AN ANALYSIS OF COMPLEX TRAIT VARIATION USING *DROSOPHILA MELANOGASTER*

AN ANALYSIS OF COMPLEX TRAIT VARIATION: WING MORPHOLOGY AND GENE EXPRESSION VARIATION IN *DROSOPHILA MELANOGASTER*

By AMANDA BURKHARDT NEVES, B.H.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University, Master of Science (2022), Hamilton, Ontario (Biology)

Title: An Analysis of Complex Trait Variation: Wing Morphology and Gene Expression Variation in *Drosophila melanogaster*

Author: Amanda Burkhardt Neves, B.H.Sc. Biomedical Science, University of Calgary

Supervisor: Dr. Ian Dworkin

Number of pages: 1- 135

LAY ABSTRACT

Complex traits, like height for example, are difficult to study. Their variation is influenced by the small effects of many genes throughout the genome, by gene-environment interactions, and by gene-gene interactions. In this thesis, I explore complex trait variation using *Drosophila melanogaster*. First, I examine the relationship between variation in gene expression in developing wing tissue and variation in adult wing shape, where the complex trait in question is wing morphology. I also examine gene expression itself as a complex trait and study how its variation is affected by gene-gene interactions and genetic perturbation. Overall, I present a novel way of modelling the relationship between gene expression variation and wing shape variation. I also show that global gene expression variation is in large part correlated not with genotype, as expected, but with genetic background, and that changes in cell size or shape may underlie background-dependent phenotypic effects.

ABSTRACT

Complex traits are traits which vary quantitatively along a normal distribution. Their variation is influenced by many loci throughout the genome, each contributing a small fraction to trait heritability. Complex traits are also shaped by gene-gene interactions between focal alleles with genetic modifiers, as well as gene-environment interactions. In this thesis, I study complex traits from multiple angles using *Drosophila melanogaster* as a model system. First, the relationship between variation in gene expression in developing wing tissue and variation in adult wing shape is assessed, where the complex trait in question is wing shape. Using gene wise multivariate linear models, I show that at the level of natural variation, no single gene's variation in expression has a "significant" effect on variation in wing shape. When genes are grouped into functional categories using Gene Ontology (GO) terms, I show that not only can the signal of effect be recovered, but also that genes from within a GO term group have similar effects on wing shape even when accounting for correlations in gene expression between genes. I also study gene expression and trait expressivity and penetrance under a complex trait framework. An *sd/vg* allelic series is used to study the joint effect of genetic background and the magnitude of allelic perturbation on global gene expression in the developing wing tissue. I show that global transcriptional variation is largely correlated with wildtype genetic background, and not with strength of perturbation as might have been expected. Further, variation in cell shape or cell size are shown to be candidate mechanisms contributing to background-dependent phenotypic variation, and specific genes are suggested for follow-up functional analysis.

ACKNOWLEDGEMENTS

This thesis would not have happened without the support of those I met during my time in Hamilton. Thank you to Katie, Tyler, Brandon, and Arteen who gave me so much support and friendship as well as their scientific insight. Thank you for helping me to be a better scientist and colleague. Thank you to Dr. Ian Dworkin for guidance and support and for always getting me to push further with my analyses. Thank you to my committee for their insight throughout this process. Thank you to my lovely roommates Alex, Garrett, and Mazin for making home always somewhere to come back to and relax from whatever the day brought. Thank you to my CUPE 3906 friends for bringing out the best in me and for showing endless solidarity and compassion during the most stressful times.

Thank you to my parents for everything. Thank you to Daniel for everything.

Thank you to Hannah for the love and support and partnership throughout this process.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

CHAPTER 2

CHAPTER 3

LIST OF ALL ABBREVIATIONS AND SYMBOLS

DGRP – Drosophila Genetics Reference Panel

eQTL – expression quantitative trait loci

GBE – genetic background effect

GP – genotype-phenotype

GWAS – genome-wide association study

ORE – Oregon-R

PCA – Principal Componenta Analysis

sd – *scalloped*

vg – *vestigial*

DECLARATION OF ACADEMIC ACHIEVEMENT

RNA sequencing data from wing imaginal discs (for chapter 2) was provided by Dr. Jason Mezey at Cornell and collected by Yuxin Shi. Wing shape data was provided by Dr. Will Pitchers and collected by several members of the Dworkin lab and the Houle lab. I performed quality control of the reads, trimming of the reads, generated gene counts, and performed statistical analysis of the RNA seq and wing shape data including running linear models and vector correlations. I created all figures.

RNA sequencing data from wing imaginal discs (for chapter 3) was collected by Dr. Chris Chandler. Wings were imaged by various past members of the Dworkin lab. Katie Pelletier generated perturbation values for the wings. Arteen Torabi-Marashi and I developed iterations of the linear models used in this chapter. I performed quality control of the RNA seq data, trimmed the data, generated gene counts, performed statistical analysis of the RNA seq and perturbation data including running linear models and vector correlations. I created all figures.

CHAPTER 1: The motivation underlying studying variation in complex traits and the connection between gene expression variation, complex trait variation, and variation in trait expressivity.

A major interest for many biologists is understanding what makes individual organisms, as well as whole species, different from one another. Specifically, how does variation in genotype contribute to the morphological diversity that is essential for evolution by natural selection? To study this relation ship, the way in which variation in genotype translates to variation in phenotype through a black box of developmental processes is conceptualized in terms of what is called a genotype-phenotype (GP) map (Orgogozo et al. 2015). The simplest GP map one can imagine is a situation where a change in a focal allele results in a dramatic, qualitative change in a phenotype. An example of this scenario is the genetics of tiger coat colour. A single amino acid change in the transporter protein SLC45A2 results in Bengal tigers with white fur and black stripes. This missense mutation leads to blocked transporter channel activity, affecting pheomelanin production and inhibiting the synthesis of red and yellow pigments (Xu et al. 2013). While it makes for an appealing model system to study the GP relationship, cases such as this one where a single mutation has a large, observable effect are rare (Hoekstra 2006).

The genetic basis of trait variation can be thought of as a spectrum. On one end are traits affected by single large effect alleles, such as tiger coat colour. At the other end of the spectrum, trait variation is the result of thousands of loci of small effect spread out across the genome, each contributing a small proportion to heritable trait variance. To understand the full range of phenotypic diversity in nature, we need to also understand the genetic variation underlying the variation at this second extreme, where quantitative, complex traits are found.

Complex traits are influenced by variants spread across the genome (Robinson et al. 2014). Not only are these traits affected by genomic changes, but also by the environment and by gene-gene interactions (epistasis) (Phillips 2008). These traits are difficult to study because of the many factors that influence their variation, and genetic effects can be difficult to quantify using traditional genome-wide association study (GWAS) or expression quantitative trait loci (eQTL) methods because genetic effects are often too small to reach significance thresholds unless extremely large sample sizes are used (Visscher et al. 2017). Complex traits have large mutational target sizes, meaning that there is a large portion of the genome for which mutations can lead to trait variation (Haygood 2006). A consequence of these factors is that studies of complex trait variation often fall very short of accounting for expected trait heritability. For example, over the past 10 years thousands of SNPs have been associated with human height variation (Yang et al. 2013; Yengo et al. 2018, 2022). Together these variants explain only a fraction (roughly 10%) of the heritable height variation that is predicted by twin studies (upwards of 80%) (Manolio et al. 2009). There is a need to develop methods that can utilize other levels of biological variation to understand the variation in complex traits.

In the following chapter of this thesis, I use gene expression as an intermediate trait between genotype and phenotype and study the relationship between variation in gene expression in developing tissue and variation in adult phenotype. The wing of *Drosophila melanogaster* is used as a model complex trait. Wing shape has a large mutational target size, is the result of the combined effect of many developmental processes, and shape can be easily quantified to a high resolution (Carreira et al. 2011; Pitchers et al. 2019). Using gene expression data from developing wing tissue and phenotyped adult wings, I relate variation in mean gene expression across 83 strains of flies to variation in mean wing shape using gene wise multivariate linear models. The sample of *Drosophila* used in this chapter come from 83 isogenic strains from the Drosophila Genetics Reference Panel (DGRP) (Mackay and Huang 2018). By grouping genes into functional categories using GO terms, I can leverage known functional information from these genes to assess the extent to which effects on wing shape of genes within GO term groups are correlated. This also shows how signal that could not be detected at the gene wise level due to small effect sizes can be recovered by grouping genes. Ultimately, this chapter represents the development of multivariate approaches aimed at understanding the relationship between variation in gene expression and variation in complex traits.

In the chapter 3, I continue to explore complex trait variation but rather than using fly wing morphology as the trait of interest, gene expression and trait expressivity are the focal traits. Variation in gene expression can be studied in similar ways as variation in other “classic” complex traits such as height or wing. Like wing shape, gene expression variation is heritable, and genetic factors influence gene expression levels (Nica and Dermitzakis 2013). There is abundant variation in gene expression levels within populations, and variation is affected by the environment and by gene-gene interactions (Oleksiak et al. 2002; Sackton and Hartl 2016; Fournier and Schacherer 2017). Many complex traits display variable penetrance and expressivity. The joint influence of gene-gene interactions and strength of allelic perturbation on gene expression variation and variable penetrance and expressivity is the focus of the third chapter.

To study the effects of genetic background and perturbation on variation in gene expression, I use an established model for studying genetic background effects (GBE) (Dworkin et al. 2009). In *Drosophila melanogaster*, perturbations to alleles of *scalloped (sd)* and *vestigial (vg)* show markedly different wing phenotypes depending on if the allele is introduced in the Samarkand (SAM) or Oregon-R (ORE) wildtype genetic background, two commonly used laboratory strains (Dworkin et al. 2009; Chandler et al. 2017). *Sd* and *vg* are transcription factors which form a heterodimer necessary for the gene expression program that dictates wing development (Halder et al. 1998). Titration of the function of *sd* and *vg* across these two backgrounds has shown that the degree of background dependence is influenced by the severity of the perturbation (Chandler et al. 2017). I use developing wing tissue RNA sequencing data from an allelic series of *vg* and *sd*, representing a spectrum of perturbation effect on wing size. For each gene, I model the effect of genetic background, perturbation strength, and their interaction on gene expression to study the response of global transcription to genetic background and perturbation. In doing so, I highlight the sources of global gene expression variation and suggest genes and mechanisms which may lead to the background-dependent phenotypes observed in this model system.

1.2 REFERENCES

Carreira, V. P., I. M. Soto, J. Mensch, and J. J. Fanara. 2011. Genetic basis of wing morphogenesis in Drosophila: sexual dimorphism and non-allometric effects of shape variation. BMC Dev Biol 11:32.

Chandler, C. H., S. Chari, A. Kowalski, L. Choi, D. Tack, M. DeNieu, W. Pitchers, A. Sonnenschein, L. Marvin, K. Hummel, C. Marier, A. Victory, C. Porter, A. Mammel, J. Holms, G. Sivaratnam, and I. Dworkin. 2017. How well do you know your mutation? Complex effects of genetic background on expressivity, complementation, and ordering of allelic effects. PLOS Genetics 13:e1007075. Public Library of Science.

Cooper, D. N., M. Krawczak, C. Polychronakos, C. Tyler-Smith, and H. Kehrer-Sawatzki. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet 132:1077–1130.

Dworkin, I., E. Kennerly, D. Tack, J. Hutchinson, J. Brown, J. Mahaffey, and G. Gibson. 2009. Genomic consequences of background effects on scalloped mutant expressivity in the wing of Drosophila melanogaster. Genetics 181:1065–1076.

Fournier, T., and J. Schacherer. 2017. Genetic backgrounds and hidden trait complexity in natural populations. Curr Opin Genet Dev 47:48–53.

Halder, G., P. Polaczyk, M. E. Kraus, A. Hudson, J. Kim, A. Laughon, and S. Carroll. 1998. The Vestigial and Scalloped proteins act together to directly regulate wing-specific gene expression in Drosophila. Genes Dev 12:3900–3909.

Haygood, R. 2006. Mutation Rate and the Cost of Complexity. Molecular Biology and Evolution 23:957–963.

Hoekstra, H. E. 2006. Genetics, development and evolution of adaptive pigmentation in vertebrates. Heredity 97:222–234. Nature Publishing Group.

Mackay, T. F. C., and W. Huang. 2018. Charting the genotype-phenotype map: lessons from the Drosophila melanogaster Genetic Reference Panel. Wiley Interdiscip. Rev.-Dev. Biol. 7:e289. Wiley, Hoboken.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753. Nature Publishing Group.

Nica, A. C., and E. T. Dermitzakis. 2013. Expression quantitative trait loci: present and future. Philos Trans R Soc Lond B Biol Sci 368:20120362.

Oleksiak, M. F., G. A. Churchill, and D. L. Crawford. 2002. Variation in gene expression within and among natural populations. Nat Genet 32:261–266. Nature Publishing Group.

Orgogozo, V., B. Morizot, and A. Martin. 2015. The differential view of genotype-phenotype relationships. Front Genet 6:179.

Phillips, P. C. 2008. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9:855–867.

