be useful to specifically identify translational control elements that explained the species divergence in translation of a gene.

The temperature-controlled experimental design allowed us to investigate the impact of environmental factors on gene expression regulation. Differences in gene expression and translation efficiency observed between the two temperature conditions provided insights into how temperature influences regulatory divergence in hybrids. We subjected parental lines and hybrids to low temperature (20°C) treatment to investigate transcriptional and translational response as well as whether the allelic effects can be modified by different environments (genotype by environment interaction). However, we did not find strong temperature effect and found no evidence of GxE. This is in stark contrast to a previous study in adult flies that found strong evidence of GxE for temperature treatment in gene expression within *Drosophila melanogaster* ((Huang, Carbone, Lyman, Anholt, & Mackay, 2020)). There could be many reasons. First, the divergence between the two species may be too large such that small changes

induced by temperature may not be significant. Second, the temperature treatment (20 versus 25℃) may not be strong enough for 3rd instar larvae. Further investigation is needed to demonstrate the usefulness of this approach to study GxE.

Drosophila crossing, sample collection and sequencing

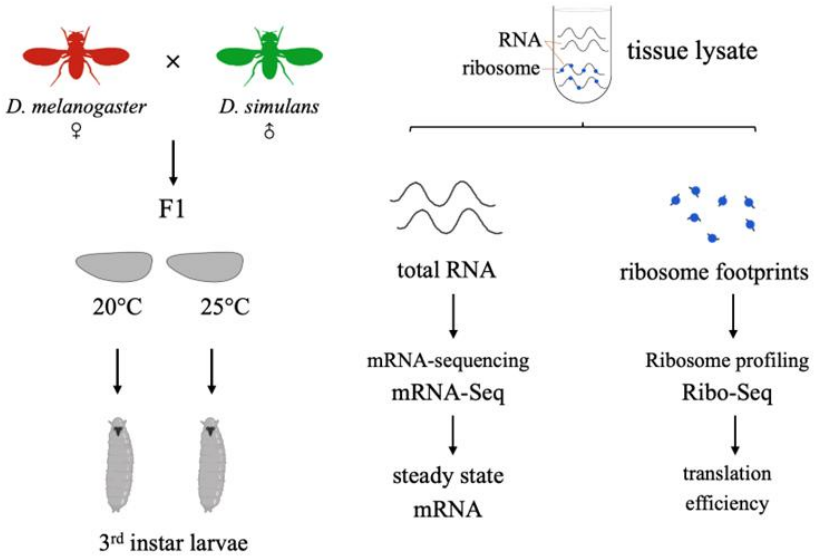


Figure 3.12 Experimental design of Drosophila crossing and sample collecting

Our experimental design, illustrated in **Figure 3.12**, details the temperature-dependent hybrid crosses and sampling of third-instar larvae for RNA-Seq and Ribo-seq. Two Drosophila lines were purchased from National Drosophila Species Stock Center (Cornell College Agriculture and Life Science): *Drosophila simulans* (simC167.4, SKU: 14021-0251.199), *Drosophila melanogaster* (Genome project WGS strain, SKU: 14021-0231.36). Virgin females of *Drosophila melanogaster* were picked every 12 hours for newly hatched flies and then were raised in vials for 3 days to allow them to mature, and to verify that they were virgins. Adult males of *Drosophila simulans* were selected and crossed to *Drosophila melanogaster* virgin females with a ratio of 90 males to 30 females in a vial. For each group, two crosses were conducted in temperature conditions 25 ℃ and 20 ℃. The flies were allowed to mate and lay eggs for 3 days in 25 ℃ and 5 days 20 ℃ before they were removed from the vials. The same routine was performed for both parental lines. After 3-6 days of development, we collected 3rd instar larvae every 2 hours, 3rd instar larvae were identified as those individuals who climbed to the top of vial wall and kept still. The collected larvae were frozen in liquid nitrogen and stored

in -80℃ refrigerator until lysis. Around 30 larvae of each sample were pooled together and added to 2ml tubes with 1ml frozen lysis buffer inside, tissue lysates were obtained following this method (Dunn, Foo, Belletier, Gavis, & Weissman, 2013). For each sample, 600 µl of lysate was sent to Advanced RNA Profiling Core in Case Western Reserve University for ribosome profiling libraries construction, and another 70 µl of lysate was used to extract total RNA. Ribosome profiling libraries along with total RNA were sent to the Genomic Core in Michigan State University to do library construction and sequencing. Stranded RNA-Seq libraries were sequenced on NovaSeq 6000 using paired-end 150-pb sequencing, Small RNA Ribo-Seq libraries were sequenced on NovaSeq 6000 using single-end 100-bp sequencing.

