The evolution of gene regulation in Drosophila

by

Kraig Ryan Stevenson

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Bioinformatics) in The University of Michigan 2014

Doctoral Committee:

Associate Professor Trisha Wittkopp, Chair Professor Margit Burmeister Assistant Professor James D. Cavalcoli Assistant Professor Maureen A. Sartor Professor Jianzhi Zhang "Nature has provided us a spectacular toolbox. The toolbox exists. An architect far better and smarter than us has given us that toolbox, and we now have the ability to use it." - Barry Schuler, from Genomics 101 TED Talk at Taste3 2008 -

For my mother, who continues to inspire me and since my early childhood has encouraged me to pursue my passion and curiosity for science.

ACKNOWLEDGEMENTS

A very special thank-you to my wife Jinyi for being incredibly patient with me while I finished my doctoral work, and for giving me, in addition to love, the greatest gift of all: our wonderful son Declan. I truly couldn't have done any of this without you, and I look forward to a lifetime of experiences with our new family.

Also, I'd like to thank my mentor Trisha for a wonderful five years of research, and plenty of laughs and great times along the way. And to Joe, thank you for allowing me to work on some great projects with you, we truly were the "dynamic duo" of the lab. Both of you, along with the entire lab, made my thesis work a real pleasure, and I'll definitely miss our many wonderful interactions.

Lastly, to my family members and friends who have been following me along this journey, I say "Thank you!" for believing in me. And no, you don't have to call me "Dr. Stevenson"...

TABLE OF CONTENTS

DEDICATIO	N	ii
ACKNOWLE	DGEMENTS	iii
LIST OF FIG	URES	vii
CHAPTER		
I. Intro	duction	1
1.1	Dissertation outline	11
	es of bias in measures of allele-specific expression derived from -seq data aligned to a single reference genome	14
2.1	Abstract	14
2.2	Introduction	15
2.3	Results and Discussion	18
	of differentiating sites	18
	2.3.2 Read length and the amount of sequence divergence can also affect	
	allelic bias	20
	2.3.3 Allele-specific differences in mappability and insertions/deletions	
	affect measurements of ASE	22
	2.3.4 Aligning real sequencing data to a single genome can produce reli-	
	able measures of relative ASE	23
	2.3.5 Excluding selected differentiating sites maintains ability to measure	
	relative ASE for most exons	25
2.4	Conclusions	26
2.5	Methods	27
	2.5.1 Generating allele-specific short reads comparing <i>D. melanogaster</i>	
	genotypes in silico	27
	2.5.2 Quantifying allelic abundance in simulated RNA-seq data	28
	2.5.3 Measuring number of neighboring differentiating sites and mappa-	
	bility across genomes	30
	2.5.4 Quantifying relative ASE in an F1 hybrid between <i>D. melanogaster</i>	
	and <i>D. simulans</i>	30
2.6	Acknowledgements	31
III. Geno	mic imprinting absent in Drosophila melanogaster adult females	41
3.1	Summary	41
3.2	Introduction	42
3.3	Results and Discussion	43
0.0		

	3.3.1		46
	3.3.2	Low-frequency deletion(s) responsible for some cases of apparent	
		imprinting	46
	3.3.3	Non-clustered PIGs have higher-than-normal intrinsic noise	48
	3.3.4	What role does imprinting play in regulating <i>D. melanogaster</i> gene	
			49
3.4	Genomi		50
3.5			52
	3.5.1		52
	3.5.2		52
	3.5.3	Quantifying total and allele-specific expression from sequencing reads	
	3.5.4	Sliding-window analyses with Monte Carlo sampling and approxi-	-
	0.0.1		54
	3.5.5		54
3.6			54 56
3.7			56
3.8		•	56
	3.8.1		56
	3.8.2		57
	3.8.3	Quantifying allele-specific expression from sequencing reads	58
I V T			771
IV. 1emp	o and m	ode of regulatory evolution in <i>Drosophila</i>	71
4.1	Abstract	t	71
4.2			72
4.3			74
1.0	4.3.1		74
	4.3.2		75
	4.3.3		76
	4.3.3 4.3.4		77
	4.3.4 $4.3.5$		79
	4.3.6	e e e e e e e e e e e e e e e e e e e	80
4.4			83
	4.4.1		85
	4.4.2	· · · · ·	86
	4.4.3		87
4.5			88
			88
	4.5.2		89
	4.5.3	1 0/0 1 0	90
	4.5.4		90
	4.5.5		91
	4.5.6	Comparing total expression among genotypes	91
	4.5.7	Inferring the mode of inheritance	92
	4.5.8	Normalizing allele-specific RNA-seq read counts among comparisons	93
	4.5.9	Evaluating <i>cis</i> - and <i>trans</i> -regulatory changes	95
	4.5.10	Scripts and software used	96
4.6			97
4.7			97
4.8		<u> </u>	97
	4.8.1		97
	4.8.2		99
	4.8.3		.01
4.9		nental note	

pseudo	pobscura and its closely-related subspecies $D. p. bogotana \ldots \ldots 1$	33
5.1	Abstract	33
5.2	Introduction	34
5.3	Results and Discussion	36
	5.3.1 Measuring Drosophila pseudoobscura and D. p. bogotana sex- and	
	tissue-specific gene expression	36
	5.3.2 Inheritance patterns for gene expression are most similar within	
		37
	5.3.3 Carcass tissues are enriched for genes regulated only in <i>trans</i> , while	
	gonads are enriched for genes regulated only in cis	38
	5.3.4 No faster-X effect, but sex-biased genes show increased levels of	
	gene expression divergence	40
5.4	Conclusions	42
5.5	Materials and Methods	43
	5.5.1 Fly strains, rearing, and collections	43
	5.5.2 Library preparation and Illumina sequencing	44
	5.5.3 Building parental exomes from $D.$ pseudoobscura and $D.$ p. bo-	
	gotana gDNA	44
	5.5.4 Quantifying total and allele-specific gene expression	
	5.5.5 $$ Normalizing total and allele-specific read counts across samples $$. $$ 1	
	5.5.6 Mode of inheritance and regulatory divergence classification 1	
	5.5.7 Determining sex-biased and sex tissue-specific genes	
	5.5.8 Total and <i>cis</i> -regulatory expression divergence	
5.6	Supplemental Methods	50
VI. Discu	ssion	62
6.1	Dissertation summary	62
	6.1.1 Chapter II	62
	6.1.2 Chapter III	63
	6.1.3 Chapter IV	64
	6.1.4 Chapter V	64
6.2	Reflections	65
BIBLIOGRA	PHY	69

LIST OF FIGURES

Figure

2.1	Simulating an allele-specific RNA-seq experiment	33
2.2	The density of differentiating sites affects relative allelic abundance when simulated	
	reads are mapped to only one genome	34
2.3	Imperfect mappability causes inaccurate measures of relative allelic abundance	35
2.4	Insertions and deletions (indels) cause biased allele-specific assignment when reads	
	are aligned to a single reference genome	36
2.5	Real reads aligned to a single reference genome produce reliable measures of allelic	
	abundance after excluding problematic differentiating sites	37
2.6	Relative allelic abundance can be estimated for most exons after excluding sites	
	problematic sites	38
2.7	The density of differentiating sites affects measures of relative ASE when simulated	
	reads are mapped to the alternative genome	39
2.8	36- and 50-base sequence reads produced comparable measures of relative ASE	
	when a similar ratios of mismatches to bases in a sequence read is allowed	40
3.1	Allelic Expression from Reciprocal Crosses Suggests that $<2\%$ of Genes in the	~ .
	Genome Might Be Imprinted	61
3.2	Putatively Imprinted Genes Clustered Significantly on Chromosomes	62
3.3	Replicate Pools of Flies Showed Different Allele Frequencies in Genomic DNA for	60
0.4	Putatively Imprinted Genes Located in a Cluster	63
3.4	Putatively Imprinted Genes Have High Intrinsic Noise	64
3.5	Experimental Method for Investigating Imprinting, Related to Figures 3.1 - 3.4	65
3.6	Coverage of Genes Tested for Imprinting, Related to Figure 3.1	66
3.7	Pyrosequencing Validation of RNA-seq Data, Related to Figure 3.1	67
3.8	RNA-seq Data Plotted with Filtered PIGs Highlighted, Related to Figure 3.4	68
3.9	Descriptive statistics of Illumina sequencing, Related to Figure 3.1	69 70
$3.10 \\ 4.1$	Genotyping individual MZ and ZM progeny, Related to Figure 3.3	70
	Studying regulatory evolution in the melanogaster group of Drosophila	
$\begin{array}{c} 4.2 \\ 4.3 \end{array}$	Expression divergence between genotypes and in F1 hybrids	
4.5 4.4	Effects of <i>cis</i> -regulatory divergence	
4.4 4.5	Expression divergence in mammals, <i>Drosophila</i> and yeast	
4.5 4.6	Methodological overview	
4.0 4.7	Independent confirmation of relative allelic expression levels inferred from RNA-seq	110
4.7	data	111
4.8	Total expression levels were similar between reciprocal hybrids	
4.9	Most significant expression differences between reciprocal hybrids are small in mag-	112
4.0	nitude	112
4.10	Overall expression differences increase with divergence time	
4.10	Many small expression differences are statistically significant between genotypes .	
4.11	Expression differences between F1 hybrids and parental species increase with di-	110
4.14	vergence time	116
		тт0

4.13	The proportion of genes showing misexpression in F1 hybrids increased with diver-
	gence time of the parental genotypes
4.14	The frequency of large expression differences between F1 hybrids and parental
	species increases with divergence time
4.15	Allele-specific sequence reads are accurately assigned to genotypes
4.16	Relative allelic expression was similar between reciprocal hybrids
4.17	Evolution of <i>cis</i> - and <i>trans</i> -regulation
4.18	Differences in <i>cis</i> -regulatory activity increase with divergence time more rapidly
	than differences in total expression
4.19	Differences between gene sets used to analyze total and allele-specific expression data123
4.20	Evolutionary trajectories for expression divergence of individual genes $\ldots \ldots \ldots 124$
4.21	The relative contributions of <i>cis</i> - and <i>trans</i> -regulatory changes to expression diver-
	gence change with divergence time
4.22	Effects of <i>cis</i> -regulatory divergence
4.23	Summary of sequencing depth for RNA-seq and gDNA
4.24	Number of genes suitable for quantifying total expression in each genotype 128
4.25	Number of genes suitable for quantifying allele-specific expression in each genotype 129
4.26	Accuracy of mapping maternally inherited mitochondrial alleles in interspecific F1
	hybrids
4.27	Criteria for assigning genes to regulatory evolution classes
4.28	Pyrosequencing assays for quantification of allelic expression ratios
5.1	Experimental design
5.2	Mode of inheritance differs between sexes and tissues
5.3	Regulatory divergence classification differs between sexes and tissues
5.4	% cis for additively and non-additively inherited gene expression consistent with
	expectation in gonad but not carcass tissue
5.5	X-linked and autosomal female-biased genes show higher total expression divergence
	than male-biased and non biased genes in testes
5.6	Male-biased autosomal genes show elevated gene expression divergence driven by
	cis in carcass samples but not gonads
5.7	D. pseudoobscura and D. p. bogotana gDNA mismapping rates are very low across
	chromosomes
5.8	Testis- and ovary-specific genes show slightly higher levels of sequence divergence $\ . \ 159$
5.9	neo-X shows slower-X effect in male carcass samples but not testes
5.10	No elevated cis -regulatory expression divergence on X chromosomes in female samples 161

CHAPTER I

Introduction

In a Nobel Prize-winning series of experiments conducted in the early 20th Century, Thomas H. Morgan used *Drosophila melanogaster* to study the inheritance of particular characteristics. He was the first to identify a mutant factor whose presence determined the white eye phenotype in otherwise red-eyed wild-type D. melanogaster, which he named white eye (Morgan, 1910). One of the most remarkable aspects of this work is that it was done without any knowledge of DNA as the unit of heritability; their discoveries were made solely by crossing many individuals and observing the characteristics of their progeny. A few years later, Morgans colleagues Alfred H. Sturtevant and Calvin B. Bridges came one step closer to identifying the heritable unit by generating the first genetic maps linking factors, which we now know as genes, to their chromosomal locations (Sturtevant, 1913; Bridges, 1914). Their work, along with Morgan's early observation that the white eye phenotypes inheritance pattern was sex-specific, helped bolster the chromosomal theory of inheritance (Bridges, 1916). A couple decades later, Bridges successfully overlaid his genetic maps onto polytene chromosomes, which could easily be visualized in the salivary glands of *D. melanogaster* after undergoing many rounds of DNA replication without cellular division (Bridges, 1935). Around that same time, the work of Hermann J. Muller showed that heavy doses of X-rays were sufficient to cause heritable genetic abnormalities (Muller, 1927). With Bridges physical maps linking genes to their chromosomal locations and Mullers mutagenesis method, many scientists were able to systematically mutate chromosomal segments and observe how this affected certain characteristics of D. melanogaster. This later became one of the most important tools in genetics: by observing the phenotypic consequences of mutating a gene or genetic locus, one could infer its function. The capacity to successfully breed them in large numbers, their outwardly visible phenotypes, and the ability to screen many mutations for their phenotypic outcomes helped designate D. melanogaster as a premier model organism for genetic research (Rubin and Lewis, 2000).

It was not until decades after this initial work that the biochemical nature of DNA was discovered and established as the primary unit of heritability (Watson and Crick, 1953). We now know that chromosomes contain genes in a particular order and that these genes are made up of DNA. Although it is sometimes difficult to define what it means to be a gene, one of the more popular working definitions is a region of DNA that encodes for a particular protein product. These enzymes collectively perform the most basic functions in living things, and their expression is one of the most important determinants of an organisms characteristics. The concept of gene expression was firmly established in the central dogma of molecular biology: the transcription of DNA into RNA and the translation of RNA into protein (Crick, 1970). While perhaps understated, Francis Crick also summarized in his 1970 Nature article that, The principal problem could then be stated as the formulation of the general rules for information transfer from one polymer with a defined alphabet to another. These observations helped cement DNA as the primary unit of heritability and provided generations of scientists the tools to study how an organisms genotype determines

its phenotype. A well-characterized example directly linking the expression of a gene to its phenotypic outcome is that of the PAX6 gene. Initially discovered in mice, the protein product of this gene has been shown to be necessary and sufficient to develop the complex eye (Hill et al., 1991). Orthologs of this gene have been found in fruit fly (Quiring et al., 1994) and human where it has also been shown to be required for normal eye development, with mutations in humans causing a condition known as aniridia (Glaser et al., 1992). To demonstrate even further the conserved function of this gene, Halder and colleagues ectopically expressed the *D. melanogaster* ortholog ey in wings, legs, and antennae and found that they developed eye-like structures (Halder et al., 1995). This and many other examples helped emphasize the importance of gene expression in determining phenotypes as well as the conservation of this process across all living things.

After establishing the dependence of the genotype-to-phenotype relationship on gene expression, one important area of biological research became focused on explaining the heterogeneity of phenotypes within and between individuals and how these differences have evolved. Consider the fact that each cell of complex eukaryotes contains the same DNA, yet groups of neighboring differentiated cell types in the form of tissues only express certain genes important to their function (Levine, 2010). This is largely due to differences in where and when the expression of certain genes takes place during development, also known as spatiotemporal patterns of gene expression (Ong and Corces, 2011). Decades after the work of T.H. Morgan, Christiane Nüsslein-Vollhard and Eric Wieschaus demonstrated the importance of certain genes through mutational screening in the developing *D. melanogaster* embryo (Nüsslein-Volhard and Wieschaus, 1980), which we now know are responsible for establishing spatiotemporal gradients in gene expression (Frohnhöfer and Nüsslein-Volhard, 1986). Specifically, they identified 15 loci that, when mutated, caused disruptions in the segmental patterning of the embryo by either duplicating each segment, deleting alternating segments, or deleting groups of adjacent segments. It is easy to imagine that subtle changes in the expression of any one of these patterning genes could cause a range of phenotypic differences, thus providing nature the raw material on which to select particular characteristics. In addition to those within species (Whitehead and Crawford, 2006), changes in gene expression are also common between closely- and distantly-related species (Rifkin et al., 2003; Khaitovich et al., 2004; Rifkin et al., 2005; Lemos et al., 2005).

The orchestration of these spatiotemporal and overall gene expression differences is broadly known as gene regulation and occurs at the level of transcription (Wray et al., 2003). One of the earliest examples of gene regulation is that of the *lac* operon in *Escherichia coli*, whereby the expression of lacZ is normally repressed, but the presence of lactose alters this repression and allows for its activation (Jacob and Monod, 1961). A repressor that is bound to the upstream DNA sequence prevents RNA polymerase from transcribing this gene, but the presence of lactose alters the repressor in such a way as to prevent it from binding to the regulatory sequence and allowing RNA polymerase to transcribe this gene which, when translated, produces the enzyme β -galactosidase, which is necessary for the cell to utilize lactose as source of energy. One can also imagine a situation whereby a single mutation in this regulatory sequence prevents the binding of the repressor, thus the lacZ gene would be constitutively expressed even in the absence of lactose. Although this example is fairly simple, it highlights one of the main components of transcriptional regulation and one whose effects will be discussed throughout this dissertation: the complex interplay between *trans*-acting factors and the *cis*-regulatory sequences they target to alter gene expression (Wittkopp et al., 2004; Wittkopp, 2005).

As differences in gene expression function to drive phenotypic diversity, so then must the regulatory mechanisms governing them; it has been shown that *cis*- and *trans*-regulatory changes cause gene expression differences both within and between species (Wittkopp et al., 2008a,b; Carroll, 2008). Although early work seemed to indicate higher-than-expected levels of conservation of regulatory sequences between species of *Drosophila* (Bergman and Kreitman, 2001), sequencing of the *D. pseudoob*scura genome showed a general lack of conservation in regulatory sequences. We also know that this lack of conservation can be attributed to transcription factor binding site turnover, whereby changes in the underlying *cis*-regulatory sequence do not affect its function across different species (Ludwig et al., 2005; Venkataram and Fay, 2010; Bradley et al., 2010). These observations offered support to the hypothesis that changes in gene regulation, like changes in the coding sequences of individual genes, drive phenotypic variation between species (Richards et al., 2005; Wittkopp, 2006). In fact, it has been shown that the proportion of total regulatory divergence attributable to differences in *cis*-regulation is greater when comparing between different species to within species (Wittkopp et al., 2008b; Emerson et al., 2010; Coolon et al., 2014). Finally, prior to the sequencing of the human genome, it was hypothesized that gene content increased linearly with organismal complexity. However, the human genome sequence revealed roughly 20,000 genes, a number not much different from that in *D. melanoqaster* or the roundworm *Caenorhabditis elegans*. This affectionately became known as the G-value paradox and required scientists to consider more than genome size or gene content when evaluating the level of complexity of an organism (Hahn and Wray, 2002). It is mostly accepted that the size and complexity of the proteome, driven by a gene regulatory mechanism known as alternative splicing, correlates well with our perception of organismal complexity (Schad et al., 2011). This further highlights the importance of gene regulation and its contribution to phenotypic variation.

Having established the importance of gene expression and its regulation, it is equally important to discuss how this entity is measured, as the scientific process is dependent on our ability to make observations. One of the earliest methods of measuring gene expression was the northern blot, whereby a labelled DNA probe is hybridized to a sample of RNA that has undergone electrophoresis on an agarose gel (Alwine et al., 1977). The RNA that was targeted by the labelled DNA probe can then be visualized on the gel and used to compare gene expression between a variety of samples. Another method of measuring gene expression that relies on a visual output involved placing a genes regulatory sequence in front of a reporter gene whose expression could then easily be identified. One such example of a reporter gene that transformed the study of gene expression is the green fluorescent protein (GFP) which, as the name suggests, fluoresces a brilliant green when exposed to ultraviolet light (Chalfie et al., 1994). Putting GFP under the control of the regulatory sequence for a particular gene of interest allows that genes expression output to be measured as a level of fluorescence; it also serves to indicate the precise location of its expression. One of the more sensitive methods to measure gene expression is that of quantitative reverse-transcription polymerase chain reaction (qRT-PCR), which is a technique to measure the amount of reverse-transcribed sample RNA, or complementary DNA (cDNA), as it is amplified by PCR (Becker-Andre and Hahlbrock, 1989; Heid et al., 1996).

All of the methods listed above are limited to the study of one or several genes at any given time, as they are expensive and inefficient assays to measure even a dozen or more genes. This makes their applicability to studying broad levels of gene expression very limited. However, in the mid-1990s, methods started to become available that allowed for efficient, sensitive, and cost-effective quantification of dozens of genes simultaneously. In one of the earliest examples of this technology, known as a microarray, 48 cDNAs from Arabidopsis thaliana were amplified and placed in a 96-well plate, and fluorescent probes were added to each well (Schena et al., 1995). Using a laser, their hybridization intensity was measured as the output of emitted light, which was used to estimate relative levels of gene expression. As with other new technologies, what followed the gene expression microarray was a host of background correction, normalization, and analysis techniques so that gene expression estimates could be correctly estimated and compared between samples (Quackenbush, 2002). This technology has since evolved to measure nearly complete sets of known genes in many model organisms. Although this represented a significant advance in the study of gene expression, microarrays were not without their own pitfalls. Due to the nature of the biochemical methods used in microarrays, the cDNA present in a sample could only hybridize to as many probes as were fixed to the array, which led to a saturation of the signal of hybridization intensity (Scott et al., 2009). This made it difficult to accurately measure highly- and lowly-expressed genes concurrently, limiting the range of gene expression that could be detected. Nevertheless, when used correctly, the gene expression microarray has proven to be a powerful technique to generate lists of candidate genes whose expression differences between samples are potentially driving observed biological variability.

As the costs of sequencing continued to plummet during the last decade, the use of microarray technology to measure gene expression was slowly replaced by what became known as next-generation sequencing (NGS) technologies (Wang et al., 2009; Metzker, 2009). Like microarrays, NGS is largely dependent on the detection of light, the difference being that microarrays measure hybridization intensities, whereas NGS detects the incorporation of individual nucleotides. When NGS is applied to a sample of RNA that has been reverse-transcribed into cDNA and then fragmented to a particular size, it is most popularly referred to as RNA-seq (Wilhelm et al., 2008). Unlike gene expression microarrays, which depend on probes for a known set of genes, RNA-seq allows the highly-parallelized sequencing of short fragments from a sample, which could comprise the known set of genes as well as those that have yet to be characterized. Some key advantages of RNA-seq over gene expression microarrays include capturing uncharacterized variant transcripts from alternative splicing, identifying transcribed sequence variants, a low signal-to-noise ratio, and a greater dynamic range of expression levels (Wang et al., 2009).

The analysis of RNA-seq data presented a unique set of challenges requiring a variety of interdisciplinary approaches. As opposed to microarrays, where each datum represents the level of expression of a particular gene, transcript, or exon, RNA-seq produces sequence reads, each representing a small fragment of a larger piece of sequence randomly drawn from an even larger pool of combined sequences in a given sample. In order for this type of data to be useful, bioinformaticians needed to determine the precise genomic location of each short sequence read which, in the current realm of high-throughput sequencing technologies, easily reach total numbers on the order of 10⁸ (Boland et al., 2013). It quickly became clear that conventional string-searching methods applied to data of this magnitude were not sufficient. Methods were developed that took advantage of algorithms that could search for strings in a compressed version of the highly-redundant genome (Li and Durbin, 2009; Langmead et al., 2009). This search could be done quickly and with a modest amount of compu-

tational resources, which was key to the success of such methods. By aligning a set of sequence reads derived from RNA-seq to a genome, the number of reads falling within annotated genes could be used as a proxy for that genes relative level of expression (Mortazavi et al., 2008; Marioni et al., 2008). While this depends on the availability of an assembled genome with annotated genes, creating a genome or transcriptome using de novo assembly in non-model organisms is becoming more commonplace and will someday no longer be a limitation in gene expression studies (Grabherr et al., 2011; Martin and Wang, 2011). Following closely behind this bioinformatic challenge was the statistical challenge of analyzing complex count data (Bullard et al., 2010), including methods to assign statistical significance to differences in expression (Robinson et al., 2010; Anders and Huber, 2010; Hardcastle and Kelly, 2010; Tarazona et al., 2011) as well as normalizing read counts across samples (Robinson and Oshlack, 2010; Dillies et al., 2012; Hashimoto et al., 2014). This explosion of gene expression data produced many statistical packages related to quantifying total gene expression as well as evaluating differences among samples, which have been rigorously tested (Oshlack et al., 2010; Rapaport et al., 2013; Soneson and Delorenzi, 2013).

While the ability to quantify total levels of gene expression is important to establish differences between samples, it does not provide enough information to characterize how individual genes are being regulated. As alluded to earlier, gene expression at the level of transcription is typically regulated by diffusible *trans*-acting molecules such as transcription factors that bind to *cis*-regulatory DNA sequences near a gene. It is very difficult to characterize the effects on gene expression of changing a particular transcription factor by either altering its expression pattern or its coding sequence. It is even more challenging to identify the nucleotide(s) in *cis*-regulatory sequences causing changes in gene expression for a single gene, not to mention for many genes. Instead of searching directly for these individual changes in *cis*-regulatory sequences or *trans*-acting factors, we can observe their downstream regulatory effects on gene expression. In F1 hybrids of diploid, sexually-reproducing organisms, gene expression represents the contribution from each parental allele. Because each allele experiences the same set of *trans*-acting factors as contributed by each parent, differences in allele-specific expression (ASE) serve as a proxy for *cis*regulatory activity (Cowles et al., 2002; Yan et al., 2002; Wittkopp et al., 2004). Inherent to this logic is the assumption that they share common regulatory sequence and that their set of *trans*-acting factors can interact equally well with each allele, but this assumption is quite reasonable, even for divergent species that can still interbreed and form viable F1 hybrid offspring.

Although gene expression levels affected by *cis*-regulatory changes can be directly inferred from differences in ASE, identifying those affected by *trans*-acting changes is more difficult. The total level of gene regulatory variation, measured by directly comparing parental levels of expression, can be thought of as the sum of *cis*- and *trans*-regulatory components (Wittkopp et al., 2004). In this respect, to determine the *trans*-regulatory component, one needs only to subtract the *cis*-regulatory component from the total level of gene expression variation between samples. That is, if a gene is differentially expressed between parents, but there are no allele-specific differences attributing this change to *cis*-regulation, then we infer that a *trans*-acting change is affecting a genes expression level. By applying bioinformatic strategies mentioned above, RNA-seq data has been used to understand how genes are being regulated across the transcribed genome both within (Emerson et al., 2010) and between species (McManus et al., 2010).

1.1 Dissertation outline

In this dissertation, I characterize how the regulation of gene expression has evolved both within and between several species in the *Drosophila* lineage. As these inferences are made across the expressed genome, they require extensive analysis of RNA-seq data from individual parental strains and/or species as well as their intraand interspecific F1 hybrids. Compared to methods to accurately quantify total gene expression, such methods for allele-specific expression (ASE) are lacking. This is likely because experimental conditions for such analyses are often more specialized than for measuring total gene expression. Because of this, I developed bioinformatics tools and pipelines to analyze gene expression data in an allele-specific manner. As I will later show, such analyses also depend on adequate genomic resources for all strains and/or species considered here, requiring me to develop expertise in classifying sequence variation and incorporating such variation into custom-built genomes.

In chapter II, I expand on the bioinformatic strategies I developed. I begin by describing current methodology to accurately quantify ASE from RNA-seq data. It had previously been shown that when aligning RNA-seq data to a single reference genome to capture allelic variation, estimates of ASE were systematically biased toward the allele represented by the reference genome (Degner et al., 2009). However, the cause of this bias was largely ignored, as were loci harboring such bias. In addition to comparing two different methods of quantifying ASE, I highlight the primary sources of bias that can have undesirable effects when estimating ASE and include recommendations for improving the accuracy of these estimates.

In chapter III, I use this methodology to measure allele-specific differences in F1 hybrids made by reciprocally crossing two strains of *D. melanogaster*. These two sets of genetically-identical hybrids differed only by which strain contributed the maternal or paternal allele, allowing me to compare their ASE profiles and test the hypothesis that D. melanogaster do not imprint their genome. Genomic imprinting is a phenomenon whereby either the maternal or paternal allele is epigenetically silenced, and can be detected by comparing ASE profiles between these hybrids. Imprinting has been shown to be important in mice and humans and is involved in X-chromosome inactivation. Improper silencing of alleles can also cause certain diseases, which makes it an important aspect of human health. Previous work had shown that D. melanoque do not undergo genomic imprinting, but this was shown for a relatively small number of genes (Wittkopp et al., 2006). Using RNA-seq to test for this phenomenon across the expressed genome, I found marginal evidence against this hypothesis, although it turned out to be an artifact of the samples used that had retained a lowly-segregating heterozygous deletion, whose clustered patterns of differential ASE appeared as genomic imprinting. This clearly demonstrates the caution one must use when inferring patterns of genomic imprinting from RNA-seq data.

In chapter IV, using the same intraspecific comparison as well as two interspecific comparisons between D. simulans and D. sechellia and between D. melanogaster and D. simulans, I measure total and allele-specific gene expression to categorize *cis*-and *trans*-regulatory differences across divergence times ranging from 0.01-2.5 million years ago. This allowed me to test the hypothesis that *cis*-regulatory differences account for more of the total regulatory differences driving expression divergence between inter- and intraspecific comparisons (Wittkopp et al., 2006), as well as determine how gene expression patterns have been inherited differently across a range of evolutionary times.

In chapter V, I test the hypothesis that sex- and tissue-specific differences in gene expression variation are prevalent in *Drosophila*. All of the comparisons in previous chapters were made using only female whole fly tissues. One of the main functions of gene regulation is to differentiate tissues within organisms, so it is reasonable to infer that, depending on the tissue, patterns of regulatory divergence between species may differ. Using gene expression data from female and male carcass and gonad tissues between *D. pseudoobscura* and its closely-related subspecies *D. p. bogotana*, I determined that patterns of inheritance of gene expression as well as regulatory divergence differ between sexes and across tissues. I also found that gene expression though not for non sex-biased genes. One must be careful when inferring patterns of regulatory divergence in whole organisms, as the integration over all different tissue types can hide the complexity of gene regulation.

The work presented in this dissertation has greatly expanded our knowledge of how the regulation of gene expression differs across a well-characterized lineage and will continue to drive further studies of these phenomena in even more distantlyrelated species.

CHAPTER II

Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome

2.1 Abstract

RNA-seq can be used to measure allele-specific expression (ASE) by assigning sequence reads to individual alleles; however, relative ASE is systematically biased when sequence reads are aligned to a single reference genome. Aligning sequence reads to both parental genomes can eliminate this bias, but this approach is not always practical, especially for non-model organisms. To improve accuracy of ASE measured using a single reference genome, we identified properties of differentiating sites responsible for biased measures of relative ASE. We found that clusters of differentiating sites prevented sequence reads from an alternate allele from aligning to the reference genome, causing a bias in relative ASE favoring the reference allele. This bias increased with greater sequence divergence between alleles. Increasing the number of mismatches allowed when aligning sequence reads to the reference genome and restricting analysis to genomic regions with fewer differentiating sites

Official citation:

Stevenson, K.R., Coolon, J.D., Wittkopp, P.J. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome, *BMC Genomics* 2013, 14:536 doi:10.1186/1471-2164-14-536 Copyright 2014 BioMed Central Ltd

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

than the number of mismatches allowed almost completely eliminated this systematic bias. Accuracy of allelic abundance was increased further by excluding differentiating sites within sequence reads that could not be aligned uniquely within the genome (imperfect mappability) and reads that overlapped one or more insertions or deletions (indels) between alleles. After aligning sequence reads to a single reference genome, excluding differentiating sites with at least as many neighboring differentiating sites as the number of mismatches allowed, imperfect mappability, and/or an indel(s) nearby resulted in measures of allelic abundance comparable to those derived from aligning sequence reads to both parental genomes.

2.2 Introduction

During the last five years, massively parallel sequencing of cDNA libraries synthesized from RNA samples (known as "RNA-seq") has largely replaced the use of microarrays for comparative studies of gene expression (Marioni et al., 2008; Metzker, 2009; Wang et al., 2009). Advantages of RNA-seq over microarrays include a greater dynamic range and the ability to survey expression in new strains and species without the set-up costs of microarrays and without complications from hybridization differences among genotypes (Mortazavi et al., 2008; Brawand et al., 2011). In addition, because RNA-seq provides full sequence information for the transcriptome, it is better suited for discovering novel transcripts and splice isoforms and for quantifying allelic abundance in heterozygous and mixed genotype samples than microarrays. Measures of allele-specific expression (ASE) are particularly important for studying the regulation of gene expression because they can be used to distinguish *cis*- and *trans*-regulatory changes (Cowles et al., 2002; Wittkopp et al., 2004) and to detect genomic imprinting (Coolon et al., 2012; DeVeale et al., 2012).

To quantify transcript abundance using RNA-seq, each short sequence read (hereafter simply called a "read") is compared to an annotated reference genome. Assignment of a read to a specific gene is made by finding the region of the genome with the highest sequence similarity, and the number of reads aligning to a gene is used as a proxy for its relative expression level (Mortazavi et al., 2008). Mapping reads to specific genes is relatively straightforward with the bioinformatics tools available today (Li et al., 2008; Langmead et al., 2009; Li et al., 2009; Li and Durbin, 2009), but using these tools to distinguish between reads derived from alternative alleles of the same gene remains challenging (DeVeale et al., 2012). This challenge was most clearly demonstrated by Degner et al., who simulated reads from a heterozygous human genotype and assigned them to specific alleles after mapping to a reference human genome (Degner et al., 2009). Reads perfectly matching the reference genome were assigned to the reference allele, whereas reads containing mismatches to the reference genome were assigned to the alternative allele. Despite simulating an equal number of reads from each allele, a bias was observed causing reads to be assigned more often to the reference allele than the alternative allele. Controlling for sites known to be polymorphic in humans prior to aligning the simulated reads produced symmetrical measures of relative ASE, showing that the differentiating sites themselves caused this bias.

Recently, two alternative strategies for aligning reads have been shown to eliminate the systematic bias in measures of relative ASE favoring the reference allele. In the first, RNA-seq reads are aligned separately to maternal and paternal genomes. These allele-specific genomes can be generated either by sequencing inbred lines with the maternal and paternal genotypes (Coolon et al., 2012; McManus et al., 2010; Graze et al., 2012; Shen et al., 2012) or by inferring the maternal and paternal haplotypes using phased genotype information such as that available for humans from the 1000 Genomes Project Consortium (Rozowsky et al., 2011; Rivas-Astroza et al., 2011). However, researchers interested in measuring relative ASE in organisms for which parent-specific genomes cannot be readily obtained will struggle to use this approach. The second strategy is to consider all possible phasings of variants that can occur in the same sequence read and either supplement the reference genome with these haplotypes (Satya et al., 2012) or use this information during alignment with a polymorphism-aware aligner, such as GSNAP (Wu and Nacu, 2010; Skelly et al., 2011). This is a viable strategy for both model and non-model species, but will likely be most effective for intraspecific studies of species like humans with relatively low levels of polymorphism because the number of possible haplotypes increases exponentially with the number of polymorphic sites.

To better understand the source(s) of biased measures of relative ASE, we identified properties of sites showing inaccurate measures of relative ASE using simulated *Drosophila* sequencing data with known values of relative allelic abundance. Simulated datasets contained either \sim 10-fold or \sim 100-fold more differentiating sites than the human genotypes used to validate other methods for measuring relative ASE (Degner et al., 2009; Rozowsky et al., 2011; Satya et al., 2012). We also examined the impact of these factors on measures of relative ASE derived from real sequencing data. Reads from simulated and real sequencing data were aligned to a single reference genome, varying the number of mismatches allowed, as well as aligned to separate maternal and paternal genomes with no mismatches allowed. We found that limiting analysis of relative ASE to regions of the genome with no more differentiating sites than the number of mismatches allowed eliminated the systematic bias toward the reference allele and produced measures of ASE similar to those inferred from aligning reads separately to the maternal and paternal genomes. Excluding differentiating sites contained within reads that cannot be aligned uniquely or that overlap an insertion or deletion (indel) further improved measures of relative allelic abundance.

2.3 Results and Discussion

2.3.1 The systematic bias in measures of ASE correlates with the density of differentiating sites

As described above, Degner *et al.* found that allele-specific reads mapped preferentially to the reference allele when using a single reference genome to quantify ASE (Degner et al., 2009). The alignment parameters they used allowed two or fewer bases within each read to differ from the reference genome. Reads perfectly matching the reference genome were assigned to the reference allele, while reads with at least one difference from the reference genome were assigned to an alternative allele. We hypothesized that the inability to map reads with more differences from the reference genome than mismatches allowed underestimated the abundance of the alternative allele and caused measures of ASE to be biased toward the reference allele.

To test this hypothesis, we generated an equal number of reads from two genotypes in silico, combined them, and measured the relative abundance of allele-specific reads. These sequences were derived from 52,370 non-overlapping constitutivelyexpressed exons in *Drosophila melanogaster* (McManus et al., 2010). The annotated D. melanogaster genome (dm3) was used as the "reference" allele, and an edited version of this genome with 93,781 coding sites altered to match alleles in a line of D. melanogaster from the Drosophila Genetic Reference Panel (Ayroles et al., 2009; Mackay et al., 2012) was used as an "alternative" allele. We generated 36-base reads from each allele starting at every possible position in each exon and repeated this process for both strands of DNA because RNA-seq is usually performed using double-stranded cDNA (Figure 2.1). This process generated 93,395,272 reads, representing 3.4 Gb of sequencing data. Importantly, this approach guaranteed that reads from each allele were present in equal amounts. To quantify relative allelic abundance as a proxy for relative ASE, we aligned each read to the reference genome using Bowtie (Langmead et al., 2009), excluding reads that mapped to multiple locations, and evaluated the number of reads assigned to the reference and alternative alleles at each differentiating site using SAMtools (Li et al., 2009).

Initially, we allowed one mismatch to the reference genome during the alignment step, which is the minimum number required to align a read from the alternative allele. We found that 50.9% of differentiating sites had unequal measures of allelic abundance, 99.3% of which were biased toward the reference allele. To determine whether this bias was influenced by the density of differentiating sites, we calculated the maximum number of sites that differed between the two alleles among all possible 36-base reads overlapping each differentiating site (Figure 2.1). Of all sites considered, 49.8% had at least one neighboring differentiating site (i.e., at least one other differentiating site within an overlapping read). Of these sites, 99.8% showed more reads assigned to the reference allele than to the alternative allele. Furthermore, the extent of bias toward the reference allele increased with the number of neighboring differentiating sites (Figure 2.2A). This bias was caused by the failure of reads simulated from the alternative allele to align to the reference genome more often than those simulated from the reference allele. Aligning reads to only the alternative allele produced complementary results (Figure 2.7). These findings are consistent with our hypothesis that the density of differentiating sites complicates the mapping of reads and leads to biased measures of relative ASE.