Pitchers, W., J. Nye, E. J. Márquez, A. Kowalski, I. Dworkin, and D. Houle. 2019. A Multivariate Genome-Wide Association Study of Wing Shape in Drosophila melanogaster. Genetics 211:1429–1447. Genetics.

Robinson, M. R., N. R. Wray, and P. M. Visscher. 2014. Explaining additional genetic variation in complex traits. Trends Genet 30:124–132.

Sackton, T. B., and D. L. Hartl. 2016. Genotypic Context and Epistasis in Individuals and Populations. Cell 166:279–287.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 101:5–22.

Xu, X., G.-X. Dong, X.-S. Hu, L. Miao, X.-L. Zhang, D.-L. Zhang, H.-D. Yang, T.-Y. Zhang, Z.-T. Zou, T.-T. Zhang, Y. Zhuang, J. Bhak, Y. S. Cho, W.-T. Dai, T.-J. Jiang, C. Xie, R. Li, and S.-J. Luo. 2013. The Genetic Basis of White Tigers. Current Biology 23:1031–1035.

Yang, J., T. Lee, J. Kim, M.-C. Cho, B.-G. Han, J.-Y. Lee, H.-J. Lee, S. Cho, and H. Kim. 2013. Ubiquitous Polygenicity of Human Complex Traits: Genome-Wide Analysis of 49 Traits in Koreans. PLOS Genetics 9:e1003355. Public Library of Science.

Yengo, L., J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, M. N. Weedon, T. M. Frayling, J. Hirschhorn, J. Yang, and P. M. Visscher. 2018. Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. Hum Mol Genet 27:3641–3649.

Yengo, L., S. Vedantam, E. Marouli, J. Sidorenko, E. Bartell, S. Sakaue, M. Graff, A. U. Eliasen, Y. Jiang, S. Raghavan, J. Miao, J. D. Arias, S. E. Graham, R. E. Mukamel, C. N. Spracklen, X. Yin, S.-H. Chen, T. Ferreira, H. H. Highland, … J. N. Hirschhorn. 2022. A saturated map of common genetic variants associated with human height. Nature 610:704–712. Nature Publishing Group.

CHAPTER 2: Examining the relationship between natural variation in gene expression in the developing wing imaginal disc with variation in adult wing shape in *Drosophila melanogaster*

2.1 INTRODUCTION

The relationship between variation in genotype and variation in phenotype is of interest across multiple disciplines, including in the study of plant and animal genetics, human disease, and trait evolution. Many phenotypes of interest are complex traits. That is that they are highly polygenic, with individual allelic effects usually of small magnitude (Robinson et al. 2014). Untangling the sources of complex trait variation is tricky because complex traits can also be influenced by environmental effects, as well as gene-gene (epistatic effects) and genotype-by-environment interactions, meaning that many of these small effects are context dependent (Koch and Sunyaev 2021). How variation in the genotype causes complex trait variation has been of interest since the use of the word genotype to conceptualize organismal variation unseen in the phenotype, and many methods have been developed to study this relationship (Johannsen 1911).

The Advantages and Limitations of Genome-Wide Association Analysis

Towards the goal of identifying segregating variants associated with complex trait variation, genome-wide association studies (GWAS) have been incredibly useful and successful over the past decade and a half (Visscher et al. 2017). A major, and originally surprising, finding of early GWAS was that complex trait variation is associated with variation in many loci, spread throughout the genome, most often of very small phenotypic effect (Visscher et al. 2017). These associated variants have led to unprecedented insight into disease susceptibility, clinical care, and the genetic architecture of heritable traits (Tam et al. 2019). When performed on large cohorts, GWAS results tend to be robust, replicating across multiple studies with rates around ~40% (Marigorta et al. 2018), an especially important feature given the emphasis on reproducibility in science. Though originally expensive and laborious, sequencing costs have decreased over the years allowing for increasingly larger sample sizes. What has followed is a revolution not only in how trait variation is studied, but also in data collection and world-wide scientific collaboration.

As useful as they are, GWAS are not without their limitations. Though they are well suited to detect highly penetrant mutations as well as common variants with small effect, they struggle to detect rare variants with small or moderate effects, unless sample sizes are extremely large (Tam et al. 2019). Of the many variants detected by GWAS, the variants explaining most trait heritability in are individually very small effect, and in aggregate, still explain only a modest portion of trait heritability predicted by twin studies (Manolio et al. 2009). Furthermore, it is likely that many associated variants are not themselves causal, but in linkage disequilibrium with causal variants, resulting in many false positives (Marigorta et al. 2018; Tam et al. 2019). Due to the burden of multiple testing, GWAS require extremely large sample sizes, which can be prohibitive to the study of rare traits within small populations, or with traits which are expensive or complicated to quantify (Tam et al. 2019). GWAS are undoubtedly useful and important to the study of complex trait variation, but there is a need to investigate what makes organisms phenotypically diverse that looks to other levels of biological variation.

The genetics of gene expression, and variation in gene expression as an intermediate trait between genotype and phenotype

As early as 1975, it has been hypothesized that gene expression differences among species may account for substantial phenotypic differences that are left unaccounted for when examining gene sequence differences (King and Wilson 1975). King and Wilson (1975) reviewed the genetic differences between humans and chimpanzees and concluded that the incredibly high sequence similarity between the two species suggested that amino acid changes could not account for their phenotypic diversity. Thus, the authors suggested that changes to regulatory mechanisms underlying gene expression changes might account for the variation in physiology and behavior that span the two closely related species. Since then, variation in gene expression within and between species, and how this leads to phenotypic variation, has been of great interest (Wray 2007). Initial studies of how gene expression variation contributes to complex trait variation established that there is a heritable genetic basis to variation in gene expression. For example, it has been shown that two polymorphisms in the promoter-proximal transcriptional regulatory region of an allele in *DQB1* alters expression of this gene. It was suggested that the presence of these polymorphisms in several species represent selection for alternate regulatory mechanisms (Beaty et al. 1995). Studies have also shown that there is considerable variation in gene expression, within and among populations. For example 18% of 907 genes within a population of *Fundulus* fish vary significantly among individuals (Oleksiak et al. 2002). In a study of inter-individual variation in genes along human chromosome 21, many of which have been associated with Down syndrome phenotypes, 61% of genes of showed substantial differential expression between "normal" individuals (Deutsch et al. 2005). It is evident that there exists much genetic diversity within and between populations.

Importantly, the ample natural variation in gene expression that exists among individuals is heritable. Expression quantitative trait loci (eQTL) or eGWAS analyses operate under the same framework as GWAS, the difference being that the quantitative trait under investigation is gene expression itself. These studies demonstrate that global gene expression is subject to genetic control, and that gene expression can be considered a complex trait (Brem et al. 2002; Schadt et al. 2003; Rockman and Kruglyak 2006). Just like with other complex traits, gene expression is influenced by many loci and interactions among loci and the environment, underlying a complex inheritance pattern (Brem et al. 2002; Rockman and Kruglyak 2006; Potokina et al. 2008). Given the substantial variation of gene expression within and among populations, as well as its heritability, change to the regulatory structure underlying variation in gene expression has been suggested as a mechanism for rapid evolution among organisms. This is in part because this may be a way to overcome constraints to evolution posed by severe pleiotropic effects of deletions in vital genomic regions (Stamatoyannopoulos 2004; Romero et al. 2012; Hamann et al. 2021). For example, changes to the gene regulatory systems in mammalian tissues has accumulated more rapidly over time across lineages than changes to DNA sequences (Brawand et al. 2011). Further, a study of gene expression profiles of a population of *Brassica rapa* from 1997 to 2014 during a time of rapid climate change in Southern California shows differential gene expression across generations (Hamann et al. 2021).

Gene expression variation has also been shown to be correlated with phenotypic variation. Using a knock-down approach (via weak mutations in non-coding regions of genes), gene expression variation of individual genes in *Drosophila melanogaster* was shown to be significantly (albeit weakly) correlated with changes in wing shape (Dworkin et al. 2011). In mice, changes in gene dosage of *Fgf8* is associated with change in craniofacial shape, where shape changes in response to dosage in a non-linear manner (Green et al. 2017). The Cockayne Syndrome complementation group B (CSB) protein is involved with transcription-coupled nucleotide excision repair, and it has been found to regulate the expression of thousands of neuronal genes (Wang et al. 2014). The majority of patients with Cockayne syndrome, a disorder with severe neurological symptoms, carry mutations in CSB. Gene expression changes have been found in the brains of patients with Cockayne Syndrome, suggesting that dysregulation within gene regulatory networks may be the main cause of the neurological symptoms of Cockayne Syndrome (Wang et al. 2014). Finally, eQTL analysis of GWAS hits reveal that most SNPs identified in GWAS are likely to be associated with the regulation of gene expression (Nicolae et al. 2010; Porcu et al. 2021).

Gene expression variation is abundant, heritable, contributes to trait evolution, and has been correlated to phenotypic variation. Together this suggest that variation in gene expression may provide an alternative way of examining sources contributing to complex trait variation, that remain difficult to explain using standard GWAS. Given that current GWAS methods only explain a fraction of trait heritability predicted by twin studies, and that sample sizes are largely prohibitive, it would be useful to develop alternate approaches to investigate the genetics of complex trait variation. Investigating the role of gene expression as an intermediate trait between variation in genotype and variation in phenotype may be a fruitful next step (Ritchie et al. 2015).

To this end, a few studies have attempted to integrate gene expression information with GWAS approaches, with varying degrees of success. Jin et al. (2016) explored the relationship between extreme variation in gene expression and trait variation. By using information about gene expression across a panel of 369 maize inbred lines as a quantitative phenotype to perform GWAS on, the authors searched for correlations among what they termed expression presence and absence variation (ePAV) and over 600 quantitative traits (Jin et al. 2016). The authors of this study used presence/absence of gene expression as their measure of gene expression variation. This approach, which treats gene expression as a polymorphism with an on or off state, is useful in increasing study power and understanding how large gene expression changes correlate to phenotypic changes. However, this method cannot get at the nuances of variation in gene expression, which is fundamentally quantitative, and thus is blind to the magnitude of differences in expression.

Based on the evidence that GWAS signals are enriched near transcription factor (TF) binding sites, Lin et al. (2017) investigated the degree to which variation in the expression of TFs and other genes across genotypes contributes to phenotypic variation. Using variation in gene expression from 27 genetically diverse maize tissues as an explanatory variable in a GWAS of 13 quantitative traits in the same panel of 369 inbred maize lines, they found that phenotypes predicted by associations correlated highly with the empirically measured traits (Lin et al. 2017). In this study, gene expression was treated as a quantitative trait rather than as expressed or not as was the case in (Jin et al. 2016). This approach, while likely leading to more interpretable and accurate results than treating gene expression as binary, still cannot capture the effect of correlated expression among genes.

Multivariate approaches to examine the relationship between variation in gene expression and trait variation

One important feature to recognize with regards to gene expression is that there is every expectation that many transcripts within and between genes will show a high degree of correlation in expression. Whether due to physical proximity in a shared “operon” like feature, domains of open chromatin, or shared regulatory architecture (i.e. shared transcription factors), many genes will be co-expressed (Sabatti et al. 2002; Lercher et al. 2003). As such, examining co-variation in expression remains essential (van Dam et al. 2018). Multivariate approaches that can account for such correlation structures, provide a promising avenue to evaluate changes. In addition to the consideration of co-variation, one can consider both the magnitude and direction of effect vectors – as opposed to treating gene expression as a binary trait or investigating only the magnitude of the effect. A way to do this utilizing multivariate statistics is to analyze the vector of the effect of change in gene expression on change in phenotype (Kuruvilla et al. 2002; Zinna et al. 2018). The magnitude of this vector gives information on the strength of this relationship on a continuum. Additionally, the directions of these vectors can be compared, and one can compute a vector correlation coefficient (or equivalently an angle between vectors) to determine similarity of effects based on the correlation in the direction of the effect (Kuruvilla et al. 2002; Zinna et al. 2018).

Analysis of these vectors, as well as the ability to detect associations between complex traits, greatly benefits from deeply detailed phenotyping

(Houle 2010). This can be accomplished through the analysis of multivariate traits (Shriner 2012; Topp et al. 2013; Pitchers et al. 2019). Natural selection, as an ultimate causation of variation in complex traits, takes place in a broad and multidimensional space. To get at a richer understanding of complex trait variation, studies can go beyond univariate, or over-simplified phenotypic descriptions, and embrace a phenomic approach, capturing multidimensional phenotype effects (Bilder et al. 2009; Houle 2010). Examining wing shape in multiple dimensions, rather than collapsing multiple dimensions of shape variation into the first few principal components (PCs), has been shown to increase the power to detect associations between polymorphisms and trait variation (Pitchers et al. 2019). Additionally, using a subset of PCs rather than all dimensions of shape variability limit interpretability as this removes the ability to meaningfully think about the directions of effect in relation to the actual phenotype. Characterizing a phenotype through many dimensions may be a more powerful approach for dissecting complex trait variation than relying on few key measures of a phenotype.