High molecular weight DNA (HMW DNA) of hybrids were extracted from F1 larvae using PacBio Nanobind HMW DNA extraction kit. HMW DNA of hybrids were then sent out to commercial sequencing company for PacBio HiFi-Seq and ONT ultra-length-Seq. Genomic DNA of parental lines were extracted and sent to MSU Genomics Core for libraries construction and Sequencing. Stranded DNA-Seq libraries were sequenced on NovaSeq 6000 using paired-end 150-pb sequencing (**Error! Reference source not found.**a).

Genome assembly and annotation

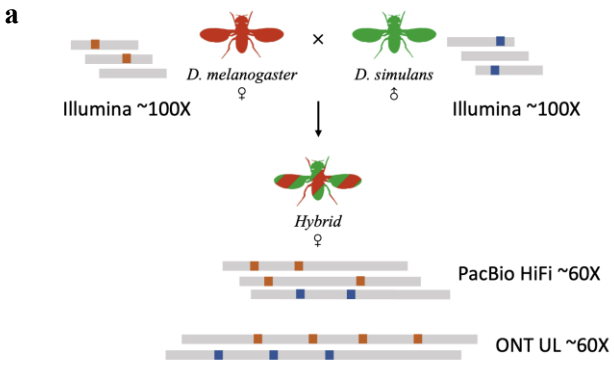
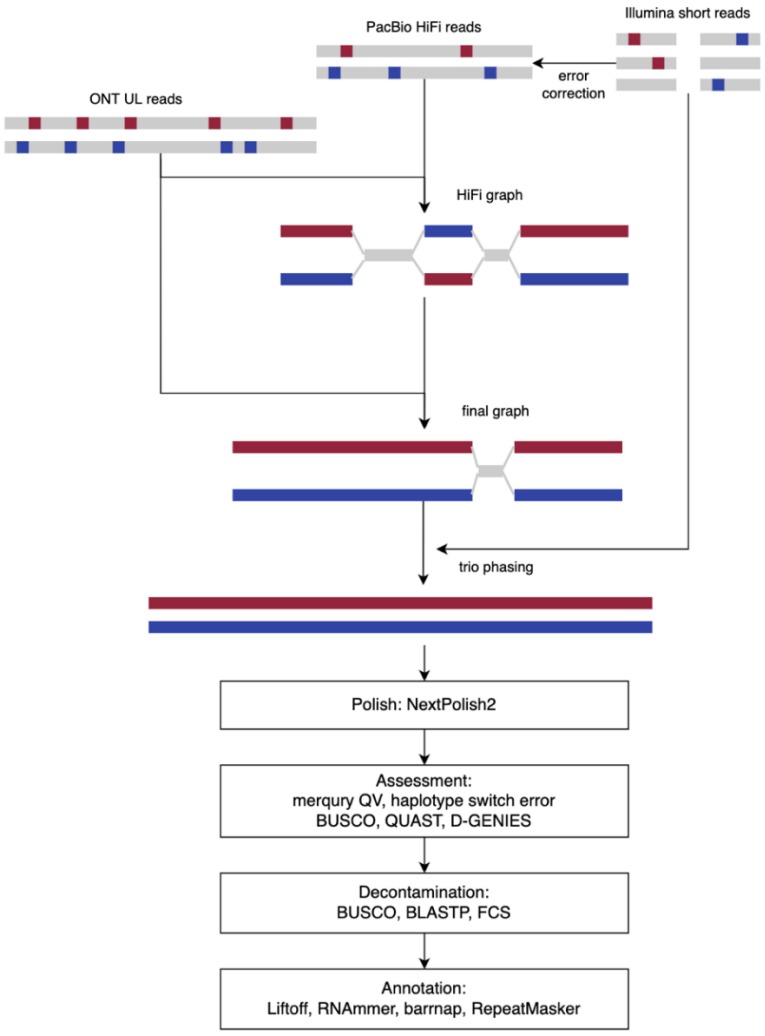