To decrease the impact of neighboring differentiating sites on allelic assignment, we allowed two or three mismatches when aligning our simulated reads to the reference genome. We found that increasing the number of mismatches improved measures of allelic abundance: 80.2% and 91.9% of differentiating sites were inferred to be equally abundant when two and three mismatches, respectively, were allowed. A bias toward the reference allele was still observed, but only for sites where the number of neighboring differentiating sites was greater than or equal to the number of mismatches allowed during the alignment step (Figure 2.2B,C). Increasing the number of mismatches allowed reduced the bias toward the reference allele, but increased the percentage of reads that failed to map uniquely: allowing one, two, and three mismatches, 2.2%, 2.5%, and 2.9% of all reads failed to map uniquely, respectively.

For comparison, we aligned the simulated reads independently to the reference and alternative genomes with the same parameters used when aligning reads to the single reference genome except that zero mismatches were allowed. This is analogous to aligning reads to the maternal and paternal genomes, which is a strategy that has previously been shown to produce unbiased measures of relative ASE (Coolon et al., 2012; Rozowsky et al., 2011; Rivas-Astroza et al., 2011; Satya et al., 2012; Graze et al., 2009). We found that 99.0% of differentiating sites showed equal representation of the two alleles, with the rest showing no systematic bias toward either allele (Figure 2.2D). Only 1.9% of all reads were excluded because they failed to map uniquely to at least one genome.

2.3.2 Read length and the amount of sequence divergence can also affect allelic bias

Given the observed impact of neighboring differentiating sites on allelic assignments, we hypothesized that longer reads might produce less accurate measurements of allele-specific abundance because they should overlap more neighboring differentiating sites. To test this hypothesis, we repeated our simulation with 50-base reads, determining the maximum number of sites that differed between the two alleles among all possible 50-base reads overlapping each differentiating site. We found that 40.6%, 73.0%, and 88.9% of differentiating sites showed equal representation of the two alleles when aligned to a single reference genome with one, two or three mismatches allowed (Figure 2.2E-G). Increasing the number of mismatches allowed when aligning the 50-base sequence reads to be more similar to the ratio of mismatches allowed for the 36-base sequence reads eliminated this difference, however. 91.9% and 92.1% of differentiating sites showed equal allelic abundance for 36- and 50-base reads when three and four mismatches, respectively, were allowed (Figure 2.8). By contrast, 98.8% of differentiating sites showed equal representation when reads were aligned to the maternal and paternal genomes with zero mismatches allowed (Figure 2.2H).

Increased sequence divergence is also expected to affect measures of relative allelic abundance because it should increase the average number of neighboring differentiating sites within each read. To test this hypothesis, we simulated 36-base reads from two different *Drosophila* species (*D. melanogaster* and *D. simulans*) (Graze et al., 2012) and analyzed them as described above, using the *D. melanogaster* exome as the single reference genome. Sequences from 60,040 orthologous exons with 1,130,435 differentiating sites were used for this simulation, which is an order of magnitude more differentiating sites than between the two strains of *D. melanogaster* analyzed. As predicted, we found that the bias toward the reference allele was higher for the interspecific comparison than for the intraspecific comparison when reads were aligned to a single reference genome (Figure 2.2, compare I-K with A-C). When aligning reads to both parental genomes, however, sequence divergence had a negligible impact: the intra- and interspecific datasets produced nearly identical results (Figure 2.2, compare L with D).

2.3.3 Allele-specific differences in mappability and insertions/deletions affect measurements of ASE

Differences between alleles in sequences that appear more than once in the genome can also cause reads to be excluded for one allele but not the other (Degner et al., 2009). Assuming the number of such differentiating sites is similar between alleles, differences in allele-specific mappability should not systematically favor one allele or the other, but will still cause errors in relative ASE. To examine the impact of mappability on measures of relative allelic abundance derived from our simulated data, we used software from the GEM library (Derrien et al., 2012) to calculate a mappability score for each differentiating site by averaging the mappability scores of all possible reads that included that site. In each case, mappability scores were calculated using the same number of mismatches allowed during read alignment. Differentiating sites with an average mappability score <1 were considered to have imperfect mappability when using a single reference genome. When using parental genomes, we summed the average mappability scores for each allele, and mappability scores <2 were considered to have imperfect mappability.

We then compared relative allelic abundance for sites with perfect and imperfect mappability in all three simulated datasets (Figure 2.3), excluding sites with more neighboring differentiating sites than the number of mismatches allowed when aligning to a single reference genome. For both the 36- and 50-base reads simulated from the two *D. melanogaster* genotypes, >97.9% of sites with perfect mappability showed the expected equal abundance of the reference and alternative alleles under all mapping conditions (Figure 2.3A-H). For the 36-base reads simulated from the D. melanogaster and D. simulans genomes, 99.9% of sites with perfect mappability showed equal abundance when reads were aligned to both parental genomes (Figure 2.3L), but only \sim 94% of sites with perfect mappability showed such equal abundance when reads were aligned to a single D. melanogaster reference genome (Figure 2.3I-K).

We hypothesized that this decrease in accuracy after aligning D. melanogaster and D. simulans reads to a single reference genome might be caused by the presence of insertions or deletions (indels) between D. melanogaster and D. simulans that are located near differentiating sites (i.e., within the length of a read from the differentiating site). Such indels can prevent the alignment of D. simulans reads to the D. melanogaster genome. Consistent with this hypothesis, we found that sites with perfect mappability that had an indel nearby showed more reads assigned to D. melanogaster than D. simulans allele when reads were aligned to only the D. melanogaster genome, whereas sites with perfect mappability that lacked such an indel did not (Figure 2.4A-C). When reads were aligned to both parental genomes, sites with perfect mappability showed equal representation of the two alleles regardless of the presence or absence of nearby indels (Figure 2.4D). Indels were not a factor in our comparisons of the two D. melanogaster strains because the alternative allele was constructed by changing only single nucleotides in the reference allele.

2.3.4 Aligning real sequencing data to a single genome can produce reliable measures of relative ASE

Assessing the accuracy of relative ASE measurements derived from RNA-seq data is challenging because the true value of relative ASE is rarely known. Independent empirical methods for measuring relative ASE such as Pyrosequencing and qPCR can be used to validate RNA-seq data for individual genes, but they are not suitable for quantifying relative ASE on a genomic scale. Therefore, instead of using real RNAseq data to evaluate factors affecting measures of relative ASE, we used sequence data that was collected in a comparable manner from genomic DNA extracted from F1 hybrids, in which all maternal and paternal alleles are expected to be present in equal amounts.

Specifically, we used 36-base reads from genomic DNA extracted from female F1 hybrids that were produced by crossing inbred strains of D. melanogaster and D. simulans (Graze et al., 2012). These strains had the same genotypes as the D. melanogaster and D. simulans sequences used for the interspecific simulation described above. Reads were aligned to the D. melanogaster exons allowing one, two, or three mismatches, as well as to both the D. melanogaster and D. simulans exons allowing zero mismatches. Because real sequencing data involves stochastic sampling, the proportion of the reference allele observed was not always expected to be 0.5. Therefore, after aligning reads, we excluded differentiating sites with fewer than 20 overlapping reads and used binomial exact tests with a false discovery rate threshold of 0.05 to test each differentiating site for a statistically significant difference in relative allelic abundance (McManus et al., 2010; Fontanillas et al., 2010a).

As described above, our simulated datasets showed that reads containing (1) as many or more neighboring differentiating sites as mismatches allowed during alignment, (2) imperfect mappability, and/or (3) an indel(s) between alleles can cause inaccurate measures of relative allelic abundance. Differentiating sites with an excess of neighboring differentiating sites were the most common of these three types of problematic sites in both intra- and interspecific simulations (Figure 2.5A). To determine the relative impact of each of these factors on measures of allele-specific abundance derived from real sequencing data, we filtered the differentiating sites based on each factor sequentially and determined the percentage of differentiating sites retained that had no statistically significant difference in abundance between alleles (hereafter referred to as "equal allelic abundance") for each alignment strategy.

Prior to excluding any sites, 70.4%, 88.9%, and 93.3%, respectively, of all differentiating sites showed equal allelic abundance when reads were aligned to a single genome with one, two, or three mismatches allowed. After aligning reads to both parental genomes, 96.9% showed evidence of equal allelic abundance. Excluding differentiating sites with at least as many neighboring differentiating sites as the number of mismatches allowed increased this percentage to 96.3%-96.6% when aligning to a single reference genome (Figure 2.5B). Further restricting the set of differentiating sites to those with perfect mappability increased these percentages ~0.1%, and subsequently excluding differentiating sites with indels nearby increased the percentage of genes with equal allelic abundance an additional ~0.1% (Figure 2.5B). After filtering out these problematic sites, measures of relative allelic abundance derived from aligning reads to a single reference genome were similar to those produced by aligning sequence reads separately to the maternal and paternal genomes (Figure 2.5C-E).

2.3.5 Excluding selected differentiating sites maintains ability to measure relative ASE for most exons

We focused on measures of relative ASE for individual sites in this study, but most researchers are more interested in relative ASE for individual exons and/or genes. The major consequence of excluding sites based on the density of differentiating sites, mappability, and/or indels is that fewer allele-specific reads will be successfully mapped for each exon and for each gene. After filtering based on the number of neighboring differentiating sites, we found that 46.6%-86.9% and 8.3%-50.5% of differentiating sites were retained in the 36-base intra- and interspecific simulations, respectively, when the reads were aligned to a single reference genome and one, two, or three mismatches were allowed (Figure 2.6). By comparison, 81.8%-91.8% and 66.3%-95.2% of exons contained at least one of these reliable differentiating sites when the same alignment conditions were used in the intra- and interspecific simulations, respectively. Excluding additional differentiating sites with imperfect mappability in both datasets, as well as sites with one or more nearby indels in the intraspecific dataset, had little effect on the proportion of differentiating sites and exons retained (Figure 2.6). The retention of more differentiating sites and exons in the intraspecific simulation than in the interspecific simulation (Figure 2.6) is consistent with the lower sequence divergence within than between species. Analyses using real and simulated reads to compare the same sets of alleles retain the same sites and exons when aligned to the same reference genome because differentiating sites are excluded based only on the genome sequence(s).

2.4 Conclusions

RNA-seq is a powerful tool for measuring ASE on a genomic scale; however, a systematic bias occurs when reads from a heterozygous individual are aligned to a single reference genome (Degner et al., 2009). We found that this systematic bias is predominantly caused by additional differentiating sites located near the focal differentiating site that interfere with read alignment. A similar bias toward the reference allele is caused by the presence of an indel near the focal differentiating site. Differences between alleles in mappability (i.e. the ability to align a read uniquely within the genome) also contribute to inaccuracy of ASE, but do not systematically favor one allele or the other across the genome.

Using both simulated and real sequencing data, we found that sites affected by the systematic bias toward the reference allele could be identified and excluded prior to estimating ASE based on the density of differentiating sites. The precise density at which neighboring differentiating sites became problematic depended on the number of mismatches allowed during the alignment of sequencing reads. After excluding these biased sites, as well as those affected by imperfect mappability and/or an indel(s) nearby, we found that RNA-seq data aligned to a single reference genome produced measures of relative ASE that were comparable to those resulting from separately aligning the same reads to allele-specific maternal and paternal genomes. Furthermore, we showed that excluding these problematic sites did not preclude measuring relative ASE for most exons, although the most rapidly evolving exons are expected to be preferentially eliminated. By identifying the specific factors causing erroneous measures of relative allele-specific expression reported in prior work and determining the relative impact of these factors on these measures, results from this study are expected to foster further improvements in methods for quantifying relative allele-specific expression.

2.5 Methods

2.5.1 Generating allele-specific short reads comparing *D. melanogaster* genotypes in silico

Simulating an allele-specific RNA-seq experiment requires variability to differentiate alleles and a set of clearly defined transcriptional units from which to generate allele-specific reads. Using data from the Drosophila Genetic Reference Panel (DGRP), we examined site-specific sequence information from a single highly-inbred line ("line_40") isolated from an outbreeding population of *Drosophila melanogaster*. This specific line was chosen because it had the fewest sites with evidence of residual heterozygosity. Sequence information from this line was compared to the current build of the *D. melanogaster* genome (dm3), and sites that differed from this reference genome were retained as sites differentiating the dm3 and "line_40" alleles, referred to as the reference and alternative alleles, respectively.

Because RNA-seq experiments collect sequence information from the transcribed genome, we chose to generate reads from constitutive exons in *D. melanogaster* (Mc-Manus et al., 2010). These constitutive exons are defined as those present in all alternatively-spliced transcripts for a particular gene. We filtered out overlapping regions of exons located on opposite strands to avoid ambiguity. Starting from the 5' end of each exon, we generated 36- and 50-base reads offset by a single base in the 3' direction, for the reference and alternative alleles and in each strand orientation, creating a complete set of all possible allele-specific and strand-specific reads. This ensured that reads from each allele were present in equal abundance. Because the reference and alternative alleles differed only at these predefined differentiating sites, only reads overlapping these sites had the possibility to be informative for relative ASE.

2.5.2 Quantifying allelic abundance in simulated RNA-seq data

All alignments were performed using Bowtie v0.12.7 (Langmead et al., 2009), requiring that reads align uniquely to the genome (bowtie -f -m 1 -v [0,1,2,3] –best). Alignments were processed using SAMtools v0.1.18 (Li et al., 2009) (samtools view -S -b -T; samtools sort; samtools mpileup -f), which generates site-specific allele frequencies using overlapping reads (read pileup). ASE was quantified using custom Perl and R scripts (available upon request), and any deviation from equal allelic abundance was considered allelic imbalance.

Initially, we aligned the simulated reads to the D. melanogaster (dm3) reference

genome. Since reads generated from the alternative allele overlapping a differentiating site will have at least a single base mismatch to the reference genome, we successively allowed one (-v 1), two (-v 2), or three (-v 3) mismatches, but still required unique alignment to the reference genome (-m 1). Although the -v parameter assesses mismatches for the length of the entire read, and has an upper limit of three, an alternative parameter -n allows additional mismatches outside of a specified region at the beginning of each read, called a seed. To allow a fourth mismatch for the 50-base reads, we specified a 36-base seed region with up to three mismatches and increased the maximum sum of mismatch quality scores across the entire read to 161, since base quality scores for FASTA reads are assumed to equal 40 (bowtie -f -n 3 -e 161 -l 36 -m 1 -best). After each alignment was performed, we considered only reads overlapping the previously defined differentiating sites. We then quantified relative allelic abundance by determining whether or not each overlapping read at these sites matched the reference or the alternative alleles. These summed counts represented our measures of relative allelic abundance at each differentiating site.

Next, we aligned the same allele-specific reads independently to the aforementioned reference genome and the edited copy of the reference genome representing the alternative allele (bowtie -f -m 1 -v 0 -best). As described above, this alternative genome was obtained by editing the bases at differentiating sites to match the fixed genotypes from the DGRP "line_40" sequencing data. No mismatches were allowed when aligning simulated reads to either allele-specific genome. This allowed us to determine, for any read, whether or not it aligned uniquely to one or the other allele-specific genome. We posited that reads aligning uniquely to one or the other allele-specific genome was evidence that that read was allele-specific, while reads aligning equally well to both genomes was not. To measure relative ASE at each differentiating site, we counted the number of reads overlapping differentiating sites that aligned uniquely to only one of the allele-specific genomes and summed these counts for each allele.

2.5.3 Measuring number of neighboring differentiating sites and mappability across genomes

After quantifying allelic abundance at each differentiating site, we calculated the maximum number of other sites showing differences between alleles contained within any of the possible k-base reads, where k = simulated read length (either 36- or 50-bases). For each genome, we used the GEM-mappability tool from the GEM library build 475 (Derrien et al., 2012) to measure genome mappability, or the ability for a read from a particular location to uniquely align to a genome. For the simulated and real data, we measured mappability for the appropriate read length (either 36 or 50 bases), allowing zero, one, two, or three mismatches, with default parameters (gemmappability -1 [36,50] -m [0,1,2,3]). Mappability for individual sites was calculated using the reciprocal frequency of the number of locations a read beginning at that site would align to in the genome. To calculate mappability scores for differentiating sites, we averaged mappability for all read positions that overlapped each differentiating site (Derrien et al., 2012).

2.5.4 Quantifying relative ASE in an F1 hybrid between D. melanogaster and D. simulans

To assess the accuracy of allele-specific abundance inferred from real sequencing data, we used published 36-base Illumina reads from genomic DNA extracted from a pool of female F1 hybrids between laboratory strains of *D. melanogaster* and *D. simulans* (Berlin: BDSC 8522 and C167.4: BDSC 4736, respectively) (Graze et al., 2012). We restricted our analysis to the first mate of this set of paired-end reads, combining reads from all three technical replicates. We used the custom set of 60,040

orthologous exon sequences (exomes) between D. melanogaster and D. simulans developed in Graze et al. (Graze et al., 2012) for the reference and alternative genomes. We also used these sequences to simulate and analyze 36-base reads comparing D. melanogaster and D. simulans alleles in the same manner outlined above for the two D. melanogaster genotypes.

We first performed a pairwise alignment for each orthologous pair of exons using the Fast Statistical Alignment v1.15.7 software (Bradley et al., 2009) with default parameters (fsa –stockholm). We used custom Perl scripts to identify 1,130,435 sites that could differentiate these two alleles as well as to identify regions of the exome present in one allele but not the other (indels).

We then aligned the Illumina reads to the D. melanogaster exome, requiring unique alignment to a single location and allowing one, two, or three mismatches. We also aligned the same reads independently to the D. melanogaster- and D. simulansspecific exomes, masking indels identified by the pairwise alignments. After each of these alignments, we quantified ASE, measured the density of differentiating sites, and determined the mappability to each genome using the same strategies described above for the simulated data. We performed binomial exact tests for differentiating sites with 20 or more overlapping reads, controlling the false discovery rate at 0.05 to correct for multiple comparisons.

2.6 Acknowledgements

We thank members of the Wittkopp laboratory, especially Richard Lusk, Brian Metzger, Fabien Duveau, and Bing Yang, for helpful discussions and comments on the manuscript. This work was supported by a grant from the National Science Foundation (MCB-1021398) to PJW, a postdoctoral fellowship from the National Institutes of Health (1F32-GM089009-01A1) to JDC, and a position on National Science Foundation Training Grant No. 0903629 for KRS.

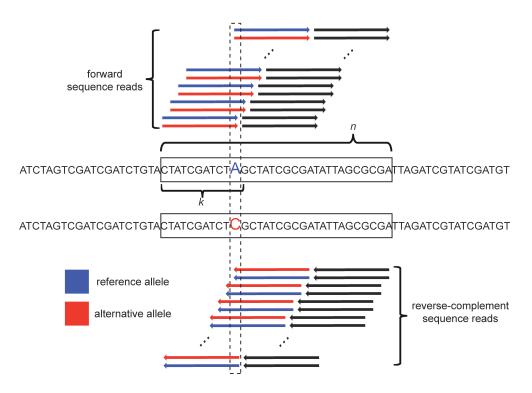


Figure 2.1: Simulating an allele-specific RNA-seq experiment. Reads were generated from the reference D. melanogaster (dm3) allele (blue) and from an alternative allele (red) that contained all homozygous single nucleotide variants found in the DGRP strain line_40. For each exon, one read (arrow) was generated starting at each position for each allele from 1 to n-k, where n is the length of the exon and k is the length of the read, both in bases. This process was repeated for the reverse complement of each exon. The black arrows indicate reads with no allele-specific information.

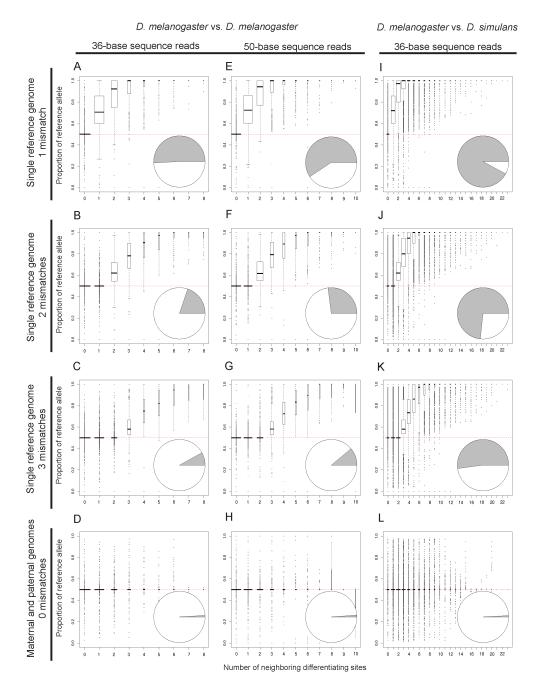


Figure 2.2: The density of differentiating sites affects relative allelic abundance when simulated reads are mapped to only one genome. Relative allelic abundance was measured using the 36-base (A-D) and 50-base (E-H) reads simulated from the two D. melanogaster genotypes as well as using the 36-base reads simulated from D. melanogaster and D. simulans (I-L) aligned to a single reference genome, allowing either one mismatch (A, E, I), two mismatches (B, F, J), or three mismatches (C, G, K), as well as by aligning reads to both allele-specific genomes allowing no mismatches (D, H, L). The number of neighboring differentiating sites is shown on the x-axis of each panel for each differentiating site and describes the maximum number of other sites that differ between the two alleles in any potential read overlapping the focal differentiating site. The y-axis shows the proportion of reads that were assigned to the reference allele for each differentiating site, summarized in box plots where the width of each box is proportional to the number of sites in that class. A proportion of 0.5 (indicated with a red dotted line in each panel) is expected if all reads overlapping a differentiating site are correctly assigned to alleles. The pie chart inset in each panel shows the total number of differentiating sites with equal (white) and unequal (grey) abundance of reads assigned to each allele.

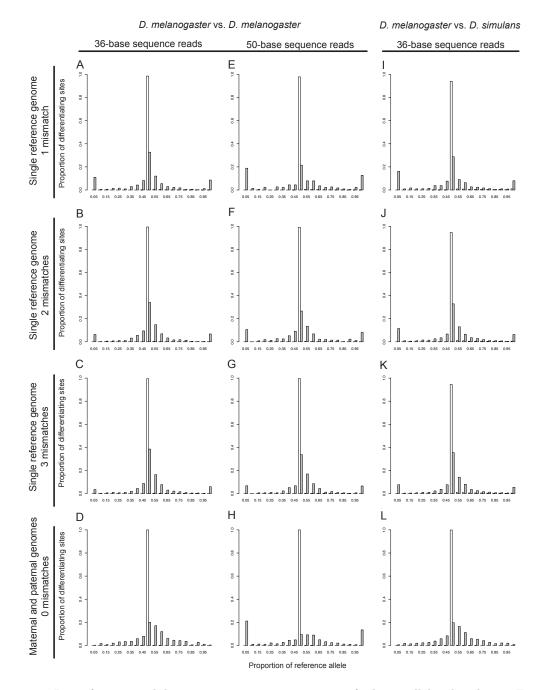


Figure 2.3: Imperfect mappability causes inaccurate measures of relative allelic abundance. For unbiased differentiating sites (i.e., those with fewer neighboring differentiating sites than the number of mismatches allowed) with either perfect (white) or imperfect (grey) mappability, the distribution of relative allelic abundance (measured as the proportion of mapped reads assigned to the reference allele) is shown for the 36-base (A-D) and 50-base (E-H) reads simulated from the two D. melanogaster genotypes as well as for the 36-base reads simulated from D. melanogaster and D. simulans (I-L) aligned to a single genome, allowing one (A, E, I),two (B, F, J),or three (C, G, K) mismatches. The distribution of relative allelic abundance for unbiased differentiating sites with perfect (white) and imperfect (grey) mappability is also shown for all three simulated datasets after aligning reads to both the reference and alternative genomes, allowing no mismatches (D, H, L).

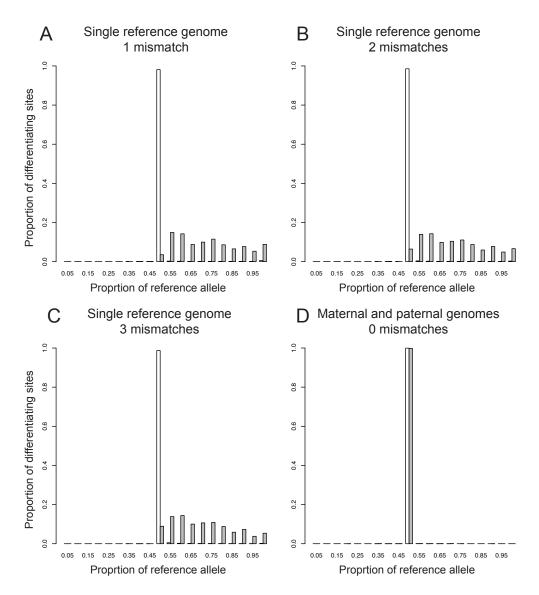


Figure 2.4: Insertions and deletions (indels) cause biased allele-specific assignment when reads are aligned to a single reference genome. For differentiating sites with perfect mappability and fewer neighboring differentiating sites than the number of mismatches allowed, the distributions of relative allelic abundance are shown for differentiating sites with (grey) and without (white) one or more nearby indel(s) after aligning the 36-base reads simulated from D. melanogaster and D. simulans to either the D. melanogaster genome with one (A), two (B), or three (C) mismatches allowed or to both the D. melanogaster and D. simulans genomes with no mismatches allowed (D).

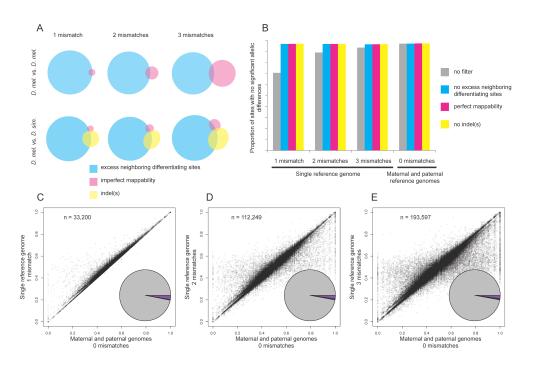


Figure 2.5: Real reads aligned to a single reference genome produce reliable measures of allelic abundance after excluding problematic differentiating sites. (A) The relative proportions of sites with an excess of neighboring differentiating sites (cyan), imperfect mappability (magenta), an indel(s) nearby (yellow), or more than one of these properties are shown for the simulated 36-base intra- (mel-mel) and interspecific (mel-sim) datasets allowing one (1 mm), two (2 mm), or 3 (3 mm) mismatches during alignment to a single reference genome. (B) The proportion of differentiating sites with no statistically significant difference in relative allelic expression is shown for the real reads from F1 hybrids between D. melanogaster and D. simulans after aligning to either a single reference genome with one, two, or three mismatches allowed or to both the maternal and paternal genomes with zero mismatches allowed before excluding any sites (grey) and after sequentially excluding differentiating sites with an excess of neighboring differentiating sties (cyan), imperfect mappability (magenta), or an indel(s) nearby (yellow). (C-E) For each differentiating site retained after filtering based on neighboring differentiating sites, mappability, and indels, the proportion of reads assigned to the reference allele is plotted after aligning reads to a single reference genome (y-axis) or to separate allele-specific genomes (x-axis), allowing one (C), two (D), or three (E) mismatches. The pie chart insets reflect the total number of differentiating sites that showed either no statistically significant difference in relative allelic abundance using either alignment strategy (grev), a statistically significant difference when reads were aligned to either a single reference genome (blue) or both the maternal and paternal genomes (red), or a significant difference with both alignment methods (purple). Binomial exact tests and a false discovery rate of 0.05 were used to assess statistical significance in all cases.

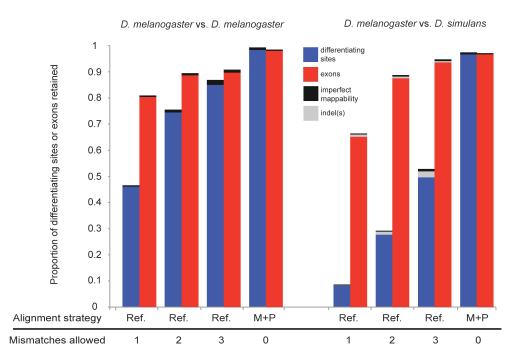


Figure 2.6: Relative allelic abundance can be estimated for most exons after excluding sites problematic sites. The proportion of differentiating sites (blue) and exons with at least one differentiating site (red) suitable for quantifying ASE after excluding sites with an excessof neighboring differentiating sites, imperfect mappability (black) and an indel(s) nearby (grey) are shown for the 36-base reads simulated from the two D. melanogaster genotypes (left) and from the D. melanogaster and D. simulans exomes (right). Each pair of bars results from aligning reads to either a single reference genome (Ref) or both the maternal and paternal genomes (M+ P) with zero (0), one (1), two (2), or three (3) mismatches allowed. The two D. melanogaster genotypes compared did not include any indels, as described in the main text.

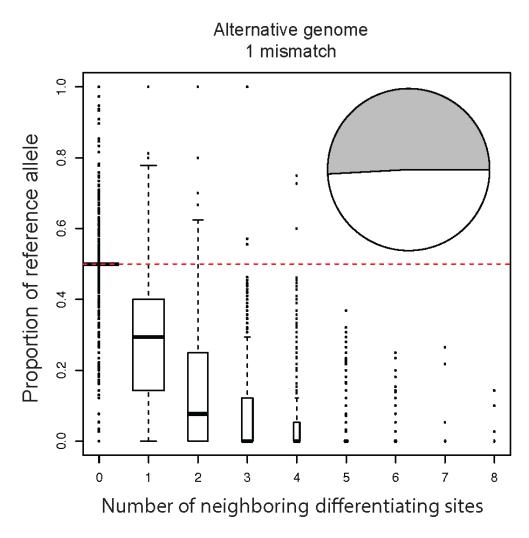


Figure 2.7: The density of differentiating sites affects measures of relative ASE when simulated reads are mapped to the alternative genome. Relative ASE was measured by aligning simulated reads to an alternative genome ("line_40") allowing one mismatch. The number of neighboring differentiating sites is shown on the x-axis, describing the maximum number of other sites that differ between the two alleles in any potential 36-base read overlapping the focal differentiating site. The y-axis shows the proportion of reads that were assigned to the reference allele for each differentiating site, summarized in box plots where the width of each box is proportional to the number of sites in that class. A proportion of 0.5 (indicated with a red dotted line in each panel) is expected if all reads overlapping a differentiating site are correctly assigned to alleles. The pie chart inset reflects the total number of differentiating sites that showed equal (white) and unequal (grey) abundance of reads assigned to each allele.

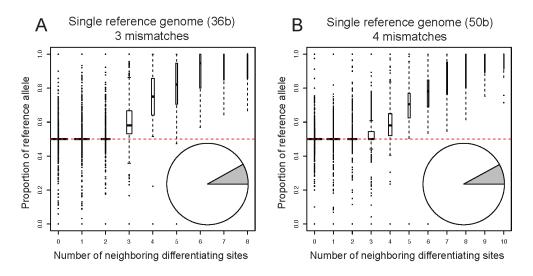


Figure 2.8: 36- and 50-base sequence reads produced comparable measures of relative ASE when a similar ratios of mismatches to bases in a sequence read is allowed. Relative ASE was measured for 36- and 50-base reads simulated from the two D. melanogaster genomes by aligning simulated reads to the single reference D. melanogaster genome. Three mismatches were allowed for 36-base reads (A), which is 0.083 mismatches per base, and four mismatches were allowed for 50-base reads (B), which is 0.080 mismatches per base. The number of neighboring differentiating sites is shown on the x-axis, describing the maximum number of other sites that differ between the two alleles in any potential 36-base (A) or 50-base (B) read overlapping the focal differentiating site. The y-axis shows the proportion of reads that were assigned to the reference allele for each differentiating site, summarized in box plots where the width of each box is proportional to the number of sites in that class. A proportion of 0.5 (indicated with a red dotted line in each panel) is expected if all reads overlapping a differentiating site are correctly assigned to alleles. The pie chart inset reflects the total number of differentiating sites that showed equal (white) and unequal (grey) abundance of reads assigned to each allele.

CHAPTER III

Genomic imprinting absent in Drosophila melanogaster adult females

3.1 Summary

Genomic imprinting occurs when expression of an allele differs based on the sex of the parent that transmitted the allele. In D. melanogaster, imprinting can occur, but its impact on allelic expression genome-wide is unclear. Here, we search for imprinted genes in D. melanogaster using RNA-seq to compare allele-specific expression between pools of 7- to 10-day-old adult female progeny from reciprocal crosses. We identified 119 genes with allelic expression consistent with imprinting, and these genes showed significant clustering within the genome. Surprisingly, additional analysis of several of these genes showed that either genomic heterogeneity or high levels of intrinsic noise caused imprinting-like allelic expression. Consequently, our data provide no convincing evidence of imprinting for D. melanogaster genes in their na-

Official citation:

Coolon, J.D., Stevenson, K.R., McManus, C.J., Graveley, B.R., Wittkopp, P.J. Genomic imprinting absent in *Drosophila melanogaster* adult females, *Cell Reports* 2012, 2, 69-75 doi:10.1016/j.celrep.2012.06.013 Copyright 2014 Elsevier Inc.

This is an open-access article distributed under the terms of the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported License (CC-BY-NC-ND; http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode).

In this collaborative work, I developed the custom bioinformatics pipeline to quantify total and allele-specific expression (see 3.5.3 and 3.8.3) from sequence reads of the MZ and ZM directional crosses that had been aligned to newly-built *D. melanogaster* genomes (zhr and z30; see 3.8.1). After putatively-imprinted genes (PIG) were identified, I tested the hypothesis that they are clustered in the genome by using Monte Carlo permutation tests. I then corrected these p-values by adjusting to a false-discovery rate (FDR) of 0.05 (see 3.5.4). Although I did not perform the statistical analysis of significant genomic imprinting, I participated in all discussions related to these tests as well as normalizing read counts both between alleles and between whole samples.

tive genomic context. Elucidating sources of false-positive signals for imprinting in allele-specific RNA-seq data, as done here, is critical given the growing popularity of this method for identifying imprinted genes.

3.2 Introduction

More than 50 years ago, Helen Crouse coined the term "imprinting" to describe a case in Sciard flies in which the sex of the parent influenced the inheritance of a chromosome (Crouse, 1960). Since that time, the definition of imprinting has been expanded to include any parent-of-origin-dependent chromosome marking, especially those causing differential gene activity or expression (Ferguson-Smith, 2011). Recently, genomic scans for imprinting at the level of RNA abundance in plants and mammals have shown that (1) only a small percentage of genes (typically on the order of 100 genes) appear to be imprinted (Babak et al., 2008; Gehring et al., 2011; Hsieh et al., 2011; Luo et al., 2011; Wang et al., 2008b, 2011; Waters et al., 2011; Wolff et al., 2011); (2) these genes are sometimes found in clusters within the genome (Ferguson-Smith, 2011; Wood and Oakey, 2006); and (3) their imprinting is often required for normal development (Mcgrath and Solter, 1984; Surani et al., 1984) and physiology (Buiting et al., 1995; Weksburg et al., 1993).

In *Drosophila melanogaster*, studies of imprinting have yielded conflicting results. Euchromatic genes inserted onto the heterochromatic Y chromosome and genes located on chromosomes with deletions, duplications, rearrangements, and/or translocations can show differences in their activity depending on the parent from which they are inherited, demonstrating that *D. melanogaster* is capable of imprinting (Anaka et al., 2009; Golic et al., 1998; Haller and Woodruff, 2000; Joanis and Lloyd, 2002; Lloyd et al., 1999; Macdonald et al., 2010; Maggert and Golic, 2002; Menon and Meller, 2009). However, when Wittkopp et al. (Wittkopp et al., 2006) tested for evidence of imprinting by analyzing allele-specific expression of eight genes that showed strong parent-of-origin effects on total gene expression in a genomic survey of D. melanogaster (Gibson et al., 2004), no evidence of imprinting was observed. Furthermore, gynogenetic and androgenetic D. melanogaster, which inherit all of their genetic information from a single parent, are viable, suggesting that imprinting is not essential in this species (Fuyama, 1984; Komma and Endow, 1995). Consequently, even though it is clear that D. melanogaster can form parent-of-origin-specific imprints that affect gene activity, the prevalence of imprinted genes in their native genomic context within the D. melanogaster genome remains unclear (Menon and Meller, 2009).

3.3 Results and Discussion

To search for imprinting genome-wide, we used Illumina sequencing in conjunction with a novel bioinformatics pipeline to infer allele-specific RNA transcript abundance in progeny from reciprocal crosses. This method uses transcribed sequence polymorphisms to distinguish sequencing reads derived from each of the two parental alleles in F1 offspring. To maximize the proportion of sequencing reads informative for allele-specific expression, we used a cosmopolitan (M-type) and an African (Z-type) line of *D. melanogaster* (Hollocher et al., 1997). The M-type line used was the zygotic hybrid rescue line (zhr) first described by Sawamura and colleagues (Sawamura et al., 1993) and the Z-type line was a Zimbabwean isofemale line (z30) isolated in 1990 (Begun and Aquadro, 1993; Wu et al., 1995). To improve the accuracy of allele assignments, we sequenced the M-type (zhr) and Z-type (z30) genomes to 23.2X and 21.5X coverage (Figure 3.9) and used these data to assemble line-specific genomic sequences (see Extended Experimental Procedures).

M-type females were crossed to Z-type males, producing F1 hybrids hereafter referred to as MZ. Likewise, Z-type females were crossed to M-type males, producing F1 hybrids hereafter referred to as ZM (Figure 3.5A). MZ and ZM hybrid flies were collected 7-10 days after eclosion, and total RNA was extracted from a pool of 20 hybrid females for each genotype. MZ and ZM RNA samples were used to make cDNA sequencing libraries, which were sequenced using an Illumina GAIIx machine. The resultant paired-end ($2\times76bp$) sequencing reads (Figure 3.9) were aligned to the strain-specific M-type and Z-type genomes. Using two strain-specific genome sequences for mapping avoids mapping biases introduced by using only a single reference genome (Degner et al., 2009; Graze et al., 2012). Of the reads from the MZ and ZM samples, 86% and 87%, respectively, were aligned without mismatches to unique genomic loci (Figure 3.9). In each case, 21% of the uniquely mapping reads aligned perfectly to only one genome and were used to infer allele-specific expression (Figure 3.5B; Figure 3.9).

The power to infer allele-specific expression using RNA-seq data (which is necessary to test for imprinting with this method) depends upon the expression level of a gene, as well as the density of transcribed polymorphisms within it (Fontanillas et al., 2010a). Prior work has shown that obtaining at least 20 allele-specific reads for a gene results in reproducible measures of relative allelic expression (McManus et al., 2010). Retaining only genes with 20 or more allele-specific reads (allele 1 + allele 2 \geq 20) in both the MZ and ZM samples, 7,206 genes were tested for allelic expression patterns consistent with imprinting (Table S2). This includes 3% of the 4,875 genes with a number of fragments per kilobase per million mapped reads (FPKM) less than 1, 51% of the 1,706 genes with an FPKM between 1 and 5, and 83% of the 7,430 genes with an FPKM greater than 5 (Figure 3.6). The modENCODE consortium used a threshold of FPKM = 1 to classify *D. melanogaster* genes as "expressed" or "not expressed" (Graveley et al., 2011) and according to this definition, we tested 77% of the 9,136 genes expressed (in the 7- to 10-day-old adult females we examined) for imprinting.