In this chapter, I explore gene expression as an intermediate trait between genotype and phenotype using *Drosophila melanogaster* wing shape as a model system. *Drosophila melanogaster* wing shape is a complex trait that is shaped by various genetic signalling pathways and through multiple cellular and developmental processes such as wing patterning and vein specification (Matamoro-Vidal et al. 2015), with a very large mutational target size (Birdsall et al. 2000; Zimmerman et al. 2000; Mezey and Houle 2005; Weber et al. 2005). The study sample is comprised of ~10,000 measured wings from 83 strains of flies and line-matched RNA sequencing data from 3rd instar wing imaginal discs (sacs of cells set aside in the embryo that ultimately go on to form the wing blade, and parts of the body wall). Using gene wise multivariate linear models to estimate the vector of the effect of variation in mean gene expression across strains on variation in mean wing shape across strains, I examined the individual impact gene expression has on wing shape variation. Using both biological and molecular gene ontology (GO) terms, I grouped genes into functional categories, and ask if the direction of the effect vectors for genes in a GO term group are more correlated with each other than expected if genes were grouped randomly, and if this effect holds true when correlation in gene expression is accounted for.

2.2 METHODS

Drosophila strains

The DGRP, a set of inbred lines derived from iso-female lines collected at a farmer's market in Raleigh, NC (Mackay et al. 2012) was used for this study. 83 strains from the DGRP were used which contained both RNA sequencing data from imaginal wing disc tissue as well as morphometric data from the adult fly wings. The 83 different lines each represent a snapshot of the natural genetic variation in the population, and as such are an ideal tool for understanding how natural variation in gene expression affects natural variation in wing shape.

Rearing, fly handling, wing imaging, and morphometric data

For further details, see Pitchers et al. (2019) (from which all adult phenotypic data was derived). Flies from the Dworkin lab were reared at 24°C in bottles on a cornmeal-molasses-yeast-based medium, using carrageenan as a gelling agent and propionic acid and methyl paraben as preservatives. After adult flies eclosed and completed sclerotization, they were preserved in 70% ethanol prior to dissection, mounting, and image analysis.

As per Pitchers et al. (2019), a modified protocol from Houle et al. (2013) was used to landmark and semi landmark data. Nine cubic B-spline functions were fit to the wing veins and margins of the imaged wings using Wings 3.72 (Van der Linde 2004-2014) CPR (Marquez 2012-2014) was used to obtain 14 landmark and 34 semi landmark positions from fitted splines. Generalized Procrustes superimposition (Rohlf and Slice 1990) was performed on combined shape data from the Dworkin and Houle labs, validation, and replication data sets. This process scales wings to a common centroid size to minimize the effect of size, translates wings to the origin to remove the effect of location, and iteratively rotates wings to the orientation of a selected configuration until the sum of squared landmark distances is no longer significantly reduced. This is important to remove nuisance parameters from the data such that useful size and shape information remains. We are then left with 58 dimensions of shape to analyze. For each DGRP line, replicates were averaged. A male and female mean wing shape was calculated for each line, and the sex means were averaged (since RNAseq data was not sexed).

Wing Imaginal Disc RNA Sequencing Data

Flies for RNA sequencing were reared in the lab of Dr. Jason Mezey at Cornell University. All data collection and submission of samples for sequencing was done by Yuxin Shi, in the lab of Dr. Mezey. Wing imaginal discs were dissected from late 3rd instar larvae, collected just prior to the initiation of pupation (i.e. wandering larval stage). Pools of about 20 wing imaginal discs from each DGRP line was used for sequencing for each sample. Tissue was sequenced to a depth of 10-12M single-end reads per samples. Fastq files were trimmed using the BBduk tool from BBMap (ver. 38.90). To determine the quality of RNA from each sample, a transcript integrity number (TIN) was calculated for each transcript in each sample using RseQC (ver.4.0.0) on STAR (ver. 2.7.9a) aligned reads, and for each sample a median TIN was computed. Most samples had a median TIN above 74, and no sample had a TIN of below 65, where 60 can be considered a threshold for acceptable quality (Figure S1) (Wang et al. 2016). Therefore, all samples were kept in the study. Transcript counts were generated using Salmon (ver. 1.4.0) with the option to include a decoy. The decoy uses genomic sequences, as well as the transcriptome, to reduce spurious mapping of reads from unannotated genomic loci that are sequence-similar to an annotated transcriptome. Transcript counts were collapsed to length-scaled transcript per million gene-level counts using tximport (ver. 1.20.0) in R. Where there were replicates for the same DGRP line, a line mean expression for each gene was calculated. However, for most samples there was only one sequence file (but note that each sequence file is pooled and represents RNA from upwards of 20 imaginal discs). Genes with 0 expression across all samples were excluded from the analysis, resulting in a total of 12936 genes used for the linear models. See Table 1 for more parameter information for each tool.

Multivariate Linear Models and Gene Expression Vectors

To detect associations between variation in gene expression and variation in wing shape, a multivariate linear model was run for each of the 12936 genes. For a schematic of the workflow of analysis, see Figure 1.

To account for the influence of size allometry on wing shape, $\log_2$ centroid size (strain means) was used as a predictor. Some pairs of lines in the DGRP are more closely related to each other than other DGRP lines (Huang et al. 2014). To account for this, two approaches were used. First, freeze 2 genotypes from

February 2013, publicly available from (ftp://ftp.hgsc.bcm.edu/DGRP/freeze2_Feb_2013/vcf_files/freeze2.vcf.gz) (Huang et al. 2014), were used to calculate group structure among the DGRP for use as a covariate in the linear models. PLINK 2.0 (ver. 2.0.a3) was used to prune variants in the freeze 2 genotypes in high linkage disequilibrium (LD). A PCA on the genotypes was run to analyze major axes of variation in allele frequencies, and the first two eigenvectors from this analysis was used in the linear models. Second, the three most common chromosomal inversions (In(2L)t, In(3R)K, and In(3R)K) were also accounted for, as there is evidence that they influence variation in wing shape (Pitchers et al. 2019). Although Wolbachia (a common intracellular bacteria) infection status has not been shown to affect wing shape, information on whether each strain of the DGRP was infected with Wolbachia or not was included as a variable in the models.

Thus, the final multivariate linear model used for each gene was:

$$\boldsymbol{y}_i = \boldsymbol{\beta_0} + \boldsymbol{\beta}_{h,1} x_{h,i1} + \sum_{j=2}^{8} \boldsymbol{\beta}_{h,j} x_{ij} + \boldsymbol{\epsilon}_{h,i}$$

Where $\boldsymbol{y}_i$ is the vector containing the mean shape for the *ith* DRGP strain. $\boldsymbol{\beta}_0$ is the model intercept, and $\boldsymbol{\beta}_{h,1}$is the vector of estimated effects for expression of gene*h*. All remaining model coefficients ($\boldsymbol{\beta}_{2,h} - \boldsymbol{\beta}_{8,h}$) are associated with the effects of the 7 predictors common to all models as described above. These estimates are likely similar across models (gene-to-gene), except in the presence of strong correlations between gene expression for the h^{th} gene, and one or more of these predictors. $\boldsymbol{\epsilon}_{i,h}$represents the vector of unmodeled (residual) variation.

The vector of estimated coefficients associated with the predictor for gene expression, $\hat{\beta}_{h,1}$, was extracted from each model. This vector represents the effect that a gene's change in mean expression across DGRP strains has on mean wing shape across the strains. The magnitude of this vector (l^2 norm) was computed for each gene, describing the overall magnitude of this effect across all landmarks. The correlation in direction of these vectors was also examined to assess how similar the effect of gene expression on wing shape is for genes of the same group (more about this below). Analyses based on these vectors are split into two parts: in the first part the gene wise effects of change in expression on change in wing shape are considered, and in the second part the correlation of effect across groups of genes is studied.

Detecting gene expression effects on wing shape at the single gene level

To calculate the magnitude of the effect of the change of a gene's expression on the change in wing shape, the l^2 norm of each vector of model coefficients was calculated, where the magnitude is equal to square root of the sum of squared model coefficients, for each vector.

To assess whether any gene had a "significantly" large magnitude of effect on wing shape, compared to what is expected under the null of no effect, a permutation test was done to calculate a null distribution, after (Churchill and Doerge 1994). Briefly, 1000 permutations were run, where for each iteration, the vector of gene expression measures for a given strain ($\boldsymbol{x}_i$) were randomly assigned to different strains, while all other components (response $\boldsymbol{y}_i$ and other predictors $\boldsymbol{x}_2 - \boldsymbol{x}_8$) were left unchanged. As such the correlation structure among genes was maintained, while gene expression would no longer have any association with shape (or other predictors). After trait values were re-assigned, new linear models were fit for all genes. For each permutation, the highest vector magnitude was recorded (across all genes from that iteration of the permutation). At the end of the 1000 permutations, the highest values from each permutation are ordered, and the 95th percentile of these values is used as the threshold for significance (analogous to an alpha of 0.05, corrected for multiple comparisons, but accounting for correlational structure among genes). As the shuffling of the trait values retains the covariance structure among gene expression, this method was considered appropriate.

Detecting gene expression effects on wing shape for groups of genes

Genes do not act in isolation, and especially in the case of a sample without a pronounced genetic perturbation (i.e. a severe mutation, or RNAi), it is not expected that a single gene should exert a strong effect on a complex trait such as wing shape. Thus, genes were grouped based on GO terms extracted from FlyBase (ver. FB2022_05) to assess if effects among genes from this grouping were correlated, allowing for leveraging of existing knowledge of the functions of groups of genes to elucidate the causes of variation in wing shape at a more holistic level (Aponte et al. 2021).

Groups were chosen to represent a variety of number of genes in each group (ranging from 11 to 107 genes) and both molecular and biological group ontologies related to wing development were considered. The biological groups were chosen based on prior evidence of an effect on wing shape of their constituents. The groups considered were “cell elongation involved in imaginal disc-derived wing morphogenesis ” (cell_elong), “imaginal disc-derived wing vein morphogenesis” (wing_vein), “regulation of imaginal disc-derived wing size” (disc_size). The molecular groups considered were “BMP Signalling Pathway” (bmp), “Hedgehog Signalling Pathway” (hh), “Hippo Signalling Pathway” (hippo), “Insulin-Like Receptor Signalling Pathway” (insulin), “EGFR Signalling Pathway” (egfr), and “Wnt-TCF Signalling Pathway” (wnt). See Table 2 for more information on each GO term group.

To assess whether the effect on wing shape of genes in a group would be more correlated than expected for genes not grouped by GO terms, the pairwise correlations between vectors of effect of gene expression on wing shape for genes in a GO term group were calculated. A correlation matrix was computed for each group, and the absolute value of the off diagonals (upper diagonal), of this matrix were used to compute a mean correlation value for each group. Absolute values were chosen as genes were considered to have a correlated effect regardless of whether this correlation was negative or positive.

To compare the mean correlation values for each group to an appropriate control distribution, we took two approaches. The mean correlation values for each group were compared to two comparison distributions. The first was a random gene distribution, where for each GO term group, a group of random genes was constructed such that the new pathway had the same number of genes. The reason for this control group was to account for correlations that might occur due to a low number of genes in a group, as it is expected that higher correlations will be seen in the groups with fewer genes due only to this low number of genes rather than true correlations. First, all genes involved in the GO term groups were removed from the pool of genes from which sampling was to occur. Where *n* is the number of genes in that GO term group, *n* random genes were selected 1000 times, so that for each GO term group there were 1000 random gene comparison groups. For each of the 1000 permutations, a mean vector correlation value was calculated (as above), and the 97.5th and 2.5th percentile value of these 1000 permutations for each gene-length matched group was used as what we will refer to as the "random distribution".

To account for patterns of gene co-expression (i.e. covariation in gene expression) among the GO term gene groups, which is expected to account for at least some of the correlations seen in the effect vectors, a “matched gene-expression correlation group” distribution was also created for each GO term group. A mean gene *expression* correlation value was calculated for each GO term group, where pairwise correlations in gene expression for the genes in each GO term group were calculated and used to compute a correlation matrix. The absolute value of the upper diagonals from this matrix was considered the mean gene expression correlation value. For each GO term group, this mean gene expression correlation value was the seed value. A random gene (from the pool of genes excluding those genes in the GO term groups) was selected, and a correlation matrix of this gene’s expression with all other genes was computed. If there are multiple genes which are correlated at a value of the seed or higher, one was selected from random. This random gene becomes the next seed, and the process is repeated until the group is of length *n* genes. This was done 1000 times for each GO term group. For each of the 1000 new groups, the mean vector correlation was calculated and the 97.5th and 2.5th percentile values were selected to create the distribution referred to as the “matched gene expression distribution”.

The magnitude of the effect of change in gene expression on change in wing shape (l^2 norm of the effect vector) was calculated for each GO term group, and for each comparison distribution (where the 97.5th and 2.5th percentile values were chosen to form the distribution).

To compare the directions of the effect of change in gene expression on change in wing shape to the directions of natural variation in shape change, a linear model was fit to the landmark data with $\log_2$ centroid size as the sole predictor. A principal components analysis (PCA) was conducted on the residuals of the model, which represent shape variation after accounting for the influence of shape~size allometry. The scores from PCs one through five were then correlated to the effect vector from each gene in each GO term group. The random gene group method described above was used to create a comparison distribution.