Figure 3.13 Diagram of DNA sequencing and T2T genome assembly

Figure 3.13 (cont'd)

b



Genome assembly pipeline was show in **Error! Reference source not found.** b. Illumina s hort reads were cleaned by fastp (v0.23.2) (Chen, 2023). PacBio HiFi reads were filtered as average reads quality higher than 20. ONT ultra-length reads were firstly committed to adapter detection and trim by Porechop (v0.2.4) and then filtered with lowest average reads quality of 10 and shortest read length of 40k. Hybrid HiFi reads along with parental short reads were used to build 19-mer merqury (v1.3) (Rhie, Walenz, Koren, & Phillippy, 2020) database to estimate sequencing depth and heterozygosity. Three kinds of sequencing data were subjected to HiFiasm (v0.19.9) (Cheng, Concepcion, Feng, Zhang, & Li, 2021) with trio binning mode and verkko

(v2.2.1) (Rautiainen et al., 2023). Haplotype assemblies from HiFiasm have better continuity and used for next assessment. Bandage (Wick, Schultz, Zobel, & Holt, 2015) was used to visualize haplotype assemblies and calculate assemble statistics. Merqury was used to estimates correctness (quality value, QV) of assemblies and to estimate haplotype switch error. BUSCO (Manni, Berkeley, Seppey, & Zdobnov, 2021) was used to evaluate assemble completeness. Quast (v5.2.0) (Mikheenko, Prjibelski, Saveliev, Antipov, & Gurevich, 2018) compares assemblies with respective reference to estimate fragment coverage of reference. D-Genies (Cabanettes & Klopp, 2018) was used to draw co-linearity plot of assemblies with reference genome. Assemblies' decontaminations were performed by combined BUSCO, FCS (NCBI) and Quast searching/mapping results that indicate bacterial genome. Ribosomal RNA genes were predicted by RNAmmer (Lagesen et al., 2007), and repeat regions in assemblies are annotated and masked by RepeatMasker (v4.1.7). Annotation of new assemblies were done by lift over reference gene annotation by Liftoff (v1.6.3) (Shumate & Salzberg, 2021). At last, ortholog gene pairs between *D.melanogaster* and *D.simulans* were identified by SynGAP (v1.2.5) (Wu, Mai, Chen, & Xia, 2024).

RNA-Seq and Ribo-Seq data analysis

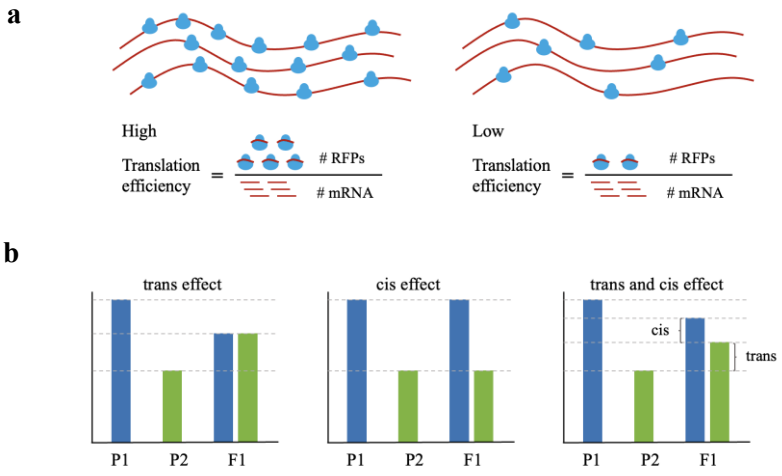


Figure 3.14 Translation efficiency calculation and cis-/trans- effects identification.

RNA-Seq reads was cleaned by fastp (v0.23.2) (Chen, 2023) to remove low quality reads or excessive N contained reads. Ribosome profiling reads were first subject to cutadpt (v4.9) (Martin, 2011) to trim adapter introduced through libraries construction, then the short reads was mapped to a combined sequences of rRNA, tRNA, snRNA and snoRNA download from Rfam (Kalvari et al., 2021) database and flybase by Bowtie (v1.3.1) (Langmead, Trapnell, Pop, & Salzberg, 2009) software. Failed to mapped reads were then filtered length from 25nt to 34 nt considered as clean RPF reads.