To assess the accuracy of our allele-specific expression measurements, we compared the allelic expression ratios determined by RNA-seq to estimates from pyrosequencing (Ahmadian et al., 2000) of individual genes. Ten genes selected at random were used for pyrosequencing of the same MZ and ZM samples used for RNA-seq (Table S3). Pyrosequencing measurements were highly correlated (R2 = 0.88) with estimates from RNA-seq (Table S3; Figure 3.7A), suggesting that RNA-seq produces reliable measures of relative allelic expression. This is consistent with previous comparisons of RNA-seq and pyrosequencing measures of allelic expression that used distinct bioinformatic pipelines (McManus et al., 2010; Emerson et al., 2010).

To identify genes that might be imprinted, we tested for differences in relative allele-specific expression between MZ and ZM using the Fisher's exact test (FET). This test evaluates whether differential allelic expression (when present) is equal in magnitude and direction in the two genotypes. At a false discovery rate (FDR) of 5%, 119 (1.65%) of the 7,206 genes analyzed had significant differences (FET, q <0.05) in relative allelic expression between the two types of F1 hybrid progeny (Figure 3.1; Table S2). To evaluate the accuracy of RNA-seq measurements of allele-specific expression specifically for putatively imprinted genes (PIGs), we used pyrosequencing to independently measure allele-specific expression for four genes in this class using the same ZM and MZ samples as those used for RNA-seq. We again observed strong concordance (R2 = 0.85, Figure 3.7B) between pyrosequencing and RNA- seq measures of allele-specific expression, suggesting that inaccurate quantification of expression levels in cDNA pools by RNA-seq is unlikely to explain the observed differences in relative allelic expression between hybrid genotypes.

3.3.1 Putatively imprinted genes are clustered in the genome

In mammals, imprinted genes are often found in clusters throughout the genome (Ferguson-Smith, 2011; Wood and Oakey, 2006), and this clustering might relate to the mechanism by which they are regulated (Caspary et al., 1998; Mancini-Dinardo et al., 2006; Lewis et al., 2004; Lopes et al., 2003). To determine if this was also true for the PIGs in the *D. melanogaster* genome, we used a sliding-window Monte Carlo sampling approach with FDR-corrected approximate permutation tests to investigate potential clustering. We found that there were four regions in the *D. melanogaster* genome that showed significant clustering (permutation test, q 0.05) of PIGs (Figure 3.2). Interestingly, all four significant clusters were found on chromosome 3, with two on the left arm (3L) and two on the right arm (3R) of the chromosome. Together, these four regions contain 27% (32/119) of the PIGs, with one cluster located on chromosome arm 3R (6,550,000-8,280,000) containing 17% (20/119) of all PIGs (Figure 3.2). Clustering of PIGs in the genome is consistent with previously described mechanisms of imprinting, but it could also be caused by other factors.

3.3.2 Low-frequency deletion(s) responsible for some cases of apparent imprinting

To further test for evidence of imprinting, we more closely examined 12 genes within the largest and most significant cluster of PIGs (3R 6.5-8.3 MB region, Figure 3.2). Seven of these genes were PIGs and five were genes that showed no significant differences in relative allelic expression between ZM and MZ. Pyrosequencing was again used to obtain an independent measure of relative allelic expression, except that instead of testing the same biological sample used for RNA-seq (as described above), we analyzed four independent biological replicate pools of ZM and MZ flies, each containing twenty 7- to 10-day-old adult females (Table S3). From each pool, we sequentially extracted genomic DNA (gDNA) and RNA.

F1 flies produced by crossing two highly inbred lines are expected to be genetically identical; thus, analysis of gDNA serves as a control for differential amplification of the two alleles during PCR prior to pyrosequencing (Landry et al., 2005; Wittkopp, 2011; Wittkopp et al., 2004, 2008a). Surprisingly, and unlike the case for the 34 genes located outside of clustered PIGs that we analyzed (data not shown), relative allelic abundance differed greatly for the gDNA samples among the biological replicates between the MZ and ZM genotypes as well as among replicate MZ or ZM samples (Figure 3.3). When present, deviations from equal allelic abundance in gDNA were similar for genes throughout the cluster within a replicate pool but differed among pools. The M-type (zhr) allele was always the allele underrepresented (Figure 3.3).

A polymorphic deletion(s) in the M-type (zhr) strain or a polymorphic duplication(s) in the Z-type (z30) strain could account for the differences in gDNA content observed among replicate pools of F1 flies. To directly test for evidence of a deletion or duplication, we used pyrosequencing to genotype 48 individual F1 progeny (24 MZ and 24 ZM) at four loci within the 3R 6.5-8 MB region that showed a cluster of PIGs (indicated with asterisks in Figure 3.3), as well as at two loci on other chromosomes. All but two of the 48 hybrid flies showed evidence of one M-type and one Z-type allele at all six loci tested, as expected. The remaining two hybrids showed evidence of only the Z-type (z30) allele at the four loci within the cluster, but both flies showed both alleles at the two loci tested on other chromosomes (Figure 3.10); the presence of these heterozygous sites demonstrates that these two flies are in fact F1 hybrids and not contaminating flies with parental genotypes. Based on these data, we conclude that the M-type (zhr) strain contains one or more deletion(s) in this region on 3R that remains heterozygous despite years of inbreeding followed by 10 generations of pair mating immediately prior to the start of this experiment. Residual heterozygosity such as this has also been reported in *D. melanogaster* following extensive inbreeding in lines used for genomic sequencing (Mackay et al., 2012).

The presence of this deletion haplotype at low frequency in the zhr line used to produce MZ and ZM hybrids suggests that differences in its frequency in the pools of 20 MZ and 20 ZM hybrid flies used for RNA-seq are more likely than imprinting to be responsible for the observed difference in relative allelic expression. Indeed, after controlling for differences in the alleles present in gDNA among the replicate pools analyzed by pyrosequencing (see Experimental Procedures), relative allelic expression in cDNA samples was not significantly different (p >0.05 for all tests). It remains to be seen whether genotypic differences between the MZ and ZM pools of flies used for RNA-seq are also responsible for differences in relative allelic expression observed for other clustered PIGs, but we believe it is likely.

3.3.3 Non-clustered PIGs have higher-than-normal intrinsic noise

Our initial RNA-seq survey for imprinting identified as PIGs all genes with significant differences in relative allelic expression between F1 hybrid progeny from reciprocal crosses; however, imprinting is often defined in a more limited way, such that only one allele of a gene (either the maternally or paternally inherited allele) accounts for the majority (or all) of the expression of the imprinted gene. Among the original set of 119 PIGs, only 18 showed patterns of allelic expression consistent with this more strict definition (Table S2; Figure 3.8), and none of these were located in the clusters described above (Figure 3.2). To further test these 18 genes for evidence of imprinting, we analyzed allelic expression for each gene in the MZ and ZM biological replicates described in the preceding section (Table S3). Unlike for clustered PIGs examined in these samples, no significant differences in allele frequency were found among replicate gDNA samples for any of these 18 genes.

The relative allelic expression for these genes in the four MZ and four ZM biological replicates was still not typical; however, these 18 genes showed greater variance in relative allelic expression among the biological replicate pools than most genes that we have analyzed with pyrosequencing. Indeed, a Wilcoxon rank-sum test showed that the standard errors of \log_2 -transformed allelic expression ratios were significantly greater for the 18 PIGs than for 16 genes selected at random (W = 260, p = 2.68e-7; Figure 3.4). Additional statistical tests showed no evidence for imprinting of these genes (q>0.05 for all tests). Given (1) the high degree of variability we observed for these genes among replicate pools with the same genotype (MZ or ZM), (2) the lack of evidence for imprinting found by pyrosequencing, and (3) that we only analyzed one pool of flies for each genotype by RNA-seq, we conclude that significant differences observed between MZ and ZM for relative allelic expression in the RNA-seq data are most likely caused by sampling error.

3.3.4 What role does imprinting play in regulating D. melanogaster gene expression?

As described above, RNA-seq analysis (validated by pyrosequencing) identified 119 of 7,206 genes as having differences in relative allele-specific expression in reciprocal hybrids; however, analysis of gDNA and cDNA from additional replicate biological samples identified other factors (the presence of a polymorphic deletion(s) and using a single measurement to represent a highly variable phenotype) that are more likely than imprinting to be responsible for the differences in allelic expression observed in our RNA-seq data. Consequently, we conclude that these data provide no convincing evidence that imprinting affects expression of endogenous D. *melanogaster* genes in their native genomic contexts at least in the 7- to 10-day-old adult females we examined.

Given the evidence of imprinting in other studies of *D. melanogaster*, why do we fail to find evidence of it in our genomic analysis? We cannot rule out the possibility that imprinting affects allelic activity in males, at other developmental stages, in limited tissues (with the signal masked by the absence of imprinting in the majority of cells sampled), or for genes with expression and/or polymorphism levels that cause them to be below our detection threshold, but there is also no evidence suggesting that imprinting is occurring under any of these conditions. In addition, as described by Menon and Meller (Menon and Meller, 2009), evidence of imprinting in *D. melanogaster* comes from studying particular genotypes, and imprinting might not impact gene expression in all genotypes: "In Drosophila, imprints are detected by alteration in expression of genes on rearranged chromosomes, but there is little to suggest that expression of any gene in karyotypically normally (sic) flies is governed by imprinting". We tested 77% of the expressed genes in the *D. melanogaster* genome for imprinting in this study, and evidence that imprinting affects the expression of genes in their native genomic context is still lacking.

3.4 Genomic surveys for imprinting using RNA-seq: proceed with caution

In addition to providing insight into imprinting in *D. melanogaster*, this study identifies important considerations for using RNA-seq to test for imprinting in any species. RNA-seq has been used to search for imprinted loci in both plants and animals, including mouse (Babak et al., 2008; Wang et al., 2008b, 2011), *Arabidopsis* (Gehring et al., 2011; Hsieh et al., 2011; Wolff et al., 2011), maize (Waters et al., 2011), and rice (Luo et al., 2011); but this approach is not without its pitfalls. For example, a study using RNA-seq to identify imprinted genes in various mouse tissues reported over 1,000 imprinted loci (Gregg et al., 2010b,a), but most of these loci were subsequently shown to be false positives caused by biased sequencing and the failure to measure and account for technical and biological variability (DeVeale et al., 2012).

Data presented here and in DeVeale et al. (DeVeale et al., 2012) clearly show the importance of validating putatively imprinted genes identified by RNA-seq with independent techniques (and, ideally, independent biological samples) prior to concluding that they are imprinted. To focus validation efforts on the loci most likely to be imprinted, RNA-seq experiments should include both biological and technical replicates, as well as, whenever possible, the analysis of gDNA extracted from the same tissue homogenate as the RNA. This final control is particularly important when working with small organisms (e.g., flies), for which multiple inbred individuals (that could have residual heterozygosity) are typically pooled prior to RNA extraction and cDNA sequencing, but it can also detect and control for differences in genomic content that might exist among cells from the same individual due to somatic mutations. For example, Shibata et al. (Shibata et al., 2012) have recently shown that microdeletions can cause genomic heterogeneity among mouse and human cells. Sequencing gDNA and cDNA derived from the same tissue sample can also allow corrections for bias introduced during the library preparation and sequencing. With more and more researchers turning to RNA-seq to study genomic imprinting, it is important to keep these caveats in mind.

3.5 Experimental Procedures

3.5.1 Fly strains, rearing, and collections

The *D. melanogaster* strain zhr carrying the hybrid rescuing Zhr1 chromosome (full genotype, XYS.YL.Df(1)Zhr; (Ferree and Barbash, 2009; Sawamura et al., 1993)) and the Zimbabwean isofemale line z30 (Begun and Aquadro, 1993; Wu et al., 1995) were used for this study. All flies were reared on cornneal medium on 16:8 light:dark cycle at 20°C. Prior to crossing, both strains were subjected to 10 generations of sibling pair matings to reduce genome-wide heterozygosity, and this was followed by three generations of population expansion to generate the quantity of flies needed for crosses. For each reciprocal cross performed, 10 vials were set up with 3 female and 3 male flies. Virgin female progeny were allowed to mate from the time of eclosion to 3 days posteclosion, then males and females were separated and females aged to 7-10 days post-eclosion. All flies were collected during the same time of day to minimize the effects of circadian rhythm, and flies were snap-frozen in liquid N₂.

3.5.2 Library preparation and Illumina sequencing

Pools of 20 female flies were used for total RNA extraction with TRIzol reagent according to manufacturer instructions (Invitrogen). Illumina sequencing libraries were prepared (see Extended Experimental Procedures) as previously reported (Mc-Manus et al., 2010). Two lanes of paired-end (2×76 bp) Illumina GAIIx sequencing were performed.

3.5.3 Quantifying total and allele-specific expression from sequencing reads

We developed a bioinformatics pipeline to quantify gene expression from the Illumina sequencing output (Figure 3.5B; Extended Experimental Procedures). Briefly, we aligned each mate of the paired-end RNA-seq reads separately to the newly built *D. melanogaster* genomes (zhr and z30; Extended Experimental Procedures), keeping only those reads that aligned to one genomic location. Reads that did not map were trimmed by 13 bases and realigned in three iterations. Reads that did not align were then discarded. We then converted zhr and z30 genomic coordinates of aligned reads to sequenced *D. melanogaster* genomic coordinates using the liftOver utility from the UCSC Genome Browser (Kent et al., 2002). Aligned sequence reads were then filtered based on their alignment to a previously identified set of overlap filtered constitutively expressed exons within the *D. melanogaster* genome (McManus et al., 2010) using the intersectBed module of BedTools (Quinlan and Hall, 2010) (Version 2.12.0).

Remaining sequencing reads that aligned to only one of the two line-specific genomes were used for quantification of allele-specific gene expression. Down-sampling followed by rounding to the nearest integer was used to account for differences in overall sequencing output between MZ and ZM and differences in mappability between zhr and z30 alleles. For each gene, allele-specific expression levels are reported (Table S2). To reduce the effect of sampling error (Fontanillas et al., 2010a; Mc-Manus et al., 2010), we analyzed only genes that had more than 20 allele-specific reads (allele 1 + allele $2 \ge 20$) in both ZM and MZ. To test for unequal allelic expression between ZM and MZ, we performed Fisher's exact tests using zhr and z30 allelic counts. Due to the multitude of tests performed, a false discovery rate (FDR) significance threshold of 5% was used to determine significance (Benjamini and Hochberg, 1995). Statistical analyses were performed in R (version 2.12.2, CRAN).

FPKM values reflecting total expression levels for individual genes were calculated by dividing the total number of paired-end reads mapped to a gene (including reads that were and were not informative for allele-specific expression) by the length of the sequence representing that gene in kilobases and then dividing this value by the number of millions of mapped reads from that sample.

3.5.4 Sliding-window analyses with Monte Carlo sampling and approximate permutation tests

Genomic clustering of putatively imprinted genes was analyzed using a slidingwindow approach where we divided the genome into 11,726 overlapping 500 kb windows and moved stepwise, offsetting by 10 kb with each step. For each window, we counted the number of total genes and PIGs within each region. To test whether the observed clustering was significant, we used a Monte Carlo sampling approach to approximate the null distribution of imprinted genes randomly scattered along the genome. A Monte Carlo sampling approach was used to approximate the null distribution, because the number of permutations required for an exact test in this case was exceedingly large (7.8e261). From the total set of 7,206 genes, we randomly sampled 119 genes without replacement, assigned them imprinting status, and aggregated new imprinting counts for each window. This was done 10,000 times, resulting in an approximate null distribution of the number of imprinted genes expected by chance in each window. To calculate an approximate p-value for each window, we summed the number of occurrences where the permuted value exceeded the observed value. Due to the multitude of tests performed, an FDR-corrected significance threshold of 5% was used to determine significance (q < 0.05). Significant windows were collapsed to four regions based on overlap (Figure 3.2).

3.5.5 Pyrosequencing

To evaluate the accuracy of allelic expression measurements derived from our RNA-seq data and analysis, new cDNA pools were synthesized from the same RNA samples used for Illumina sequencing and used for pyrosequencing. cDNA was synthesized from total RNA using T(18)VN primers and Superscript II (Invitrogen) according to manufacturer recommendations. Both cDNA and gDNA were analyzed using pyrosequencing. For each gene assayed, PCR was performed in triplicate on both the cDNA and gDNA samples (separately) and followed by pyrosequencing (QIAGEN). The genomic DNA was extracted from an independent pool of F1 flies and was used to normalize cDNA measurements (Wittkopp, 2011). log₂-transformed cDNA allelic expression ratios from Illumina and pyrosequencing were compared after normalization using type 2 regressions in R.

To investigate allelic expression within a cluster of genes on chromosome 3R, we constructed four new replicate pools of 20 individuals each for both ZM and MZ samples and coextracted RNA and gDNA from a single tissue homogenate of each pool of flies using the Promega SV total RNA extraction system with modified protocol (Wittkopp, 2011). cDNA was made from total RNA as above, and both gDNA and cDNA were used for PCR followed by pyrosequencing. To account for differences in gDNA allelic abundance among replicate pools of flies, the log₂ allelic expression ratio for gDNA from a particular pool was subtracted from the log₂ allelic expression ratios for cDNA samples derived from the same pool of flies (Wittkopp et al., 2004, 2006, 2008a; Wittkopp, 2011).

The four biological replicates were used to investigate variation in allelic expression for a set of randomly chosen genes, and this was compared to a set of putatively imprinted genes. The standard error for the log₂ allelic expression ratio was calculated for each assay-sample combination for the randomly chosen genes and nonclustered PIGs, and these two sets were compared using a Wilcoxon rank-sum test in R.

3.6 Acknowledgements

We would like to thank all members of the Wittkopp lab, especially Brian Metzger and Bing Yang, for helpful comments on experimental design, data analysis, and manuscript presentation, Sebastian Zöllner for computational resources and Chung-I Wu for the z30 fly line. This work was supported by a National Science Foundation grant to P.J.W (MCB-1021398) and a National Institutes of Health grant to B.R.G. (5R01GM095296), as well as a National Research Service Award from the National Institutes of Health to J.C.D. (1F32GM089009-02). P.J.W. is a Research Fellow of the Alfred P. Sloan Foundation.

3.7 Accession Numbers

The sequencing data from this study have been submitted to the National Center for Biotechnology Information Sequence Read Archive under accession number SRA052065.

3.8 Extended Experimental Procedures

3.8.1 Resequencing of zhr and z30 genomes and genome assembly

Genomic DNA sequence reads were aligned to the *D. melanogaster* genome assembly (dm3; (Adams et al., 2000; Celniker et al., 2002)) using BWA ((Li and Durbin, 2009); version 0.5.6). Each read was aligned separately using default parameters, and SAM format files were generated using the bwa sampe command. For zhr, an additional SAM file was prepared from single-read Illumina data. Alignment files were converted to bam format and vcf files describing snps and indels were created using the samtools package ((Li et al., 2009); version 0.1.7a; modules view, sort, and pileup). SNP and indel calls were filtered using the samtools.pl varFilter command

(as described at http://samtools.sourceforge.net/cns0.shtml) to retain SNPs and indels with PHRED scale quality scores of 20 or higher. At some positions, SAMtools identified heterozygous sites. This creates a complication for comparative RNA-seq, as the heterozygous genotype of one strain can partially overlap with the other strain. For example, if resequencing identified an "R" (either A or G) base at a coordinate in zhr and a "G" in z30, RNA-seq reads originating from the z30 could be mapped to both strains, while "A" containing reads from zhr would be strain specific. In order to avoid using regions of partial overlap in allele-specific RNA-seq assignments, we changed both SNP calls to the most ambiguous genotype possible using a custom perl script (snp_compare_filter.pl), effectively making these sites uninformative for allele assignment.

Strain-specific genome sequences were produced using a custom Perl script (snp_--adder.pl). This script sequentially rewrites the *D. melanogaster* genome with corrected SNP calls and indels. The positions of insertions and deletions were recorded in custom liftover chain files during the rewriting process. These chain files allow the conversion of genomic features between strain and reference genomes using the UCSC genome browser liftover tool (http://genome.ucsc.edu; (Kent et al., 2002). Heterozygous indel sites (insertion in one allele in one strain) were tracked in separate genome files (mixed indel 1 and mixed indel 2). The genomes and chain files are available upon request.

3.8.2 Library preparation

cDNA libraries were prepared as in McManus *et al.* (McManus et al., 2010). Briefly, 10 μ g of total RNA from each sample was treated with DNase 1 (Invitrogen) followed by poly(A)+ selection using Dynal magnetic beads (Invitrogen) following manufacturer recommendations. Poly(A)+ RNA was then fragmented using RNA fragmentation reagent (Ambion) before cDNA synthesis. Double-stranded cDNA was primed using random hexamers and Superscript II reverse transcriptase (Invitrogen). cDNA was run on a 2% agarose gel and the region corresponding to ~ 300bp fragments was gel extracted. This size-selected double-stranded cDNA was used in the Paired-End Genomic DNA Library Preparation Kit (Illumina) according to manufacturer's recommendations. Genomic DNA libraries were prepared from pools of 20 female flies of each strain (zhr and z30) and genomic DNA was extracted using the DNeasy Blood & Tissue Kit (QIAGEN). 10 μ g of gDNA was used to make gDNA sequencing libraries using the Paired-End Genomic DNA Library Preparation Kit (Illumina) according to manufacturer's recommendations. The cDNA libraries (ZM and MZ) as well as the gDNA libraries (zhr and z30) were subjected to paired-end sequencing on an Illumina Genome Analyzer IIx on one lane each for 76 cycles per read. Images were analyzed using the Firecrest and Bustard modules to generate sequence and quality scores for each read.

3.8.3 Quantifying allele-specific expression from sequencing reads

To quantify gene expression from the Illumina sequencing output we aligned each mate of the paired-end RNA-seq reads separately to the newly built D. melanogaster genomes (zhr and z30) using the MOSAIK aligner (version 1.0.1384, http:// bioinformatics.bc.edu/marthlab/Mosaik). We used the following command line options for the alignment: -hs 13 -mm 0 -p24 -mph 100 -act 20. The 13 base hash size (-hs 13) option allowed >99% of ambiguous base containing regions to be seeded for alignment by MOSAIK. Only uniquely aligning reads with no mismatches were retained for analysis. After the initial 76 bp reads were aligned to both reference genomes, those reads that did not map to either were trimmed 13 bases from the 3' end using

a custom Perl script (fastq_trimmer.pl) and again aligned with MOSAIK. This was repeated three times (sequence lengths 76bp, 63bp, 50bp, 37bp). Any sequences that did not uniquely align after the final iteration were discarded.

Using the chain files created in the genome assembly process, we converted the respective genome coordinates (in zhr or z30 space) to the sequenced dm3 coordinates using the liftOver utility from the UCSC Genome Browser (Kent et al., 2002) (http://genome.ucsc.edu) and a custom Perl script (convert.pl). Sequence reads were then filtered based on their alignment to a previously identified set of constitutively expressed exons within the *D. melanogaster* genome (McManus et al., 2010) using intersectBed module of BedTools, with those reads not aligning to these regions discarded. Additionally, regions in the constitutive exon set found to overlap were removed using intersectBed module of BedTools and custom scripts. Constitutively expressed exon filtering was performed to reduce biases associated with isoform specific differences. The filtered set of sequencing reads was used for quantification of allele-specific gene expression. Reads were assigned to the zhr or z30 allele based on reported alignments using a custom Perl script (classify.pl). Because each paired-end read represents a single transcript, we only incremented gene counts once for each paired-end read (or once if only one end of the read mapped). For many genes, the number of reads aligning and contributing to quantification of gene expression exceeded the number of mappable positions, which means that identical sequencing reads were identified and included in our final quantification to avoid imparting maximum expression level thresholds to genes based on their length.

To correct for mappability differences between the two genomes that could lead to biases, we determined the total number of informative reads that aligned allelespecifically to the zhr and the z30 genomes for all genes in each F1 hybrid sample (ZM and MZ) with the expectation of equal representation. Because the zhr alleles were slightly more abundant across the whole genome in both MZ and ZM, we downsampled the zhr allelic counts globally by multiplying by 0.9706 in ZM and by 0.9736 in MZ followed by rounding to the nearest integer. To make comparisons between the reciprocal crosses we corrected for differences in sequencing depth between the two sequencing efforts. The ZM library had 31,432,754 reads and the MZ library had 31,439,998 reads. To correct for this minor difference, we multiplied the MZ counts by 0.9997 followed by rounding to generate integer read counts. For each gene, allele-specific expression levels are reported as the number of sequences that map to either the zhr or the z30 allele (Table S2) with no correction for gene length because all comparisons were made between alleles of equal size in the two strains. Because genes with low counts are more likely to be influenced by sampling error, we removed all genes from analyses that had less that 20 allele-specific reads used for expression quantification, retaining those that satisfy (allele 1 + allele $2 \ge 20$) for statistical analysis.

We performed Fisher's exact tests (FETs) using allelic expression counts (zhr and z30) from MZ and ZM to test for unequal allelic expression between progeny from reciprocal crosses. To correct for the multiple comparisons made (FETs), we used a false discovery rate of 5% (Benjamini and Hochberg, 1995). Statistical analyses were performed in R (version 2.12.2, CRAN).

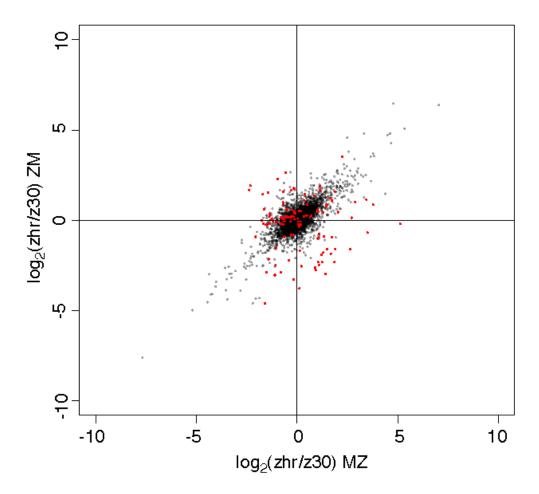


Figure 3.1: Allelic Expression from Reciprocal Crosses Suggests that <2% of Genes in the Genome Might Be Imprinted. log₂-transformed allelic expression ratios (zhr/z30) from MZ on the x axis and log₂(zhr/z30) allelic expression ratio from ZM on the y axis. Each point represents one gene. Points are color-coded by significance in false-discovery- rate-corrected Fishers exact tests, where red points indicate q <0.05. Note that the power to detect differences in allelic expression between ZM and MZ differs from gene to gene and is dependent upon the number of Illumina sequencing reads obtained that map to that gene. See also Figures 3.5-3.7 and Tables S1 and S2.

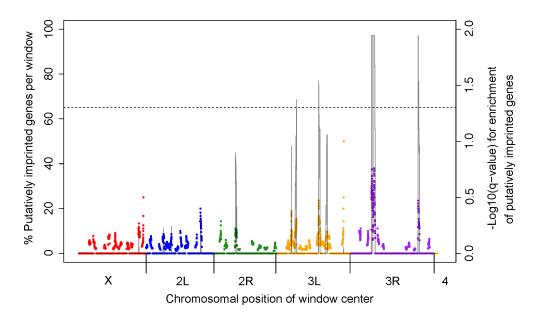


Figure 3.2: Putatively Imprinted Genes Clustered Significantly on Chromosomes. Using a slidingwindow analysis, the proportion of genes within each 500 kb window that were identified as putatively imprinted is indicated for positions across the genome. Each chromosome arm is indicated on the x-axis, with one point representing each window. Using a Monte Carlo sampling approach and approximate permutation tests that control for differences in the number of genes within each window, and following these steps with a multiple testing correction, weidentified regions of the genome that were significantly enriched for PIGs. FDR-corrected p-values are indicated by the solid line, and the dotted line indicates the threshold used to identify significant clusters (q < 0.05).

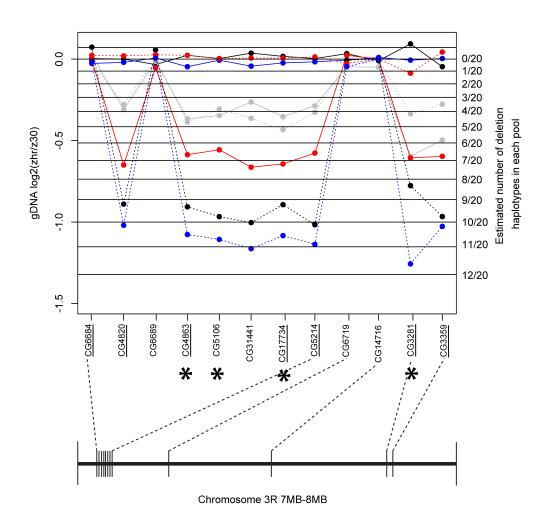


Figure 3.3: Replicate Pools of Flies Showed Different Allele Frequencies in Genomic DNA for Putatively Imprinted Genes Located in a Cluster. For 12 genes in the region containing the largest cluster of putatively imprinted loci (7,000,0008,000,000 on chromosome 3R), seven that were identified as putatively imprinted (underlined) and five that were not, we used pyrosequencing to determine the relative abundance of zhr and z30 alleles in genomic DNA in additional biological replicate pools containing 20 F1 heterozygous flies each. The $\log_2(zhr/z30)$ ratio is plotted for gDNA from each biological replicate pool, with the four ZM pools indicated by solid lines and the four MZ pools indicated by dotted lines. Replicates are arbitrarily colored blue, gray, red and black. The genomic arrangement of these genes is shown below the plot. Genes labeled with an asterisk were also genotyped in individual flies (Table S4). Note that CG6684 is underlined because it showed significant evidence of allelic expression differences between MZ and ZM in the RNA-seq data; however, this gene does not appear to be included within the deleted region(s). Pyrosequencing analysis of CG6684 showed no evidence of differential allelic expression between MZ and ZM and normal variance among replicate biological samples, suggesting that it was a false positive in the RNA-seq data. CG5106 and CG31441 appear to be included within the deleted region but showed no significant evidence of an imprinting-like pattern of allelic expression in the RNA-seq data, probably due to lack of power, as these two genes had the lowest read counts of those tested. See also Tables S3 and S4.

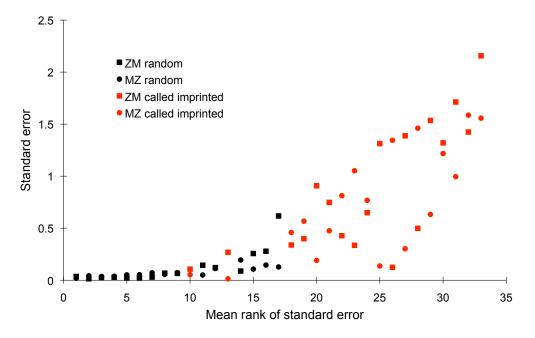


Figure 3.4: Putatively Imprinted Genes Have High Intrinsic Noise. For each gene for each sample type (ZM or MZ), the standard error for log₂- transformed allelic expression ratios from biological replicate pools of flies is shown, with black points representing genes selected at random from the genome (none of which showed significant evidence of imprinting) and red points representing PIGs. Square marks represent the ZM sample and circles represent the MZ sample, with one rank for each gene tested. The x-axis is rank of standard error for the two samples for each gene. See also Figure 3.8 and Table S3.

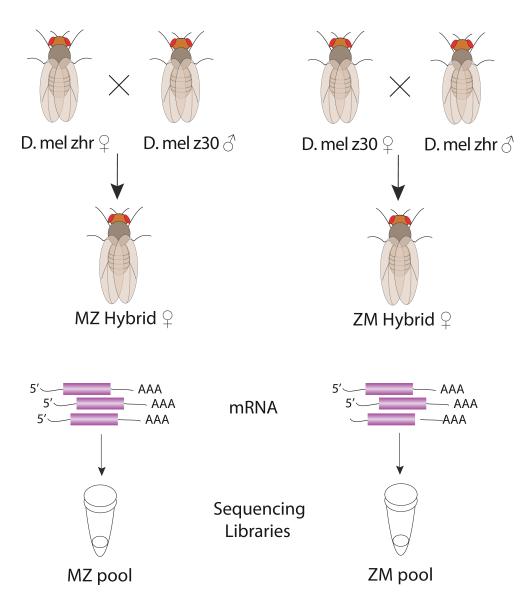


Figure 3.5: Experimental Method for Investigating Imprinting, Related to Figures 3.13.4. (A) Reciprocal crosses between M-type (zhr) and Z-type (z30) strains of D. melanogaster were performed to generate MZ (zhr females X z30 males) and ZM (z30 females X zhr males) F1 progeny. Pools of 20 female progeny were used for isolation of RNA and DNA (see Experimental Procedures). (B) A flowchart of the steps used to transform Illumina sequencing reads into allele-specific gene-expression counts for MZ and ZM cDNA libraries (see Extended Experimental Procedures) is shown.

65

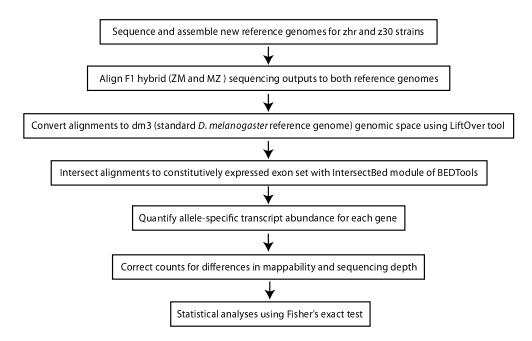


Figure 3.6: Coverage of Genes Tested for Imprinting, Related to Figure 3.1. (A) Genes were binned based on their total expression level in fragments per kilobase per million mapped reads. Total expression is plotted on the x axis and the proportion of genes in each bin is indicated on the y axis. (B) Total expression (FPKM) is plotted on the x axis, and the proportion of genes in each total expression bin tested for imprinting is shown on the y axis

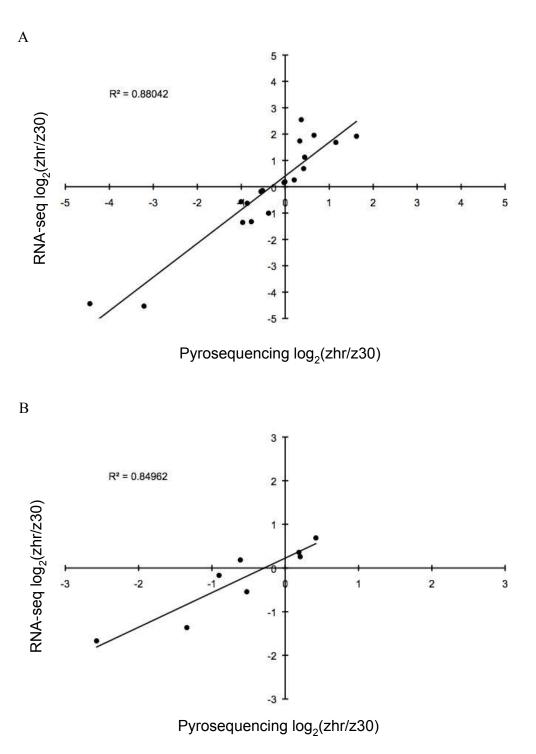


Figure 3.7: Pyrosequencing Validation of RNA-seq Data, Related to Figure 3.1. (A and B) \log_2 transformed allelic expression ratios (zhr/z30) from pyrosequencing on the x axis and $\log_2(zhr/z30)$ allelic expression ratio from RNA-seq on the y axis. Two points represent each gene, one for allelic expression measures from ZM and one for those from MZ. Data from randomly selected genes (A) and from PIGs (B) are shown. Type 2 regressions were performed, and correlation coefficients (R²) for (A) and (B) are shown.

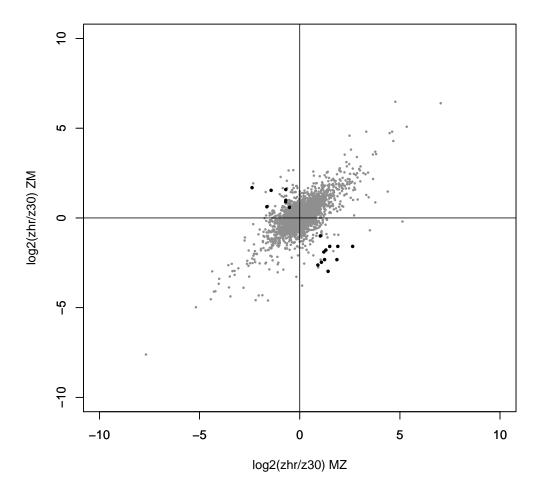


Figure 3.8: RNA-seq Data Plotted with Filtered PIGs Highlighted, Related to Figure 3.4. \log_2 transformed allelic expression ratios (zhr/z30) from MZ on the x-axis and $\log_2(zhr/z30)$ allelic expression ratio from ZM on the y-axis. Each point represents one gene. Points are color-coded with black points those that were identified with significant allelic expression differences between MZ and ZM and met a more strict definition of imprinting-like expression pattern where either the maternally- or paternally-inherited allele accounts for the majority of the expression of the imprinted gene and grey points representing all other genes quantified with RNA-seq.

Sample (gDNA)	Total number of reads (SE)	Total number of reads (PE)	X Coverage		
zhr	15,692,412	27,051,150	23.23		
z30	NA	25,863,911	21.46		
Sample	Total number	Number of uniquely	%	Number of allele-	%
(cDNA)	of reads	mapping reads		specific reads	
MZ	31,432,754	26,974,244	86	6,598,776	21
ZM	31,439,998	27,432,712	87	6,673,601	21
total	62,872,752	54,406,956	87	13,272,377	21

Figure 3.9: Descriptive statistics of Illumina sequencing, Related to Figure 3.1. Results for Illumina sequencing of gDNA and cDNA libraries are shown. The total sequencing output, number of sequencing reads that aligned and those that aligned uniquely from RNA-seq of MZ and ZM samples are listed. Details of gDNA sequencing of zhr and z30 lines including single-end and paired end sequencing of zhr and paired-end sequencing output from z30. Average coverage (X) for zhr and z30 sequencing was also determined.

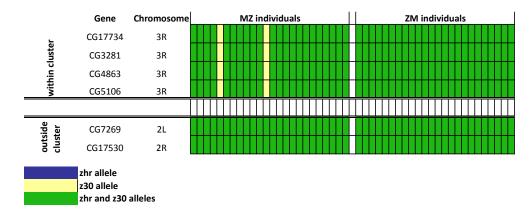


Figure 3.10: Genotyping individual MZ and ZM progeny, Related to Figure 3.3. Genotyping of individual F1 flies was performed for six loci; four within a significant clustering on chromosome 3R (between positions 7,000,000-8,000,000, see Figures 2, 3) and two were on chromosome 2 (one on 2R and one on 2L). Forty-eight individual flies were genotyped, 24 MZ and 24 ZM. Detected alleles are indicated by color (zhr in blue, z30 in yellow and when both were found green). Each individual is shown in one column.