2.3 Results

There is no evidence to suggest that the magnitude of the effect of variation in gene expression on variation in wing shape among the DGRP is significantly higher for any one gene

To assess the magnitude of the effect of variation in gene expression in developing wing tissue on variation in adult wing shape, gene wise linear models were performed and the l^2 norm from these models was used as the magnitude. A Churchill and Doerge (1994) permutation test was run on the l^2 norm from all genes, and it was found that no gene passed the significance threshold determined from the permutation test. In fact, there was almost no overlap between the actual distribution of l^2 norms and the distribution of values from the permutation test (Figure 2). This is odd given that under a state of no significance, it is expected that these two distributions be the same. However, when assessing the outcomes of distributions from seven individual permutations, the highest l^2 norm computed gives the distributions an extreme right tail (Figure S2). As these are the values from which alpha is selected, the distributions in Figure 2 make sense. Regardless, this method of assessing for strong effects suggests that under natural variation the effect of any single gene on wing shape does not stand out as extreme.

Pairwise correlations in gene expression among genes in GO term groups are higher than that of genes in both the random gene group and matched correlation in gene expression group, this is partially accounted for by the latter distribution

Due to the reasoning that genes and gene products do not act in isolation, and therefore it is expected that correlations in effect exist among certain genes, biological and molecular GO terms were used to group genes into functional groups (Table 2). As the goal of this analysis was to determine if there was correlation in the direction of the effect of variation in gene expression on variation in wing shape for genes within a GO term group, two comparison distributions were also considered. The intention of the matched correlation in gene expression group was to account for some of the correlation seen in the GO term group that might be due to genes from these groups having correlated gene expression. In the Hippo pathway for example, it is expected that as *hpo* expression increases, *wts* expression increases as *hpo* activates *wts*.

Analysis of the distribution of pairwise correlation in gene *expression* among genes in the GO term groups, random gene groups, and matched correlation in gene expression group (Figure 3) shows that the distribution of these correlations in the GO term groups is varied (seen in the violin plots that span almost the entirety of the Y-axis). As expected, the 97.5th and 2.5th percentile values for mean pairwise correlation in gene expression from 1000 permutations shows that the magnitude of correlations in gene expression for the random gene groups is low. This distribution for the pairwise correlation in gene expression group spans the mean for the GO term groups, though does not capture the higher ends of this correlation. Therefore, the matched correlation in gene expression group accounts for most of the correlation in gene expression seen in the GO term group, but not all of it.

Genes grouped by GO term have a high correlation in their direction of effect on wing shape when compared to genes grouped randomly or genes grouped by matched correlation in gene expression. This correlation is particularly high for hippo and hh signalling genes.

The mean pairwise correlation in direction of effect between genes within a GO term group is higher than the 95% of highest mean pairwise correlations from 1000 permutations of matched gene-length random gene groups (Figure 4). When compared to the top 95% from 1000 permutations of the matched correlation in gene expression groups, the genes from GO term groups are consistently on the higher end of these distributions. The mean pairwise correlation in direction of effect for genes within the *hippo* and *hh* GO term groups are particularly high, reaching the 97.5th percentile for their respective matched correlation in gene expression distributions.

The magnitude of the effect of change in gene expression on change in wing shape is not necessarily higher among genes within GO term groups when compared to the comparison distributions.

To determine if the magnitudes of effect from genes within GO term groups were higher than of the two comparison groups, the l^2 norms from genes within each group were calculated. There was no evidence to suggest that the l^2 norms from the GO term groups were high when compared to 95% highest l^2 norms from 1000 permutations of the comparison groups (Figure 5).

There is a correlation between the direction of effect vectors in a GO term group and the axes of natural shape variation, but this correlation is not elevated when compared to random sets of genes

The distribution of absolute pairwise correlations in direction of effect between genes in each GO term group with the first five PCs of shape among the DGRP reveals that there is modest correlation among the directions of effect and the directions of natural variation in shape (Figure 6). This correlation is the highest among PC 3 for all GO term groups. However, when contrasting the distributions from the GO term groups to distributions made up of the 97.5th and 2.5th percentile from 1000 matched-gene length random gene groups, the correlation among the GO term group genes is not higher than expected based on this comparison.

2.4 DISCUSSION

Given the abundance in variation in gene expression within and between populations, the documented correlations between gene expression and phenotypic variation, and the need to develop methods that look at levels other than amino acid sequence variation with high-dimensional phenotypes, the goal of this chapter was to use multivariate statistics to examine gene expression as an intermediate trait between genotype and phenotype.

Though there was a lack of signal detected at the gene wise level, this result remains inconclusive

The permutation test done on the magnitudes of the effect vectors from the gene wise multivariate linear models found that no gene passed a significance threshold of $\alpha = 0.05$ (Figure 2). There are a few important things to note about this result. First, at face value, this suggests that at the level of natural variation, the effect of variation in gene expression on variation in wing shape is not concentrated to a few genes but is likely weakly spread out throughout the transcriptome. This interpretation is in line with the many findings from GWAS and QTL over the years that SNPs throughout the genome contribute small fractions to trait variation. For traits with a large mutational target size, like wing shape, this could mean that small expression changes in many genes contribute to shape variation.

However, this finding should be tempered by the fact that although each individual RNA sequencing sample represents a pool of about ~20 individual wing discs per strain, there were replicate samples for only 4 of the 83 strains. Therefore, within line gene expression variation could not be accounted for meaningfully, confounding results and making associations more difficult to detect. The use of only 83 strains weakens power further. Though the use of geometric morphometrics allowed for wing shape to be measured multidimensionally and at an extremely high resolution, the change in wing shape and gene expression being dealt with are at the level of natural variation, and thus any changes are expected to be very subtle. Genes within GO term groups did not show a higher magnitude of effect than of the comparison groups, a possible interpretation of this result is that the magnitudes are truly too small to detect anything easily, especially without sufficient replicates or strains.

Given the above limitations, it is difficult to say whether the finding that the magnitude of the effect of change in gene expression across strains on change in wing shape across strains is not significant for any one gene is due to biological reality or statistical artifact. Nevertheless, given evidence from GWAS of complex traits and studies of wing mutational target size, I do expect that the variation in wing shape at the level of natural variation is spread across the transcriptome rather than few genes with large effect, and this finding does not contradict this. Dworkin (2011) found that relatively low changes in gene expression were very weakly correlated with changes in wing shape, like the current findings. Therefore, even given hundreds more DGRP strains and plentiful strain RNA seq replicates, I expect similar magnitudes would have been estimated. That being said, such a high sample size would lead to significance likely being observed due to smaller sampling uncertainty.

The effect of variation in gene expression is highly correlated in genes grouped by functional categories, and this correlation is only partially explained by a correlation in gene expression

As discussed above, a likely explanation for the results at the single gene level was that the sampling variation was high given the limited sample sizes. Genes were grouped according to GO terms (and as such have related effect) to assess if signal could be recovered (i.e. in aggregate, sufficient signal could be recovered from the noise). Coordinated gene effects are important in gene regulatory and signalling networks. Grouping genes by known molecular and biological processes leverages *a priori* knowledge of functional genomics with the high dimensionality of multivariate and complex phenotypes (Aponte et al. 2021). While this approach necessarily ignores genes which are not known to have obvious roles in these pathways, limiting the ability for novel gene discovery as is available in GWAS, the advantage is that the results are interpretable in terms of the developmental mechanisms and processes underlying complex trait variation (Aponte et al. 2021).

To this end, nine GO term groups were assigned based on molecular and biological GO terms (Table 2). The pairwise correlations in direction of effect vectors for each of these groups were assessed and compared to two comparison groups, with one accounting for spurious correlations due to a low number of genes in a group and the other accounting for some of the correlations due to correlated gene *expression* among genes in a GO term group (Figure 3).

The results from this analysis (Figure 4) show that the effects of variation in gene expression among genes in a GO term group are more correlated with each other than is the case for genes in random groups or genes grouped to have similar levels of correlation in gene expression. This finding underscores the utility of harnessing levels of information, in this case knowledge of gene function, gene expression, and a multivariate phenotype. This finding also suggests that genes sharing common signalling pathways or are a part of shared biological processes shape complex traits in similar ways, and that this is an effect that is not due only to having correlated patterns in gene expression.

This correlation might be high relative to the comparison groups because genes from these GO term groups are known to be important for development, and perhaps the functional redundancy in terms of direction of effect can contribute to phenotypic robustness. It could be that proteins encoded by these genes interact in ways that shape the wing similarly, as protein levels do not necessarily correlate highly with gene expression levels (Gry et al. 2009). It has been shown that protein variation among the DGRP is far lower than genetic variation, and protein levels are associated with wing size in *Drosophila melanogaster* (Okada et al. 2016), supporting this interpretation.

In an analysis of the effect of genomic variants from biological and molecular groups and multivariate mouse craniofacial shape, it was found that in some pathways, the variation in shape was loaded heavily on a few genes, while in others this variation was spread out more evenly(Aponte et al. 2021). It would be interesting to investigate the degree to which some genes may be driving high correlations within the GO term groups analyzed here through a process of removing a gene at a time and re-calculated the mean pairwise effect correlation.

Genes from the Hippo and Hedgehog signalling pathways are especially correlated in their effect on wing shape

The genes from the Hippo signalling pathway stood out as having very high mean pairwise effect vector correlation, an interesting result given that genes from the Hippo signalling pathway have been shown to have significant effects on wing shape variation through SNP analysis (Pitchers et al. 2019; Pelletier et al. 2022). Genes from this pathway were originally discovered for their roles as tumor suppressors, and have since been linked to control of organ size and cell fate (Misra and Irvine 2018). Hippo pathway regulation is complex, and multiple serine/threonine kinases along with *Hippo* can phosphorylate and activate *Warts,* leading to the cascade of effects upon which initiate transcriptional regulation of genes involved in proliferation and anti-apoptosis by the complex formed by *Yorkie* and *Scalloped* (Chen et al. 2020). Hippo pathway components have been conserved through animal evolutionary history, underscoring its importance (Chen et al. 2020). In the *Drosophila melanogaster* wing discs, Yorkie activity is associated with patterns of cell proliferation, where increased Yorkie activity has been shown to lead to wing overgrowth through increased cell proliferation (and a decrease in Yorkie activity leading to a decrease in wing growth) (Pan et al. 2018).

Like the Hippo pathway, the Hedgehog (Hh) signal transduction pathway is involved with cellular growth, division, lineage specification, and survival, and the pathway operates through a complex series of mechanisms (Sasai et al. 2019). Mutations to genes involved in the Hh pathway can lead to developmental defects and cancer (Ogden et al. 2004). In *Drosophila melanogaster*, Hh signalling is necessary for proper patterning of the embryo and adult structures including the wing. An Hh morphogen gradient forms along the wing imaginal disc and patterns wing veins, intervein space, and wing bristles (Ingham and McMahon 2001; Cohen 2003; Ogden et al. 2004). Furthermore, Hedgehog produced by the wing imaginal disc has been shown to produce cell type-specific responses in other tissues, for example inducing signal transduction in myoblasts (Hatori and Kornberg 2020).

In terms of its effects on complex traits, the Hippo pathway has been involved in craniofacial and tooth development due to the crosstalk of Hippo members and cranial neural crest cells (Wang and Martin 2017; Dema et al. 2020). The Hh pathway is also necessary for proper development in the vertebrate face, and aberrations in the pathway are involved with craniofacial disorders such as cleft lip and palate (Cobourne and Green 2012; Xavier et al. 2016). Hh signalling is also implicated in tooth development (Pan et al. 2013; Xavier et al. 2016). These findings show that both pathways are implicated in complex trait variation and that members of these pathways play a role in quantitative variation of shape and size in various organisms and tissues (Pelletier et al. 2022).

Given their evolutionarily conserved nature, importance in embryo patterning and control of organ growth, and contributions to complex trait variation, perhaps it is not surprising that genes from the Hippo and Hh pathways show up as having especially correlated effects in the current study. Both continue to show up in studies of craniofacial and *Drosophila melanogaster* wing shape (Pelletier et al. 2022). An interpretation of the finding of highly correlated direction of effect of variation in gene expression on variation in wing shape for genes within the Hippo and Hh pathways is that for these pathways to be able to exert their functions across a wide variety of traits properly, their functions must be similar. Otherwise, one might imagine that the trait will break down. This correlation might have been built up over evolutionary history in each of these pathways such that it is integral, and the correlated effect of gene expression is a by-product of this. Investigating the role of Hh signalling with regards to wing shape will be an interesting follow up. Genetic screens of the members of this pathway could be conducted and assessed for their individual effects on wing shape. The direction of the effect of each mutant or RNAi knockdown on wing shape could be calculated relative to wildtype, and the correlation between these could be assessed to determine whether this recapitulates the result from the current study, adding a functional component to this work (directions of effect are a powerful way to assess study repeatability, as in Pelletier et al. (2022)). This might also reveal whether certain members are driving the observed variation or whether it is spread out throughout all members. This would also determine whether the direction of effect of mutant variation is correlated to the direction of the effect of natural variation, as has been reported by others (Aponte et al 2021). This would be an interesting approach to follow-up with the results from this study that the direction of the effect vectors from GO term groups were somewhat aligned with the first five PCs of natural variation in wing shape, but that this was not elevated with respect to random gene groups.