RNA-Seq reads and RPFs were mapped to combined assemblies of *D.melanogaster* and *D.simulans* by HISAT2 (v2.2.1) (Kim, Paggi, Park, Bennett, & Salzberg, 2019). Only uniquely mapped gene were retained that considered as allele-specific. Readcount of genes were then counted by HTSeq (v2.0.5) (Putri, Anders, Pyl, Pimanda, & Zanini, 2022; Danecek et al., 2021). Translation efficiency, defined as the ratio of RPFs to mRNA abundance, was calculated for each gene as TPM of RPFs divided by TPM of RNA-Seq (**Figure 3.14**a). A negative binomial generalized log-linear model was fit to the read counts for each gene by edgeR package (Robinson, McCarthy, & Smyth, 2010) to identify species source effect, temperature effect and their interaction effect. For TE values, a generalized linear model was fitted by Ribodiff for all gene to identify source effect and temperature effect. To identify cis- and trans-effects, Source effect of each gene assessed between parental lines and between hybrid two alleles were compared as **Figure 3.14**b illustrated. For RNA-Seq, RPFs and TEs, source effect affected genes and temperature effect affected genes were subjected to Gene ontology (GO) enrichment respectively to exploring possible pathways.

REFERENCES

Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., . . . Phillippy, A. M. (2018). De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol*. doi:10.1038/nbt.4277

Huang, W., Carbone, M. A., Lyman, R. F., Anholt, R. R. H., & Mackay, T. F. C. (2020). Genotype by environment interaction for gene expression in *Drosophila melanogaster*. *Nat Commun, 11*(1), 5451. doi:10.1038/s41467-020-19131-y

Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R., & Weissman, J. S. (2013). Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *Elife, 2*, e01179. doi:10.7554/eLife.01179

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta, 2*(2), e107. doi:10.1002/imt2.107

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol, 21*(1), 245. doi:10.1186/s13059-020-02134-9

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods, 18*(2), 170-175. doi:10.1038/s41592-020-01056-5

Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., . . . Koren, S. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol, 41*(10), 1474-1482. doi:10.1038/s41587-023-01662-6

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics, 31*(20), 3350-3352. doi:10.1093/bioinformatics/btv383

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing Genomic Data Quality and Beyond. *Curr Protoc, 1*(12), e323. doi:10.1002/cpz1.323

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics, 34*(13), i142-i150. doi:10.1093/bioinformatics/bty266

Cabanettes, F., & Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ, 6*, e4958. doi:10.7717/peerj.4958

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res, 35*(9), 3100-3108. doi:10.1093/nar/gkm160

Shumate, A., & Salzberg, S. L. (2021). Liftoff: accurate mapping of gene annotations. *Bioinformatics, 37*(12), 1639-1643. doi:10.1093/bioinformatics/btaa1016

Wu, F., Mai, Y., Chen, C., & Xia, R. (2024). SynGAP: a synteny-based toolkit for gene structure annotation polishing. *Genome Biol, 25*(1), 218. doi:10.1186/s13059-024-03359-8

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal, 17*(1), 10-12.

Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., . . . Petrov, A. I. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res, 49*(D1), D192-D200. doi:10.1093/nar/gkaa1047

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol, 10*(3), R25. doi:10.1186/gb-2009-10-3-r25

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol, 37*(8), 907-915. doi:10.1038/s41587-019-0201-4

Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E., & Zanini, F. (2022). Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics, 38*(10), 2943-2945. doi:10.1093/bioinformatics/btac166

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., . . . Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience, 10*(2). doi:10.1093/gigascience/giab008

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi:10.1093/bioinformatics/btp616

CHAPTER 4: OVERALL CONCLUSION AND DISCUSSION

Gene expression regulation, encompassing transcriptional, post-transcriptional, and translational processes, represents a cornerstone of phenotypic diversity and evolutionary adaptation. This dissertation offers a comprehensive exploration of regulatory mechanisms within and between Drosophila species, elucidating the roles of cis- and trans-regulatory effect in shaping gene expression. Through the application of sequencing technologies such as RNA-Seq and ribosome profiling (Ribo-Seq), this work has provided new insights into the genetic and environmental influences on gene expression dynamics.