CHAPTER IV

Tempo and mode of regulatory evolution in Drosophila

4.1 Abstract

Genetic changes affecting gene expression contribute to phenotypic divergence; thus, understanding how regulatory networks controlling gene expression change over time is critical for understanding evolution. Prior studies of expression differences within and between species have identified properties of regulatory divergence, but technical and biological differences among these studies make it difficult to assess the generality of these properties or to understand how regulatory changes accumulate with divergence time. Here, we address these issues by comparing gene expression among strains and species of Drosophila with a range of divergence times and use F1 hybrids to examine inheritance patterns and disentangle *cis*- and *trans*-regulatory

Official citation:

Coolon, J.D., McManus, C.J., Stevenson, K.R., Graveley, B.R., Wittkopp, P.J. Tempo and mode of regulatory evolution in *Drosophila*, *Genome Research* 2014, 24, 797-808 doi:10.1101/gr.163014.113

Copyright 2014 Cold Spring Harbor Laboratory Press

This is an open-access article distributed under the terms of the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported License (CC-BY-NC-ND; http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode).

In this collaborative work, I developed the custom bioinformatics pipeline to quantify total and allele-specific expression (as previously described; see 4.5.4 and 4.8.2) from sequence reads of each comparison (mel-mel, sim-sec, and mel-sim) that had been aligned to their respective newly-built genomes (zhr, z30, tsimbazaza, and/or droSec1; see 4.8.1). To determine the level of sequence divergence for each comparison, I created reverse chain files to convert exonic coordinates annotated in the *D. melanogaster* genome to their respective zhr, z30, tsimbazaza, and droSec1 coordinates. I used these species-specific coordinates to extract the coding sequence of each exon, and for each comparison, the two strains or species sequences were pairwise aligned. Using contiguous sequence for each alignment, I identified the number of sites that differentiated each allele. I used these estimates with the contiguous sequence to determine the percentage of sequence divergence for each comparison by averaging across all coding regions (see 4.5.3). I also participated extensively in discussions as to how to appropriately normalize (downsample) sequence read counts between alleles and across comparisons (see 4.9).

changes. We find that the fixation of compensatory changes has caused the regulation of gene expression to diverge more rapidly than gene expression itself. Specifically, we observed that the proportion of genes with evidence of *cis*-regulatory divergence has increased more rapidly with divergence time than the proportion of genes with evidence of expression differences. Surprisingly, the amount of expression divergence explained by *cis*-regulatory changes did not increase steadily with divergence time, as was previously proposed. Rather, one species (*Drosophila sechellia*) showed an excess of *cis*-regulatory divergence that we argue most likely resulted from positive selection in this lineage. Taken together, this work reveals not only the rate at which gene expression evolves, but also the molecular and evolutionary mechanisms responsible for this evolution.

4.2 Introduction

Understanding the relationship between tempo (the rate at which a trait evolves) and mode (the manner in which a trait evolves) is essential for understanding the evolutionary process (Simpson, 1944). This is true not only for organismal phenotypes, but also for the molecular phenotypes that produce organismal traits. Gene expression is one such molecular phenotype (Barrière et al., 2012); it is essential for organismal form, fitness, and function, and frequently varies within and between species. Comparative studies using genomic surveys of gene expression in yeast (Busby et al., 2011), *Drosophila* (Rifkin et al., 2003), and mammalian species (Brawand et al., 2011) with a range of divergence times have provided insight into the tempo of gene expression evolution, but the mode and its relationship to tempo remain less well understood.

Elucidating the mode of gene expression evolution includes identifying the types

of regulatory changes that have evolved as well as how interactions among divergent regulatory alleles affect gene expression. F1 hybrids, in which divergent regulatory alleles interact in the same cellular environment, can be used to investigate these issues. Allele-specific expression in F1 hybrids separates the effects of *cis*- and *trans*regulatory changes affecting a gene's expression by providing a readout of relative *cis*regulatory activity in a common *trans*-regulatory environment (Cowles et al., 2002). Expression differences between genotypes not attributed to *cis*-regulatory changes are inferred to be caused by *trans*-regulatory divergence (Wittkopp et al., 2004). In addition, the net effects of interactions among divergent regulatory alleles are revealed by comparing levels of total expression in F1 hybrids to parental genotypes.

This approach was initially used to separate *cis*- and *trans*-regulatory effects of divergence affecting expression of dozens of genes. These studies suggested that (1) *cis*-regulatory changes are more common than *trans*-regulatory changes between species (Wittkopp et al., 2004); (2) genes with *cis*- and *trans*-acting changes favoring expression of opposite alleles are more likely than other types of changes to cause misexpression in F1 hybrids (Landry et al., 2005); (3) environmental factors modulate relative *cis*-regulatory activity (Meaux et al., 2006); (4) *cis*-regulatory variation is abundant in natural populations (Osada et al., 2006; Genissel et al., 2008; Campbell et al., 2008; Gruber and Long, 2009); and (5) the amount of expression divergence attributable to *cis*-acting changes is greater between than within species (Wittkopp et al., 2008b).

More recently, microarrays and RNA-seq have been used to extend these analyses to the genomic scale (Wang et al., 2008a; Graze et al., 2009; Tirosh et al., 2009; Zhang and Borevitz, 2009; Fontanillas et al., 2010a; McManus et al., 2010; He et al., 2012; Shi et al., 2012; Coolon and Wittkopp, 2013; Levy et al., 2013; Schaefke et al., 2013). In some cases, relationships seen in the smaller scale studies were replicated. For example, *cis*- and *trans*-regulatory changes with effects in opposite directions were overrepresented among misexpressed genes (Tirosh et al., 2009; McManus et al., 2010; Schaefke et al., 2013) and *cis*-regulatory changes explained more of the expression differences between than within species (Tirosh et al., 2009; Emerson et al., 2010). Other observations, such as the relative proportion of genes with evidence of *cis*and/or *trans*-regulatory changes, were much more variable among studies. Finally, novel patterns, such as the relationship between dominance and *cis/trans*-regulatory changes (Lemos et al., 2008; McManus et al., 2010) and the frequency of compensatory *cis*- and *trans*-regulatory variants (Tirosh et al., 2009; Goncalves et al., 2012; Shi et al., 2012), were identified.

Despite this growing collection of case studies examining the types of changes responsible for expression differences within and/or between particular pairs of species, the use of different organisms (flies, yeast, plants, and mice), techniques (pyrosequencing, microarrays, RNA-seq), and analysis methods (linear models, exact tests, and Bayesian approaches) among these studies precludes the type of meta-analysis needed to determine how the mode of regulatory evolution changes with divergence time and to robustly assess the generality of relationships reported in previous studies. To address these issues, we examined the tempo and mode of regulatory evolution in concert using strains and species of *Drosophila* with a range of divergence times.

4.3 Results

4.3.1 Experimental overview

mRNA abundance was compared among (1) African and non-African strains of *Drosophila melanogaster* (mel-mel), which have been geographically isolated for \sim 10,000 yr and show evidence of behavioral isolation (David and Capy, 1988; Lachaise et al., 1988; Wu et al., 1995; Hollocher et al., 1997) and expression divergence (Hutter et al., 2008); (2) *D. simulans* and *D. sechellia* (sim-sech), which diverged \sim 250,000 yr ago (Garrigan et al., 2012); and (3) *D. melanogaster* and *D. simulans* (mel-sim), which diverged \sim 2.5 million yr ago (Figure 4.1A; (Cutter, 2008)). For each of these genotypes, we derived a strain-specific genome sequence and used RNA-seq to measure mRNA abundance (hereafter referred to as expression) in a pool of 20 adult female flies. Reciprocal crosses were performed for each of the three pairs of genotypes (mel-mel, sim-sech, and mel-sim), and RNA-seq was used to measure both total and allele-specific expression in pools of 20 female F1 hybrids from each cross (Figure 4.1B). Sequence divergence observed in transcribed regions of these strains correlated with published estimates of divergence time (Figure 4.1C) as well as the number of RNA-seq reads informative for allele-specific expression (Figure 4.1D). Gene-specific and allele-specific read counts were used to investigate regulatory evolution as shown in Figure 4.5.

4.3.2 Quantifying gene expression levels

For each comparison (mel-mel, sim-sech, and mel-sim), RNA-seq reads from the two strains or species and their F1 hybrids were aligned to the relevant genomes and mapped to specific genes. Differences in sequencing depth among libraries (Figure 4.23) were eliminated by using random sampling without replacement to produce a data set with the same number of mapped reads for each sample. After excluding genes with fewer than 20 mapped reads in any sample (Figure 4.24), 7587 genes were deemed suitable for comparing total expression levels between all pairs of genotypes and their F1 hybrids (Data set 1), which is 83% of the genes classified as expressed in D. melanogaster adult females by modENCODE (Graveley et al., 2011).

Measures of relative gene expression derived from these mapped and normalized

RNA-seq data correlated well with estimates of relative gene expression derived from independent pyrosequencing experiments (Figure 4.7A; (Ahmadian et al., 2000)). Genome-wide, expression levels between F1 hybrids from reciprocal crosses were also highly correlated (Figure 4.2A; Figure 4.8). Despite this similarity, Fisher's exact tests (FETs) with a false discovery rate (FDR) of 0.05 identified significant expression differences between reciprocal hybrids for 26%-49% of individual genes (Figure 4.2B). Most of these significant expression differences were small in magnitude (median expression difference = 1.20- to 1.25-fold) (Figure 4.9), however, they reflect the sensitivity of the Fisher's exact test for detecting differences in relative expression from RNA-seq data when read counts are high. These differences in expression differences detected in the mel-mel, sim-sech, and mel-sim comparisons because they include variance from technical and biological replication as well as parent-of-origin effects.

4.3.3 Evolution of expression differences

To determine the tempo of regulatory divergence, we compared total expression levels in the mel-mel, sim-sech, and mel-sim comparisons for the set of 7587 genes in Data set 1 described above. First, we analyzed overall expression divergence (1-Spearman's ρ , see Methods) and found that it increased consistently and significantly with divergence time (Figure 4.2A; Figure 4.10). We then used FETs to compare expression levels for individual genes and determine whether the increased overall expression divergence resulted from more genes with divergent expression or more divergent expression of similar numbers of genes. Surprisingly, we found that the proportion of genes with significant expression differences did not increase consistently with divergence time (Figure 4.2B), suggesting that increasing magnitudes of expression differences rather than increasing numbers of genes with divergent expression drive the overall increase in expression differences with divergence time observed.

We also examined the evolutionary trajectories of individual genes by assigning each of the 7587 genes in Data set 1 to one of nine classes depending on whether its expression difference increased, decreased, or remained similar between mel-mel and sim-sech and between sim-sech and mel-sim. Expression differences less than 1.25-fold were considered similar for this analysis to minimize the impact of small but statistically significant expression differences (Figure 4.11). Despite observing that expression differences increased with divergence time on a genomic scale (Figure 4.2A), this pattern was only seen for 2% of individual genes (Figure 4.2C, class I). Expression differences of similar magnitude in all three comparisons were much more common (43%) of all genes examined) and tended to be small in magnitude (median expression difference = 1.18-fold) (Figure 4.2C, class II). The remaining 55% of genes fell into one of seven categories in which two of the three comparisons showed similar expression differences (Figure 4.2C, class III). Interestingly, nearly half (45%) of such genes showed similar expression differences in mel-mel and mel-sim but larger or smaller expression differences in the sim-sech comparison (Figure 4.2C, IIIc and IIId), which has an intermediate divergence time.

4.3.4 Evolution of regulatory incompatibilities

Divergence of the regulatory networks controlling gene expression can cause misexpression in F1 hybrids that can contribute to speciation (Meiklejohn et al., 2003; Michalak and Noor, 2004; Ranz et al., 2004; Haerty and Singh, 2006; Moehring et al., 2007; Maheshwari and Barbash, 2012). This can occur, for example, when proteins and/or DNA with sequence-specific interactions coevolve such that divergent alleles of the interacting molecules do not function properly together in F1 hybrids. To determine the rate at which misexpression resulting from such regulatory incompatibilities evolves, we compared expression levels in mel-mel, sim-sech, and mel-sim F1 hybrids to expression levels in the corresponding parental genotypes. We found that overall expression differences between parents and F1 hybrids increased with divergence time, most dramatically in the mel-sim comparisons (Figure 4.2A; Figure 4.12). A similar increase was seen in the proportion of genes showing misexpression in F1 hybrids (Figure 4.2B). The much more extensive misexpression seen in mel-sim F1 hybrids compared with mel-mel or sim-sech F1 hybrids is consistent with mel-sim F1 hybrid females having morphological defects that cause sterility (Dickinson et al., 1984) and mel-mel and sim-sech F1 hybrid females being completely fertile (Lachaise et al., 1986; Hollocher et al., 1997).

To further investigate the inheritance of gene expression levels and how inheritance patterns change over evolutionary time, we considered each gene separately and classified its expression in F1 hybrids as dominant, additive, misexpressed (i.e., over- or under-dominant), or similar (Figure 4.13). To minimize the impact of small but statistically significant expression differences on this analysis (Figure 4.14), we considered expression similar between genotypes if the expression difference was less than 1.25-fold. In the mel-mel F1 hybrids, we found that 7% of genes showed additivity, 14% showed misexpression, and 43% showed dominant inheritance. The remaining 36% of genes showed similar expression in both strains of *D. melanogaster* and in their F1 hybrids. The proportions of genes with additive and dominant inheritance decreased consistently with divergence time, whereas the proportion of genes showing misexpression increased dramatically with divergence time (Figure 4.2D).

4.3.5 Using allele-specific RNA-seq reads to study regulatory evolution

Differences in gene expression can be caused by changes in *cis*- and/or *trans*regulation. Understanding the relative contribution of these two types of changes is critical for understanding the mode of regulatory evolution (Barrière et al., 2012). To separate the effects of *cis*- and *trans*-regulatory divergence, we analyzed allelespecific expression in F1 hybrids and contrasted it with comparable measures of total expression differences between parental genotypes derived from allele-specific reads in "mixed parental" samples. These mixed parental samples were constructed in silico by combining equal numbers of mapped RNA-seq reads from each parental genotype and subjected to the same bioinformatic analysis as the reads from F1 hybrids. Expression differences between alleles in F1 hybrids were attributed to *cis*-regulatory differences, and differences in relative expression between parental genotypes that were not explained by differences in *cis*-regulatory activity were attributed to *trans*regulatory divergence (Wittkopp et al., 2004).

For each F1 hybrid and mixed parent sample, RNA-seq reads that aligned perfectly and uniquely to one parental genome but not the other were considered allelespecific. Genes with low confidence allele assignments (see Supplemental Material), fewer than 20 total allele-specific reads, or expression consistent with genomic imprinting in any comparison were excluded from analysis (Figure 4.25). For each of the remaining 4851 genes, differences in the number of allele-specific reads among comparisons were eliminated by using hypergeometric sampling to produce a data set with the same number of allele-specific reads in all comparisons (Data set 2). Measures of relative total expression derived from allele-specific reads in the mixed parental samples were strongly correlated with measures of relative total expression derived from the full RNA-seq data set (Figure 4.15) and pyrosequencing (Figure 4.7B). Relative allele-specific expression in F1 hybrids also showed a strong correlation between the RNA-seq and pyrosequencing data (Figure 4.7C) and was similar in F1 hybrids from reciprocal crosses (Figure 4.16). In the analyses described below, hybrids from reciprocal crosses were considered separately, with results from one hybrid for each comparison presented in the main text and results from the other hybrid presented in the Supplemental Material. With few exceptions (noted below), results were similar between reciprocal hybrids.

4.3.6 Evolution of cis- and trans-regulation

To determine the rate of *cis*-regulatory divergence and compare it with the rate of total expression divergence for the same genes, we contrasted overall differences in relative allelic abundance between the F1 hybrid and mixed parental samples for the 4851 genes deemed suitable for measuring allele-specific expression (Data set 2). Compared with the 7587 genes discussed above (Data set 1), this set of genes showed more similar levels of overall expression differences among the three comparisons (Figure 4.3A; Supplemental Figs. S12A, S13A-C), resulting from increased expression divergence in mel-mel and sim-sech in Data set 2 relative to Data set 1 (Figure 4.19). Despite this similarity in total expression differences among comparisons, we found that *cis*-regulatory differences were greater between than within species, with similar differences in relative *cis*-regulatory activity observed in sim-sech and mel-sim (Figure 4.3A; Figures 4.17A, 4.18D-I). Comparing the proportions of genes with statistically significant differences in total expression and *cis*-regulatory activity showed a similar pattern, except that the proportion of genes with evidence of a *cis*-regulatory difference increased consistently and significantly with divergence time (Figure 4.3B; Figure 4.17B). This suggests that the greater overall *cis*-regulatory divergence observed in the sim-sech comparison for these 4851 genes results from large differences in relative *cis*-regulatory activity for some genes rather than an excess of genes with divergent *cis*-regulatory activity.

We also compared the evolutionary trajectories of individual genes for total expression differences (Figure 4.3C) and relative *cis*-regulatory activity (Figure 4.3D; Figure 4.20) by dividing the 4851 genes in Data set 2 into the same nine classes described above for Data set 1 (Figure 4.2C). Compared with total expression, we found that more genes showed consistent and small (median = 1.16-fold) differences in relative *cis*-regulatory activity in all three comparisons (Figure 4.3C, II, 3D, II). We also observed more genes with unique differences in *cis*-regulatory activity in simsech (Figure 4.3D, IIIc,d) and mel-sim (Figure 4.3D, IIIe,f) that were greater in these comparisons than the other two comparisons. In other words, genes with a similar difference in *cis*-regulatory activity in mel-mel and mel-sim but not sim-sech were more likely to show increased than decreased divergence in sim-sech relative to the other two comparisons. Such asymmetry was much less pronounced for levels of total expression (Figure 4.3C), suggesting that *trans*-acting changes have compensated for differences in *cis*-regulatory activity in many cases.

Differences between divergent *cis*-regulatory activity and total gene expression are caused by the divergence of *trans*-regulatory factors. We found that significantly more genes showed evidence of *trans*-regulatory differences in the mel-mel and melsim comparisons than in the sim-sech comparison (Figure 4.3B; Figure 4.17B). This suggests that *cis*-regulatory divergence accounts for a larger proportion of overall expression divergence in sim-sech than in mel-mel or mel-sim. Consistent with this inference, a regression analysis showed that *cis*-regulatory differences explained more of the expression differences between *D. simulans* and *D. sechellia* than between either of the other two pairs of genotypes (Figure 4.21).

As overall sequence divergence increases, the number of loci with variation affecting expression of each gene is also expected to increase. Consistent with this expectation, we found that the proportion of genes with regulatory changes showing evidence of both *cis*- and *trans*-regulatory changes increased with divergence time, although the increase between the mel-mel and sim-sech comparisons was only statistically significant for one of the two hybrids (Figure 4.3E; Figure 4.17C). For the majority of these genes, the *cis*- and *trans*-regulatory changes favored expression of alternative alleles (Figure 4.3F; Figure 4.17D), suggesting that stabilizing selection has favored regulatory mutations that reduce expression differences. As described above, this type of developmental systems drift (True and Haag, 2001) is thought to cause misexpression in F1 hybrids (Michalak and Noor, 2004; Ranz et al., 2004; Landry et al., 2005; McManus et al., 2010; Barrière et al., 2012; Maheshwari and Barbash, 2012). The frequency of genes with compensatory *cis*- and *trans*-regulatory changes did not increase steadily with divergence time; however, *cis*- and *trans*-regulatory changes favoring expression of opposite alleles were observed least often in the sim-sech comparison (Figure 4.3F; Figure 4.17D). Contrary to prior studies (Landry et al., 2005; Tirosh et al., 2009; McManus et al., 2010), we found that genes affected by *cis*- and *trans*-regulatory changes with opposing effects on total expression levels were not more likely to show misexpression in F1 hybrids (Figure 4.3G; Figure 4.17E).

To determine how the relative effects of *cis*- and *trans*-regulatory changes vary with divergence time, we calculated the percentage of total regulatory divergence attributable to *cis*-regulatory changes for each gene. This value is referred to as "percent *cis*" (%*cis*), and prior studies of flies (Wittkopp et al., 2008b; McManus et al., 2010) and yeast (Emerson et al., 2010) found it to be larger between than within species. We also found that %*cis* was larger between than within species;

however, in contrast to prior predictions (Wittkopp et al., 2008b), %*cis* did not increase systematically with divergence time. Rather, it was largest for the sim-sech comparison with intermediate divergence time (Figure 4.4A; Figure 4.22A). A correlation between %*cis* and total expression divergence for individual genes was previously reported between *D. melanogaster* and *D. sechellia* (McManus et al., 2010), but we did not observe this pattern for any of the three comparisons (Figure 4.22B-G). Finally, two prior studies (Lemos et al., 2008; McManus et al., 2010) reported that %*cis* was higher for genes showing additive than nonadditive (i.e., dominant, over-dominant, or under-dominant) inheritance. We observed this relationship only for the comparison of *D. simulans* and *D. sechellia* in one hybrid (Figure 4.4B; Figure 4.22H), suggesting that it is also not a general feature of regulatory evolution.

4.4 Discussion

Researchers have been comparing genomic patterns of expression divergence among species for over a decade using microarrays, but sequence divergence between microarray probes and RNA samples often complicates comparisons among species and differences in normalization and statistical analyses can complicate comparisons among studies. Here, we use RNA-seq data to determine the tempo and mode of regulatory evolution among four divergent strains and species of Drosophila. This technique is better suited for interspecific comparisons than microarrays because it uses full sequence information instead of hybridization signals to determine gene expression levels, allowing more direct comparisons among species and studies.

RNA-seq was also recently used to compare expression levels in six different tissues among nine mammalian species and a bird (Brawand et al., 2011). Using Spearman's rank correlation coefficient ρ to compare overall expression differences in each pair of species, this study showed that expression similarity decreased quickly over shorter divergence times and then slowed. Patterns of expression divergence were strikingly similar among RNA samples from brain (cerebral cortex or whole brain without cerebellum), cerebellum, heart, kidney, and liver, with accelerated expression divergence in RNA samples from testes (Brawand et al., 2011). By combining our data with data from three previous studies (McManus et al., 2010; Meisel et al., 2012; Suvorov et al., 2013), we found that expression divergence among Drosophila species showed a similar pattern to that of mammals, but on a different timescale (Figure 4.5A,B). The Drosophila data showed greater expression divergence (lower values of ρ) than the mammalian data, which could be due to differences in tissue size among *Drosophila* species given that whole bodies rather than single tissues were used to generate these data. RNA-seq has also been used to compare expression divergence among four species of yeast (Busby et al., 2011), but it is difficult to compare the tempo in yeast to that of Drosophila and mammals because only three divergence time points were sampled (Figure 4.5C).

For each gene, interspecific expression differences can be caused by *cis*- and/or *trans*-regulatory changes. When F1 hybrids can be made between species, measures of allele-specific expression can be used to disentangle the net effects of these two types of changes (Wittkopp et al., 2004). Such analyses have been reported for closely related pairs of strains or species in yeast (Tirosh et al., 2009; Emerson et al., 2010; Schaefke et al., 2013), flies (Graze et al., 2009; McManus et al., 2010; Coolon et al., 2012), plants (Shi et al., 2012; Bell et al., 2013), fishes (Murata et al., 2012; Shen et al., 2012), and mice (Goncalves et al., 2012). To the best of our knowledge, this is the first genomic study collecting data on *cis*- and *trans*-regulatory divergence for more than one pair of genotypes. As such, it provided unprecedented insight into

the rate at which *cis*- and *trans*-regulatory changes evolve and allowed us to better assess the generality of relationships reported in other studies.

4.4.1 Compensatory cis- and trans-regulatory changes are common

We found that the number of genes with evidence for *cis*-regulatory divergence increased linearly with divergence time, but the number of genes with differences in total expression did not (Figure 4.3B; Figure 4.S12D). This suggests that transregulatory factors might often compensate for *cis*-regulatory differences at the level of total gene expression, either by fixing compensatory *trans*-regulatory variants or by feedback mechanisms affecting availability or activity of *trans*-acting factors (Mc-Manus et al., 2014). Consistent with this interpretation, *cis*- and *trans*-acting changes affecting expression of the same gene had opposite effects on expression levels 79%, 73%, and 87% of the time in the mel-mel, sim-sech, and mel-sim comparisons, respectively (Figure 4.3F). The exponential accumulation of genes that are misexpressed in F1 hybrids (Figure 4.2D) is also consistent with compensatory changes playing an important role in maintaining gene expression levels over evolutionary time (Landry et al., 2005). Such compensation can result from stabilizing selection acting to maintain similar expression levels in the face of new mutations, and has been seen not only in flies, but also in yeast (Tirosh et al., 2009), mice (Goncalves et al., 2012), and plants (Shi et al., 2012).

Compensation for *cis*-regulatory divergence resulting from the fixation of *trans*acting changes could evolve by fixing *cis*-acting mutations first and then compensating *trans*-acting mutations, or vice versa. We favor the latter model because *trans*-acting mutations appear to arise more frequently than *cis*-acting mutations for individual genes (Gruber et al., 2012) and most *trans*-acting mutations that compensate for *cis*-regulatory divergence of one gene are expected to have deleterious pleiotropic effects on expression of other genes (Wray et al., 2003; Carroll, 2008; Orgogozo and Stern, 2009). Goncalves *et al.* (Goncalves et al., 2012) favored a similar explanation for the extensive compensatory *cis*- and *trans*-regulatory changes they observed between strains of mice. An example of such *trans*-regulatory divergence subsequently compensated for by *cis*-regulatory changes has been described in yeast (Kuo et al., 2010). Regardless of which type of regulatory mutation is usually fixed first, it is clear that the regulatory networks controlling gene expression evolve more rapidly than the output from these networks.

4.4.2 Relative impact of selection and drift on regulatory evolution

A common goal for comparative studies of gene expression is identifying the selective and nonselective forces responsible for patterns of divergence and conservation, but this is not straightforward (Gilad et al., 2006a; Fay and Wittkopp, 2008; Emerson et al., 2010). Without the biological replication needed to make statistically robust inferences based on alternative evolutionary models (e.g., (Rifkin et al., 2003; Fay and Wittkopp, 2008; Bedford and Hartl, 2009; Brawand et al., 2011)), we can only make speculative statements about the evolutionary processes responsible for each of the nine different trajectories of expression divergence we observed (Figures 4.2C, 4.3C,D). For example, genes with similar (and typically small) expression differences in all three comparisons (class II in Figures 4.2C and 4.3C,D) may either have low mutation-drift variance or be subject primarily to stabilizing selection. This is the most abundant class of genes for both total expression and *cis*-regulatory activity with 43% and 34% of genes showing this pattern for total expression in Data sets 1 and 2, respectively, and 48% of genes showing this pattern for differences in *cis*regulatory activity in Data set 2. This is consistent with prior work suggesting that stabilizing selection has had a larger impact on the evolution of gene expression than genetic drift (Hsieh et al., 2003; Rifkin et al., 2003; Lemos et al., 2005; Gilad et al., 2006b; Xing et al., 2007; Kalinka et al., 2010). Indeed, <2.2% of genes in each comparison showed the increasing differences in total expression and/or *cis*-regulatory activity with divergence time (class I in Figures 4.2C, 4.3C,D, and Figure 4.20) that are expected when expression evolves primarily due to genetic drift (Khaitovich et al., 2004; Gilad et al., 2006b). The remaining genes fell into one of seven categories consistent with variable selection pressures among lineages (class III in Figures 4.2C and 4.3C,D).

4.4.3 Lineage-specific regulatory changes in D. sechellia

Gene-specific patterns of total expression divergence consistent with lineage-specific selection were more abundant in sim-sech than mel-mel or mel-sim for both Data sets 1 and 2 despite the sim-sech comparison having an intermediate divergence time (Figures 4.2C, 4.3C). This is consistent with *D. sechellia* being an island endemic species with a small effective population size that has evolved many novel phenotypes relative to D. melanoque and D. simulans (Orgogozo and Stern, 2009), including adaptation to a new host plant (Jones, 2005). As a consequence of this evolutionary history, D. sechellia might have fixed more deleterious mutations than the other two species by drift as well as more adaptive substitutions by positive selection. We observed an apparent excess of *cis*-regulatory divergence between *D. simulans* and *D.* sechellia (Figures 4.3A, 4.4A; Figure 4.21) that we believe is more likely to result from positive selection than drift because (1) *trans*-acting variation contributes more than cis-acting variation to polymorphic expression within species (Lemos et al., 2008; Wittkopp et al., 2008b; Emerson et al., 2010), suggesting that drift is more likely to fix trans-acting than cis-acting variants; (2) cis- and trans-regulatory changes affecting expression of the same gene were most likely to act in the same direction in the

sim-sech comparison (Figure 4.3F), which is consistent with positive, directional selection; and (3) simulation studies have shown that *cis*-regulatory divergence is more likely to be driven by natural selection than *trans*-regulatory divergence (Emerson et al., 2010). These results emphasize the importance of considering not only divergence time, but also the demographic and ecological history of individual species when studying the tempo and mode of evolution.

4.5 Methods

4.5.1 Fly strains, rearing, and collections

Four Drosophila genotypes were used for this study: the *D. melanogaster* North American zhr strain [full genotype: XYS.YL.Df(1)Zhr] (Sawamura et al., 1993; Ferree and Barbash, 2009), the *D. melanoqaster* Zimbabwean isofemale strain z30 (Begun and Aquadro, 1993; Wu et al., 1995), the sequenced D. sechellia strain (droSec1 [14021-0428.25]), and an isofemale strain of D. simulans (Tsimbazaza) that mates well with D. melanogaster (Hollocher et al., 1997). All flies were reared on cornmeal medium using a 16:8 light:dark cycle at 20°C. Just prior to the start of the experiment, all strains were subjected to 10 generations of sibling pair matings to reduce genome-wide heterozygosity, followed by three generations of population expansion to generate the quantity of flies needed for crosses. For each cross between strains of D. melanoqueter, 10 vials were set up with three female and three male flies each. For each interspecific cross, 30 vials were set up with three female and three male flies each. Virgin female progeny were allowed to mate from the time of eclosion to 3 d post-eclosion, then males and females were separated and females aged to 7-10 d post-eclosion. All flies were collected between 9 and 10 am to minimize the effects of circadian rhythm and snap-frozen in liquid nitrogen.

89

4.5.2 Sample preparation and sequencing

For each genotype analyzed, a pool of 20 female flies was used for total RNA extraction with TRIzol reagent according to manufacturer instructions (Invitrogen). This incorporates variation from biological replication into a single sample. Prior work has shown that expression for most genes is similar among replicate pools constructed in this way (Wittkopp et al., 2004, 2008a; Coolon et al., 2012). Genomic DNA (gDNA) was extracted from a separate pool of 20 flies for each genotype using the DNeasy Blood & Tissue Kit (Qiagen). Illumina sequencing libraries for RNA-seq were prepared as previously reported (McManus et al., 2010; Coolon et al., 2012). Briefly, 10 μ g of total RNA from each sample was treated with DNase I (Invitrogen) followed by poly(A)+ selection using Dynal magnetic beads (Invitrogen). Poly(A) + RNA was fragmented using RNA fragmentation reagent (Ambion) before cDNA synthesis. Double-stranded cDNA was produced using random hexamers and SuperScript II reverse transcriptase (Invitrogen). cDNA was run on a 2% agarose gel and the region corresponding to ~ 300 -bp fragments was extracted. The size-selected double-stranded cDNA extracted from this gel slice was used in the Paired-End Genomic DNA Library Preparation Kit (Illumina) according to manufacturer's recommendations. For the gDNA sequencing libraries, 10 μ g of gDNA was used with the Paired-End Genomic DNA Library Preparation Kit (Illumina), following manufacturer's recommendations. Each cDNA and gDNA library was subjected to a full lane of paired-end sequencing on an Illumina Genome Analyzer IIx using 76 cycles. On average, 24 million 76-bp, paired-end sequence reads were generated from each sequencing library (Figure 4.23). The zhr gDNA sample was also sequenced from a single end on an additional lane for 76 cycles per read. Images were analyzed using the Firecrest and Bustard modules to generate sequence and quality scores for each read.

4.5.3 Resequencing, genome assembly, and sequence divergence

Using the gDNA sequences, we constructed a strain-specific genome sequence for each genotype as described in the Supplemental Material. To determine percent sequence divergence in each comparison (mel-mel, sim-sec, mel-sim), we created reverse chain files to liftOver coordinates from *D. melanogaster* dm3 space to each of the other strain or species genomic space (zhr, z30, Tsimbazaza, droSec1) using the chainSwap utility from the UCSC Genome Browser (Kent et al., 2002). Using these chain files, we converted the dm3 genomic coordinates for each exon used for quantification in this study into their respective strain- or species-specific genomic coordinates. Using these coordinates, sequences for each exon were extracted from each strain- or species-specific genome. These sequences were aligned in pairs using Fast Statistical Alignment (FSA) (Bradley et al., 2009), and the number of divergent sites per gene was determined using custom perl scripts (pairwise_aln_FSA.pl, compare_pairwise.pl, seq_div_from_set.pl). Strain-specific genomes and chain files are provided in Supplemental File 1, and all custom perl scripts are included in Supplemental File 2.

4.5.4 Mapping sequencing reads to genes and alleles

We built a bioinformatics pipeline to measure total and allele-specific expression from Illumina sequencing outputs similar to those reported previously (McManus et al., 2010; Coolon et al., 2012). This pipeline, as well as the pyrosequencing methods used to validate measures of total and allele-specific expression derived from this pipeline, is described in the Supplemental Material.

4.5.5 Normalizing RNA-seq read counts among comparisons

Different numbers of sequence reads were recovered for each of the 10 cDNA libraries sequenced. These differences in read counts caused the Fisher's exact tests used to identify significant changes in gene expression between pairs of genotypes to have differences in power among the mel-mel, sim-sech, and mel-sim comparisons. To equalize power in all three comparisons, we considered exactly 12,704,991 mapped reads from each RNA-seq data set by down-sampling mapped reads randomly without replacement in all but the *D. sechellia* data set, which already had exactly 12,704,991 mapped reads (Figure 4.24). A similar down-sampling strategy was recently used to investigate the power of different bioinformatic tools for identifying expression differences (Rapaport et al., 2013). We then excluded genes with fewer than 20 reads in any of the RNA-seq data sets, resulting in the same 7587 "expressed" genes being analyzed in each comparison (Figure 4.24). Simulations confirmed that a larger data set down-sampled in this way has the same power to detect significant expression differences with a Fisher's exact test as a data set originally collected at the smaller sample size (data not shown). The exact data analyzed are provided in Supplemental Material as Data set 1.

4.5.6 Comparing total expression among genotypes

Spearman's correlation coefficients (ρ) were used to measure overall expression differences between pairs of genotypes, following Brawand *et al.* (Brawand *et al.*, 2011) and Meisel *et al.* (Meisel et al., 2012). Unlike Pearson's r, Spearman's ρ makes no assumptions about normality, linearity, or homoscedasticity. It is also less sensitive to outliers. Bootstrapping was used to test for statistically significant differences in ρ between mel-mel and sim-sech and between sim-sech and mel-sim by sampling with replacement 7587 gene-specific read counts from the observed 7587 genes 10,000 times using R, calculating ρ in each case, and determining the 2.5% and 97.5% percentiles. Significant differences were inferred when these 95% quantiles did not overlap.

We also tested for significant differences in expression level of individual genes by comparing the number of reads mapping to the focal gene to the number of reads mapping to the other 7586 genes between parental types, between reciprocal hybrids, and between each hybrid and parent using Fisher's exact tests with a null hypothesis of equal expression in both samples. This test was used instead of other methods for detecting differential expression because it recovers a similar proportion of true positives with fewer false positives without requiring replicates (Tarazona et al., 2011). Fisher's exact tests were also used to test for significant differences in the proportion of genes with significant differences between mel-mel and sim-sech and between sim-sech and mel-sim.

4.5.7 Inferring the mode of inheritance

To determine the mode of inheritance for each gene in each comparison, we followed the logic outlined in Gibson *et al.* (Gibson et al., 2004) and used previously for RNA-seq data in McManus *et al.* (McManus et al., 2010). Using a 1.25-fold expression difference cutoff and total expression levels in the F1 hybrids and corresponding parental genotypes, we classified each gene as either "similar", "additive", "parent 1 dominant", "parent 2 dominant", "under-dominant", or "over-dominant". Dominant inheritance was inferred when total expression in the F1 hybrid was similar to expression in one of the parental genotypes but different from the other parental genotype. Such genes were classified as either "parent 1 dominant" or "parent 2 dominant" depending on which parent the F1 hybrid resembled. Additive inheritance was inferred when F1 hybrid expression was different from, and intermediate to, both parents; and misexpression was inferred when the total expression in the F1 hybrid was different from both parental genotypes and greater than (over-dominant) or less than (under-dominant) the more extreme parental expression level. Genes with similar expression in both parents and F1 hybrids were classified as similar. Fishers exact tests were used to test for significant differences in the proportion of genes in each category between mel-mel and sim-sech and between sim-sech and mel-sim.

4.5.8 Normalizing allele-specific RNA-seq read counts among comparisons

To equalize power when testing for *cis*-regulatory divergence in mel-mel, sim-sech, and mel-sim, as well as when comparing tests for *cis*-regulatory and total expression divergence, we created a second data set with the same number of allele-specific reads for each gene in all comparisons. This data set was constructed by (1) combining the equal numbers of mapped reads for each genotype used in the first data set to make a "mixed parental" sample for each comparison (e.g., reads from zhr and z30 were combined for the mel-mel comparison); (2) counting allele-specific reads (i.e., reads that mapped perfectly and uniquely to only one of the parental genomes) in all mixed parental and F1 hybrid samples; and (3) equalizing allele-specific read counts for each gene in all mixed parental and hybrid samples by identifying the sample with the fewest allele-specific reads for that gene and using hypergeometric sampling of the observed allele-specific read counts to randomly reduce the number of allele-specific reads considered in each of the other samples. Simulations confirmed that this downsampling approach produced data sets with the same power to detect significant expression differences with Fishers exact tests as data sets originally collected at the smaller sample sizes (data not shown), and a similar method was recently used for allele-specific RNA-seq data from humans (Lappalainen et al., 2013).

Prior to analysis, genes with low confidence allele-assignments in the mel-mel, sim-sech, or mel-sim comparisons, defined as having >10% of the mapped reads from one parent aligned solely to the genome of the other parent, were excluded. Genes with less than 20 total allele-specific reads (allele 1 + allele 2 < 20) in any mixed parental or hybrid sample were also excluded from all comparisons; this threshold was based on prior theoretical and empirical work (Fontanillas et al., 2010b; McManus et al., 2010). Finally, nine more genes were excluded because they showed significant differences in relative allelic expression between reciprocal hybrids using Fisher's exact tests with a null hypothesis of equal expression and an FDR of 0.05. Such differences in relative allelic expression can result from parent-of-origin effects such as mitochondrial inheritance or genomic imprinting; imprinting seems rarely, if ever, responsible for this pattern of expression in *Drosophila*, however (Wittkopp et al., 2006, 2008a; Coolon et al., 2012). After applying these filters, 4851 genes were deemed suitable for allele-specific analysis in all comparisons, with most of the genes excluded from this data set because they had too few allele-specific reads in the mel-mel comparison (Figure 4.25).