Study Limitations and considerations

As mentioned above, effect sizes dealt with in this study are small. A consequence of this is that vector directions are more difficult to estimate than if the effects were larger. This means that there might be many false negatives with respect to correlation in direction of effect vectors, but also that there is a good degree of confidence in the higher correlations found. This could also explain the inability to find meaningful effects when considering magnitude (via l^2 norms of effect vectors), rather than direction of effects. There is also the fact that this study does not have enough within strain replicate RNA sequence data to meaningfully account for within strain variation, limiting the ability to estimate within strain gene expression variation relative to between strain variation. Finally, though this study provides interesting avenues for future exploration, such as investigation of the Hippo and Hh pathways, and of correlated effects in gene expression on complex traits broadly, it remains to be seen how these results hold up to functional analysis and what the mechanisms driving this correlated effect are. This will be a key future direction.

Conclusion

The goal of this chapter was to investigate the effect of variation in gene expression on variation in complex trait variation, and the extent to which gene expression serves as an intermediate trait between genotype and phenotype, using RNA sequencing data from developing wing tissue and phenotyped adult wings. Though it was found that the effect of variation in gene expression on variation in wing shape was not significant for any one gene (perhaps due to small effect magnitudes and limited sample size), when genes were grouped by GO terms it was found that the pairwise correlations in direction of effect vectors were higher for genes within GO term groups than compared to randomly grouped genes or genes grouped by correlation in gene expression. Genes from the Hippo and Hh signalling pathways were especially correlated and given the involvement of these pathways in complex trait expression including wing and craniofacial shape, these would be interesting future candidates for study from a gene group perspective.

2.4 TABLES AND FIGURES

Software(version)	Purpose	Parameters
BBMap (38.90)	Data-quality-related trimming (filtering, adapter trimming, and contaminant filtering via kmer matching)	Settings: threads = 8, ktrim = r, k = 23, mink = 10, hdist = 1, qtrim = rl, trimq = 15, minlength = 36
Salmon (1.4.0)	Generate counts data for gene expression analysis	*D. melanogaster* transcript version = r6.38, *D. melanogaster* genome version (for decoy) = r6.38

		Settings: -p = 16, --validateMappings, --rangeFactorizationsBins = 4, --seqBias, --gcBias, --recoverOrphans
STAR (2.7.9a)	Align sequences for RseQC quality control	Default
RseQC (4.0.0)	Compute transcript integrity number to assess RNA quality across samples	Default
Tximport(1.20.0)	Collapse transcript counts to gene-level	countsFromAbundance = lengthScaledTPM
PLINK 2.0 (2.00a3)	Prune for variants in linkage disequilibrium, run principal components analysis to obtain eigenvectors of shared	Pruning: --indep-pairwise 50 10 0.5 PCA: Default with output of pruning step as the input for --extract

	population stratification across DGRP	
R (4.1.0)	Data analysis and creating figures	

Table 1. Additional information for tools used in data processing and analysis pipeline

Gene Group	Shorthand	Number of genes	Description
Cell elongation involved in imaginal disc-derived wing morphogenesis (biological process)	cell_elong	11	The process in which a cell elongates and contributes to imaginal disc-derived wing morphogenesis

Regulation of imaginal disc-derived wing size (biological process)	disc_size	13	Any process that modulates the size of an imaginal disc-derived wing (Lee et al 2011)
Imaginal disc-derived wing vein morphogenesis (biological process)	wing_vein	44	The process in which anatomical structures of the veins on an imaginal disc-derived wing are generated and organized
Bone mineral protein (BMP) signalling pathway (molecular process)	bmp	57	The series of molecular signals initiated by the binding of the BMP family to a receptor on the surface of a target cell, and ending with the regulation of a downstream cellular process (Cordero et al 2007); BMPs regulate cell shape and induce

			multiple cell types at distinct positions which then go on to form the adult wing veins (Raftery and Umulis 2012)
Hippo signalling pathway (molecular process)	hippo	75	The series of molecular signals mediated by the serine/threonine kinase Hippo or one of its orthologs (Zeng & Hong 2008, Pan et al. 2018, Yu & Guan 2013)
Insulin-like Receptor signalling pathway (molecular process)	insulin	76	ILPs are important regulators of metabolism, growth, reproduction, and lifespan (Shingleton et al. 2005)
Hedgehog signalling pathway	hh	89	Initiated by hh ligand binding to the extracellular domain of pathed receptor, leading to the

			depression of smoothened activity. Signalling is required for the survival of cells in the wing disc. (Ingham 2016, Chen & Jiang 2013, Hatori & Kornberg 2020, Lu et al. 2017)
Wnt-TCF signalling pathway	wnt	107	Activation of the pathway leads to the inhibition of arm degradation and its subsequent accumulation in the nucleus, where it regulates the transcription of target genes. In wing imaginal discs, wingless signalling affects cell shape (Swarup & Verheyen 2012, Widmann & Dahmann 2009)

Table 2. Table of GO term gene group information, including shorthand used in figures, number of genes in each group, and brief description.

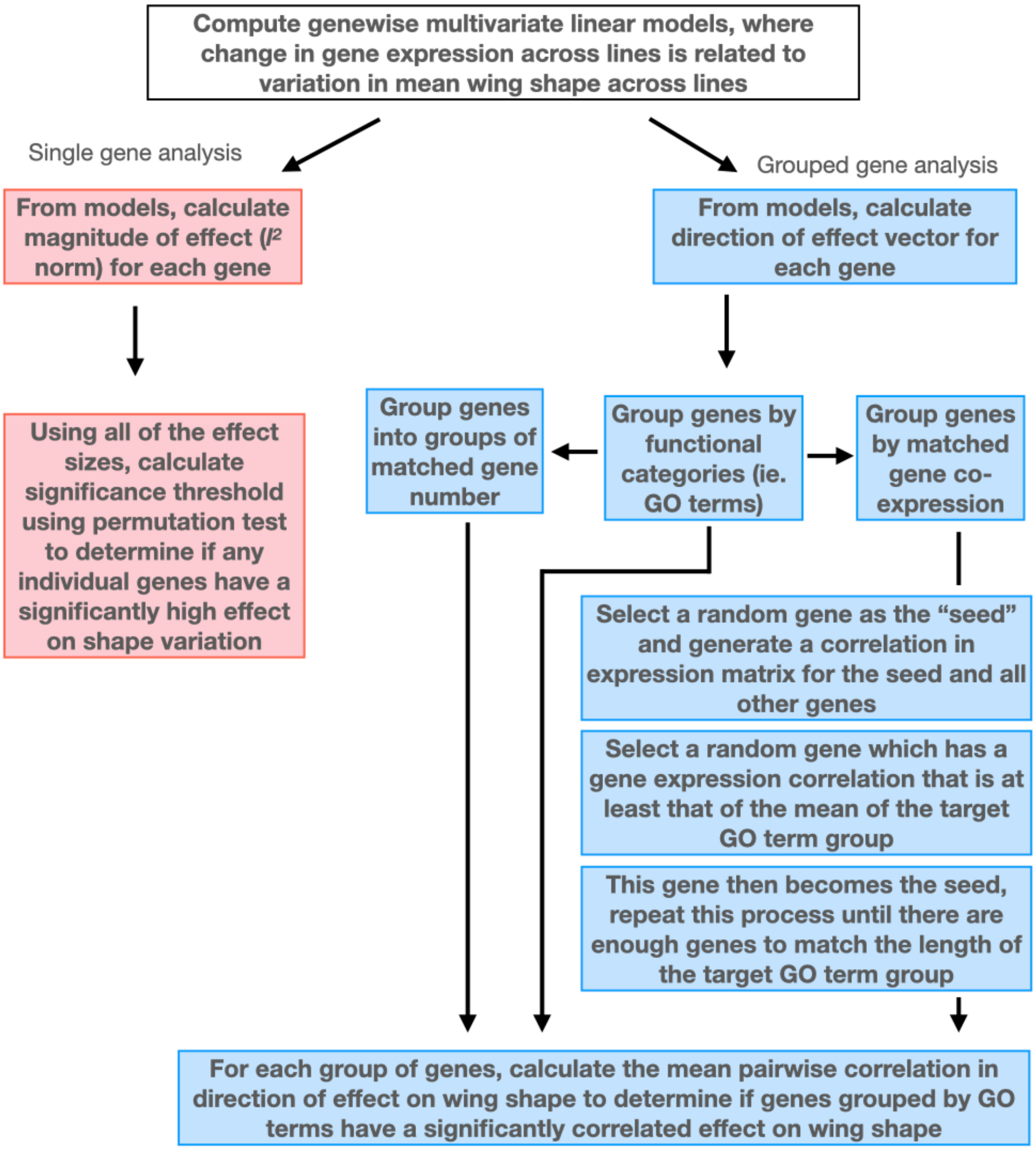
Compute genewise multivariate linear models, where change in gene expression across lines is related to variation in mean wing shape across lines
Single gene analysis
Grouped gene analysis
From models, calculate magnitude of effect (l^2 norm) for each gene
From models, calculate direction of effect vector for each gene
Using all of the effect sizes, calculate significance threshold using permutation test to determine if any individual genes have a significantly high effect on shape variation
Group genes into groups of matched gene number
Group genes by functional categories (ie. GO terms)
Group genes by matched gene co-expression
Select a random gene as the "seed" and generate a correlation in expression matrix for the seed and all other genes
Select a random gene which has a gene expression correlation that is at least that of the mean of the target GO term group
This gene then becomes the seed, repeat this process until there are enough genes to match the length of the target GO term group
For each group of genes, calculate the mean pairwise correlation in direction of effect on wing shape to determine if genes grouped by GO terms have a significantly correlated effect on wing shape

Figure 1. Methods outline for examining the relationship between variation in gene expression in developing wing tissue and variation in adult wing shape.

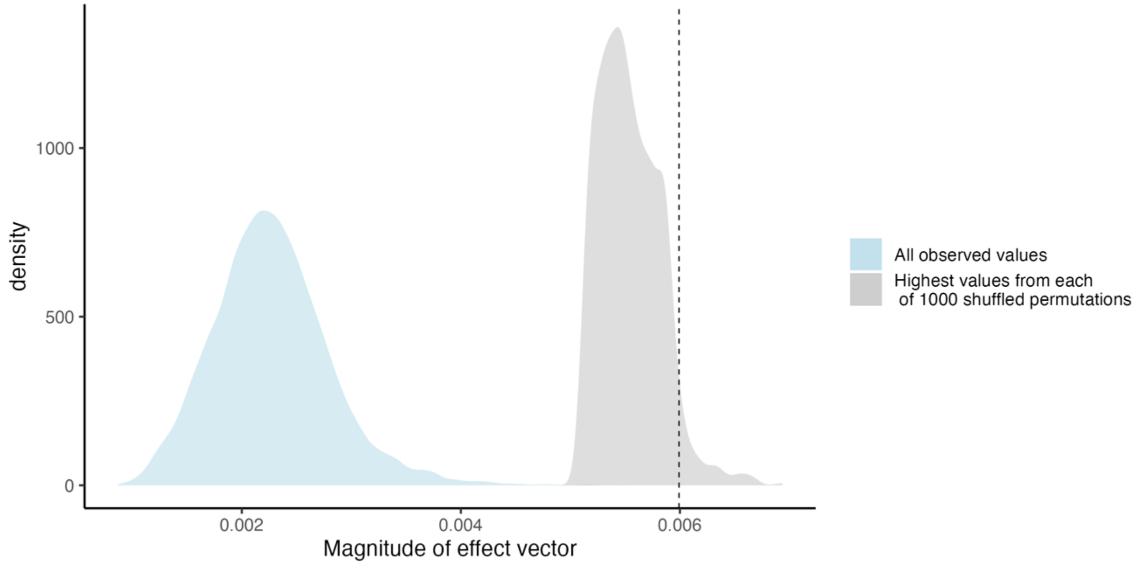

Figure 2. No individual gene's magnitude of effect on wing shape reaches significance. Density plot depicting the density of the magnitude of the effect size vector from the multiple multivariate linear models (blue) and the highest values from the C&D permutations (grey). The dashed line represents significance (α = 0.05). Note the minimal overlap between the blue and grey densities, as well as no blue values reaching the dashed line.