Summary of Key Findings

The first study characterized regulatory variation within *Drosophila melanogaster*, highlighting the predominance of cis-regulatory effects. These findings underscore the localized influence of genetic variants near their target genes in driving transcriptional diversity. While trans-regulatory effects were less common, their presence suggests an essential role in maintaining network-level coherence in gene expression.

The second study also revealed a dominant contribution of cis-regulatory effect in interspecies comparisons between *Drosophila melanogaster* and *Drosophila simulans*. Notably, we found many cis- and trans- affected genes at the translational level in addition to the transcriptional level, suggesting that these regulatory effects modulate translation efficiency to buffer or amplify transcriptional differences.

Discussion on Methodological Advances

One of the contributions of this dissertation is the innovative application of allele-specific RNA-Seq and Ribo-Seq in F1 hybrids to dissect cis- and trans-regulatory contributions at both transcriptional and translational levels. These methods allowed for high-resolution analysis of allele-specific expression and translation, capturing subtle regulatory differences that might be masked in bulk analyses.

However, it is essential to acknowledge the limitations of these methodologies. For example, allele-specific analyses depend heavily on differences of coding sequences between parental lines and require inbred lines. This method will have limited power to capture genes that are only bearing non-coding region variants.

Moreover, the use of temperature-controlled experiments provided valuable insights into environmental modulation of regulatory dynamics. This approach can be extended to study other environmental variables, such as nutrient availability or oxidative stress, which are known to influence gene expression through both cis- and trans-regulatory mechanisms.

Evolutionary Implications

This dissertation highlights how the interplay between cis- and trans-regulatory elements contributes to evolutionary processes at multiple levels. Cis-regulatory elements, being tightly linked to their target genes, are more likely to evolve under purifying selection, ensuring the conservation of essential gene functions. In contrast, trans-regulatory factors, such as transcription factors or RNA-binding proteins, often have broader effects across gene networks, making them more prone to diversifying selection.

The observation that translational regulation often buffers transcriptional variation suggests a protective mechanism during evolutionary transitions. This buffering could provide populations with the flexibility to tolerate potentially deleterious mutations at the transcriptional level while gradually adapting to new environmental conditions. Such a mechanism may be particularly relevant in hybridization events, where divergent regulatory networks must integrate and maintain organismal fitness.

Gene-Environment Interactions

Environmental modulation of gene expression is an important aspect of this dissertation. The results from temperature-dependent experiments revealed environmental factors can act differently across different layers of gene expression process between species.

Future research could employ multi-omics approaches to explore how environmental changes influence gene expression at multiple regulatory levels simultaneously. For instance, integrating ATAC-Seq for chromatin accessibility, ChIP-Seq for transcription factor binding, and metabolomics for cellular state profiling could provide a holistic view of gene-environment interactions.

Translational Applications

Beyond basic research, the insights gained from this dissertation have potential translational applications in fields such as agriculture, medicine, and biotechnology. Understanding the mechanisms underlying cis- and trans-regulatory variation can inform breeding programs aimed at improving stress tolerance or productivity in crops and livestock. Similarly, identifying regulatory elements that contribute to disease susceptibility could lead to novel therapeutic targets.

In the context of evolutionary biology, the findings from this dissertation can help predict how populations might respond to rapid environmental changes, such as those driven by climate change. By identifying genes and regulatory networks that are highly plastic or robust under environmental stress, researchers can better anticipate the adaptive potential of natural populations.

Future Directions

While this dissertation has provided significant insights into the regulatory dynamics of gene expression, several questions remain open for future investigation:

Integration of Post-Transcriptional and Post-Translational Layers: While this work focused on transcriptional and translational regulation, other regulatory layers, such as RNA modification (e.g., m6A) and protein post-translational modifications, remain underexplored. Integrating these layers into the current framework could provide a more comprehensive understanding of gene regulation.

Functional Validation: The regulatory effects identified here were inferred based on allele-specific expression and translation data. Functional validation using CRISPR/Cas9 genome editing or other experimental approaches such as reporter genes could confirm the causal relationships between specific regulatory elements and their target genes.

Population-Level Studies: Extending these analyses to population-level datasets could reveal how regulatory variation contributes to phenotypic diversity and adaptive evolution within and between populations.

Non-Model Organisms: Expanding the methodologies and insights gained from this dissertation to non-model organisms could validate the generality of the findings and uncover unique regulatory mechanisms in different evolutionary contexts.