Mitochondrial genes were excluded from our allele-specific data set; however, allele assignments for F1 hybrid reads that mapped to mitochondrial genes were used as one metric to evaluate the reliability of our bioinformatic allele assignments. In the absence of sequencing and allele-assignment errors, all of these reads should map to the maternal allele. We found that 99.5% and 99.8% of reads from mitochondrial genes mapped to the maternal allele in F1 hybrids between D. simulans and D. sechellia and between D. melanogaster and D. simulans, respectively (Figure 4.26). Additional validation of allele assignments is described in the main text. The exact data analyzed are provided in the Supplemental Material as Data set 2.

4.5.9 Evaluating cis- and trans-regulatory changes

Spearman's ρ was used to measure *cis*-regulatory divergence on a genomic scale in the mel-mel, sim-sech, and mel-sim comparisons by assessing the correlation between allele 1 and allele 2 read counts from F1 hybrids. It was also used to repeat the analysis of overall expression divergence in each comparison using the mixed parental samples. To test for statistically significant differences in ρ between mel-mel and sim-sech and between sim-sech and mel-sim, we used bootstrapping. Specifically, we sampled with replacement 4851 gene-specific read counts from the observed 4851 genes 10,000 times using R, calculated ρ in each case, and determined the 2.5% and 97.5% percentiles. Significant differences were inferred when these 95 percentiles did not overlap.

Binomial exact tests with a null hypothesis of equal expression were used to identify significant expression differences between genotypes in the mixed parental pools as well as significant differences in relative allelic expression in the F1 hybrid samples that indicate differences in relative *cis*-regulatory activity. An FDR of 5% was used to determine statistical significance despite the fact that the P-values produced by binomial exact tests when the null hypothesis is true are not uniformly distributed as assumed by the FDR correction for multiple tests (Skelly et al., 2011). This is because our simulations showed that the violation of this assumption had no effect on the number of genes called significant in this study (Supplemental Material). To test for the unequal allelic abundance between mixed parental and F1 hybrid samples that would indicate *trans*-regulatory divergence, we performed Fisher's exact tests with a null hypothesis of equal expression by comparing read counts from genotype 1 and genotype 2 in the mixed parental sample to allele 1 and allele 2 in the corresponding F1 hybrid samples. Each gene in each comparison was classified as "conserved", "all cis", "all trans", "cis+trans", "cis×trans", "compensatory", or "ambiguous" based on the results of the Fisher's and binomial exact tests using the criteria described in Figure 4.27. These same classifications were used previously in Landry *et al.* (Landry et al., 2005) and McManus *et al.* (McManus et al., 2010). Fisher's exact tests were also used to test for significant differences in the proportion of genes with significant differences between mel-mel and sim-sech and between sim-sech and mel-sim.

For each gene in each comparison, the total expression difference was calculated as $\log_2(\text{genotype 1 read count/genotype 2 read count})$ from the mixed parental sample, and the *cis*-regulatory difference ("cis") was calculated as $\log_2(\text{allele 1 read}$ count/allele 2 read count) from each of the F1 hybrid samples. The *trans*-regulatory difference ("trans") for each gene in each comparison was calculated as the difference between the total expression and *cis*-regulatory differences: $\log_2(\text{genotype 1 read}$ count/genotype 2 read count) - $\log_2(\text{allele 1 read count/allele 2 read count})$. %*cis* (proportion of total regulatory divergence attributable to *cis*-regulatory changes) was then calculated as $\frac{|cis|}{|cis|+|trans|} \times 100$.

4.5.10 Scripts and software used

All statistical analyses, down-sampling, and simulations were performed in R (version 2.12.2 or version 3.0.1, CRAN). This code includes the use of fisher.test for Fisher's exact tests, binom.test for binomial exact tests, corr.test for Spearman's ρ , sample to randomly down-sample mapped reads and simulate mapped read counts from a multivariate distribution, rhyper to randomly down-sample allele-specific read counts, rbinom to simulate allele-specific read counts. Custom perl and R scripts used in this work are included in Supplemental File 2.

4.6 Data access

The sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under accession numbers SRA052065 and SRP023274.

4.7 Acknowledgements

We thank Sebastian Zöllner and the University of Michigan LSA High Performance Computing for computational resources, Hope Hollocher, Chung-I Wu, and the Bloomington and UCSD stock centers for Drosophila strains, Laura Sligar for pyrosequencing assistance, Bing Yang for computational assistance, Sebastian Zöllner, J.J. Emerson, and Manolis Dermitzakis for statistical advice, and Brian Metzger, Richard Lusk, Fabien Duveau, and Alisha John for comments on the manuscript. Funding for this work was provided by the National Institutes of Health (5F32GM089009 -02 to J.D.C. and 5R01GM095296 to B.R.G.), the National Science Foundation (NSF 0903629 to K.R.S. and MCB-1021398 to P.J.W.), and the Alfred P. Sloan Research Foundation (fellowship to P.J.W.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Institutes of Health, National Science Foundation, or Sloan Foundation.

4.8 Supplemental methods

4.8.1 Resequencing and genome assembly

Production of the zhr and z30 genomes was previously described in Coolon et al. (Coolon et al., 2012). To construct the *D. simulans* tsimbazaza and *D. sechellia* droSec1 genomes, gDNA sequence reads from *D. sechellia* droSec1 and *D. simulans*

Tsimbazaza were aligned to the *D. simulans* and *D. sechellia* genome assemblies respectively using BWA (Li and Durbin, 2009) (version 0.5.6). Each read was aligned separately using default parameters, and SAM format files were generated using the BWA sampe command. Alignment files were converted to BAM format, and VCF files describing single nucleotide polymorphisms (SNPs) and indels were created using the SAMtools package (Li et al., 2009) (version 0.1.7a; modules view, sort, and pileup). SNP and indel calls were filtered using the samtools.pl varFilter command (as described at http://samtools.sourceforge.net/cns0.shtml) to retain SNPs and indels with phred scale quality scores of 20 or higher.

The public reference genomes of D. simulans and D. sechellia were originally sequenced at a coverage depth of 3- and 5-fold, respectively. This low coverage left large genomic regions unfinished. To close these gaps, we realigned gDNA sequences to the SNP and indel corrected genomes. Unmapped read-pairs were assembled into contigs with Velvet (Zerbino and Birney, 2008). Contigs whose ends both aligned to the genomes were considered "gap spanning", and extended 100 bp in each direction. Velvet assembled contigs (including gap-spanning) were aligned to the D. melanogaster reference genome (dm3) using LASTZ (Harris, 2007). Contigs that aligned uniquely to D. melanogaster were retained as the "extra genome", and comprised 12.6 and 1.6 Mb of sequence from D. simulans and D. sechellia, respectively. liftOver coordinate files were assigned to extra-genome contigs using the axtChain, chainNet, and netChainSubset utilities from the UCSC Genome Browser (Kent et al., 2002). gDNA sequence reads were then remapped to identify SNPs and indels in the extra genome, and genome sequences were rewritten accordingly

Despite our initial inbreeding, SAMtools identified residual heterozygosity at some sites in each genotype. This complicates allele-specific RNA-seq when one of the alleles segregating in strain 1 matches the allele that is invariant in strain 2. For example, consider a site polymorphic for A and C in zhr, but fixed for C in Tsimbazaza. RNA-seq reads originating from the Tsimbazaza allele would align to genome sequences for both strains, whereas reads originating from the zhr allele would be align only to the zhr genome. To eliminate the impact of such sites on measures of allele-specific expression, we changed such sites to the polymorphic genotype in both strains using a custom Perl script (snp_compare_filter.pl), effectively making these sites uninformative for allele assignment and producing comparison and strain specific genomes.

These comparison- and strain-specific genome sequences were produced using a custom Perl script (snp_adder.pl). This script sequentially rewrites the corresponding genome with corrected SNP calls and indels. The positions of insertions and deletions were recorded in custom liftOver chain files during the rewriting process. These chain files allowed the conversion of genomic features between strains and reference genomes using the UCSC Genome Browser liftOver tool (http://genome.ucsc.edu) (Kent et al., 2002). All genome and chain files are available upon request.

4.8.2 Mapping sequencing reads to genes and alleles

We aligned each mate of the paired-end RNA-seq reads separately to the strainor species-specific genomes specific to each comparison using the MOSAIK aligner (version 1.0.1384, http://bioinformatics.bc.edu/marthlab/Mosaik). For example, in the mel-mel comparison, reads derived from zhr, z30, and the F1 hybrids from reciprocal crosses between zhr and z30 were each aligned to both the zhr and z30 genomes that had been created specifically for the mel-mel comparison. Aligning reads to both parental genomes prevents the biased mapping described in prior RNA-seq studies of allele-specific expression (Degner et al., 2009; Stevenson et al., 2013). The following command line options were used for these alignments: -hs 13 -mm 0 -p 24 -mph 100 -act 20). The 13 base hash size (-hs 13) option allowed >99% of ambiguous base containing regions to be seeded for alignment by MOSAIK. Reads aligning uniquely with no mismatches to one or both genomes were considered "mapped reads" and retained for analysis. After the initial 76 bp reads were aligned to both reference genomes, reads that did not map to either genome were trimmed 13 bases from the 3 end using a custom Perl script (fastq_trimmer.pl) and again aligned using MO-SAIK. This was repeated three times (sequence lengths 76bp, 63bp, 50bp, 37bp). Any sequences that did not uniquely align after the final iteration were discarded.

Using the chain files created in the genome assembly process, we converted the genome coordinates from the zhr, z30, droSec1, and Tsimbazaza genomes to the sequenced D. melanogaster dm3 coordinates using the liftOver utility from the UCSC Genome Browser (Kent et al., 2002) (http://genome.ucsc.edu) and a custom Perl script (convert.pl). Sequence reads were then filtered based on their alignment to a previously identified set of constitutively expressed exons within the *D. melanogaster* genome (McManus et al., 2010) using the intersectBed module of BEDTools (Quinlan and Hall, 2010), with reads not aligning to these regions discarded. Additionally, overlapping regions in the constitutive exon set were removed using intersectBed module of BEDTools and custom scripts. Gap files were produced for each comparison and combined using the mergeBed module of BEDTools to create one gap file used for all comparisons. Sequences containing gaps in one or more genotypes were excluded. Reads were assigned to alleles based on alignments to strain-specific genomes using a custom Perl script (classify.pl). Because paired-end reads are derived from a single transcript, each set of paired-end reads was treated as a single read count regardless of whether one or both reads were mapped successfully.

4.8.3 Pyrosequencing

To evaluate the reproducibility of expression measurements derived from our RNA-seq data, we used pyrosequencing (Qiagen) to independently measure differences in total and allelic expression in the mel-mel and sim-sech comparisons. We focused our validation efforts on the mel-mel and sim-sech comparisons because they contained fewer divergent sites than the mel-sim comparison, making allele assignments more challenging. For the mel-mel comparison, we analyzed new F1 hybrid and mixed parental cDNA libraries synthesized from the same RNA samples used for Illumina sequencing, incorporating variation from technical replication. These mixed parental libraries were constructed by pooling equal amounts of RNA prior to cDNA synthesis. For the sim-sech comparison, we used RNA extracted from 4 biological replicate mixed parental (each containing 10 *D. simulans* and 10 *D. sechellia* flies) and F1 hybrid (each containing 20 F1 hybrid flies) samples to synthesize cDNA pools, incorporating variation from both technical and biological replication.

Pyrosequencing assays were developed for 10 genes in the mel-mel comparison and 18 genes in the sim-sech comparison (Table S6). Custom dispensation orders were used for the pyrosequencing reactions and custom formulas were developed for calculating relative allelic abundance (Table S6). Both gDNA and cDNA were analyzed in mixed parental and F1 hybrid samples in each case. cDNA was always synthesized from total RNA using T(18)VN primers and SuperScript II (Invitrogen) according to manufacturer recommendations. gDNA was extracted from an independent pool of F1 flies for mel-mel and sequentially from the same homogenate of flies as the RNA for each biological replicate of sim-sech using the protocol described in Wittkopp (2011). For mel-mel, pyrosequencing reactions were performed in triplicate for both the cDNA and gDNA samples. For sim-sech, single pyrosequencing reactions were performed on the cDNA and gDNA samples from each biological replicate.

Relative allelic abundance observed in the gDNA samples was used to normalize measurements from the corresponding cDNA samples, as described in (Wittkopp, 2011). Following normalization, the mean log₂-transformed ratio of relative allelic expression reported by pyrosequencing for each gene was compared to the log₂transformed ratio of relative allelic expression derived from the RNA-seq data using a type 2 regression in R.

4.9 Supplemental note

When using binomial exact tests to identify significant differences in relative allelic expression from RNA-seq data, p-values are not uniformly distributed when the null hypothesis (p = 0.5) is true, violating an assumption of the widely-used false discovery rate (FDR) correction for multiple testing (Skelly et al., 2011). To better understand the non-uniformity of p-values when the null hypothesis (p = 0.05) is true, we repeated and extended the simulations reported in Skelly *et al.* (Skelly et al., 2011). We followed the same simulation strategy except that we did not add a noise parameter. Specifically, we simulated allele-specific read count data by (i) sampling total allelespecific read counts (allele 1 +allele 2) for 4851 genes (the number of genes we examined for allele-specific expression) from a Poisson distribution with a mean (lambda) of 10, 100, 200, 500, 1000, or 2000 reads; (ii) determining the number of reads from one allele for each gene by drawing from a binomial distribution with p = 0.5, and (iii) determining the p-value from a binomial exact test comparing the reads from one allele to the total number of allele-specific reads in each case. As shown in the figure at right, the uniformity of the null p-value distribution increased with increasing total read counts.

In our Dataset 2, the total number of read counts per gene ranged from 20 to 16165, with a median of 142. We repeated the simulation described above using the observed number of total allele-specific read counts for the 4851 genes and found that the distribution of p-values assuming the null hypothesis was true did indeed violate uniformity. This is important because the FDR correction uses the assumption of a uniform distribution when the null hypothesis is true to model the FDR based on the observed distribution of p-values. To determine the impact of this non-uniform distribution on the proportion of genes called significant using the q-value = 0.05 cutoff we used, we artificially eliminated the over-representation of large, non-significant p-values from the null model in the observed p-value distribution by replacing all p-values>0.05 with an equal number of values drawn from a uniform distribution (min = 0.05, max = 1). We did this for the dataset simulated assuming the null hypothesis was true (top panels) as well as for the p-values observed in the mel-mel comparison (bottom panels).

Next, we applied the FDR correction to the list of p-values from each dataset. As shown in the figure at right for the observed p-values, q-values before (black) and after (red) the redistribution of non-significant p-values are quite similar. The largest difference in q-value was for p-values between 0.2 and 0.8. With the q-value threshold (0.05) used in the paper, as well as for other qvalue thresholds up to 0.1, the number of genes called significant using the FDR correction was unchanged by the redistribution of non-significant p-values (see Table below). This suggests that the non-uniformity of the BET null distribution has a negligible effect on the FDR in our study. Furthermore, even if the FDR determined by the correction model was slightly higher or lower than the true FDR, it is expected to deviate similarly in all comparisons because the same number of total allele-specific read counts were considered in each case, resulting in the same null distribution. Based on these results, we conclude that the non-uniformity of the BET null distribution does not affect the comparisons of regulatory evolution across divergence times that we report.

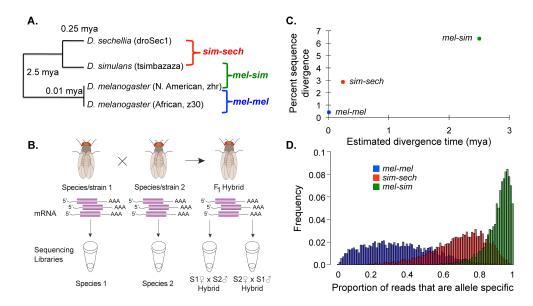


Figure 4.1: Studying regulatory evolution in the melanogaster group of Drosophila. Phylogenetic relationships and estimated divergence times for the strains and species analyzed are shown. B, Sequencing libraries for RNA-seq data were derived from mRNA isolated from each species and strain as well as F1 hybrids from reciprocal crosses, in which the maternal and paternal genotypes were reversed (e.g., S1S2 and S2S1). C, The percent sequence divergence observed in the regions of the genome used to map RNA-seq reads (Y-axis) is compared to published estimates of divergence time (X-axis). D, The proportion of reads from each gene that are allele-specific are shown for the mel-mel (blue), sim-sech (red), and mel-sim (green) comparisons. Figure

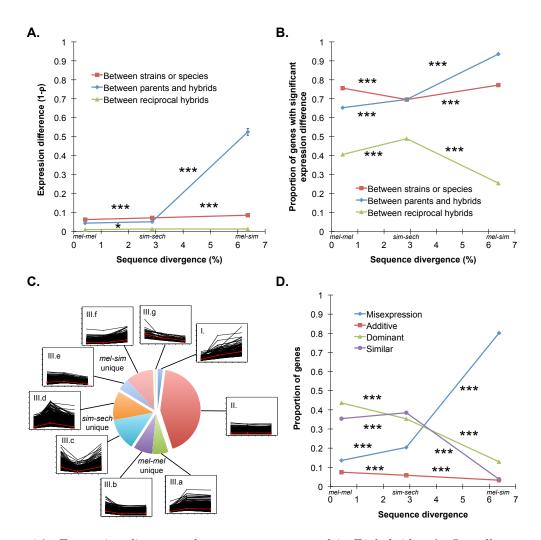


Figure 4.2: Expression divergence between genotypes and in F1 hybrids. A, Overall expression divergence $(1 - \rho)$ is shown for the mel-mel, sim-sech, and mel-sim comparisons in red, with the data used for these calculations shown in Figure 4.10. Average differences in expression between F1 hybrids and each of the parental species are shown in blue, with the data used for these calculations shown in Figure 4.12. Expression divergence between reciprocal F1 hybrids is included as a baseline in green with the data used for these calculations shown in Figure 4.8. In this and all other figures, results from each comparison are plotted using the genomic sequence divergence observed between the genotypes involved (Fig. 4.1C). B, The proportion of genes showing evidence of significant expression differences between genotypes (red), the average proportion of genes showing significant expression differences between F1 hybrids and each parental species (blue), and the proportion of genes with significant expression differences between reciprocal F1 hybrid genotypes (green) are shown. C, The line-plots show expression differences for individual genes in the melmel, simsech, and mel-sim comparisons plotted according to divergence time, with the 7,587 genes included in Dataset 1 classified into nine groups depending on whether they showed increased, decreased, or similar expression differences between mel-mel and sim-sech and between sim-sech and melsim. The red line in each plot shows the median expression difference for genes in that class for each comparison. The pie chart shows the relative frequency of genes in each class. D, The proportion of genes showing expression levels in F1 hybrids consistent with additive inheritance (red), dominant inheritance (green), misexpression (blue), or similar expression (purple) are shown for each comparison. Data used to calculate these proportions are shown in Fig. 4.13. Error bars in panel A show the 95% quantiles from 10,000 bootstrap replicates in which differences in 1 - ρ between mel-mel and sim-sech as well as between sim-sech and mel-sim were calculated for each bootstrap replicate. The significance of the observed deviation from zero was determined by comparing the observed value to the distribution of bootstrap values. In panels B and D, significance was determined using Fishers exact tests. Significance of each comparison is indicated with one (p ≤ 0.05), two (p ≤ 0.001), or three (p $\leq 1e-4$) asterisks.

106

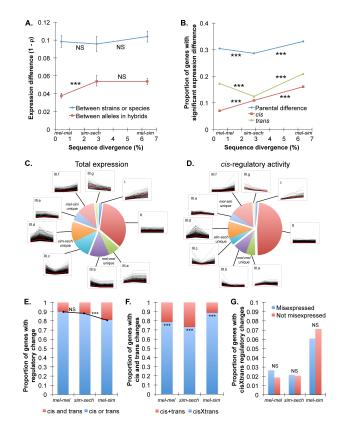


Figure 4.3: Evolution of *cis*- and *trans*-regulation. A, Overall differences $(1 - \rho)$ in total expression between genotypes (blue) and allele-specific expression in F1 hybrids (red) are shown for each comparison, with data used for these calculations shown in Fig. 4.18. Relative allelic expression in F1 hybrids provides a readout of relative cis-regulatory activity. B, For each comparison, the proportions of genes with evidence of significant differences in total expression (blue), cis-regulation (red), and *trans*-regulation (green) are shown. Data used to determine these proportions are shown in Fig. 4.21. Significance tests used to identify differences in trans-regulation had different power than those used to identify differences in total expression and cis-regulation, thus only the evolutionary trends, not the proportions of significant genes, should be compared among these classes. Power was comparable, however, in the tests for differences in total expression and relative cis-regulatory activity summarized in this figure. C and D, The line-plots show expression differences (C) and differences in relative cis-regulatory activity (D) for individual genes in the mel-mel, sim-sech, and mel-sim comparisons plotted according to divergence time, with the 4,851 genes included in Dataset 2 classified into nine groups depending on whether they showed increased, decreased, or similar expression differences between mel-mel and sim-sech and between sim-sech and mel-sim. The red line in each plot shows the median expression difference for genes in that class for each comparison. The pie-charts show the relative frequency of genes in each class. E, The proportion of genes with evidence of significant *cis*- and *trans*-regulatory changes (red) is compared to the proportion of genes with evidence of cis- or trans-regulatory changes (blue). F, For genes with evidence of both cis- and trans-regulatory changes, the frequency of genes with cis- and trans-regulatory changes affecting gene expression in the same (cis+trans, red) and opposite (cistrans, blue) directions are compared. G, The relative frequencies of genes with cis- and trans-regulatory changes in opposite directions that do (blue) and do not (red) show evidence of misexpression in F1 hybrids are compared. Error bars in panel A show the 95% quantiles from 10,000 bootstrap replicates in which differences in 1 - ρ between mel-mel and sim-sech as well as between sim-sech and mel-sim were calculated for each bootstrap replicate. The significance of the observed deviation from zero was determined by comparing the observed value to the distribution of bootstrap values. Significance was determined using Fishers exact tests in panels B, E and G and using binomial exact tests in panel F. Significance of each comparison is indicated with either NS for non-significant (p > 0.05) or one $(p \le 0.05)$, two $(p \le 0.001)$, or three $(p \le 1e-4)$ asterisks. Comparable analyses for reciprocal hybrids are shown in Fig. 4.17 and 4.20.

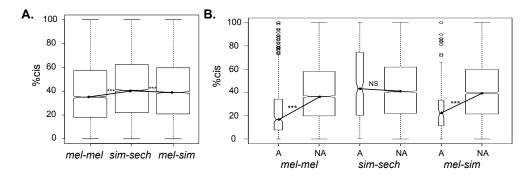


Figure 4.4: Effects of *cis*-regulatory divergence. A, The percentage of total regulatory divergence attributable to *cis*-regulatory divergence (%*cis*) is shown for the mel-mel, sim-sech, and mel-sim comparisons. B, %*cis* is compared for sets of genes showing additive (A) and non-additive (NA, dominant or misexpression) inheritance for each comparison. In all panels, notched boxplots show the full range of values as well as the 25, 50, and 75th percentiles. Within both panels, the width of the boxes are proportional to the number of genes represented. Statistical significance of differences between median values connected with solid lines were determined using Mann- Whitney U tests (* indicates $p \le 0.05$, ** indicates $p \le 0.001$, and *** indicates $p \le 1e-4$). Comparable analyses for reciprocal hybrids are shown in Fig. 4.22.

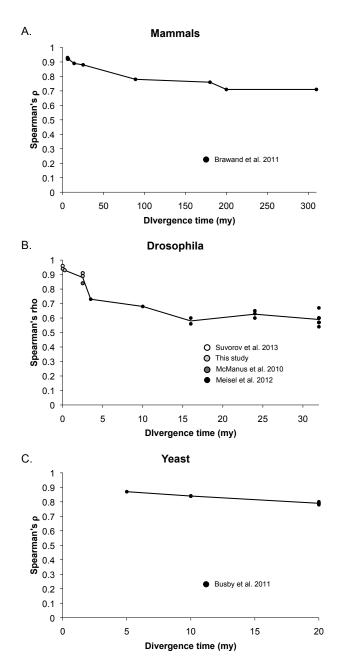


Figure 4.5: Expression divergence in mammals, *Drosophila* and yeast. A, Expression similarity (Spearmans ρ) was calculated using RNA-seq data from kidneys published in Brawand *et al.* (2011) comparing human samples to those of eight other mammalian species and one bird. We chose to analyze the data from kidney because it was most representative of all the tissues examined (excluding testes). Divergence times in millions of years are as reported in Brawand *et al.* (2011). B, Expression similarity (Spearmans ρ) was calculated for data described in this paper (light grey circles circles) as well as data published in Suvorov *et al.* (2013) (open circles), McManus *et al.* (2010) (grey circles), and Meisel *et al.* (2012) (black circles). Divergence times for mel-mel, simsech, and mel-sim are as described in Fig. 4.1A. For all other comparisons, estimated divergence times from Obbard *et al.* (2012) were used. C, Expression similarity (Spearman's ρ) was calculated using the data reported in Busby *et al.* (2011) for all pairwise comparisons of four yeast species. Divergence times for these species are from *et al.* (2003). In all three cases, the black line connects the average value of ρ for each divergence time sampled.

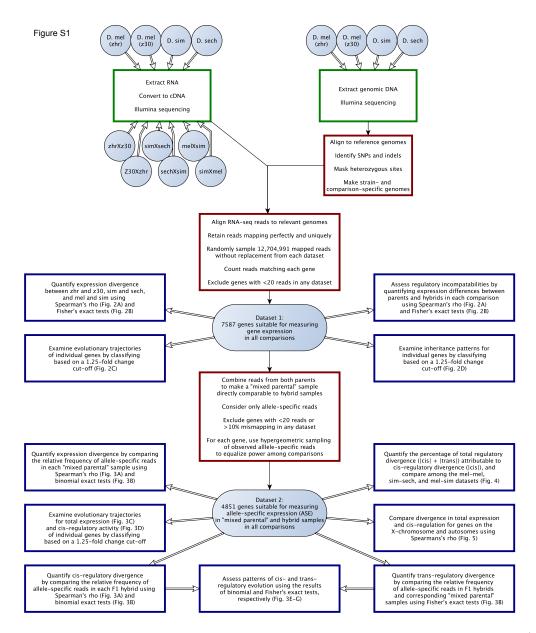


Figure 4.6: Methodological overview. This figure summarizes the biological samples analyzed (blue circles), the production of the raw sequencing data (green boxes), the bioinformatic methods (red boxes) used to convert the raw data into datasets 1 and 2 (blue ovals), and analyses performed on each of these datasets to examine patterns of regulatory evolution (blue boxes).

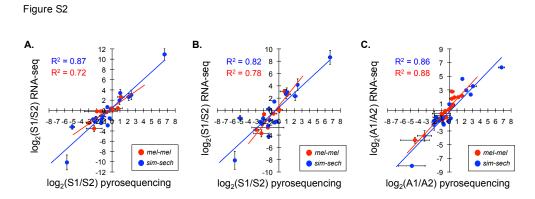


Figure 4.7: Independent confirmation of relative allelic expression levels inferred from RNA-seq data. Measures of total expression (A, B) and relative allelic expression (C) derived from RNA-seq read counts in Dataset 1 (A) and Dataset 2 (B, C) were compared to measures of total expression (A, B) and relative allelic expression (C) determined using pyrosequencing. For the mel-mel comparisons (red), cDNA samples analyzed by pyrosequencing were derived from the same RNAs used for Illumina sequencing (i.e., technical replicates). For the sim-sech comparison (blue), RNA samples extracted from new pools of flies (i.e., biological replicates) were analyzed by pyrosequencing. Coefficients of determination (\mathbb{R}^2) from type 2 regressions were used to compare expression measurements based on RNA-seq and pyrosequencing.



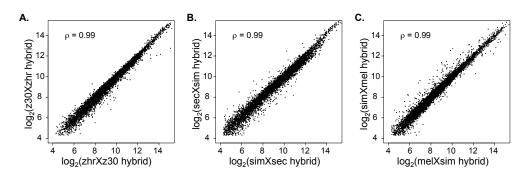


Figure 4.8: Total expression levels were similar between reciprocal hybrids. Total expression levels, plotted as log2(total read count) for each gene, were compared between F1 hybrids from reciprocal crosses for the mel-mel (A), sim-sech (B), and mel-sim (C) comparisons. Hybrid genotypes are written as maternal genotype x paternal genotype. Spearmans ρ , which makes no assumptions about normality, was used to compare the strength of the correlation in each case.

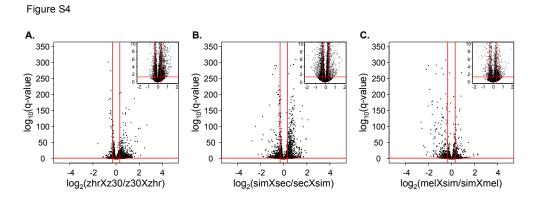


Figure 4.9: Most significant expression differences between reciprocal hybrids are small in magnitude. Volcano plots are shown for the comparison of total expression between reciprocal hybrids in the mel-mel (A), sim-sech (B), and mel-sim (C) comparisons. Statistical significance, represented by $\log_{10}(q$ -value), is plotted on the Y-axis and the expression difference, plotted as $\log_2(\text{reads from}$ hybrid 1/reads from in hybrid 2), is plotted on the X-axis. Hybrid genotypes are written as maternal genotype x paternal genotype. The vertical red lines correspond a 1.25-fold expression difference, whereas the horizontal red lines correspond to q-values with a false discovery rate of 0.05. Insets show genes with the smallest q-values in more detail.



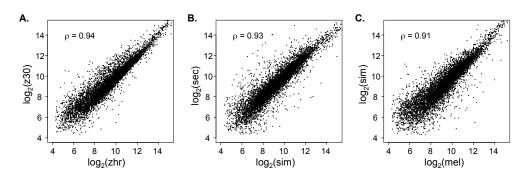


Figure 4.10: Overall expression differences increase with divergence time. For each gene, expression levels of individuals genes are compared between the zhr and z30 strains of *D. melanogaster* (A), between *D. simulans* and *D. sechellia* (B), and between *D. melanogaster* (zhr) and *D. simulans* (C). Expression levels are plotted as $\log_2(\text{read count})$. Spearmans ρ , which makes no assumptions about normality, was used to measure the overall expression similarity between genotypes in each case.

Figure S6

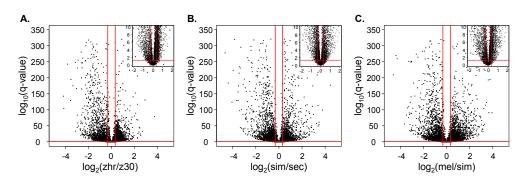


Figure 4.11: Many small expression differences are statistically significant between genotypes. Volcano plots are shown for the comparison of total expression between the zhr and z30 strains of D. melanogaster (A), between D. simulans and D. sechellia (B), and between D. melanogaster (zhr) and D. simulans (C). Statistical significance, represented by $\log_{10}(q$ -value), is plotted on the Y-axis, and the expression difference, plotted as $\log_2(\text{reads from genotype 1/reads from in genotype 2})$, is plotted on the X-axis. The vertical red lines correspond to a 1.25-fold expression difference, whereas the horizontal red lines correspond to q-values with a false discovery rate of 0.05. Insets show genes with the smallest q-values in more detail.

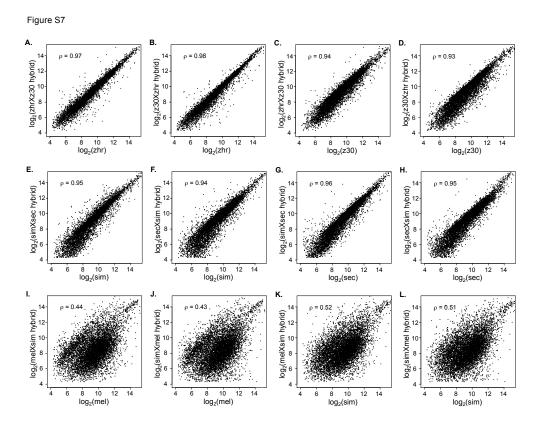


Figure 4.12: Expression differences between F1 hybrids and parental species increase with divergence time. For each gene, expression levels, plotted as $\log_2(\text{read counts})$, are compared between F1 hybrids (Y-axis) and each of the parental species (X-axis) for the mel-mel (A-D), sim-sech (E-H) and mel-sim (I-L) comparisons. Hybrid genotypes are written as maternal genotype x paternal genotype. Spearmans ρ , which makes no assumptions about normality, was used to compare overall expression differences in each case.

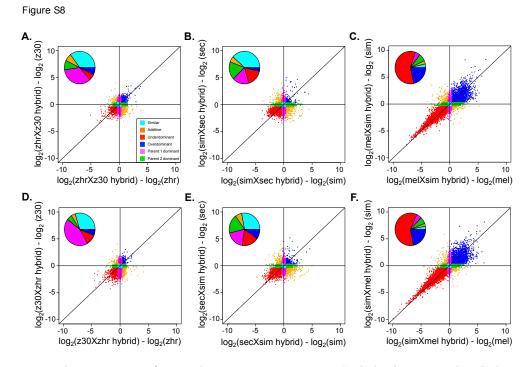


Figure 4.13: The proportion of genes showing misexpression in F1 hybrids increased with divergence time of the parental genotypes. Differences in total expression between the F1 hybrid and each parental species are compared in each panel. The difference in expression level is plotted for each gene as $\log_2(\text{reads from hybrid}) - \log_2(\text{reads from parental genotype})$, with the difference from parent 1 shown on the X-axis and the difference from parent 2 shown on the Yaxis. Hybrid genotypes are written as maternal genotype x paternal genotype. The Y=X line shown indicates an equal difference between expression in the F1 hybrid and both parents. Each gene was categorized as showing either conserved (light blue), additive (orange), underdominant (red), overdominant (blue), or dominant with expression similar to parent 1 (purple) or parent 2 (green), as described in the Methods. The pie-chart insets show the proportion of genes in each class. Interestingly, in the melmel comparison (A, D), dominant expression patterns resembled the North American zhr strain (parent 1) more than twice as often as they resembled the African z30 strain (parent 2). In the sim-sec and mel-sim F1 hybrids, dominant regulatory alleles were distributed more evenly between the two parental genotypes (B,E and C,F, respectively).

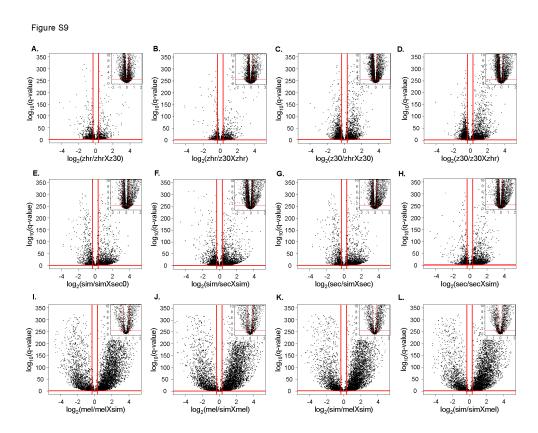


Figure 4.14: The frequency of large expression differences between F1 hybrids and parental species increases with divergence time. Volcano plots are shown for the comparison of expression levels between F1 hybrids and each parent in the mel-mel (A-D), sim-sech (E-H), and mel-sim (I-L) comparisons. Statistical significance, represented by $\log_{10}(q$ -value), is plotted on the Y-axis and the expression difference, plotted as $\log_2(\text{reads from parental genotype/reads from in hybrid genotype)}$, is plotted on the X-axis. Hybrid genotypes are written as maternal genotype x paternal genotype. The vertical red lines correspond to a 1.25-fold expression difference, whereas the horizontal red lines correspond to q-values with a false discovery rate of 0.05. Insets show genes with the smallest q-values in more detail.

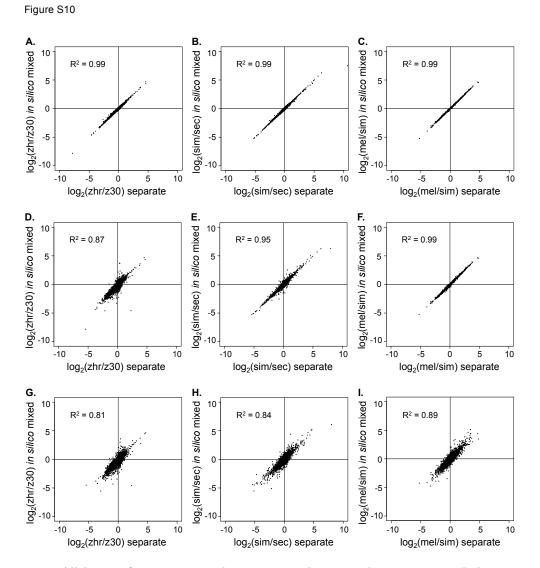


Figure 4.15: Allele-specific sequence reads are accurately assigned to genotypes. Relative expression levels inferred from in silico mixed parental samples after computational assignment of reads to specific alleles (Y-axis) were compared to relative expression levels determined using the separately sequenced samples (X-axis) for the mel-mel (A, D, G), sim-sech (B, E, H), and mel-sim (C, F, I) comparisons. In panels A-C, both allele-specific and shared reads from the mixed parental samples were included. In panels D-F, only allele-specific reads from the mixed parental samples were included. In panels G-I, the allele-specific read counts in Dataset 2 (i.e., after using hypergeometric sampling to equalize power among comparisons) were used. Relative expression is plotted as $\log_2(\text{reads from genotype 1/reads from genotype 2})$ in all cases. Coefficients of determination (\mathbb{R}^2) from linear models were used to compare relative expression between true values and those determined for in silico mixed samples.

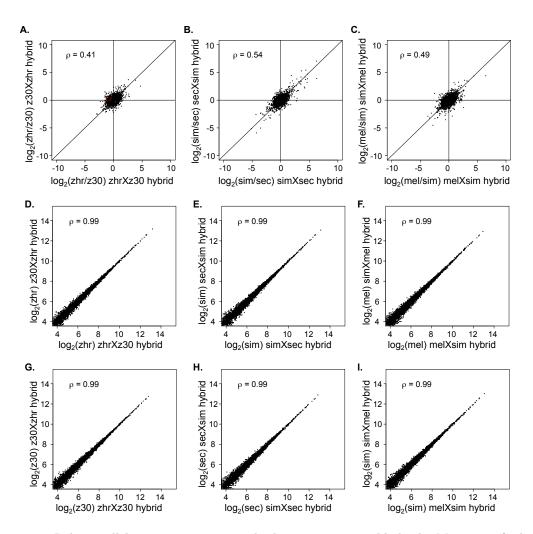


Figure 4.16: Relative allelic expression was similar between reciprocal hybrids. Measures of relative allelic expression were compared for each gene between reciprocal hybrids for the mel-mel (A), simsech (B), and mel-sim (C) comparisons. Coefficients of determination (\mathbb{R}^2) from linear models were used to compare relative allelic expression between reciprocal hybrids. The nine genes identified as having a statistically significant difference in relative allelic expression between reciprocal hybrids are shown in red. log₂(reads from allele 1/reads from allele 2) is plotted for each hybrid genotype. D-I, Allelic expression levels, plotted as log₂(allele-specific read counts) for each gene, are compared for each allele in reciprocal hybrids for all three comparisons. Spearmans ρ , which makes no assumptions about normality, was used to compare overall expression differences in each case. Genotypes of reciprocal hybrids are written as maternal genotype x paternal genotype.