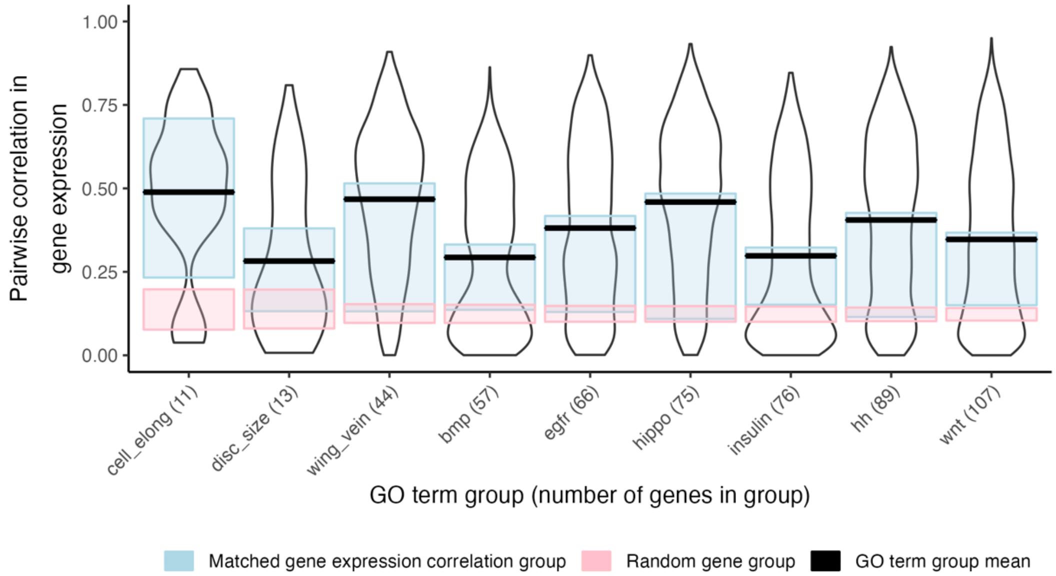

Figure 3. Pairwise correlation of gene expression among genes in GO term groups is varied. Violin plots show the distributions of pairwise gene expression correlations among the genes in each GO term group, with the number of genes in brackets. The black crossbar represents the mean of the pairwise correlation in gene expression for each GO term group. The blue and red shaded boxes represent the spread of mean pairwise absolute correlation in gene expression for the matched gene expression correlation and random gene groups, respectively. The bottom of each box marks the 2.5% percentile value from 1000 permutations and the top of each box marks the 97.5% percentile value.

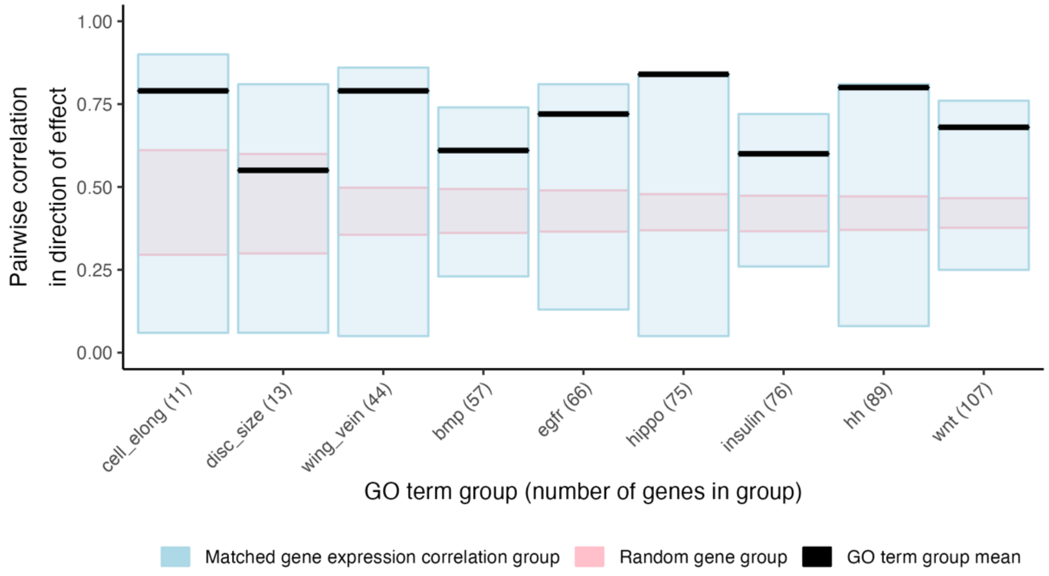

Figure 4. The mean group vector correlation of the GO term groups is higher than expected based on the distribution of random gene groups and matched gene expression groups of the same length. The black crossbar represents the mean of the pairwise correlation in direction of effect. The blue and red shaded boxes represent the spread of mean pairwise absolute correlation in direction of effect for the matched gene expression correlation and random gene groups, respectively. The bottom of each box marks the 2.5% percentile value from 1000 permutations and the top of each box marks the 97.5% percentile value.

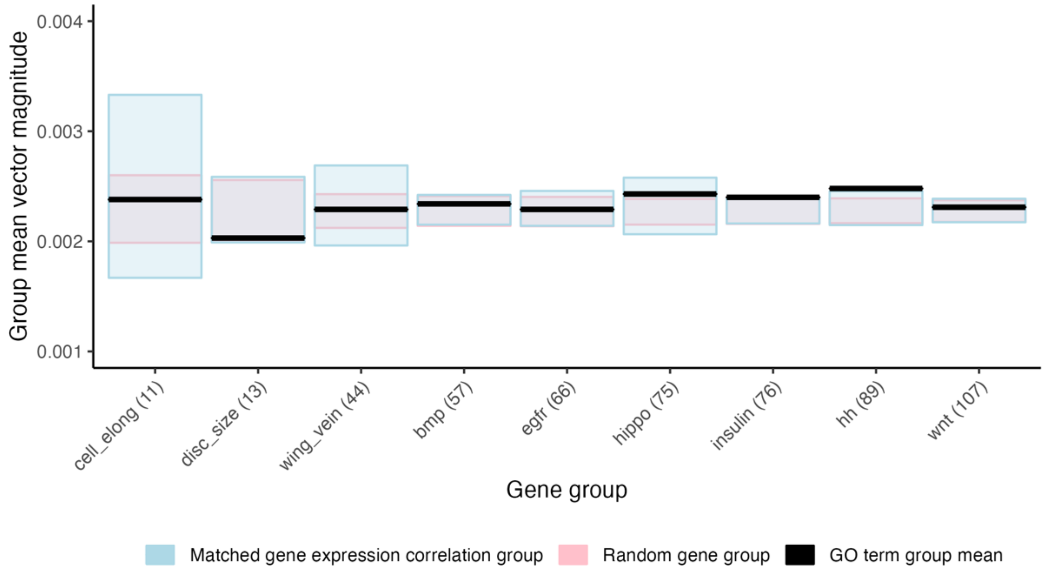

Figure 5. Magnitude of effect of genes in GO term groups not elevated in comparison genes in the random gene group or matched gene expression correlation group. The black crossbar represents the group mean magnitude of effect. The blue and red shaded boxes represent the spread of mean magnitudes for the matched gene expression correlation and random gene groups, respectively. The bottom of each box marks the 2.5% percentile value from 1000 permutations and the top of each box marks the 97.5% percentile value.

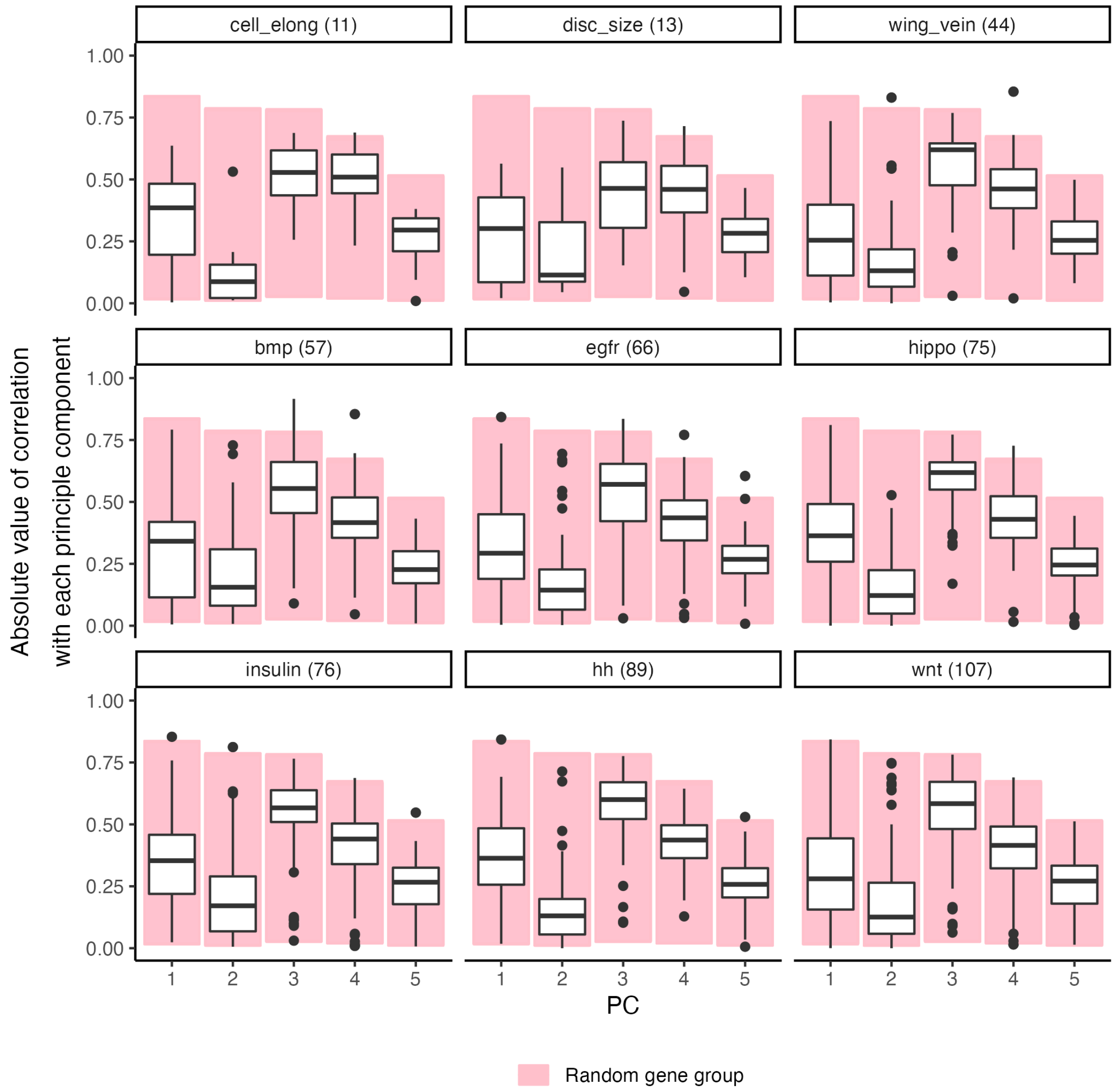

Figure 6. The correlation between major axes of natural shape variation and the direction of effect vectors is not elevated when compared to the random gene group.