Figure S11

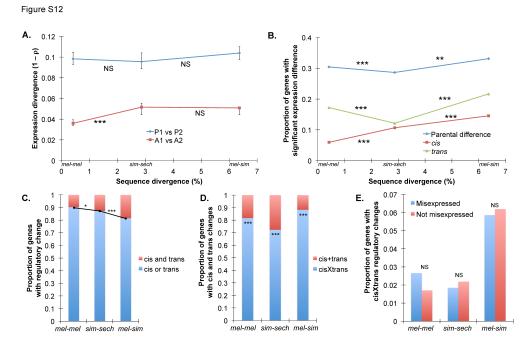


Figure 4.17: Evolution of *cis*- and *trans*-regulation. A, Overall differences $(1 - \rho)$ in total expression between genotypes (blue) and allele-specific expression in F1 hybrids (red) are shown for each comparison. Relative allelic expression provides a readout of relative *cis*-regulatory activity. Data used to calculate Spearmans ρ are summarized in Figure 4.18. B, For each comparison, the proportion of genes with evidence of significant differences in total expression (blue), *cis*-regulation (red), and *trans*-regulation (green) are shown. Data used to determine these proportions is shown in Figure 4.21. C, The proportion of genes with evidence of significant *cis*- and *trans*-regulatory changes (red) is compared to the proportion of genes with evidence of *cis*- or —it trans-regulatory changes (blue). D, For genes with evidence of both *cis*- and *trans*-regulatory changes, the frequency of genes with *cis*- and *trans*-regulatory changes affecting gene expression in the same (cis+trans, red) and opposite (cisXtrans, blue) directions are compared. E, The relative frequencies of genes with *cis*and *trans*-regulatory changes in opposite directions that do (blue) and do not (red) show evidence of misexpression in F1 hybrids are compared. Error bars in panel A show the 95% quantiles from 10,000 bootstrap replicates. Comparable analyses for reciprocal hybrids are shown in Figure 4.3.

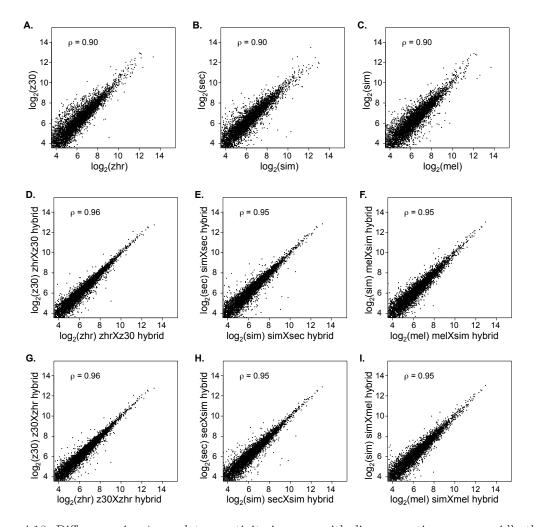


Figure 4.18: Differences in *cis*-regulatory activity increase with divergence time more rapidly than differences in total expression. For each of the 4851 genes in Dataset 2, total expression levels, plotted as $\log_2(\text{read count})$, are compared between the zhr and z30 strains of *D. melanogaster* (A), between *D. simulans* and *D. sechellia* (B), and between *D. melanogaster* (zhr) and *D. simulans* (C). Levels of allele-specific expression, plotted as $\log_2(\text{allele-specific read count})$, are compared between allele 1 (X-axis) and allele 2 (Y-axis) for each F1 hybrid from the mel-mel (D,G), sim-sech (E,H) and mel-sim (F,I) comparisons. Hybrid genotypes are written as maternal genotype x paternal genotype. Spearmans ρ , which makes no assumptions about normality, was used to measure the overall expression similarity between genotypes in each case.

Figure S13



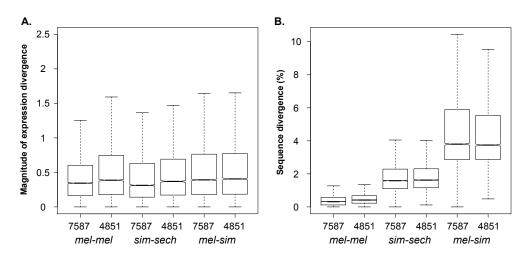


Figure 4.19: Differences between gene sets used to analyze total and allele-specific expression data. Box-plots summarize the distributions of total expression differences (A) and sequence divergence (B) for the 7587 genes in Dataset 1 and the 4851 genes in Dataset 2 for the mel-mel (MM), sim-sech (SS), and mel-sim (MS) comparisons. The notched box-plots show the full range of values as well as the 25, 50, and 75th percentiles.

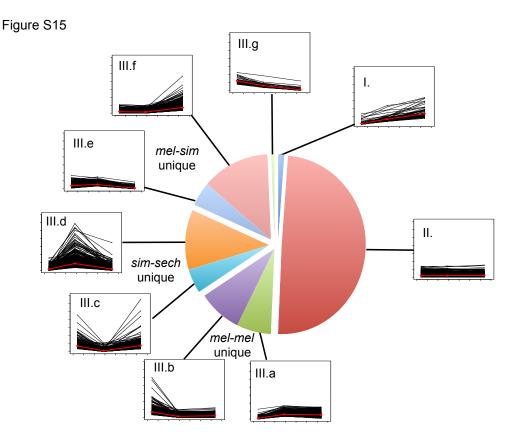


Figure 4.20: Evolutionary trajectories for expression divergence of individual genes. The line-plots show differences in total expression (A) and *cis*-regulatory activity (B) derived from Dataset 2 for individual genes in the mel-mel, sim-sech, and mel-sim comparisons plotted according to divergence time. As described in the main text, genes were classified into nine groups depending on whether they showed increased, decreased, or similar allele-specific expression differences between mel-mel and sim-sech and between sim-sech and mel-sim. The red line in each plot shows the median difference in *cis*-regulatory activity for genes in that class for each comparison. The pie-chart shows the relative frequency of genes in each class.



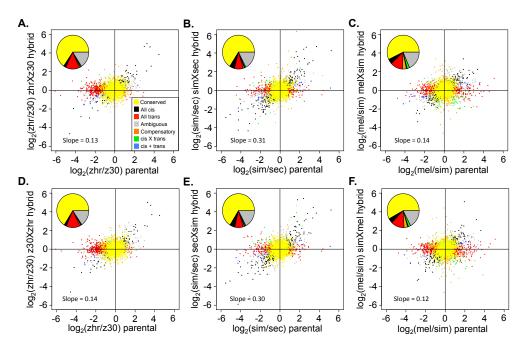


Figure 4.21: The relative contributions of *cis*- and *trans*-regulatory changes to expression divergence change with divergence time. For each gene, relative expression between parental genotypes, plotted as $\log_2(\text{reads from parent 1/ reads from parent 2})$ on the X-axis, is compared to relative allele-specific expression in F1 hybrids, plotted as $\log_2(\text{reads from allele 1/ reads from allele 2})$ on the Y-axis. Hybrid genotypes are written as maternal genotype x paternal genotype. Each gene was categorized as showing either conserved *cis*- and *trans*-regulation (yellow, conserved), only *cis*-regulatory differences (black, all cis), only *trans*-regulatory differences (red, all trans), cis-and *trans*-regulatory differences with no expression difference between parental genotypes (orange, compensatory), or *cis*- and *trans*-regulatory differences with no expression differences with effects on expression in the same (blue, cis + trans) or opposite (green, cis X trans) directions, as described in the Methods and Table S5. Genes with results from significance tests inconsistent with any of these categories (see Methods) were classified as ambiguous (grey). The pie-chart insets show the proportion of genes in each class, for each comparison. The slopes reported in each panel are from a linear regression model fitted to the corresponding dataset. A larger slope indicates a larger contribution of *cis*-regulatory divergence to total expression divergence.

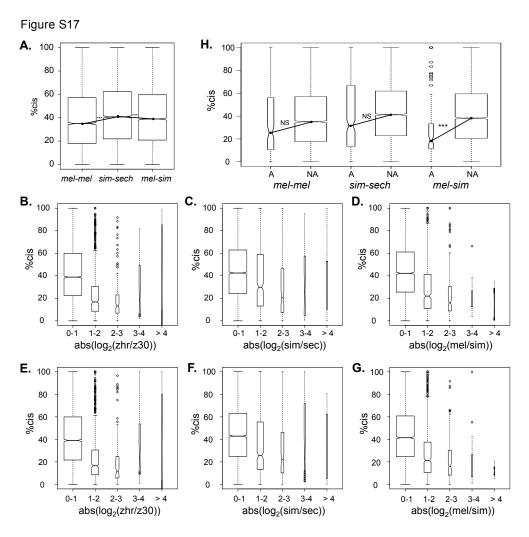


Figure 4.22: Effects of *cis*-regulatory divergence. A, The percentage of total regulatory divergence attributable to *cis*-regulatory divergence (%*cis*) is shown for the mel-mel, sim-sech, and mel-sim comparisons. B-D, %*cis* is compared among sets of genes with different levels of total expression differences, reported as the absolute value of the $\log_2(\text{reads from genotype 1}/\text{ reads from genotype 2})$ ratio, for the mel-mel (B), sim-sech (C), and mel-sim (D), comparisons. E, %*cis* is compared for sets of genes showing additive (A) and non-additive (NA, dominant or misexpression) inheritance for each comparison. In all panels, notched box-plots show the full range of values as well as the 25, 50, and 75th percentiles, with the width of the box-plots proportion to the number of genes in each class. Analyses comparable to panels A and H using reciprocal hybrids are shown in Figure 4.4.

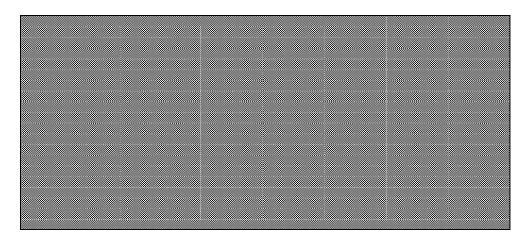


Figure 4.23: Summary of sequencing depth for RNA-seq and gDNA.

Table S2. Accuracy of mapping maternally inherited mitochondrial alleles in interspecific F1 hybrids.							
Hybrid cross	# correct	Total #	% correct				
D. simulans Tsimbazaza X D. sechellia droSec1	6141	6173	99.48				
D. sechellia droSec1 X D. simulans Tsimbazaza	11776	11837	99.48				
D. melanogaster zhr X D. simulans Tsimbazaza	41367	41414	99.89				
D. simulans Tsimbazaza X D. melanogaster zhr	21561	21599	99.82				

Figure 4.24: Number of genes suitable for quantifying total expression in each genotype.

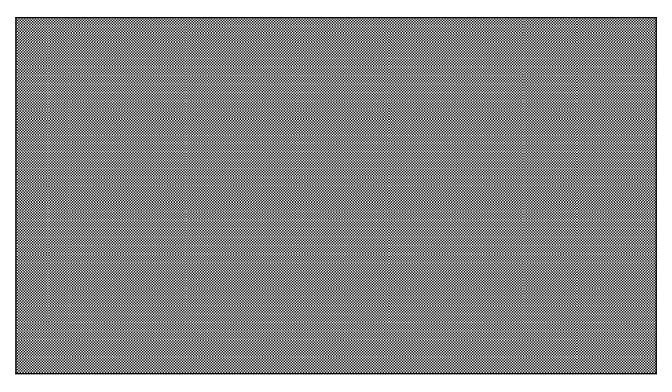


Figure 4.25: Number of genes suitable for quantifying allele-specific expression in each genotype.

Table S4. Number of genes suitable for quan	citying total expression	in each genotype.
Dataset	# genes > 20 reads	
Parental genotypes		
D. melanogaster zhr ¹	8928	
D. melanogaster zhr ²	8924	
D. melanogaster z30	9239	
D. simulans Tsimbazaza ³	9262	
D. simulans Tsimbazaza ⁴	9254	
D. sechellia droSec1	8981	
F ₁ hybrids		
D. melanogaster zhr X D. melanogaster z30	9223	
D. melanogaster z30 X D. melanogaster zhr	9050	
D. simulans Tsimbazaza X D. sechellia droSec1	8541	
D. sechellia droSec1 X D. simulans Tsimbazaza	8546	
D. melanogaster zhr X D. simulans Tsimbazaza	9314	
D. simulans Tsimbazaza X D. melanogaster zhr	9239	
Mixed parental samples		
D. melanogaster zhr and D. melanogaster z30	8700	
D. simulans Tsimbazaza and D. sechellia droSec1	8703	
D. melanogaster zhr and D. simulans Tsimbazaza	8594	
All samples	7660	
¹ when compared to <i>z30</i>		
² when compared to <i>D. simulans</i>		
when compared to <i>D. sechellia</i>		
when compared to <i>zhr</i>		

Figure 4.26: Accuracy of mapping maternally inherited mitochondrial alleles in interspecific F1 hybrids

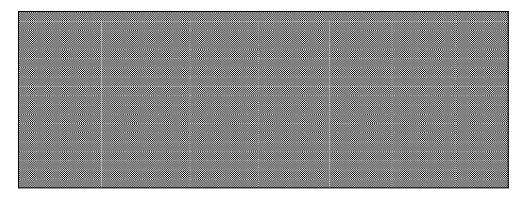


Figure 4.27: Criteria for assigning genes to regulatory evolution classes.

Classification	Fisher exact test ¹	Binomial exact test ²	Fisher exact test	Additional criteria	
conserved	P1 = P2	A1 = A2	P1/P2 = A1/A2	N/A	
all cis	$P1 \neq P2$	$A1 \neq A2$	P1/P2 = A1/A2	N/A	
all trans	$P1 \neq P2$	A1 = A2	$P1/P2 \neq A1/A2$	N/A	
cis+trans	$P1 \neq P2$	$A1 \neq A2$	$P1/P2 \neq A1/A2$	log2(P1/P2)/log2(A1/A2) > 1	
cisXtrans	$P1 \neq P2$	$A1 \neq A2$	$P1/P2 \neq A1/A2$	log2(P1/P2)/log2(A1/A2) < 1	
compensatory	P1 = P2	$A1 \neq A2$	$P1/P2 \neq A1/A2$	N/A	
ambiguous	$P1 \neq P2$	A1 = A2	P1/P2 = A1/A2	N/A	
ambiguous	P1 = P2	$A1 \neq A2$	P1/P2 = A1/A2	N/A	
ambiguous	P1 = P2	A1 = A2	$P1/P2 \neq A1/A2$	N/A	

Figure 4.28: Pyrosequencing assays for quantification of allelic expression ratios.

CHAPTER V

Sex- and tissue-specific differences in gene regulation between Drosophila pseudoobscura and its closely-related subspecies D. p. bogotana

5.1 Abstract

Gene expression is regulated by a complex interplay between cis-regulatory sequences and trans-acting factors. This process not only determines overall levels of gene expression, but is also responsible for governing gene expression patterns that result in tissue differentiation. To date, several studies in *Drosophila* have shown various properties of regulatory divergence both within and between species, but much of this work is restricted to whole fly tissues. Furthermore, most of these studies have used females so that cis-regulatory divergence could be assessed on the X-chromosome. Here we measure total and allele-specific gene expression between *D. pseudoobscura* and its closely-related subspecies *D. p. bogotana* for female and male carcass and gonad tissues. Studying regulatory divergence in female and male carcass and gonad tissues is revealing not only for sex-specific differences but also for tissue-specific differences. Because of their incipient speciation, these differences are particularly relevant for gonad tissues whose gene expression patterns may drive

Manuscript:

Stevenson, K.R., Nyberg, K.G., Coolon, J.D., Machado, C.A., Wittkopp, P.J. Sex- and tissue-specific differences in gene regulation between *Drosophila pseudoobscura* and its closely-related subspecies *D. p. bogotana, in prep*

hybrid incompatibilities. Overall we find extensive differences in inheritance patterns and regulatory divergence between sexes and tissues. We also find that gene expression diverges more quickly for sex-biased genes, although the strength of this divergence is highly dependent on the tissue type. This work illustrates how measures of regulatory divergence using whole flies can mask complexity in individual tissues.

5.2 Introduction

Understanding differences in gene expression and the regulatory mechanisms governing them is essential in studying the emergence of phenotypic variation both within and between species. Gene expression at the level of transcription is regulated by *trans*-acting factors and the *cis*-regulatory sequences to which they bind. *cis*and *trans*-acting changes affecting gene expression can be elucidated by comparing gene expression in parental samples and their F1 hybrids in which *trans*-regulatory variation is controlled by the presence of both parental alleles (Cowles et al., 2002). Differences in allele-specific expression in F1 hybrids indicate *cis*-regulation of a gene, whereas differences in gene expression between parental samples not present in F1 hybrids indicate *trans*-regulation of a gene (Wittkopp et al., 2004). Recently, many studies have expanded these inferences to include all expressed genes using a technique known as RNA-seq, which has allowed genome-wide patterns of regulatory divergence to be identified (Wilhelm et al., 2008; Graze et al., 2009; McManus et al., 2010; Coolon et al., 2012; Graze et al., 2012; Coolon et al., 2014).

Differences in the regulation of gene expression are responsible not only for phenotypic diversity between species but also for the differentiation of tissue types within an organism. Although inheritance patterns of gene expression and regulatory divergence have been studied extensively in *Drosophila*, much of this work measured gene expression from whole fly samples. While gene expression divergence has been studied in different tissues in *Drosophila* (Assis et al., 2012), the regulation of gene expression across different tissues remains poorly understood. Additionally, female samples have been the primary focus of this work because the presence of both parental copies of the X-chromosome allows regulatory divergence to be assessed for X-linked genes. Thus our understanding of how genes are regulated differentially in females and males remains is lacking.

As one of the classical species of *Drosophila* studied by Theodosius Dobzhansky, D. pseudoobscura is an ideal model organism in which to study gene expression (Dobzhansky, 1936). Dobzhansky also discovered a subspecies of *D. pseudoobscura* in the highlands of Bogota, Columbia, which he named D. p. boqotana. Because of its distal and very high location, it is geographically isolated from the nearest population of *D. pseudoobscura* (Ayala and Dobzhansky, 1974). These species diverged 80,000-230,000 years ago (Schaeffer and Miller, 1991; Jenkins et al., 1996; Wang et al., 1997; Alvarez et al., 2002), which represents an intermediate divergence time point to two pairs previously reported in a study of regulatory divergence within and between species of *Drosophila* (Coolon et al., 2014). Despite this divergence and little or no gene flow between these species (Powell, 1983; Wang and Hey, 1996; Wang et al., 1997; Machado et al., 2002; Machado and Hey, 2003), they can interbreed and produce viable F1 hybrids, although D. p. boqotana females crossed with D. pseudoobscura males produces sterile F1 hybrid males (Prakash, 1972). This hybrid incompatibility, when taken together with almost no evidence of gene flow and geographic isolation from the most closely related species D. pseudoobscura, suggests that these species may be in the process of speciation.

In the work presented here, we expand on previous studies by inferring both the mode of gene expression inheritance as well as gene regulatory mechanisms between D. pseudoobscura and the closely-related subspecies D. p. boqotana. This inference is drawn not only for females and males of these species, but also for sex-specific carcass and gonad samples, allowing us to test sex- and tissue-specific differences in gene expression variation. We find an enrichment of *trans*-regulation in carcass samples compared to gonads, which is consistent with previous work in comparable whole fly tissues (McManus et al., 2010; Suvorov et al., 2013; Coolon et al., 2014). We also measured total and *cis*-regulatory gene expression divergence to determine whether or not expression of X-linked genes is diverging faster than autosomal genes, known as the faster-X effect (Charlesworth et al., 1987). While we found no evidence of this, the neo-X chromosome in male carcass samples showed lower than expected gene expression divergence, and sex-biased genes showed elevated levels of gene expression divergence, consistent with previously reported results from microarray and RNA-seq experiments (Jiang and Wong, 2009; Assis et al., 2012). These data show important differences between sexes and tissue types, and illustrate the extent to which whole body experiments can mask the complexity of gene expression variation present in different tissues.

5.3 Results and Discussion

5.3.1 Measuring *Drosophila pseudoobscura* and *D. p. bogotana* sex- and tissue-specific gene expression

To study the effects that both different sexes and tissue types have on inferences of regulatory divergence, we sequenced poly(A)-selected RNA from female and male carcass and gonad tissues for parental species *D. pseudoobscura* (TL) and *D. p. bogotana* (Toro1) and their F1 hybrids resulting from crossing TL females×Toro1 males (H6) (see Materials and Methods). We also sequenced genomic DNA from the parental TL and Toro1 strains, identified single nucleotide variants and insertions/deletions (indels) relative to the reference genome of *D. pseudoobscura* (Fly-Base), and incorporated them into this genome to create TL- and Toro1-specific genomes (Figure 5.1; see Materials and Methods).

Annotated exons from the sequenced strain of *D. pseudoobscura* were then extracted from the TL- and Toro1-specific genomes and aligned, resulting in 137,784 coding-sequence differences between these species, or roughly 0.6% coding sequence divergence. This level of sequence divergence is intermediate to within and between species data previously reported, making it possible to distinguish between the TL and Toro1 alleles (Coolon et al., 2014). gDNA realigned to these species-specific exomes showed high levels of mapping to the correct species (Figure 5.7), and levels of sequence divergence were also consistent between chromosomes (Figure 5.8A). These exomes were then used to quantify total and allele-specific gene expression from the RNA-seq data using methods previously outlined (Coolon et al., 2012, 2014) as well as those described in the Materials and Methods.

5.3.2 Inheritance patterns for gene expression are most similar within male tissues but differ significantly between sexes and tissue types

To assess the level to which the inheritance of gene expression differed between sexes and tissue types, estimates of total expression from parents and F1 hybrids were used to categorize genes into five mode of inheritance classes: similar (consistent expression between parents and F1 hybrids), dominant (F1 hybrid expression more closely resembling TL or Toro1 expression), additive (F1 hybrid expression intermediate to both parental levels of expression), overdominant (F1 hybrid expression greater than both parental levels of expression), and underdominant (F1 hybrid expression less than both parental levels of expression) (Figure 5.2). We used 1.25fold differences in total expression as a threshold to determine these classes (Gibson et al., 2004) (see Materials and Methods).

To determine the degree of similarity between this categorization and sex-by-tissue type, Crámers V was calculated comparing female and male carcass (Figure 5.2A,B), female carcass and ovaries (Figure 5.2A,C), male carcass and testes (Figure 5.2B,D), and ovaries and testes (Figure 5.2C,D). Bound between [0,1], high levels of Crámers V indicate greater differences between samples. While *G*-tests rejected the indistinguishability of these relationships at a high level of statistical significance, likely due to large sample sizes in the contingency tables (p-value <2.2e-16), comparing female carcass to ovaries showed the most striking differences (V = 0.62), while male carcass and testes were the most similar (V= 0.10). Between-sex comparisons for carcass and gonad tissues showed intermediate levels of similarity, reflecting sex-specific differences in gene expression inheritance (V = 0.37 and 0.31 for carcass and gonads, respectively). These results show that inferences of the patterns of inheritance for gene expression are moderately different in sex-specific carcass and gonad tissues, least different between male carcass and testes, and most different between female carcass and ovaries.

5.3.3 Carcass tissues are enriched for genes regulated only in *trans*, while gonads are enriched for genes regulated only in *cis*

Differences in gene expression are caused by *cis*- or *trans*-acting changes, or a combination of both. To determine the extent that differences in gene regulation can be explained by sex and tissue type, estimates of allele-specific expression were used to categorize each gene into five types of regulatory change impacting their expression: *cis*, *trans*, *cis&trans*, conserved, and ambiguous (Figure 5.3). These

differences were determined by a series of hierarchical binomial and Fishers exact tests as previously outlined (McManus et al., 2010; Coolon et al., 2012, 2014) and described in Materials and Methods.

As was the case for the mode of inheritance, the strength of the association between sex or tissue type and the different classes of regulatory divergence was determined using Crámers V. Unlike mode of inheritance, regulatory divergence categorization seems to be more similar between sexes (V = 0.12 and 0.15 when comparing sexes across carcass and gonads, respectively). However, within each sex, the type of tissue had a larger effect on the regulatory divergence classification (V = 0.33 and 0.28 when comparing tissues across females and males, respectively). Like mode of inheritance, the indistinguishability of these tested relationships was rejected at a high level of statistical significance, indicating the clearest differences between tissue types (*G*-test; p-values <2.2e-16).

Interestingly, the proportion of genes experiencing *trans*-acting changes at the level of gene expression were enriched in female and male carcass samples, which could in part be attributable to the presence of morphological differences between tissues of the parents that cannot be observed in F1 hybrids (McManus et al 2010). This might also explain the relative depletion of such genes in the gonad samples, where regulatory divergence is dominated by *cis*-acting changes. This is also consistent with higher levels of sequence divergence in gonad-specific genes relative to other genes (Figure 5.8B; (Assis et al., 2012)).

Genes with additive gene expression inheritance have been previously shown to have increased percentages of total regulatory divergence explained by *cis*-acting changes (%*cis*; (Lemos et al., 2008; McManus et al., 2010)), although this pattern does not always seem to hold (Coolon et al., 2014). This relationship arises because of the assumption that each parental allele is independently expressed, contributing equally to total levels of expression in F1 hybrids. We observed this expected relationship in gonad tissues (Wilcoxon rank sum test: p-value = 1.11e-13 and 5.68e-23 for ovaries and testes, respectively) (Figure 5.4C,D) but not in carcass tissues (Figure 5.4A,B), although the distributions of %*cis* among additive and non-additive gene expression inheritance were significantly different in female but not male carcass tissues (Wilcoxon rank sum test: p-value = 1.85e-34 and 0.09, respectively).

5.3.4 No faster-X effect, but sex-biased genes show increased levels of gene expression divergence

To determine whether or not the expression of X-linked genes evolves more quickly than autosomal genes (faster-X effect) in a sex- and/or sex×tissue-specific manner, total expression divergence was calculated between both alleles in parents for female and male carcass, ovaries, and testes (Figure 5.9; see Materials and Methods). Overall, there appears to be no faster-X effect for total expression divergence, which is consistent with findings from Meisel *et al.*, who observed this phenomenon in head samples but not when measuring gene expression in the whole fly (Meisel et al., 2012). Because the *D. pseudoobscura* lineage has acquired a neo-X chromosome that is ancestrally an autosome, we measured expression divergence on this chromosome but found no significant faster-X effect, which is also consistent with previously reported findings (Meisel et al., 2012). On average, male samples for both tissues showed higher levels of total expression divergence, which is consistent with previous work (Jiang and Wong, 2009), while neo-X-linked genes in male carcass showed less divergence than expected. This slower-X effect disappears in testes, but we cannot rule out this slower-X effect in other tissues of the carcass, and because it does not occur in gonads, it may not be biologically relevant. Because we measured allelespecific expression in female carcass and ovaries of F1 hybrids, we can also look at the level of expression divergence being driven by *cis*-regulation for X-linked genes in female samples (Figure 5.10). While showing overall less expression divergence, there appears to be no significant expression divergence in *cis* for female carcass and ovary samples.

To test for differences in gene expression divergence between sex-biased genes, we separated genes showing at least a two-fold difference between sexes in both carcass and gonad tissues (see Materials and Methods). Sex-biased genes in carcass samples showed almost no faster-X effect compared to non-biased genes (Figure 5.5). However, sex-biased genes did show greater total expression divergence in female carcass samples compared to those in males (Figure 5.5A,B). While ovaries showed no differences in expression divergence between sex-biased and non sex-biased genes, testes showed greater total expression divergence for female-biased genes compared to male-biased and non sex-biased genes (Figure 5.5C,D). This is consistent with previous work showing that sex-biased genes have elevated levels of expression divergence in the opposite sex (Jiang and Machado 2009). As above, we also measured cis-regulatory expression divergence to determine its contribution to total expression divergence (Figure 5.6). Specifically, for the pattern of increased divergence in female-biased genes in testes, it appears that this signal is not being driven by cis-regulation (Figure 5.6B). In all cases, significant differences in expression divergence between sex-biased and non sex-biased genes were determined based on non-overlapping intervals defined by bootstrapped 2.5^{th} and 97.5^{th} percentiles (see Materials and Methods).

5.4 Conclusions

In summary, this work shows that patterns of inheritance for gene expression as well as regulatory divergence differ between sexes and across tissues. Interestingly, we observed the most striking differences in female carcass samples, which are similar to female whole body samples except lacking reproductive tissue. Female whole flies are among the most popular sex and tissue types in gene expression studies of *Drosophila*, largely due to it being an easy tissue to collect and the fact that females have both allelic copies of X-linked genes, which allows for their regulatory divergence to be classified.

We found an enrichment of *trans*-regulation in carcass compared to gonad tissues between these species. This observation has previously been reported in a study of gene regulation between species using whole fly tissues (McManus et al., 2010). Although the species studied here are outwardly morphologically indistinguishable, this signal is likely caused by differences in tissue sizes or the incomplete removal of gonad tissues from the carcass. More studies are needed that measure total and allelespecific gene expression in additional tissues such as heads, accessory glands, and fat bodies so their individual contributions to regulatory divergence can be ascertained.

We did not observe strong differences in gene expression divergence between Xlinked and autosomal genes, as has been previously reported but for more divergent species (Meisel et al., 2012). We did however find increased gene expression divergence for sex-biased genes, which showed the expression of female-biased genes to be more divergent than male- and non-biased genes in testes.

Additionally, these species could be used to determine the relative contribution of *cis*- and *trans*-acting changes to the process of speciation, which is not well understood. These species are ideal to study the speciation process due to their lack of gene flow and postzygotic reproductive isolation, and further work may help to reveal the gene regulatory basis of their hybrid incompatibility. Interestingly, we found 75 genes expressed in testes that showed misexpression consistent with genomic imprinting between F1 hybrids of reciprocal crosses, although, of the 51 that had clear *D. melanogaster* orthologs, 22 were annotated on chromosomes X and 3L in D. melanoqueter, which are the X and neo-X chromosomes in D. pseudoobscura, respectively. Such genes would necessarily show patterns of genomic imprinting because of hemizygosity of X-linked genes in males. Also, because D. melanoque are not known to imprint their genomes (Coolon et al., 2012), the remaining 29 genes are difficult to interpret and show no relevant molecular functions or biological processes involved in reproduction. This set also did not include the known hybrid sterility locus Overdrive on the neo-X chromosome, which is necessary, but not sufficient, for F1 hybrid male sterility and segregation distortion (Phadnis and Orr, 2009). We could not assess the likelihood of this gene being imprinted because it is hemizygous in males, and it was not expressed in female samples. Total expression of Overdrive in testes was consistent with both *cis*- and *trans*-acting changes, as well as dominant gene expression inheritance favoring the D. p. bogotana maternal allele in the H5 hybrid direction, which produces sterile males. More work is needed to discern the gene regulatory basis of hybrid incompatibilities involved in speciation.

5.5 Materials and Methods

5.5.1 Fly strains, rearing, and collections

Reciprocal hybrid crosses were set up between inbred lines of *Drosophila pseu*doobscura (TL) and *D. p. bogotana* (Toro1). TL was inbred five generations from stock #14011-0121.38 with the TreeLine inversion from the Drosophila Species Stock Center at the University of California, San Diego. Inbreeding of Toro1 was previously described (Machado et al. 2002). In all analyses, H5 refers to hybrid progeny of *D. pseudoobscura* males and *D. p. bogotana* females; H6 refers to hybrid progeny of *D. pseudoobscura* females and *D. p. bogotana* males. Parental and hybrid crosses were kept in incubators at 20°Con a cornmeal-molasses-yeast medium. Ten virgin individuals for each line (TL, Toro1, H5, H6) were aged seven days, and ovaries and testes (without male accessory glands) were isolated from carcasses in 1×PBS and frozen in liquid nitrogen.Frozen tissue was then ground with a pestle, and total RNA was extracted using standard Trizol protocols (Life Tech #15596-026) and cleaned via ethanol precipitation.

5.5.2 Library preparation and Illumina sequencing

Total RNA was quantified fluorometrically using a Bio-Rad Experion RNA Std-Sens kit (Bio-Rad #700-7103) and subsequently used (200ng/sample) to construct poly(A) RNA-Seq libraries using the TruSeq RNA Sample Prep Kit v2 (Illumina #RS-122-2001). Libraries were multiplexed four per lane on an Illumina HiSeq1000 to generate 101-base paired-end reads. Sequencing was performed at the University of Marylands Institute for Bioscience and Biotechnology Research Sequencing Core.

5.5.3 Building parental exomes from D. pseudoobscura and D. p. bogotana gDNA

To determine homozygous genotypes in the TL (D. pseudoobscura) and Torol (D. p. bogotana) inbred lines, gDNA from each line was independently aligned to the latest build of the D. pseudoobscura reference genome (dpse-all-chromosome-r3.1; FlyBase) using Bowtie 2 ((Langmead and Salzberg, 2012); see Supplemental Methods). Even though the gDNA sequencing effort of Torol represented paired-end reads, mates were concatenated to represent a list of single-end reads to match

the sequencing effort of TL gDNA. Reads that failed to align were trimmed in three iterations, removing 13 bases from the 3 end in an effort to rescue error-prone reads or those with low-quality ends, and the alignment results from each trimming iterative were combined to form the set of all aligned reads. To avoid possible amplification bias, PCR duplicates were removed using SAMtools (Li et al., 2009).

After formatting the alignment files using Picard (http://picard.sourceforge.net/), we applied the HaplotypeCaller walker from the Genome Analysis Toolkit (GATK; (McKenna et al., 2010)) for base quality score recalibration and local realignment of insertions and deletions (indels) to detect SNPs and indels. Although it is recommended to recalibrate base quality scores based on a set of known polymorphic sites, such a database does not exist for *D. pseudoobscura*, so instead high-quality SNPs and indels were chosen using standard hard filtering parameters (see Supplemental Methods) on which base quality scores were recalibrated. This process was repeated for three iterations, upon which the mean and accuracy of quality scores converged and a final set of variants was called based on the GATK Best Practices recommendations (DePristo et al., 2011; Auwera et al., 2013).

After homozygous SNPs and indels were called for the TL and Toro1 inbred lines, their genomic coordinates were converted into coordinates corresponding to a list of annotated exons (FlyBase) and were incorporated into the *D. pseudoobscura* reference exome using custom Perl scripts, masking heterozygous sites (convertVCF.pl and vcf2fasta.pl, respectively). Each newly-made TL and Toro1 exon was then pairwise aligned using Fast Statistical Alignment (FSA; (Bradley et al., 2009)) to find sequence gaps between these lines, which were also masked. These alignments allowed for the identification of 137,784 coding sites that differentiate the TL and Toro1 alleles, or roughly 3 differentiating sites per 500 coding bases (137,784 / 23,988,953).

5.5.4 Quantifying total and allele-specific gene expression

To determine levels of total and allele-specific gene expression in each sample, both mates from the set of paired-end reads from each sample were independently aligned to the TL- and Toro1-specific exome using Bowtie ((Langmead et al., 2009); see Supplemental Methods). For mates that failed to align uniquely and/or had at least one mismatch against both exomes, 13 bases were trimmed off the 3 end of each mate, and this process was repeated four times (get_trim_reads.pl). Upon combining the separately aligned mates after the iterative trimming, each sample would have four sets of alignments: first mates that aligned to TL, first mates that aligned to Toro1, second mates that aligned to TL, and second mates that aligned to Toro1. Combining mates created 16 different outcomes for each read, which were parsed to assign each read to the TL or the Toro1 allele if a read aligned uniquely to one but not the other allele. In the event that each read aligned equally well to both alleles, a read was assigned to a separate category called both which, when summed with the allele-specific counts, represents the total level of expression. Mate pairs that spanned multiple exons were only counted once for overall levels of gene expression, as were mate pairs that aligned within a single exon. These classifications were performed using custom shell and Perl scripts (characterize.sh and classify.pl).

Using the same logic, individual parental gDNA samples were subjected to the same pipeline and levels of total and allelic abundance were estimated. As expected, nearly all genes in each individual sample showed very low counts of sequence reads mismapping to the other species (Figure 5.7). We generated a reliable set of genes from the set of 16,756 annotated genes (FlyBase) from which to measure expression by the following criteria: (1) at least 20 allele-specific reads mapped when combining parental gDNA samples (pruning the list down to 13,695 genes), and (2) at least 99%

of all mapped reads to a gene were to the correct species (further pruning the list to 11,829 genes). This also allowed us to control for imperfect mappability within and between these species exomes (Stevenson et al 2013).

5.5.5 Normalizing total and allele-specific read counts across samples

Because samples are sequenced to different depths and can have different numbers of total mapped reads, we downsampled all samples to that with the lowest number of mapped reads. In this case, it happened to be the male carcass tissue of F1 hybrids as a result of crossing Toro1 females to TL males (H5), where we successfully mapped 16,909,573 sequence reads. The reduction as a percentage of the total mapped reads ranged from 7.6% to 52% across all samples. To normalize each sample to the one with the fewest total mapped reads, a random draw from Fishers central multivariate hypergeometric distribution was taken from the number of reads matching the TL and Toro1 alleles, as well as those mapping equally well to both, for each of the 16,524 genes with annotated exons. This is analogous to drawing a subset of marbles from an urn, where each marble is one of 49,572 (16,524×3) different colors, and then counting the number of marbles matching each of these colors. This strategy was implemented using a modified version of the R package BiasedUrn (CRAN) and is similar to that used in Coolon et al. (2014).

After equalizing the total mapped read counts across all samples, because of differing expression levels between samples, there could be different levels of allelespecific expression for any particular gene across samples. To normalize this genespecific effect, TreeLine and Toro1 parents and each individual F1 hybrid sample made from reciprocal crosses were downsampled separately for each sex- and tissuespecific sample. For example, to look at allele-specific expression in the ovary after crossing TL females with Toro1 males, the minimum number of allele-specific reads was found for each gene across the TL and Toro1 ovary samples, as well as that of the H6 ovary. To normalize allele-specific counts within this comparison, a random draw from the hypergeometric distribution was taken from the minimum number of allele-specific reads from each gene across these three samples. This process was repeated for each sex- and tissue-specific comparison and is also similar to that used in Coolon *et al.* (Coolon *et al.*, 2014). In addition, the F1 hybrids from reciprocal crosses were also normalized in the same manner for female and male carcass and gonad tissues.