2.6 SUPPLEMENTAL FIGURES

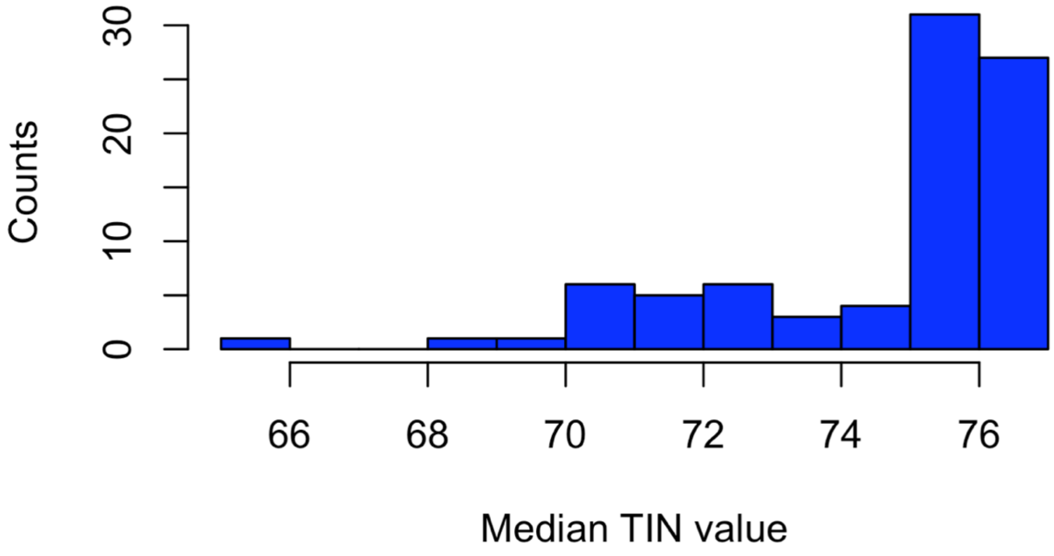

Figure S1. Median TIN values of samples

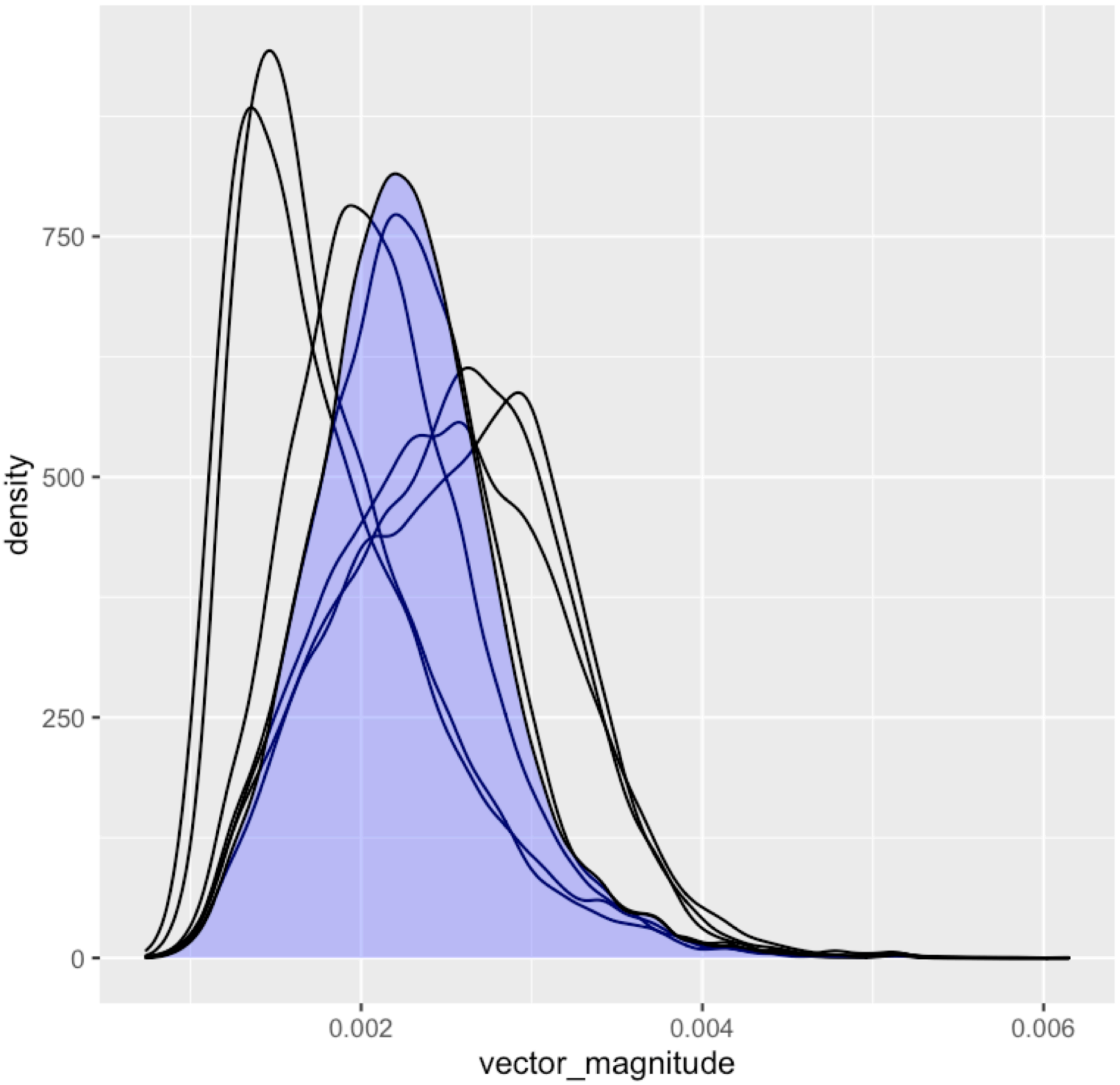

Figure S2. Densities from 7 runs of C&D permutation show strong right tail.

Black lines represent individual runs while the blue density is the mean of runs.

2.7 REFERENCES

Aponte, J. D., D. C. Katz, D. M. Roth, M. Vidal-García, W. Liu, F. Andrade, C. C. Roseman, S. A. Murray, J. Cheverud, D. Graf, R. S. Marcucio, and B. Hallgrímsson. 2021. Relating multivariate shapes to genescapes using phenotype-biological process associations for craniofacial shape. eLife 10:e68623.

Barton, N. H., A. M. Etheridge, and A. Véber. 2017. The infinitesimal model: Definition, derivation, and implications. Theor Popul Biol 118:50–73.

Beaty, J. S., K. A. West, and G. T. Nepom. 1995. Functional effects of a natural polymorphism in the transcriptional regulatory sequence of HLA-DQB1. Mol Cell Biol 15:4771–4782.

Bilder, R. M., F. W. Sabb, T. D. Cannon, E. D. London, J. D. Jentsch, D. S. Parker, R. A. Poldrack, C. Evans, and N. B. Freimer. 2009. Phenomics: the systematic study of phenotypes on a genome-wide scale. Neurosci 164:30–42.

Birdsall, K., E. Zimmerman, K. Teeter, and G. Gibson. 2000. Genetic variation for the positioning of wing veins in Drosophila melanogaster. Evol Dev 2:16–24.

Brawand, D., M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F. W. Albert, U. Zeller, P. Khaitovich, F. Grützner, S. Bergmann, R. Nielsen, S. Pääbo, and H. Kaessmann. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343–348.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–755.

Chen, Y., H. Han, G. Seo, R. E. Vargas, B. Yang, K. Chuc, H. Zhao, and W. Wang. 2020. Systematic analysis of the Hippo pathway organization and oncogenic alteration in evolution. Sci Rep 10:3173.

Churchill, G. A., and R. W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963–971.

Cobourne, M. T., and J. B. A. Green. 2012. Hedgehog signalling in development of the secondary palate. Front Oral Biol 16:52–59.

Cohen, M. M. 2003. The hedgehog signaling network. Am J Med Genet A 123A:5–28.

Dema, A., K. Yip, N. Spencer, E. Ro, D. Ehrlich, S. Padala, B. Bjork, and S. Miller. 2020. Variants in the Hippo Signaling Pathway are Associated with Craniofacial Skeletal Form and BMI. The FASEB Journal 34:1–1.

Deutsch, S., R. Lyle, E. T. Dermitzakis, H. Attar, L. Subrahmanyan, C. Gehrig, L. Parand, M. Gagnebin, J. Rougemont, C. V. Jongeneel, and S. E. Antonarakis. 2005. Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. Hum Mol Genet 14:3741–3749.

Dworkin, I., J. A. Anderson, Y. Idaghdour, E. K. Parker, E. A. Stone, and G. Gibson. 2011. The Effects of Weak Genetic Perturbations on the Transcriptome of the Wing Imaginal Disc and Its Association With Wing Shape in Drosophila melanogaster. Genetics 187:1171–1184.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52:399–433.

Green, R. M., J. L. Fish, N. M. Young, F. J. Smith, B. Roberts, K. Dolan, I. Choi, C. L. Leach, P. Gordon, J. M. Cheverud, C. C. Roseman, T. J. Williams, R. S. Marcucio, and B. Hallgrímsson. 2017. Developmental nonlinearity drives phenotypic robustness. Nat Commun 8:1970.

Gry, M., R. Rimini, S. Strömberg, A. Asplund, F. Pontén, M. Uhlén, and P. Nilsson. 2009. Correlations between RNA and protein expression profiles in 23 human cell lines. BMC Genom 10:365.

Hamann, E., C. S. Pauli, Z. Joly-Lopez, S. C. Groen, J. S. Rest, N. C. Kane, M. D. Purugganan, and S. J. Franks. 2021. Rapid evolutionary changes in gene expression in response to climate fluctuations. Mol Ecol 30:193–206.

Hatori, R., and T. B. Kornberg. 2020. Hedgehog produced by the Drosophila wing imaginal disc induces distinct responses in three target tissues. Development 147:dev195974.

Houle, D. 2010. Numbering the hairs on our heads: The shared challenge and promise of phenomics. PNAS 107:1793–1799.

Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Ràmia, A. M. Tarone, L. Turlapati, T. Zichner, D. Zhu, R. F. Lyman, M. M. Magwire, K. Blankenburg, M. A. Carbone, K. Chang, L. L. Ellis, S. Fernandez, Y. Han, G. Highnam, C. E. Hjelmen, J. R. Jack, M. Javaid, J. Jayaseelan, D. Kalra, S. Lee, L. Lewis, M. Munidasa, F. Ongeri, S. Patel, L. Perales, A. Perez, L. Pu, S. M. Rollmann, R. Ruth, N. Saada, C. Warner, A. Williams, Y.-Q. Wu, A. Yamamoto, Y. Zhang, Y. Zhu, R. R. H. Anholt, J. O. Korbel, D. Mittelman, D. M. Muzny, R. A. Gibbs, A. Barbadilla, J. S. Johnston, E. A. Stone, S. Richards, B. Deplancke, and T. F. C. Mackay. 2014. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. Genome Res 24:1193–1208.

Ingham, P. W., and A. P. McMahon. 2001. Hedgehog signaling in animal development: paradigms and principles. Genes Dev 15:3059–3087.

Jin, M., H. Liu, C. He, J. Fu, Y. Xiao, Y. Wang, W. Xie, G. Wang, and J. Yan. 2016. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. Sci Rep 6:18936.

Johannsen, W. 1911. The Genotype Conception of Heredity. The American Naturalist 45:129–159. [University of Chicago Press, American Society of Naturalists].

King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science 188:107–116.

Koch, E. M., and S. R. Sunyaev. 2021. Maintenance of Complex Trait Variation: Classic Theory and Modern Data. Front Genet 12:2198.

Kuruvilla, F. G., P. J. Park, and S. L. Schreiber. 2002. Vector algebra in the analysis of genome-wide expression data. Genome Biol 3:research0011.1-research0011.11.

Lercher, M. J., T. Blumenthal, and L. D. Hurst. 2003. Coexpression of Neighboring Genes in Caenorhabditis Elegans Is Mostly Due to Operons and Duplicate Genes. Genome Res 13:238–243.

Lin, H., Q. Liu, X. Li, J. Yang, S. Liu, Y. Huang, M. J. Scanlon, D. Nettleton, and P. S. Schnable. 2017. Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS. Genome Biol 18:192.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barrón, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L.-L. Pu, C. Qu, M. Ràmia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y.-Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman, and R. A. Gibbs. 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482:173–178.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753.

Marigorta, U. M., J. A. Rodríguez, G. Gibson, and A. Navarro. 2018. Replicability and Prediction: Lessons and Challenges from GWAS. Trends Genet 34:504–517.
Matamoro-Vidal, A., I. Salazar-Ciudad, and D. Houle. 2015. Making quantitative morphological variation from basic developmental processes: Where are we? The case of the *Drosophila* wing. Dev Dyn 244:1058–1073.

Mezey, J. G., and D. Houle. 2005. The dimensionality of genetic variation for wing shape in *Drosophila melanogaster*. Evolution 59:1027–1038.

Misra, J. R., and K. D. Irvine. 2018. The Hippo signaling network and its biological functions. Annu Rev Genet 52:65–87.

Nicolae, D. L., E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, and N. J. Cox. 2010. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLOS Genetics 6:e1000888.

Ogden, S. K., M. Ascano, M. A. Stegman, and D. J. Robbins. 2004. Regulation of Hedgehog signaling: a complex story. Biochem Pharmacol 67:805–814.

Okada, H., H. A. Ebhardt, S. C. Vonesch, R. Aebersold, and E. Hafen. 2016. Proteome-wide association studies identify biochemical modules associated with a wing-size phenotype in *Drosophila melanogaster*. Nat Commun 7:12649.

Oleksiak, M. F., G. A. Churchill, and D. L. Crawford. 2002. Variation in gene expression within and among natural populations. Nat Genet 32:261–266.

Pan, A., L. Chang, A. Nguyen, and A. W. James. 2013. A review of hedgehog signaling in cranial bone development. Front Physiol 4:61.

Pan, Y., H. Alégot, C. Rauskolb, and K. D. Irvine. 2018. The dynamics of Hippo signaling during *Drosophila* wing development. J Dev 145:dev165712.

Pelletier, K., W. R. Pitchers, A. Mammel, E. Northrop-Albrecht, E. J. Márquez, R. A. Moscarella, D. Houle, and I. Dworkin. 2022. Complexities of recapitulating polygenic effects in natural populations: replication of genetic effects on wing shape in artificially selected and wild caught populations of *Drosophila melanogaster*. bioRxiv.

Pitchers, W., J. Nye, E. J. Márquez, A. Kowalski, I. Dworkin, and D. Houle. 2019. A Multivariate Genome-Wide Association Study of Wing Shape in *Drosophila melanogaster*. Genetics 211:1429–1447. Genetics.

Porcu, E., M. C. Sadler, K. Lepik, C. Auwerx, A. R. Wood, A. Weihs, M. S. B. Sleiman, D. M. Ribeiro, S. Bandinelli, T. Tanaka, M. Nauck, U. Völker, O. Delaneau, A. Metspalu, A. Teumer, T. Frayling, F. A. Santoni, A. Reymond, and Z. Kutalik. 2021. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. Nat Commun 12:5647.

Potokina, E., A. Druka, Z. Luo, R. Wise, R. Waugh, and M. Kearsey. 2008. Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant J 53:90–101.

Ritchie, M. D., E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim. 2015. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet 16:85–97.

Robinson, M. R., N. R. Wray, and P. M. Visscher. 2014. Explaining additional genetic variation in complex traits. Trends Genet 30:124–132.

Rockman, M. V., and L. Kruglyak. 2006. Genetics of global gene expression. Nat Rev Genet 7:862–872.

Rohlf, F. J., and D. Slice. 1990. Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks. Syst Biol. 39:40–59.

Romero, I. G., I. Ruvinsky, and Y. Gilad. 2012. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet 13:505–516.