5.5.6 Mode of inheritance and regulatory divergence classification

To determine the mode by which each genes level of expression is inherited for each sample, the normalized total expression values for parents and their F1 hybrids were compared using the logic outlined in Gibson *et al.* (Gibson *et al.*, 2004). Genes were said to be expressed if the sum of the both parents and the hybrid levels of expression met or exceeded 20 sequence reads. The cutoff to claim expression differences between parents and hybrids was 1.25-fold. Genes with no expression differences when comparing each parent to their F1 hybrids were classified as similar. Genes whose F1 hybrid expression levels were similar to one parent but not the other were said to be dominant with respect to the similar parent. Genes whose F1 hybrid expression levels were different from each parent and intermediate between parents were classified as additive, while those with F1 hybrid expression levels greater than that in both parents or less than that in both parents were classified as over-dominant and under-dominant, respectively.

To characterize the manner in which each genes level of expression is regulated, the normalized allele-specific expression values for parents and their F1 hybrids were compared using a series of hierarchical exact tests. Genes were considered for this test if the sum of both alleles in parents and hybrids separately met or exceeded 20 sequence reads. Allele-specific differences between parents as well as between both alleles in F1 hybrids were determined using binomial exact tests, correcting for multiple comparisons by controlling the false discovery rate (FDR; (Benjamini and Hochberg, 1995)) to 5%. Fishers exact test was used to determine allelic differences in expression between parents and hybrids, also controlling the FDR to 5%. Because each parental allele experiences the same set of trans-acting factors in F1 hybrids, differences in allele-specific expression indicate cis-regulation, assuming that the parental alleles are also differentially expressed. This regulation occurs solely in cis when the parental and F1 hybrid allelic abundances are statistically similar. In contrast, equal allelic expression in F1 hybrids coupled with differences in expression of parental alleles indicate trans-regulation. This regulation occurs solely in trans when the parental and F1 hybrid allelic abundances are statistically different. These two types of gene regulation can also co-occur. Genes with conserved regulation occur when expression of the parental alleles and those in the F1 hybrid are all similar. Results of these tests with no clear biological interpretation of the significance tests are ambiguously regulated. G-tests and Crámers V were calculated using R 3.0.2 using the packages *Deducer* and *lsr*, respectively.

5.5.7 Determining sex-biased and sex tissue-specific genes

Genes that showed a two-fold or greater total expression (parent 1 + parent 2 + hybrid) difference between sexes were classified as sex-biased, and this was determined separately for carcass and gonad samples. Testis- and ovary-specific genes were classified as those not expressed in both testes and ovaries, with an additional filter removing genes in their respective sex-specific carcass tissues.

5.5.8 Total and *cis*-regulatory expression divergence

To determine the percentage of total regulatory divergence attributable to *cis*-regulatory changes (%*cis*), we calculated the relative contributions of both *cis*- and *trans*-regulation to total expression divergence. Total expression divergence was measured as $\log_2(\text{parental allele 1})/(\text{parental allele 2})$ for the parental samples, while *cis*-regulatory expression divergence was measured as $\log_2(\text{parental allele 1})/(\text{parental allele 2})$ for the parental allele 1)/(parental allele 2) in F1 hybrid samples. Because total expression divergence represents the sum of the respective *cis*- and *trans*-regulatory components, *trans*-regulatory expression divergence. It then follows that $\% cis = \frac{|cis|}{|cis|+|trans|} \times 100$.

To calculate total- and *cis*-regulatory expression divergence across all expressed genes within a sample, we used Spearmans ρ comparing alleles in the parental samples and F1 hybrids, respectively. Bootstrapped confidence intervals for estimates of ρ were obtained by sampling the same number of genes with replacement 10,000 times, calculating ρ each time and using the 2.5th and 97.5th percentiles from this distribution. As a representation of expression divergence, measures of $1 - \rho$ are reported.

5.6 Supplemental Methods

GATK hard filtering parameters:

(1) SNPs -filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || HaplotypeScore > 13.0 || MappingQualityRankSum < -12.5 || ReadPosRankSum < -8.0"
(2) indels -filterExpression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0" Bowtie2 parameters for gDNA-seq alignments:

bowtie 2 -f -p 12 –
very-sensitive

Bowtie parameters for RNA-seq alignmets:

bowtie -q -p 12 -v 0 -m 1

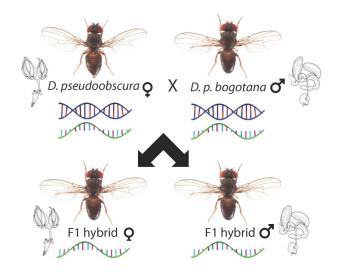


Figure 5.1: Experimental design. *D. pseudobscura* females were crossed to *D. p. bogotana* males, producing fertile female and male F1 hybrid offspring. gDNA was sequenced from parental samples for building species-specific genomes, and RNA was sequenced from female and male carcass and gonad tissues from parental samples as well as F1 hybrids. Photos of *D. pseudobscura* were taken from FlyBase (provided by Nicolas Gompel), and drawings of female and male reproductive tissue were also taken from FlyBase (Patterson 1943).

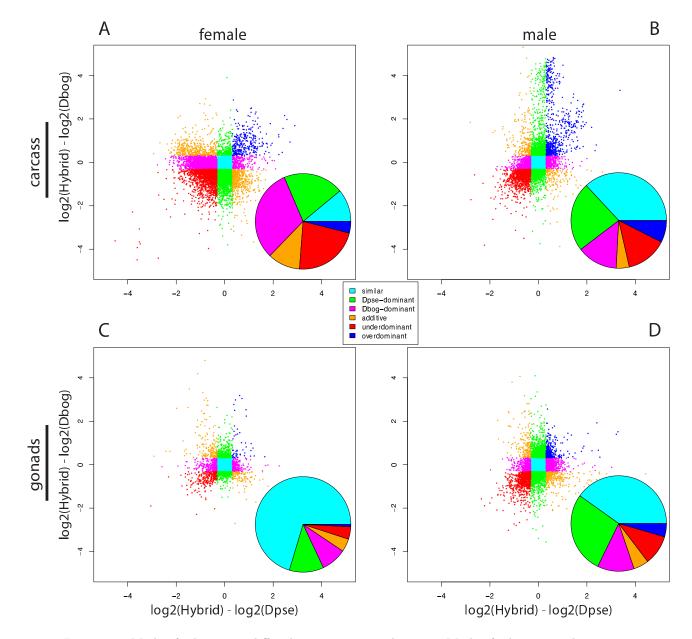


Figure 5.2: Mode of inheritance differs between sexes and tissues. Mode of inheritance classes were determined for female carcass (A), male carcass (B), ovaries (C), and testes (D). The inset pie charts show the different classes relative contribution to mode of inheritance.

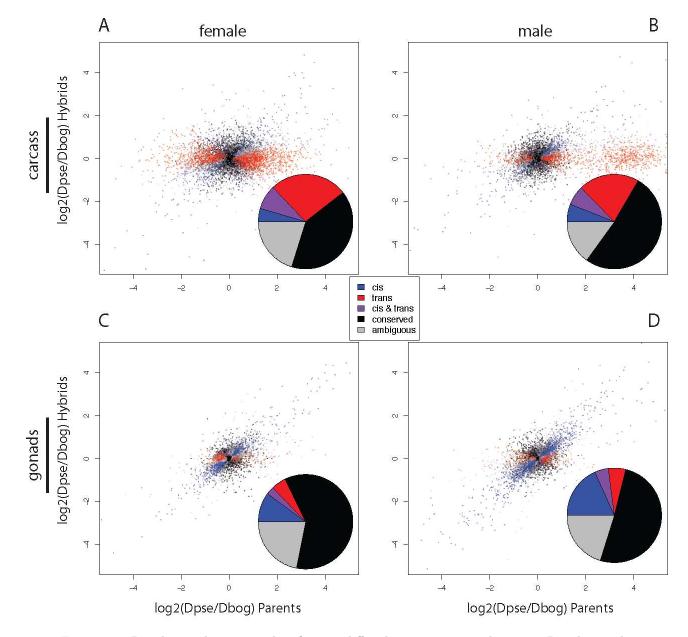


Figure 5.3: Regulatory divergence classification differs between sexes and tissues. Regulatory divergence classes were determined for female carcass (A), male carcass (B), ovaries (C), and testes (D). The inset pie charts show the different classes relative contribution to regulatory divergence.

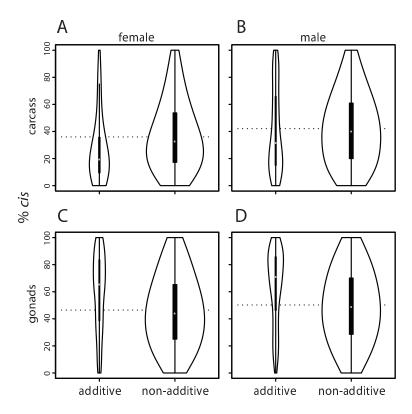


Figure 5.4: % *cis* for additively and non-additively inherited gene expression consistent with expectation in gonad but not carcass tissue. Violin plots showing distributions of % *cis* for genes with additive and non-additive gene expression inheritance for female carcass (A), male carcass (B), ovaries (C), and testes (D).

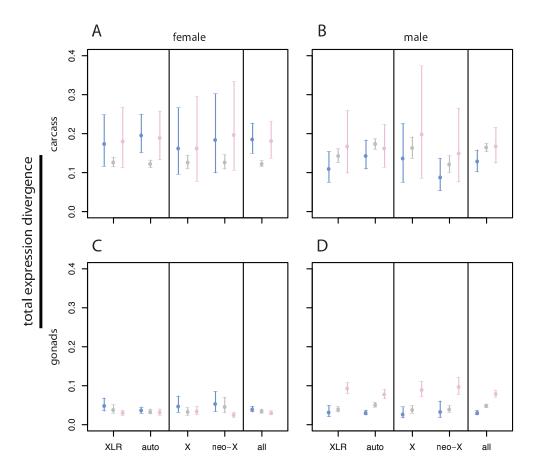


Figure 5.5: X-linked and autosomal female-biased genes show higher total expression divergence than male-biased and non biased genes in testes. Estimates of total expression divergence for female carcass (A), male carcass (B), ovaries (C), and testes (D). Male-biased (blue), female-biased (pink), and non-biased genes were determined using 2-fold differences in male vs. female total expression in respective tissues. Dots represent overall estimates of total expression divergence as measured by 1-Spearman's ρ , and whiskers represent bootstrapped 2.5th and 97.5th percentiles. Non-overlapping intervals indicate statistical significance at $\alpha = 5\%$.

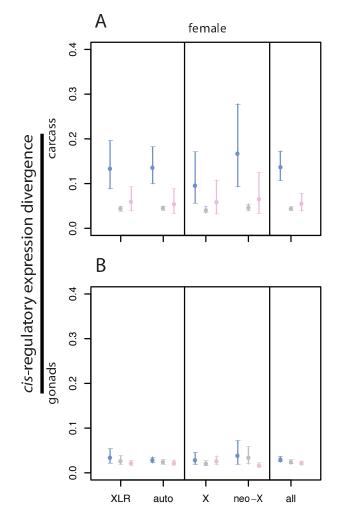


Figure 5.6: Male-biased autosomal genes show elevated gene expression divergence driven by *cis* in carcass samples but not gonads. Estimates of *cis*-regulatory expression divergence for female carcass (A) and ovaries (B). Male-biased (blue), female-biased (pink), and non-biased genes were determined using 2-fold differences in male vs. female total expression in respective tissues. Dots represent overall estimates of *cis*-regulatory expression divergence as measured by 1-Spearman's ρ , and whiskers represent bootstrapped 2.5th and 97.5th percentiles. Non-overlapping intervals indicate statistical significance at $\alpha = 5\%$.

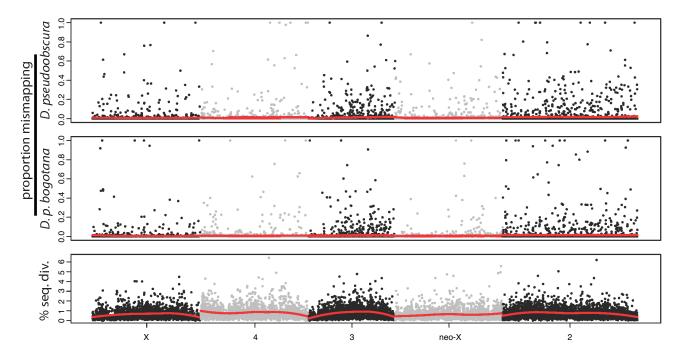


Figure 5.7: *D. pseudoobscura* and *D. p. bogotana* gDNA mismapping rates are very low across chromosomes. Genes were arranged according to their chromosomal locations and the proportion of reads mismapping to the other species was plotted. Genome-wide levels of sequence divergence, measured as the percent of coding bases that could differentiate these species in each gene, were also plotted along each chromosome. Loess lines fitting this data are plotted in red.

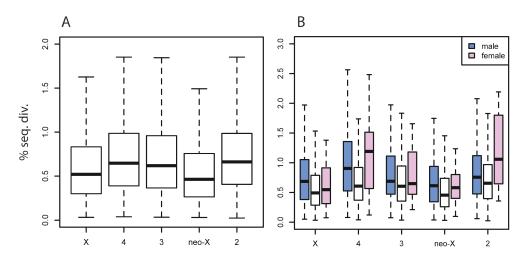


Figure 5.8: Testis- and ovary-specific genes show slightly higher levels of sequence divergence. The distribution of sequence divergence estimates separated by chromosome (A) as well as for genes expressed only in testes (blue), ovaries (pink) and all remaining genes (white) (B).

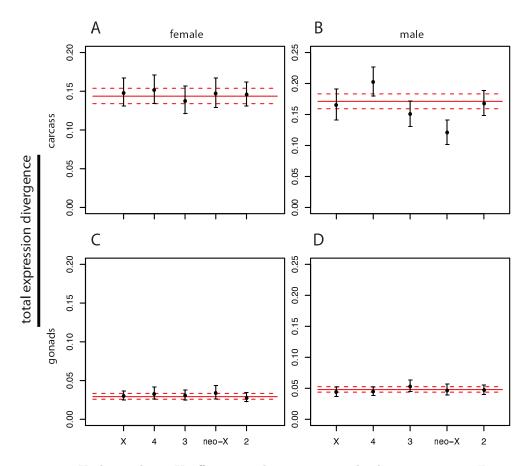


Figure 5.9: neo-X shows slower-X effect in male carcass samples but not testes. Estimates of total expression divergence for female carcass (A), male carcass (B), ovaries (C), and testes (D). Dots represent overall estimates of total expression divergence as measured by 1-Spearman's ρ , and whiskers represent bootstrapped 2.5th and 97.5th percentiles. Autosomal-wide total expression divergence is plotted as a red solid line, and its percentiles as red dashed lines. Non-overlapping intervals indicate statistical significance at $\alpha = 5\%$.

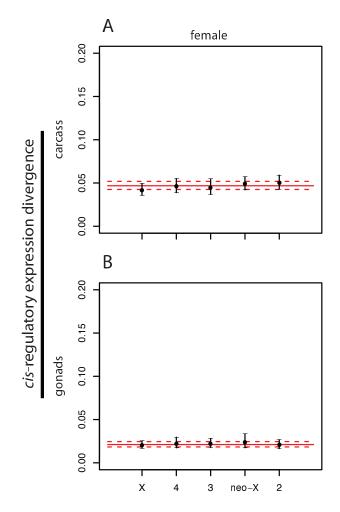


Figure 5.10: No elevated *cis*-regulatory expression divergence on X chromosomes in female samples. Estimates of *cis*-regulatory expression divergence for female carcass (A) and ovaries (B). Dots represent overall estimates of *cis*-regulatory expression divergence as measured by 1-Spearmans ρ , and whiskers represent bootstrapped 2.5th and 97.5th percentiles. Autosomal-wide *cis*-regulatory expression divergence is plotted as a red solid line, and its percentiles as red dashed lines. Non-overlapping intervals indicate statistical significance at $\alpha = 5\%$.

CHAPTER VI

Discussion

6.1 Dissertation summary

As I have shown in my dissertation, the regulation of gene expression is an important aspect of biology to study, as it is one of the main drivers of phenotypic diversity and biological complexity. Since gene expression is known to vary considerably both within and between species, deciphering the *cis-* and *trans-*regulatory mechanisms behind these differences can answer questions about how evolution has shaped gene regulation. Species of the *Drosophila* lineage represent a model system in which to study these phenomena, as many of them can interbreed to produce viable F1 hybrids that are typically easy to rear and mature relatively quickly. Perhaps more importantly for the work presented here, the genomic resources for many *Drosophila* species are among the best available. These resources, along with the advent of high-throughput sequencing technologies, has enabled studies of the regulation of gene expression to be extended from genes of interest on the order of dozens to all expressed genes in the genome.

6.1.1 Chapter II

The study of gene regulation across the expressed genome relies heavily on the ability to accurately measure both total and allele-specific levels of gene expression from next-generation sequencing data types such as RNA-seq, an endeavor to which I have dedicated much of my thesis work. The bioinformatic resources available for quantifying total levels of gene expression from RNA-seq data are abundant, whereas comparable methods to quantify allele-specific expression (ASE) currently lag far behind. Initially, RNA-seq data known to harbor allelic variation was aligned to a single reference genome, often very similar to one or both alleles represented in a sample. This was shown to produce measures of ASE that were systematically biased toward the allele closest to that represented by the reference genome. In chapter II, I identified a major source of this bias to be the inability to allow sufficient mismatches in highly-polymorphic regions of the genome. Another source of bias stemmed from the inability to uniquely map sequence reads to regions of the genome that were either highly-repetitive (low sequence complexity) or shared substantial sequence with other regions (paralogs or pseudogenes). This work advanced our understanding of bioinformatic strategies to accurately quantify ASE from RNAseq data and granted me the confidence to pursue to do so in subsequent chapters, answering interesting biological questions about how the regulation of gene expression has evolved both within and between species.

6.1.2 Chapter III

These methods, even when applied correctly, are not without caveats. Because of the extremely high-throughput nature of RNA-seq, measuring all expressed genes simultaneously and testing for statistically-significant differences between samples invariably poses the problem of false discovery; that is, the rate at which the null hypothesis no difference in gene expression is incorrectly rejected is non-negligible. A good example of this is represented by chapter III in which we putatively found genomic imprinting in *D. melanogaster*, an epigenetic phenomenon whereby the maternal or paternal allele is silenced. Although there was almost no evidence of genomic imprinting in D. melanogaster, when I compared differential ASE profiles between F1 hybrids resulting from reciprocally crossing two strains, I found a little over 100 genes whose ASE profiles were consistent with genomic imprinting. Upon further inspection, using both bioinformatic and experimental techniques, these genes were found to be clustered in the genome, and this signal was due to lowly-segregating heterozygous deletions in the strains of D. melanogaster that were used in this study. This cautionary tale underscores the importance of validating signals using alternative strategies, especially in the age of high-throughput genomics.

6.1.3 Chapter IV

With this in mind, in chapter IV I set out to test the hypothesis that, as species diverge, *cis*-regulatory divergence explains a larger proportion of the total regulatory divergence. I did this by characterizing regulatory divergence patterns across divergence times ranging from 0.01-2.5 million years ago in one intraspecific comparison in *D. melanogaster* and two interspecific comparisons between *D. simulans* and *D. sechellia* and between *D. melanogaster* and *D. simulans*. Although this hypothesis was confirmed, the proportion of expression divergence explained by *cis*-regulation did not seem to increase linearly over time. This work represented one of the few syntheses of gene regulation across multiple divergence times and laid the groundwork for studying the rate and manner in which gene regulation evolves.

6.1.4 Chapter V

Finally, knowing that gene regulation is the primary way in which tissues are differentiated within an organisms, I hypothesized that gene regulatory patterns differ across tissues. In the previous work presented here, I measured gene expression

from whole fly samples that had been ground with a pestle, thereby mixing all of the different tissue types. Such a procedure is neither possible nor defensible in eukaryotes such as mice and human, both for ethical and practical reasons, as well as that stated in my hypothesis. Additionally, the work presented earlier focused only on female samples, mainly because X-linked genes in males are hemizygous, lacking the paternal allele and precluding classification of gene regulation. In chapter V I measured total and allele-specific gene expression in female and male carcass and gonad tissues between D. pseudoobscura and its closely-related subspecies D. p. boqotana, which are outside the melanogaster subgroup of Drosophila and a popular model system to study incipient speciation. Because these species display a mild hybrid incompatibility, gene regulation in the reproductive tissue as compared to the remaining tissues is likely to reveal complex patterns. As I hypothesized, I found extensive differences in patterns of both the inheritance of gene expression as well as gene regulation. This work demonstrates that differences in gene regulation, when integrated over all tissues as in whole fly samples, can be masked, while tissue-specific inferences can reveal the complexity of gene regulation in an organism.

6.2 Reflections

All of the work presented in my dissertation was toward the effort of disentangling the evolution of gene regulation in *Drosophila*. While I feel strongly that I accomplished this, there are certainly additional improvements that could be used to refine these inferences that I would strive to implement were I to stay an additional five years. First, for simplicity as well as lack of available bioinformatics tools, we tend to simplify measures of allele-specific gene expression based on reads stemming from a simplified genetic locus such as an exon or a gene. However, we know that even in *Drosophila* gene expression is regulated by alternative splicing, producing transcripts that contain different exons of the same gene (reviewed in (Graveley, 2001). For this reason, there has been a great effort to estimate the abundance of these different transcripts, and one popular bioinformatics tool that does this is called Cufflinks (Trapnell et al., 2010). While it is beyond the scope of this dissertation to describe Cufflinks in detail, suffice it to say that it is one of the most popular tools for quantifying transcript abundance, though it lacks the ability to do this in an allele-specific manner. The ability to use sophisticated transcript abundances to quantify ASE would represent a significant advance in the field of transcriptomics. In the bioinformatics pipeline I developed for my dissertation, one stage consisted of labeling each sequence read according to the allele that it most closely resembles. I believe one could use Cufflinks in an allele-specific manner by separating sequence reads into those that are allele-specific and consider them separately in the Cufflinks pipeline, which would then give independent measures of allele-specific transcript abundance. Examining gene expression at the level of allele-specific transcripts has the potential to reveal even more complexity of gene regulation within and between species.

One area where I lack the expertise but would enjoy seeing more emphasis placed is allele-specific proteomics. A major critique of studies of gene expression at the level of transcription has been the lack of correlation between changes in transcript abundance and their translation into proteins (Foss et al., 2011). More recent work has suggested that this correlation is better than previously thought and poor normalization may have been responsible for the lack thereof (Albert et al., 2014; Li et al., 2014). This technique has already been demonstrated for 643 allele-specific proteins between *Saccharomyces cerevisiae* and *S. bayanus* analyzing liquid chromatographycoupled mass spectrometry data (Khan et al., 2012), but such analyses have yet to be performed in *Drosophila* species. Analyses of this kind would be important in determining whether or not the observed transcriptional regulation among genes translates to the level of proteins.

Lastly, as genomic resources increasingly become available for different species, I see the field of regulatory evolution expanding to study more and more divergent species pairs. Such studies could be used to test the generalizability across different taxa of the observed increase in the proportion of regulatory divergence attributable to *cis*-regulatory changes as divergence time increases. Although such work is limited by the ability of species pairs to form viable F1 hybrid offspring, a limit we have practically exhausted in the melanogaster subgroup, there are other organisms that have divergence times much greater than those in the *Drosophila* genus and are still able to interbreed. The *Saccharomyces* genus arose between 10 and 20 million years ago, spanning an evolutionary time an order of magnitude great than that studied between *Drosophila* species (reviewed in (Hittinger, 2013). All seven of the natural species in this genus are able to interbreed to form viable F1 hybrids, which would provide a great system in which to study the evolution of gene regulation on an even greater timescale.

While studying gene expression at the level of transcription provides a window into what functions are being carried out in an organism, it is certainly not the end of the line. After transcripts are exported from the nucleus and translated into proteins, those enzymes proceed to carry out cellular processes that produce metabolites, all of which contribute to an organisms physiology. In the very near future, previously low-throughput methodologies in proteomics and metabolomics will catch up to genomics, and new challenges of how to integrate all of these data, how to visualize them, how to meta-analyze them, how to interpret them, and how to store them will be at the forefront of science. How evolutionary biology will fit into this paradigm will be something I follow with great interest.

One of the most important lessons I have learned spending the last five years in the field of genomics is the importance of checking, and double-checking, and crossvalidating, and then checking again all results from high-throughput analyses, all to make sure that the findings we report are credible. It can be very tempting to report something without such rigor, especially if that finding has not been previously reported, but sometimes this can be for a very good reason: it may not actually be true. Sometimes, if one searches hard enough for a signal in ones data, it is bound to show up somewhere just by chance. In an age where the size of data increases at an alarming rate, while the cost to generate it continues to plummet, this is an especially important lesson to keep in mind.

BIBLIOGRAPHY

- Adams, M. D., Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y.-h. C., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Miklos, G. L. G., Abril, J. F., Agbayani, A., An, H.-j., Andrews-pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., Pablos, B. D., Doup, L. E., Downes, M., Dugan-rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J. and Wei, M.-h. The Genome Sequence of Drosophila melanogaster. *Science*, 287:2185–2195, 2000. doi:10.1126/science.287. 5461.2185. 56
- Ahmadian, A., Gharizadeh, B., Gustafsson, A. C., Sterky, F., Nyre, P. I. and Uhle,
 M. Single-Nucleotide Polymorphism Analysis by Pyrosequencing. *Analytical Biochemistry*, 280:103–110, 2000. doi:10.1006/abio.2000.4493. 45, 76

- Albert, F. W., Treusch, S., Shockley, A. H., Bloom, J. S. and Kruglyak, L. Genetics of single-cell protein abundance variation in large yeast populations. *Nature*, 506(7489):494–7, 2014. ISSN 1476-4687. doi:10.1038/nature12904. 166
- Alvarez, D., Noor, M. A. F. and Ruiz-Garcia, M. Comparative genetic structure between tropical Colombian and North American Drosophila pseudoobscura populations. *Biotropica*, 34(1):81–92, 2002. 135
- Alwine, J. C., Kemp, D. J. and Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *PNAS*, 74(12):5350–4, 1977. ISSN 0027-8424.
- Anaka, M., Lynn, A., Mcginn, P. and Lloyd, V. K. Genomic Imprinting in Drosophila has properties of both mammalian and insect imprinting. *Development Genes and Evolution*, 219:59–66, 2009. doi:10.1007/s00427-008-0267-3. 42
- Anders, S. and Huber, W. Differential expression analysis for sequence count data. Genome Biology, 11(10):R106, 2010. ISSN 1465-6906. doi:10.1186/ gb-2010-11-10-r106. 9
- Assis, R., Zhou, Q. and Bachtrog, D. Sex-biased transcriptome evolution in Drosophila. Genome Biology and Evolution, 4(11):1189–200, 2012. ISSN 1759-6653. doi:10.1093/gbe/evs093. 135, 136, 139
- Auwera, G. A. V. D., Carneiro, M. O., Hartl, C., Poplin, R., Angel, G., Levy-moonshine, A., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. and Depristo, M. A. From FastQ Data to High-Confidence Variant Calls : The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics*, October, pages 1–33. 2013. ISBN 0471250953. doi: 10.1002/0471250953.bi1110s43. 145

- Ayala, F. J. and Dobzhansky, T. H. A new subspecies of Drosophila pseudoobscura. The Pan-Pacific Entomologist, 50(3):211–219, 1974. 135
- Ayroles, J. F., Carbone, M. A., Stone, E. a., Jordan, K. W., Lyman, R. F., Magwire, M. M., Rollmann, S. M., Duncan, L. H., Lawrence, F., Anholt, R. R. H. and Mackay, T. F. C. Systems genetics of complex traits in Drosophila melanogaster. *Nature Genetics*, 41(3):299–307, 2009. ISSN 1546-1718. doi:10.1038/ng.332. 18
- Babak, T., Deveale, B., Armour, C., Raymond, C., Cleary, M. A., Kooy, D. V. D., Johnson, J. M. and Lim, L. P. Report Global Survey of Genomic Imprinting by Transcriptome Sequencing. *Current Biology*, 18(22):1735–1741, 2008. ISSN 0960-9822. doi:10.1016/j.cub.2008.09.044. 42, 50
- Barrière, A., Gordon, K. L. and Ruvinsky, I. Coevolution within and between Regulatory Loci Can Preserve Promoter Function Despite Evolutionary Rate Acceleration. *PLoS Genetics*, 8(9):e1002961, 2012. doi:10.1371/journal.pgen.1002961. 72, 79, 82
- Becker-Andre, M. and Hahlbrock, K. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). Nucleic Acids Research, 17(22):9437–9446, 1989. 6
- Bedford, T. and Hartl, D. L. Optimization of gene expression by natural selection. PNAS, 106(4):1133–1138, 2009. 86
- Begun, D. J. and Aquadro, C. F. African and North American populations of Drosophila melanogaster are very different at the DNA level. *Nature*, 365:548– 550, 1993. 43, 52, 88
- Bell, G. D. M., Kane, N. C., Rieseberg, L. H. and Adams, K. L. RNA-seq analysis of

allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biology and Evolution*, 5(7):1309–1323, 2013. doi:10.1093/gbe/evt072. 84

- Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):280–300, 1995. 53, 60, 149
- Bergman, C. M. and Kreitman, M. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Research*, 11(8):1335–45, 2001. ISSN 1088-9051. doi:10.1101/gr.178701. 5
- Boland, J. F., Chung, C. C., Roberson, D., Mitchell, J., Zhang, X., Im, K. M., He, J., Chanock, S. J., Yeager, M. and Dean, M. The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for wholeexome sequencing. *Human Genetics*, 132(10):1153–63, 2013. ISSN 1432-1203. doi:10.1007/s00439-013-1321-4. 8
- Bradley, R. K., Li, X.-Y., Trapnell, C., Davidson, S., Pachter, L., Chu, H. C., Tonkin, L. A., Biggin, M. D. and Eisen, M. B. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biology*, 8(3):e1000343, 2010. ISSN 1545-7885. doi:10.1371/journal.pbio.1000343. 5
- Bradley, R. K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes,
 I. and Pachter, L. Fast Statistical Alignment. *PLoS Computational Biology*, 5(5):e1000392, 2009. ISSN 1553-7358. doi:10.1371/journal.pcbi.1000392. 31, 90, 145

- Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S. and Kaessmann, H. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–8, 2011. ISSN 1476-4687. doi:10.1038/nature10532. 15, 72, 83, 84, 86, 91
- Bridges, C. B. Direct proof through non-disjunction that the sex-linked genes of Drosophila are borne by the X-chromosome. *Science*, 40:107–109, 1914.
- Bridges, C. B. Non-disjunction as a proof of the chromosome theory of heredity. Genetics, 1(2):107–163, 1916. 1
- Bridges, C. B. Salivary chromosome maps with a key to the banding of the chromosomes of Drosophila melanogaster. *Journal of Heredity*, 26(2):60–64, 1935. 1
- Buiting, K., Saitoh, S., Gross, S., Dittrich, B., Schwartz, S., Nicholls, R. D. and Horsthemke, B. Inherited microdeletions in the Angelman and Prader-Willi syndromes define an imprinting centre on human chromosome 15. *Nature Genetics*, 9:395–400, 1995. 42
- Bullard, J. H., Purdom, E., Hansen, K. D. and Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.
 BMC Bioinformatics, 11:94, 2010. ISSN 1471-2105. doi:10.1186/1471-2105-11-94.
 9
- Busby, M. A., Gray, J. M., Costa, A. M., Stewart, C., Stromberg, M. P., Barnett, D., Chuang, J. H., Springer, M. and Marth, G. T. Expression divergence measured by transcriptome sequencing of four yeast species. *BMC Genomics*, 12(1):635, 2011. ISSN 1471-2164. doi:10.1186/1471-2164-12-635. 72, 84

- Campbell, C. D., Kirby, A., Nemesh, J., Campbell, C. D., Kirby, A., Nemesh, J., Daly, M. J. and Hirschhorn, J. N. A survey of allelic imbalance in F1 mice. *Genome Research*, 18:555–563, 2008. doi:10.1101/gr.068692.107. 73
- Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1):25–36, 2008. ISSN 1097-4172. doi:10.1016/ j.cell.2008.06.030. 5, 86
- Caspary, T., Cleary, M. A., Baker, C. C., Guan, X.-j., Tilghman, S. M., Caspary, T., Cleary, M. A., Baker, C. C., Guan, X.-j. and Tilghman, S. M. Multiple Mechanisms Regulate Imprinting of the Mouse Distal Chromosome 7 Gene Cluster Multiple Mechanisms Regulate Imprinting of the Mouse Distal Chromosome 7 Gene Cluster. *Molecular and Cellular Biology*, 18(6):3466–3474, 1998. 46
- Celniker, S. E., Wheeler, D. a., Kronmiller, B., Carlson, J. W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S. P., Frise, E., Hodgson, A., George, R. a., Hoskins, R. a., Laverty, T., Muzny, D. M., Nelson, C. R., Pacleb, J. M., Park, S., Pfeiffer, B. D., Richards, S., Sodergren, E. J., Svirskas, R., Tabor, P. E., Wan, K., Stapleton, M., Sutton, G. G., Venter, C., Weinstock, G., Scherer, S. E., Myers, E. W., Gibbs, R. a. and Rubin, G. M. Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biology*, 3(12):RESEARCH0079, 2002. ISSN 1465-6914. 56
- Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. and Prasher, D. C. Green fluorescent protein as a marker gene expression. *Science*, 263:802–805, 1994. ISSN 0039-9450. 6

Charlesworth, B., Coyne, J. A. and Barton, N. H. The relative rates of evolution of

sex chromosomes and autosomes. *The American Naturalist*, 130(1):113–146, 1987. 136

- Coolon, J. D., McManus, C. J., Stevenson, K. R., Graveley, B. R. and Wittkopp,
 P. J. Tempo and mode of regulatory evolution in Drosophila. *Genome Research*, 24(5):797–808, 2014. ISSN 1549-5469. doi:10.1101/gr.163014.113. 5, 134, 135, 136, 137, 139, 148
- Coolon, J. D., Stevenson, K. R., McManus, C. J., Graveley, B. R. and Wittkopp,
 P. J. Genomic imprinting absent in Drosophila melanogaster adult females. *Cell Reports*, 2(1):69–75, 2012. ISSN 2211-1247. doi:10.1016/j.celrep.2012.06.013. 15, 16, 20, 84, 89, 90, 94, 97, 134, 137, 139, 143
- Coolon, J. D. and Wittkopp, P. J. cis- and trans-regulation in Drosophila interspecific hybrids. In *Polyploid and Hybrid Genomics*, pages 37–57. 2013. 73
- Cowles, C. R., Hirschhorn, J. N., Altshuler, D. and Lander, E. S. Detection of regulatory variation in mouse genes. *Nature Genetics*, 32(3):432–7, 2002. ISSN 1061-4036. doi:10.1038/ng992. 10, 15, 73, 134
- Crick, F. Central dogma of molecular biology. Nature, 227:561–563, 1970. 2
- Crouse, H. V. The controlling element in sex chromosome behavior in Sciara. Genetics, 45(10):1429–1443, 1960. 42
- Cutter, A. D. Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. *Molecular Biology and Evolution*, 25(4):778–86, 2008. ISSN 1537-1719. doi:10.1093/molbev/msn024. 75
- David, J. R. and Capy, P. Genetic variation of Drosophila melanogaster natural populations. *Trends in Genetics*, 4(4):106–111, 1988. 74

- Degner, J. F., Marioni, J. C., Pai, A. a., Pickrell, J. K., Nkadori, E., Gilad, Y. and Pritchard, J. K. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–12, 2009. ISSN 1367-4811. doi:10.1093/bioinformatics/btp579. 11, 16, 17, 18, 22, 26, 44, 99
- DePristo, M. a., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. a., del Angel, G., Rivas, M. a., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. and Daly, M. J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–8, 2011. ISSN 1546-1718. doi:10.1038/ng.806. 145
- Derrien, T., Estellé, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R. and Ribeca, P. Fast computation and applications of genome mappability. *PLoS ONE*, 7(1):e30377, 2012. ISSN 1932-6203. doi:10.1371/journal.pone.0030377. 22, 30
- DeVeale, B., van der Kooy, D. and Babak, T. Critical Evaluation of Imprinted Gene Expression by RNASeq: A New Perspective. *PLoS Genetics*, 8(3):e1002600, 2012. ISSN 1553-7404. doi:10.1371/journal.pgen.1002600. 15, 16, 51
- Dickinson, W. J., Rowan, R. G. and Brennan, M. D. REGULATORY GENE EVO-LUTION: ADAPTIVE DIFFERENCES IN EXPRESSION OF ALCOHOL DE-HYDROGENASE IN DROSOPHILA MELANOGAGSTER AND DROSOPHILA SIMULANS. *Heredity*, 52(2):215–225, 1984. 78
- Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant,
 N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau,
 L., Laloë, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M. and Jaffrézic, F. A
 comprehensive evaluation of normalization methods for Illumina high-throughput

RNA sequencing data analysis. *Briefings in Bioinformatics*, 2012. ISSN 1477-4054. doi:10.1093/bib/bbs046. 9

- Dobzhansky, T. H. Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. *Genetics*, 21:113–135, 1936. 135
- Emerson, J. J., Hsieh, L.-C., Sung, H.-M., Wang, T.-Y., Huang, C.-J., Lu, H. H.-S., Lu, M.-Y. J., Wu, S.-H. and Li, W.-H. Natural selection on cis and trans regulation in yeasts. *Genome Research*, 20(6):826–36, 2010. ISSN 1549-5469. doi: 10.1101/gr.101576.109. 5, 10, 45, 74, 82, 84, 86, 87, 88
- Fay, J. C. and Wittkopp, P. J. Evaluating the role of natural selection in the evolution of gene regulation. *Heredity*, 100(2):191–9, 2008. ISSN 1365-2540. doi:10.1038/sj. hdy.6801000. 86
- Ferguson-Smith, A. C. Genomic imprinting: the emergence of an epigenetic paradigm. Nature Reviews Genetics, 12(8):565–75, 2011. ISSN 1471-0064. doi: 10.1038/nrg3032. 42, 46
- Ferree, P. M. and Barbash, D. A. Species-Specific Heterochromatin Prevents Mitotic Chromosome Segregation to Cause Hybrid Lethality in Drosophila. *PLoS Biology*, 7(10):e1000234, 2009. doi:10.1371/journal.pbio.1000234. 52, 88
- Fontanillas, P., Landry, C. R., Wittkopp, P. J., Russ, C., Gruber, J. D., Nusbaum, C. and Hartl, D. L. Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. *Molecular Ecology*, 19 Suppl 1:212–27, 2010a. ISSN 1365-294X. doi:10.1111/j.1365-294X.2010.04472.x. 24, 44, 53, 73
- Fontanillas, P., Landry, C. R., Wittkopp, P. J., Russ, C., Gruber, J. D., Nusbaum, C. and Hartl, D. L. Supplemental: Key considerations for measuring allelic expression

on a genomic scale using high-throughput sequencing. *Molecular Ecology*, 2010b. 94

- Foss, E. J., Radulovic, D., Shaffer, S. a., Goodlett, D. R., Kruglyak, L. and Bedalov, A. Genetic variation shapes protein networks mainly through non-transcriptional mechanisms. *PLoS Biology*, 9(9):e1001144, 2011. ISSN 1545-7885. doi:10.1371/ journal.pbio.1001144. 166
- Frohnhöfer, H. G. and Nüsslein-Volhard, C. Organization of anterior pattern in the Drosophila embryo by the maternal gene bicoid. *Nature*, 324:120–125, 1986. 3
- Fuyama, Y. Gynogenesis in Drosophila melanogaster. Japanese Journal of Genetics, 59:91–96, 1984. 43
- Garrigan, D., Kingan, S. B., Geneva, A. J., Andolfatto, P., Clark, A. G., Thornton,
 K. R. and Presgraves, D. C. Genome sequencing reveals complex speciation in the
 Drosophila simulans clade. *Genome Research*, 22:1499–1511, 2012. doi:10.1101/gr.
 130922.111.22. 75
- Gehring, M., Missirian, V. and Henikoff, S. Genomic Analysis of Parent-of-Origin Allelic Expression in Arabidopsis thaliana Seeds. *PLoS ONE*, 6(8):e23687, 2011. doi:10.1371/journal.pone.0023687. 42, 50
- Genissel, A., McIntyre, L. M., Wayne, M. L. and Nuzhdin, S. V. Cis and trans regulatory effects contribute to natural variation in transcriptome of Drosophila melanogaster. *Molecular Biology and Evolution*, 25(1):101–10, 2008. ISSN 1537-1719. doi:10.1093/molbev/msm247. 73
- Gibson, G., Riley-Berger, R., Harshman, L., Kopp, A., Vacha, S., Nuzhdin,S. and Wayne, M. Extensive sex-specific nonadditivity of gene expression in

Drosophila melanogaster. *Genetics*, 167(4):1791–9, 2004. ISSN 0016-6731. doi: 10.1534/genetics.104.026583. **43**, **92**, **138**, **148**

- Gilad, Y., Oshlack, A. and Rifkin, S. a. Natural selection on gene expression. *Trends in Genetics*, 22(8):456–61, 2006a. ISSN 0168-9525. doi:10.1016/j.tig.2006.06.002.
 86
- Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P. and White, K. P. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature*, 440(7081):242–5, 2006b. ISSN 1476-4687. doi:10.1038/nature04559. 87
- Glaser, T., Walton, D. S. and Maas, R. L. Genomic structure, evolutionary conservation and aniridia mutations in the human PAX6 gene. *Nature Genetics*, 2:232–239, 1992. 3
- Golic, K. G., Golic, M. M. and Pimpinelli, S. Imprinted control of gene activity in Drosophila. *Current Biology*, 8(1):1273–1276, 1998. 42
- Goncalves, A., Leigh-Brown, S., Thybert, D., Stefflova, K., Turro, E., Flicek, P., Brazma, A., Odom, D. T. and Marioni, J. C. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Research*, 22(12):2376– 84, 2012. ISSN 1549-5469. doi:10.1101/gr.142281.112. 74, 84, 85, 86
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. a., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7), 2011. ISSN 1546-1696. doi:10.1038/nbt.1883. 9

- Graveley, B. R. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, 17(2):100–7, 2001. ISSN 0168-9525. 166
- Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., Baren, M. J. V., Boley, N., Booth, B. W., Brown, J. B., Cherbas, L., Davis, C. A., Dobin, A., Li, R., Lin, W., Malone, J. H., Mattiuzzo, N. R., Miller, D., Sturgill, D., Tuch, B. B., Zaleski, C., Kapranov, P., Langton, L., Perrimon, N., Sandler, J. E., Wan, K. H., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P. J., Brenner, S. E., Brent, M. R., Cherbas, P., Gingeras, T. R., Hoskins, R. A., Kaufman, T. C., Oliver, B. and Celniker, S. E. The developmental transcriptome of Drosophila melanogaster. *Nature*, 471(7339):473–479, 2011. ISSN 0028-0836. doi:10.1038/nature09715. 75
- Graze, R. M., McIntyre, L. M., Main, B. J., Wayne, M. L. and Nuzhdin, S. V. Regulatory divergence in Drosophila melanogaster and D. simulans, a genomewide analysis of allele-specific expression. *Genetics*, 183(2):547–61, 1SI–21SI, 2009. ISSN 1943-2631. doi:10.1534/genetics.109.105957. 20, 73, 84, 134
- Graze, R. M., Novelo, L. L., Amin, V., Fear, J. M., Casella, G., Nuzhdin, S. V. and McIntyre, L. M. Allelic Imbalance in Drosophila Hybrid Heads: Exons, Isoforms, and Evolution. *Molecular Biology and Evolution*, 2012. ISSN 1537-1719. doi: 10.1093/molbev/msr318. 16, 21, 24, 30, 31, 44, 134
- Gregg, C., Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G. P., Haig, D. and Dulac, C. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, 329:643–648, 2010a. doi:10.1126/science.1190830. 51
- Gregg, C., Zhang, J., Butler, J. E., Haig, D. and Dulac, C. Sex-Specific Parent-

of-Origin Allelic Expression in the Mouse Brain. *Science*, 329:682–685, 2010b. doi:10.1126/science.1190831. 51

- Gruber, J. D. and Long, A. D. Cis-regulatory variation is typically polyallelic in Drosophila. *Genetics*, 181(2):661–70, 2009. ISSN 0016-6731. doi:10.1534/genetics. 108.098459. 73
- Gruber, J. D., Vogel, K., Kalay, G. and Wittkopp, P. J. Contrasting properties of gene-specific regulatory, coding, and copy number mutations in Saccharomyces cerevisiae: frequency, effects, and dominance. *PLoS Genetics*, 8(2):e1002497, 2012. ISSN 1553-7404. doi:10.1371/journal.pgen.1002497. 85
- Haerty, W. and Singh, R. S. Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of Drosophila. *Molecular Biology and Evolution*, 23(9):1707–14, 2006. ISSN 0737-4038. doi:10. 1093/molbev/msl033. 77
- Hahn, M. W. and Wray, G. a. The g-value paradox. *Evolution & Development*, 4(2):73–5, 2002. ISSN 1520-541X. 5
- Halder, G., Callaerts, P. and Gehring, W. J. Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. *Science*, 267(5205):1788–92, 1995.
 ISSN 0036-8075. 3
- Haller, B. S. and Woodruff, R. C. Varied expression of a Y-linked P[w+] insert due to imprinting in Drosophila melanogaster. *Genome*, 43:285–292, 2000. 42
- Hardcastle, T. J. and Kelly, K. a. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11:422, 2010.
 ISSN 1471-2105. doi:10.1186/1471-2105-11-422. 9

Harris, R. S. Improved pairwise alignment of genomic DNA. Ph.D. thesis, 2007. 98

- Hashimoto, T. B., Edwards, M. D. and Gifford, D. K. Universal Count Correction for High-Throughput Sequencing. *PLoS Computational Biology*, 10(3):e1003494, 2014. ISSN 1553-7358. doi:10.1371/journal.pcbi.1003494. 9
- He, F., Zhang, X., Hu, J., Turck, F., Dong, X., Goebel, U., Borevitz, J. and de Meaux, J. Genome-wide analysis of cis-regulatory divergence between species in the Arabidopsis genus. *Molecular Biology and Evolution*, 29(11):3385–95, 2012. ISSN 1537-1719. doi:10.1093/molbev/mss146. 73
- Heid, C. a., Stevens, J., Livak, K. J. and Williams, P. M. Real time quantitative PCR.
 Genome Research, 6(10):986–994, 1996. ISSN 1088-9051. doi:10.1101/gr.6.10.986.
 6
- Hill, R. E., Favor, J., Hodan, B. L. M., Ton, C. C. T., Saunders, G. F., Hanson, I. M., Prosser, J., Jordan, T., Hastie, N. D. and Heyningen, V. v. Mouse Small eye results from mutations in a paired-like homeobox-containing gene. *Nature*, 354:522–525, 1991. 3
- Hittinger, C. T. Saccharomyces diversity and evolution: a budding model genus. Trends in Genetics, 29(5):309–17, 2013. ISSN 0168-9525. doi:10.1016/j.tig.2013. 01.002. 167
- Hollocher, H., Ting, C.-T., Wu, M.-L. and Wu, C.-I. Incipient speciation by sexual isolation in Drosophila melanogaster: extensive genetic divergence without reinforcement. *Genetics*, 147:1191–1201, 1997. 43, 75, 78, 88

Hsieh, T.-f., Shin, J., Uzawa, R., Silva, P., Cohen, S., Bauer, M. J. and Hashimoto,

M. Regulation of imprinted gene expression in Arabidopsis endosperm. *PNAS*, 108(5):1755–1762, 2011. doi:10.1073/pnas.1019273108. 42, 50

- Hsieh, W.-P., Chu, T.-M., Wolfinger, R. D. and Gibson, G. Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics*, 165(2):747–57, 2003. ISSN 0016-6731. 87
- Hutter, S., Saminadin-Peter, S. S., Stephan, W. and Parsch, J. Gene expression variation in African and European populations of Drosophila melanogaster. *Genome Biology*, 9(1):R12, 2008. ISSN 1465-6914. doi:10.1186/gb-2008-9-1-r12. 75
- Jacob, F. and Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology, 3:318–356, 1961.
- Jenkins, T. M., Basten, C. J. and Anderson, W. W. Mitochondrial gene divergence of Colombian Drosophila pseudoobscura. *Molecular Biology and Evolution*, 13(9):1266–75, 1996. ISSN 0737-4038. 135
- Jiang, H. and Wong, W. H. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics, 25(8):1026–32, 2009. ISSN 1367-4811. doi:10.1093/bioinformatics/ btp113. 136, 140
- Joanis, V. and Lloyd, V. K. Genomic imprinting in Drosophila is maintained by the products of Suppressor of variegation and trithorax group, but not Polycomb group, genes. *Molecular Genetics and Genomics*, 268:103–112, 2002. doi:10.1007/ s00438-002-0731-0. 42
- Jones, C. D. The genetics of adaptation in Drosophila sechellia. *Genetica*, 123(1-2):137–45, 2005. ISSN 0016-6707. 87

- Kalinka, A. T., Varga, K. M., Gerrard, D. T., Preibisch, S., Corcoran, D. L., Jarrells, J., Ohler, U., Bergman, C. M. and Tomancak, P. Gene expression divergence recapitulates the developmental hourglass model. *Nature*, 468(7325):811–4, 2010. ISSN 1476-4687. doi:10.1038/nature09634. 87
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, a. M. and Haussler, a. D. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, 2002. ISSN 1088-9051. doi:10.1101/gr.229102. 53, 57, 59, 90, 98, 99, 100
- Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W. and Pääbo, S. A neutral model of transcriptome evolution. *PLoS Biology*, 2(5):E132, 2004. ISSN 1545-7885. doi:10.1371/journal.pbio. 0020132. 4, 87
- Khan, Z., Bloom, J. S., Amini, S., Singh, M., Perlman, D. H., Caudy, A. a. and Kruglyak, L. Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. *Molecular Systems Biology*, 8(602):602, 2012. ISSN 1744-4292. doi:10.1038/msb.2012.34. 167
- Komma, D. J. and Endow, S. A. Haploidy and androgenesis in Drosophila. PNAS, 92(December):11884–11888, 1995. 43
- Kuo, D., Licon, K., Bandyopadhyay, S., Chuang, R., Luo, C., Catalana, J., Ravasi, T., Tan, K. and Ideker, T. Coevolution within a transcriptional network by compensatory trans and cis mutations. *Genome Research*, 20(12):1672–8, 2010. ISSN 1549-5469. doi:10.1101/gr.111765.110. 86
- Lachaise, D., Cariou, M.-L., David, J. R., Lemeunier, F., Tsacas, L. and Ashburner,

M. Historical biogeography of the Drosophila melanogaster species subgroup. *Evolutionary Biology*, 22:159–225, 1988. 74

- Lachaise, D., David, J. R., Lemeunier, F., Tsacas, L. and Ashburner, M. The reproductive relationships of Drosophila sechellia with D. mauritiana, D. simulans, and D. melanogaster from the Afrotropical region. *Evolution*, 40(2):262–271, 1986. 78
- Landry, C. R., Wittkopp, P. J., Taubes, C. H., Ranz, J. M., Clark, A. G. and Hartl,
 D. L. Compensatory cis-trans evolution and the dysregulation of gene expression
 in interspecific hybrids of Drosophila. *Genetics*, 171(4):1813–22, 2005. ISSN 00166731. doi:10.1534/genetics.105.047449. 47, 73, 82, 85, 96
- Langmead, B. and Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4):357–9, 2012. ISSN 1548-7105. doi:10.1038/nmeth.1923. 144
- Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. Ultrafast and memoryefficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009. ISSN 1465-6914. doi:10.1186/gb-2009-10-3-r25. 8, 16, 19, 28, 146
- Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. a. C., Monlong, J., Rivas, M. a., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P. J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S. E., Häsler, R., Syvänen, A.-C., van Ommen, G.-J., Brazma, A.,

Meitinger, T., Rosenstiel, P., Guigó, R., Gut, I. G., Estivill, X. and Dermitzakis,
E. T. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–11, 2013. ISSN 1476-4687. doi:10.1038/nature12531.
94

- Lemos, B., Araripe, L. O., Fontanillas, P. and Hartl, D. L. Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. *PNAS*, 105(38):14471–6, 2008. ISSN 1091-6490. doi:10.1073/pnas.0805160105. 74, 83, 87, 139
- Lemos, B., Meiklejohn, C. D., Cáceres, M. and Hartl, D. L. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution*, 59(1):126–37, 2005. ISSN 0014-3820. 4, 87
- Levine, M. Transcriptional enhancers in animal development and evolution. Current Biology, 20(17):R754–63, 2010. ISSN 1879-0445. doi:10.1016/j.cub.2010.06.070. 3
- Levy, A. A., Tirosh, I., Reikhav, S., Bloch, Y. and Barkai, N. Yeast hybrids and polyploids as models in evolutionary studies. In *Polyploid and Hybrids Genomics*, pages 3–14. 2013. 73
- Lewis, A., Mitsuya, K., Umlauf, D., Smith, P., Dean, W., Walter, J., Higgins, M., Feil, R. and Reik, W. Imprinting on distal chromosome 7 in the placenta involves repressive histone methylation independent of DNA methylation. *Nature Genetics*, 36(12):1291–1295, 2004. doi:10.1038/ng1468. 46
- Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009. ISSN 1367-4811. doi:10.1093/ bioinformatics/btp324. 8, 16, 56, 98

- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9, 2009. ISSN 1367-4811. doi:10.1093/bioinformatics/ btp352. 16, 19, 28, 56, 98, 145
- Li, H., Ruan, J. and Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–8, 2008. ISSN 1088-9051. doi:10.1101/gr.078212.108. 16
- Li, J. J., Bickel, P. J. and Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, 2:e270, 2014. ISSN 2167-8359. doi:10.7717/peerj.270. 166
- Lloyd, V. K., Sinclair, D. A. and Grigliatti, T. A. Genomic Imprinting and Position-Effect Variegation in Drosophila melanogaster. *Genetics*, 151:1503–1516, 1999. 42
- Lopes, S., Lewis, A., Hajkova, P., Dean, W., Oswald, J., Bartolomei, M., Murrell, A. and Consta, M. Epigenetic modifications in an imprinting cluster are controlled by a hierarchy of DMRs suggesting long-range chromatin interactions. *Human Molecular Genetics*, 12(3):295–305, 2003. doi:10.1093/hmg/ddg022. 46
- Ludwig, M. Z., Palsson, A., Alekseeva, E., Bergman, C. M., Nathan, J. and Kreitman, M. Functional evolution of a cis-regulatory module. *PLoS Biology*, 3(4):e93, 2005. ISSN 1545-7885. doi:10.1371/journal.pbio.0030093. 5
- Luo, M., Taylor, J. M., Spriggs, A., Zhang, H., Wu, X., Russell, S. and Koltunow,
 A. A Genome-Wide Survey of Imprinted Genes in Rice Seeds Reveals Imprinting
 Primarily Occurs in the Endosperm. *PLoS Genetics*, 7(6):e1002125, 2011. doi: 10.1371/journal.pgen.1002125. 42, 51

- Macdonald, W. A., Menon, D., Bartlett, N. J., Sperry, G. E., Rasheva, V., Meller, V. and Lloyd, V. K. The Drosophila homolog of the mammalian imprint regulator , CTCF , maintains the maternal genomic imprint in Drosophila melanogaster. *BMC Biology*, 8(105):1–14, 2010. 42
- Machado, C. a. and Hey, J. The causes of phylogenetic conflict in a classic Drosophila species group. *Proceedings of the Royal Society London B*, 270(1520):1193–202, 2003. ISSN 0962-8452. doi:10.1098/rspb.2003.2333. 135
- Machado, C. a., Kliman, R. M., Markert, J. a. and Hey, J. Inferring the history of speciation from multilocus DNA sequence data: the case of Drosophila pseudoobscura and close relatives. *Molecular Biology and Evolution*, 19(4):472–88, 2002. ISSN 0737-4038. 135
- Mackay, T. F. C., Richards, S., Stone, E. a., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., Richardson, M. F., Anholt, R. R. H., Barrón, M., Bess, C., Blankenburg, K. P., Carbone, M. A., Castellano, D., Chaboub, L., Duncan, L., Harris, Z., Javaid, M., Jayaseelan, J. C., Jhangiani, S. N., Jordan, K. W., Lara, F., Lawrence, F., Lee, S. L., Librado, P., Linheiro, R. S., Lyman, R. F., Mackey, A. J., Munidasa, M., Muzny, D. M., Nazareth, L., Newsham, I., Perales, L., Pu, L.-L., Qu, C., Ràmia, M., Reid, J. G., Rollmann, S. M., Rozas, J., Saada, N., Turlapati, L., Worley, K. C., Wu, Y.-Q., Yamamoto, A., Zhu, Y., Bergman, C. M., Thornton, K. R., Mittelman, D. and Gibbs, R. a. The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482(7384):173–8, 2012. ISSN 1476-4687. doi:10.1038/nature10811. 18, 48

Maggert, K. A. and Golic, K. G. The Y Chromosome of Drosophila melanogaster Ex-

hibits Chromosome-Wide Imprinting. *Genetics*, 162(November):1245–1258, 2002. 42

- Maheshwari, S. and Barbash, D. a. Cis-by-Trans regulatory divergence causes the asymmetric lethal effects of an ancestral hybrid incompatibility gene. *PLoS Genetics*, 8(3):e1002597, 2012. ISSN 1553-7404. doi:10.1371/journal.pgen.1002597. 77, 82
- Mancini-Dinardo, D., Steele, S. J. S., Levorse, J. M., Ingram, R. S. and Tilghman,
 S. M. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes & Development*, 20:1268–1282, 2006. doi:10.1101/gad. 1416906.chromosome. 46
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–17, 2008. ISSN 1088-9051. doi:10.1101/gr. 079558.108. 9, 15
- Martin, J. a. and Wang, Z. Next-generation transcriptome assembly. Nature Reviews Genetics, (September):1–12, 2011. ISSN 1471-0056. doi:10.1038/nrg3068.
- Mcgrath, J. and Solter, D. Completion of Mouse Embryogenesis Requires Both the Maternal and Paternal Genomes. *Cell*, 37(May):179–183, 1984. 42
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and Depristo, M. A. The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, pages 1297–1303, 2010. doi:10.1101/gr. 107524.110.20. 145

- McManus, C. J., Coolon, J. D., O'Duff, M., Eipper-Mains, J., Graveley, B. R. and
 Wittkopp, P. J. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Research*, pages 816–825, 2010. ISSN 1549-5469. doi:10.1101/gr.102491.
 109. 10, 16, 18, 24, 28, 44, 45, 52, 53, 57, 59, 73, 74, 82, 83, 84, 89, 90, 92, 94, 96,
 100, 134, 136, 139, 142
- McManus, C. J., May, G. E., Spealman, P. and Shteyman, A. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Research*, 24(3):422–30, 2014. ISSN 1549-5469. doi:10.1101/gr.164996.113. 85
- Meaux, J. D., Pop, A. and Mitchell-Olds, T. Cis-regulatory Evolution of Chalcone-Synthase Expression in the Genus Arabidopsis. *Genetics*, 174(December):2181– 2202, 2006. doi:10.1534/genetics.106.064543. 73
- Meiklejohn, C. D., Parsch, J., Ranz, J. M. and Hartl, D. L. Rapid evolution of male-biased gene expression in Drosophila. *PNAS*, 100(17):9894–9, 2003. ISSN 0027-8424. doi:10.1073/pnas.1630690100. 77
- Meisel, R. P., Malone, J. H. and Clark, A. G. Faster-X evolution of gene expression in Drosophila. *PLoS Genetics*, 8(10):e1003013, 2012. ISSN 1553-7404. doi:10.1371/ journal.pgen.1003013. 84, 91, 140, 142
- Menon, D. U. and Meller, V. H. Imprinting of the Y Chromosome Influences Dosage Compensation in roX1 roX2 Drosophila melanogaster. *Genetics*, 183:811–820, 2009. doi:10.1534/genetics.109.107219. 42, 43, 50
- Metzker, M. L. Sequencing technologies the next generation. Nature Reviews Genetics, 11(1):31–46, 2009. ISSN 1471-0056. doi:10.1038/nrg2626. 8, 15

Michalak, P. and Noor, M. a. F. Association of misexpression with sterility in hy-

brids of Drosophila simulansand D. mauritiana. *Journal of Molecular Evolution*, 59(2):277–82, 2004. ISSN 0022-2844. doi:10.1007/s00239-004-2622-y. 77, 82

- Moehring, A. J., Teeter, K. C. and Noor, M. a. F. Genome-wide patterns of expression in Drosophila pure species and hybrid males. II. Examination of multiple-species hybridizations, platforms, and life cycle stages. *Molecular Biology and Evolution*, 24(1):137–45, 2007. ISSN 0737-4038. doi:10.1093/molbev/msl142. 77
- Morgan, T. H. Sex limited inheritance in Drosophila. *Science*, 32(812):120–122, 1910. 1
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):1–8, 2008. doi:10.1038/NMETH.1226. 9, 15, 16
- Muller, H. J. Artificial transmutation of the gene. Science, 66(1699):84–87, 1927. 2
- Murata, Y., Oda, S. and Mitani, H. Allelic expression changes in Medaka (Oryzias latipes) hybrids between inbred strains derived from genetically distant populations. *PloS ONE*, 7(5):e36875, 2012. ISSN 1932-6203. doi:10.1371/journal.pone. 0036875. 84
- Nüsslein-Volhard, C. and Wieschaus, E. Mutations affecting segment number and polarity in Drosophila. *Nature*, 287:795–801, 1980. **3**
- Ong, C.-T. and Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, 12(4):283–93, 2011. ISSN 1471-0064. doi:10.1038/nrg2957. 3
- Orgogozo, V. and Stern, D. L. How different are recently diverged species? *Fly*, 3(2):117, 2009. 86, 87

- Osada, N., Kohn, M. H. and Wu, C.-I. Genomic inferences of the cis-regulatory nucleotide polymorphisms underlying gene expression differences between Drosophila melanogaster mating races. *Molecular Biology and Evolution*, 23(8):1585–91, 2006. ISSN 0737-4038. doi:10.1093/molbev/msl023. 73
- Oshlack, A., Robinson, M. D. and Young, M. D. From RNA-seq reads to differential expression results. *Genome Biology*, 11(12):220, 2010. ISSN 1465-6914. doi:10. 1186/gb-2010-11-12-220. 9
- Phadnis, N. and Orr, H. A. A single gene causes both male sterility and segregation distortion in Drosophila hybrids. *Science*, 323:376–379, 2009. 143
- Powell, J. R. Interspecific cytoplasmic gene flow in the absence of nuclear gene flow: evidence from Drosophila. *PNAS*, 80:492–495, 1983. 135
- Prakash, S. Origin of reproductive isolation in the absence of apparent genic differentiation in a geographic isolate of Drosophila pseudoobscura. *Genetics*, 72:143–155, 1972. 135
- Quackenbush, J. Microarray data normalization and transformation. Nature Genetics, 32(december):496–501, 2002. ISSN 1061-4036. doi:10.1038/ng1032. 7
- Quinlan, A. R. and Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–2, 2010. ISSN 1367-4811. doi:10. 1093/bioinformatics/btq033. 53, 100
- Quiring, R., Walldorf, U., Kloter, U. and Gehring, W. J. Homology of the eyeless gene of Drosophila to the Small eye gene in mice and Aniridia in humans. *Science*, 265:785–789, 1994. 3

- Ranz, J. M., Namgyal, K., Gibson, G. and Hartl, D. L. Anomalies in the expression profile of interspecific hybrids of Drosophila melanogaster and Drosophila simulans. *Genome Research*, 14(3):373–9, 2004. ISSN 1088-9051. doi:10.1101/gr.2019804. 77, 82
- Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D. and Betel, D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9):R95, 2013. ISSN 1465-6914. doi:10.1186/gb-2013-14-9-r95. 9, 91
- Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M. J., Chen, R., Meisel, R. P., Couronne, O., Hua, S., Smith, M. a., Zhang, P., Liu, J., Bussemaker, H. J., van Batenburg, M. F., Howells, S. L., Scherer, S. E., Sodergren, E., Matthews, B. B., Crosby, M. a., Schroeder, A. J., Ortiz-Barrientos, D., Rives, C. M., Metzker, M. L., Muzny, D. M., Scott, G., Steffen, D., Wheeler, D. a., Worley, K. C., Havlak, P., Durbin, K. J., Egan, A., Gill, R., Hume, J., Morgan, M. B., Miner, G., Hamilton, C., Huang, Y., Waldron, L., Verduzco, D., Clerc-Blankenburg, K. P., Dubchak, I., Noor, M. a. F., Anderson, W., White, K. P., Clark, A. G., Schaeffer, S. W., Gelbart, W., Weinstock, G. M. and Gibbs, R. a. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Research*, 15(1):1–18, 2005. ISSN 1088-9051. doi:10.1101/gr.3059305. 5
- Rifkin, S. a., Houle, D., Kim, J. and White, K. P. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature*, 438(7065):220–3, 2005. ISSN 1476-4687. doi:10.1038/nature04114. 4

Rifkin, S. a., Kim, J. and White, K. P. Evolution of gene expression in the Drosophila

melanogaster subgroup. *Nature Genetics*, 33(2):138–44, 2003. ISSN 1061-4036. doi: 10.1038/ng1086. 4, 72, 86, 87

- Rivas-Astroza, M., Xie, D., Cao, X. and Zhong, S. Mapping personal functional data to personal genomes. *Bioinformatics*, 27(24):3427–9, 2011. ISSN 1367-4811. doi:10.1093/bioinformatics/btr578. 17, 20
- Robinson, M. D., McCarthy, D. J. and Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, 2010. ISSN 1367-4811. doi:10.1093/bioinformatics/btp616. 9
- Robinson, M. D. and Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010. ISSN 1465-6914. doi:10.1186/gb-2010-11-3-r25. 9
- Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., Bhardwaj, N., Rubin, M., Snyder, M. and Gerstein, M. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology*, 7(522):1–15, 2011. ISSN 1744-4292. doi:10.1038/msb.2011.54. 17, 20
- Rubin, G. M. and Lewis, E. B. A Brief History of Drosophila's Contributions to Genome Research. Science, 287(5461):2216–2218, 2000. ISSN 00368075. doi: 10.1126/science.287.5461.2216. 2
- Satya, R. V., Zavaljevski, N. and Reifman, J. A new strategy to reduce allelic bias in RNA-Seq readmapping. Nucleic Acids Research, pages 1–9, 2012. ISSN 1362-4962. doi:10.1093/nar/gks425. 17, 20

Sawamura, K., Yamamoto, M.-T. and Watanabe, T. K. Hybrid Lethal Systems in

the Drosophila melanogaster species complex II. The Zygotic hybrid rescue (Zhr) gene of D. melanogaster. *Genetics*, 133(4):307–313, 1993. 43, 52, 88

- Schad, E., Tompa, P. and Hegyi, H. The relationship between proteome size, structural disorder and organism complexity. *Genome Biology*, 12(12):R120, 2011. ISSN 1465-6914. doi:10.1186/gb-2011-12-12-r120. 6
- Schaeffer, S. W. and Miller, E. L. Nucleotide sequence analysis of Adh genes estimates the time of geographic isloation of the Bogota population of Drosophila pseudoobscura. *PNAS*, 88:6097–6101, 1991. 135
- Schaefke, B., Emerson, J. J., Wang, T.-Y., Lu, M.-Y. J., Hsieh, L.-C. and Li, W.-H. Inheritance of gene expression level and selective constraints on trans- and cis-regulatory changes in yeast. *Molecular Biology and Evolution*, 30(9):2121–33, 2013. ISSN 1537-1719. doi:10.1093/molbev/mst114. 73, 74, 84
- Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270:467–470, 1995. 7
- Scott, C. P., VanWye, J., McDonald, M. D. and Crawford, D. L. Technical analysis of cDNA microarrays. *PloS ONE*, 4(2):e4486, 2009. ISSN 1932-6203. doi:10.1371/ journal.pone.0004486.
- Shen, Y., Garcia, T., Pabuwal, V., Boswell, M., Pasquali, A., Beldorth, I., Warren, W., Schartl, M., Cresko, W. a. and Walter, R. B. Alternative strategies for development of a reference transcriptome for quantification of allele specific expression in organisms having sparse genomic resources. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, 8(1):11–16, 2012. ISSN 1744117X. doi:10.1016/j.cbd.2012.10.006. 16, 84

- Shi, X., Ng, D. W.-K., Zhang, C., Comai, L., Ye, W. and Chen, Z. J. Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. *Nature Communications*, 3:950, 2012. ISSN 2041-1723. doi:10.1038/ncomms1954. 73, 74, 84, 85
- Shibata, Y., Kumar, P., Layer, R., Willcox, S., Gagan, J. R., Griffith, J. D. and Dutta, A. Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. *Science*, 336(6077):82–6, 2012. ISSN 1095-9203. doi: 10.1126/science.1213307. 51
- Simpson, G. G. Columbia University Press, 1944. 72
- Skelly, D. a., Johansson, M., Madeoy, J., Wakefield, J. and Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Research*, 2011. ISSN 1088-9051. doi: 10.1101/gr.119784.110. 17, 95, 102
- Soneson, C. and Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):91, 2013. ISSN 1471-2105. doi:10.1186/1471-2105-14-91. 9
- Stevenson, K. R., Coolon, J. D. and Wittkopp, P. J. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics*, 14(1):536, 2013. ISSN 1471-2164. doi: 10.1186/1471-2164-14-536. 99
- Sturtevant, A. H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14:43–59, 1913. 1

- Surani, M. A. H., Barton, S. C. and Norris, M. L. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature*, 308:548–550, 1984. 42
- Suvorov, A., Nolte, V., Pandey, R. V., Franssen, S. U., Futschik, A. and Schlötterer,
 C. Intra-specific regulatory variation in Drosophila pseudoobscura. *PLoS ONE*,
 8(12):e83547, 2013. ISSN 1932-6203. doi:10.1371/journal.pone.0083547. 84, 136
- Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, a. and Conesa, a. Differential expression in RNA-seq: A matter of depth. *Genome Research*, 2011. ISSN 1088-9051. doi:10.1101/gr.124321.111. 9, 92
- Tirosh, I., Reikhav, S., Levy, A. a. and Barkai, N. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science*, 324(5927):659–62, 2009.
 ISSN 1095-9203. doi:10.1126/science.1169766. 73, 74, 82, 84, 85
- Trapnell, C., Williams, B. a., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–5, 2010. ISSN 1546-1696. doi:10.1038/nbt.1621. 166
- True, J. R. and Haag, E. S. Developmental system drift and flexibility in evolutionary trajectories. *Evolution & Development*, 3(2):109–19, 2001. ISSN 1520-541X. 82
- Venkataram, S. and Fay, J. C. Is transcription factor binding site turnover a sufficient explanation for cis-regulatory sequence divergence? *Genome Biology and Evolution*, 2:851–8, 2010. ISSN 1759-6653. doi:10.1093/gbe/evq066. 5

Wang, H.-Y., Fu, Y., McPeek, M. S., Lu, X., Nuzhdin, S., Xu, A., Lu, J., Wu,

M.-L. and Wu, C.-I. Complex genetic interactions underlying expression differences between Drosophila races: analysis of chromosome substitutions. *PNAS*, 105(17):6362–7, 2008a. ISSN 1091-6490. doi:10.1073/pnas.0711774105. 73

- Wang, R.-l. and Hey, J. The speciation history of Drosophila pseudoobscura and close relatives: inferences from DNA sequence variation at the period locus. *Genetics*, 144:1113–1126, 1996. 135
- Wang, R. L., Wakeley, J. and Hey, J. Gene flow and natural selection in the origin of Drosophila pseudoobscura and close relatives. *Genetics*, 147:1091–1106, 1997. 135
- Wang, X., Soloway, P. D. and Clark, A. G. A Survey for Novel Imprinted Genes in the Mouse placenta by mRNA-seq. *Genetics*, 189(September):109–122, 2011. doi:10.1534/genetics.111.130088. 42, 50
- Wang, X., Sun, Q., Mcgrath, S. D., Mardis, E. R., Soloway, P. D. and Clark, A. G. Transcriptome-Wide Identification of Novel Imprinted Genes in Neonatal Mouse Brain. *PLoS ONE*, 3(12):e3839, 2008b. doi:10.1371/journal.pone.0003839. 42, 50
- Wang, Z., Gerstein, M. and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009. ISSN 1471-0064. doi: 10.1038/nrg2484. 7, 8, 15
- Waters, A. J., Makarevitch, I., Eichten, S. R., Swanson-Wagner, R. A., Yeh, C.-T., Xu, W., Schnable, P. S., Vaughn, M. W., Gehring, M. and Springer, N. M. Parent-of-Origin Effects on Gene Expression and DNA Methylation in the Maize Endosperm. *The Plant Cell*, 23(December):4221–4233, 2011. doi:10.1105/tpc.111. 092668. 42, 50

- Watson, J. D. and Crick, F. H. C. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953. 2
- Weksburg, R., Shen, D. R., Fei, Y. L., Song, Q. L. and Squire, J. Disruption of insulin-like growth factor 2 imprinting in Beckwith-Wiedemann syndrome. *Nature Genetics*, 5:143–150, 1993. 42
- Whitehead, A. and Crawford, D. L. Variation within and among species in gene expression: raw material for evolution. *Molecular Ecology*, 15(5):1197–211, 2006. ISSN 0962-1083. doi:10.1111/j.1365-294X.2006.02868.x. 4
- Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J. and Bähler, J. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239– 43, 2008. ISSN 1476-4687. doi:10.1038/nature07002. 8, 134
- Wittkopp, P. J. Genomic sources of regulatory variation in cis and in trans. Cellular and Molecular Life Sciences, 62(16):1779–83, 2005. ISSN 1420-682X. doi:10.1007/ s00018-005-5064-9. 5
- Wittkopp, P. J. Evolution of cis-regulatory sequence and function in Diptera. Heredity, 97(3):139–47, 2006. ISSN 0018-067X. doi:10.1038/sj.hdy.6800869. 5
- Wittkopp, P. J. Using pyrosequencing to measure allele-specific mRNA abundance and infer the effects of cis- and trans-regulatory differences. In *Molecular Methods* for Evolutionary Genetics, volume 772, pages 297–317. 2011. ISBN 9781617792281. doi:10.1007/978-1-61779-228-1. 47, 55, 102

Wittkopp, P. J., Haerum, B. K. and Clark, A. G. Evolutionary changes in cis

and trans gene regulation. *Nature*, 430(6995):85–8, 2004. ISSN 1476-4687. doi: 10.1038/nature02698. 5, 10, 15, 47, 55, 73, 79, 84, 89, 134

- Wittkopp, P. J., Haerum, B. K. and Clark, A. G. Parent-of-origin effects on mRNA expression in Drosophila melanogaster not caused by genomic imprinting. *Genetics*, 173(3):1817–21, 2006. ISSN 0016-6731. doi:10.1534/genetics.105.054684. 12, 43, 55, 94
- Wittkopp, P. J., Haerum, B. K. and Clark, A. G. Independent effects of cis- and trans-regulatory variation on gene expression in Drosophila melanogaster. *Genetics*, 178(3):1831–5, 2008a. ISSN 0016-6731. doi:10.1534/genetics.107.082032. 5, 47, 55, 89, 94
- Wittkopp, P. J., Haerum, B. K. and Clark, A. G. Regulatory changes underlying expression differences within and between Drosophila species. *Nature Genetics*, 40(3):346–50, 2008b. ISSN 1546-1718. doi:10.1038/ng.77. 5, 73, 82, 83, 87
- Wolff, P., Weinhofer, I., Seguin, J., Roszak, P., Beisel, C., Donoghue, M. T. A., Spillane, C., Nordborg, M., Rehmsmeier, M. and Köhler, C. High-Resolution Analysis of Parent-of-Origin Allelic Expression in the Arabidopsis Endosperm. *PLoS Genetics*, 7(6):e1002126, 2011. doi:10.1371/journal.pgen.1002126. 42, 50
- Wood, A. J. and Oakey, R. J. Genomic Imprinting in Mammals: Emerging Themes and Established Theories. *PLoS Genetics*, 2(11):e147, 2006. doi:10.1371/journal. pgen.0020147. 42, 46
- Wray, G. a., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V. and Romano, L. a. The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, 20(9):1377–419, 2003. ISSN 0737-4038. doi: 10.1093/molbev/msg140. 4, 86

- Wu, C.-i., Hollocher, H., T, D. J. B., Aquadro, C. F., Xu, Y. and Wu, M.-L. Sexual isolation in Drosophila melanogaster : A possible case of incipient speciation. *PNAS*, 92(March):2519–2523, 1995. 43, 52, 75, 88
- Wu, T. D. and Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–81, 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq057. 17
- Xing, Y., Ouyang, Z., Kapur, K., Scott, M. P. and Wong, W. H. Assessing the conservation of mammalian gene expression using high-density exon arrays. *Molecular Biology and Evolution*, 24(6):1283–5, 2007. ISSN 0737-4038. doi: 10.1093/molbev/msm061. 87
- Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. and Kinzler, K. W. Allelic variation in human gene expression. *Science*, 297(5584):1143, 2002. ISSN 1095-9203. doi:10.1126/science.1072545. 10
- Zerbino, D. R. and Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–9, 2008. ISSN 1088-9051. doi:10.1101/gr.074492.107. 98
- Zhang, X. and Borevitz, J. O. Global analysis of allele-specific expression in Arabidopsis thaliana. *Genetics*, 182(4):943–54, 2009. ISSN 1943-2631. doi: 10.1534/genetics.109.103499. 73