Sabatti, C., L. Rohlin, M.-K. Oh, and J. C. Liao. 2002. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res 30:2886–2893.

Sasai, N., M. Toriyama, and T. Kondo. 2019. Hedgehog Signal and Genetic Disorders. Front in Genet 10:1103.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P. S. Linsley, M. Mao, R. B. Stoughton, and S. H. Friend. 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297–302.

Shriner, D. 2012. Moving toward System Genetics through Multiple Trait Analysis in Genome-Wide Association Studies. Front Genet 3:1.

Stamatoyannopoulos, J. A. 2004. The genomics of gene expression. Genomics 84:449–457.

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

Topp, C. N., A. S. Iyer-Pascuzzi, J. T. Anderson, C.-R. Lee, P. R. Zurek, O. Symonova, Y. Zheng, A. Bucksch, Y. Mileyko, T. Galkovskyi, B. T. Moore, J. Harer, H. Edelsbrunner, T. Mitchell-Olds, J. S. Weitz, and P. N. Benfey. 2013. 3D phenotyping and quantitative trait locus mapping identify core regions of the rice genome controlling root architecture. PNAS 110:E1695–E1704.

van Dam, S., U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães. 2018. Gene co-expression analysis for functional classification and gene–disease predictions. Brief. Bioinformatics 19:575–592.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 101:5–22.

Wang, J., and J. F. Martin. 2017. Hippo Pathway: An Emerging Regulator of Craniofacial and Dental Development. J Dent Res 96:1229–1237.

Wang, L., J. Nie, H. Sicotte, Y. Li, J. E. Eckel-Passow, S. Dasari, P. T. Vedell, P. Barman, L. Wang, R. Weinshiboum, J. Jen, H. Huang, M. Kohli, and J.-P. A. Kocher. 2016. Measure transcript integrity using RNA-seq data. BMC Bioinform 17:58.

Wang, Y., P. Chakravarty, M. Ranes, G. Kelly, P. J. Brooks, E. Neilan, A. Stewart, G. Schiavo, and J. Q. Svejstrup. 2014. Dysregulation of gene expression as a cause of Cockayne syndrome neurological disease. PNAS 111:14454–14459.

Weber, K., N. Johnson, D. Champlin, and A. Patty. 2005. Many P-element insertions affect wing shape in *Drosophila melanogaster*. Genetics 169:1461–1475.

Wray, G. A. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8:206–216.

Xavier, G. M., M. Seppala, W. Barrell, A. A. Birjandi, F. Geoghegan, and M. T. Cobourne. 2016. Hedgehog receptor function during craniofacial development. Dev Biol 415:198–215.

Zimmerman, E., A. Palsson, and G. Gibson. 2000. Quantitative trait loci affecting components of wing shape in *Drosophila melanogaster*. Genetics 155:671–683.

Zinna, R., D. Emlen, L. C. Lavine, A. Johns, H. Gotoh, T. Niimi, and I. Dworkin. 2018. Sexual dimorphism and heightened conditional expression in a sexually selected weapon in the Asian rhinoceros beetle. Mol Ecol 27:5049–5072.

CHAPTER 3: The joint influence of genetic background and severity of genetic perturbation on gene expression differences in the *Drosophila* wing

3.1 INTRODUCTION

In this chapter, gene expression variation and its relationship to the expressivity of mutational effects is studied in the context of complex trait variation. As established in the last chapter, gene expression itself can be viewed as a complex trait. It varies quantitatively, has a genetic component for which loci contributing fractionally to its variation are spread across the genome, and it is influenced by gene-gene interactions. In the *Drosophila melanogaster* wing, genetic perturbation of *scalloped (sd)* and *vestigial (vg)* (two crucial genes in the gene expression program governing wing development) leads to varying background-dependent phenotypic effects in the wing. Titration of *sd* and *vg* function in the Samarkand (SAM) and Oregon-R (ORE) backgrounds has shown that the phenotypic effect is also dependant on the strength of the genetic perturbation. Here, I use this as a model system to study how genetic background (gene-gene interactions) and the magnitude of genotypic perturbation interact and affect global gene expression differences in developing wing tissue.

Genetic background effects and the context dependence of mutant alleles

Phenotypic effects mediated by a focal (mutant) allele can vary substantially depending on the wild-type background that the focal allele is imbedded within. Such effects are described as genetic background effects (GBE). The effect of wild-type genetic background on phenotypic expressivity was first documented over 100 years ago (Altenburg and Muller 1920), and is mostly observed in model genetic systems which allow for introduction of the same mutant alleles across multiple wild type strains. GBE can influence dominance patterns among alleles (Noben-Trauth et al. 1997; Johnson et al. 2006) as well as pleiotropic effects (de Belle and Heisenberg 1996).

In human disease, background effects are studied in the context of variable penetrance and expressivity. Incomplete penetrance occurs when an individual possesses a mutation, but they do not express the expected phenotype. For example, not all individuals with the same mutation in the *BRCA1* gene will develop breast or ovarian cancer. About 80% of people who have this mutation will develop the disease phenotype (Zlotogora 2003). In a study of 589,306 individuals with fully sequenced genomes, thirteen people were found with mutations for eight different severe conditions, yet none of these thirteen had ever been reported to have the implicated disease. This finding suggests that incomplete penetrance is relatively common among human populations (Chen et al. 2016).

Regardless of the level of penetrance of a mutation, it can be the case that individuals with a given mutation display heterogenous expression of the expected trait. Examples of variable expressivity in human disease include different ages of onset, progression rates, or disease severity. Variable phenotypic expressivity is the case with many so-called Mendelian diseases (Fournier and Schacherer 2017). For example, even though Huntington's Disease is caused by a single mutation in the *HTT* gene, individual age of onset varies from juvenile to over 40 years of age (Arning 2016). Some of the variation in age of onset can be explained by length of CAG repeats induced by this mutation, but not all, suggesting that there are other heritable factors interacting with the *HTT* gene deletion to influence Huntington's Disease expressivity (Di Tella et al. 2022).

Studying gene function in invariant genetic backgrounds has had severe consequences

Even though background dependence of mutant alleles and variable trait penetrance and expressivity are common in model systems and humans alike, GBE are often treated as a nuisance variable rather than an effect deserving study on its own. The current *status quo* in developmental biology is that to understanding a gene's functional consequences, one must perturb normal function of that gene while controlling for all other variables such as environment, rearing conditions, and between-sample genetic diversity (Little and Colegrave 2016). With variables not of interest controlled for, that gene's function in specific developmental processes can then be inferred through comparing the mutant phenotype to that of the wildtype. Gene knockout experiments are typically conducted on a single wildtype genetic background. The assumption in medical research is that the genetic effects seen in a single genetic background of a model organism will be the same in humans, which presupposes that the result will be generalizable to at least other inbred strains of that model organism as well as potentially to gene orthologues. Problematically, this generalization is carried through without experimental evidence.

This assumption has been shown to be wrong on several occasions. For example, the function of the *Indy* gene in *Drosophila melanogaster* was originally investigated in the Canton-S wild type genetic background (a standard lab strain). Flies heterozygous for an *Indy* knockout were reported to have a longer lifespan than those without it (but who were otherwise co-isogenic) (Rogina et al 2000). The authors of this study proposed a mechanism by which the *Indy* gene knockout leads to a higher mitochondrial density, but each individual mitochondrion works at a lower rate, thus inducing a state mimicking that of calorie restriction leading to increased lifespan (Rogina et al 2000). However, when this mutation was introduced into two other wildtype backgrounds, increased longevity was not observed (Toivonen et al 2007). It was also found that only males heterozygous for the *Indy* knockout from the original Canton-S background were long-lived. Females did not have an increased lifespan (Toivonen et al 2007).

The authors from the second study concluded that the increase in lifespan observed was not due to *Indy* itself, but likely to interactions between *Indy* and unidentified loci, as well as *Wolbachia* infection in the Canton-S flies (Toivonen et al. 2009). In another example, when null *CACNA1C* and *TCF7L2* genes were crossed into one of thirty mouse strains, it was found that most null phenotypes observed in these mice were not generalizable. In some backgrounds the mice were completely unaffected. There were several cases of completely directionally opposite allelic effects in other backgrounds (Sittig et al. 2016). These studies demonstrate that different conclusions (and sometimes opposite conclusions) can be drawn from the same experimental set-up, depending on which wildtype genetic background is used.

The polygenic nature of genetic background effects

Given the consequentiality of background effects, much work has been put into understanding their causes. Fundamentally, GBE are a result of interactions between a focal allele and other segregating alleles from within a genetic background. This phenomenon is referred to as epistasis. The term "epistasis" was first used over a century ago to describe the discrepancy between outcomes of dihybrid crosses and the predicted segregation ratios based on individual genes. Some expected phenotypes were not observed, or allelic combinations led to unexpected phenotypes (Phillips 2008; Bateson and

Mendel 2013). Epistatic interactions can lead to changes in both the magnitude and direction of allelic effects, and these effects have been shown to contribute to additive variation in quantitative traits (Mackay 2014).

In the context of GBE, researchers often attempt to locate "modifier genes", which are genes that interact with a focal allele to produce background effects (Nadeau 2001). Modifier genes can lead to phenotypic variation through mechanisms such as modifier protein products directly interacting with the mutant protein, alleles in modifier genes may produce factors that alter the timing and rate of transcription of the mutant gene, or modifier genes can produce factors that affect the degradation of mutant protein (Johnson et al. 2006). Often, GBE are highly polygenic, and many modifier genes spread out across the genome contribute to context-dependant phenotypic variation through epistatic effects. In *Drosophila melanogaster*, it has been shown that the phenotypic effect of an allele of *scalloped* (*sd*^{E3}) is profoundly different among two wildtype genetic backgrounds, SAM and ORE, and that this loss of function allele has downstream effects on gene expression (Dworkin et al. 2009). The epistatic relationship between *sd*^{E3} and *omb/bi* was also shown to be background dependent, where in one background the double mutant appeared similar to the *sd*^{E3} only mutant while in the other *omb/bi* enhanced the severity of the wing phenotype (Dworkin et al. 2009). Later, it was shown that context-dependant effects of genetic modifiers are wide-spread. A genome-wide screen of modifiers of the *sd*^{E3} allele in these two genetic backgrounds revealed that 74% of modifiers were background-dependent (Chari and Dworkin 2013). Background-dependant modifiers also contribute to human disease penetrance and expressivity. In a study of 80,928 individuals, the polygenic background of each person was shown to modify the probability of disease by age 75 from 11% to 80% for colon cancer (Fahed et al. 2020). Finally, genetic background also affects the magnitude of epistatic effects. Through crossing an allelic series of *sd* to mutants with subsets of deletions, it was shown that in general the magnitude of epistasis depended on the severity of the deletion, but also that the genetic background of individuals created variation within this trend (Henderson, 2021).

Studying global transcriptional response to the joint effect of background and perturbation strength

The *Drosophila melanogaster* wing has been used extensively as a model system for which to study genetic background effects. As mentioned above, the *sd*^{E3} allele has been found to affect wing shape differently depending on whether it is within the ORE or SAM wildtype genetic background (Dworkin et al. 2009). Further work in this system revealed that there are many background dependent modifier loci which modulate the wing phenotype (Chari and Dworkin 2013). Using this model system and expanding it to include an allelic series of *sd* and functionally related *vg*, it was shown that the background dependence of a mutation was related to the magnitude of its effect. The moderate alleles showed the most variability in effect while the weakest and strongest alleles produced similar wing phenotypes across both backgrounds (Chandler et al. 2017). This finding has been recapitulated through expansion of the *sd* alleles into many more wildtype backgrounds (Daley 2019). Additionally, the rank order of allelic effects in the series remained constant across both backgrounds. This trend remained even when a subset of the alleles was crossed into 16 additional backgrounds (Chandler et al. 2017).

Sd is a transcription factor that is a master regulator of many of the genes involved in the wing development program (Campbell et al. 1992). It encodes a TEA domain with a DNA binding region that forms a heterodimer with *yki* to regulate wing growth through the Hippo signalling pathway (Bandura and Edgar 2008). Independent of this interaction, *sd* also forms a heterodimer with the product of *vg*. Together, these genes are critically important for wing development through recruitment of additional transcription factors in the wing imaginal disc, modulating the expression of a large set of genes involved with wing determination, proper patterning, and growth (Halder et al. 1998). Loss of function mutations and RNAi knockdown of both *sd* and *vg* show a dose dependent response in loss of adult wing tissue that requires some level of stochiometric balance between Sd and Vg. The total loss of this gene expression program causes the almost complete loss of formation of wing tissue. Furthermore, ectopic expression of *vg* in non-wing tissues leads to ectopic development of wing tissue (Delanoue et al. 2004). Importantly, this only occurs in tissues where *sd* is already natively expressed (Halder et al. 1998; Simmonds et al. 1998). The results from many genetic and developmental studies of the role of *vg* and *sd* in wing development thus suggest they are necessary and sufficient to drive the wing development gene expression "program".

It remains an open question as to how global gene expression responds to changes in *sd* and *vg*, the extent to which this response is dependent on the joint effect of genetic background and the magnitude of allelic perturbation, and what mechanisms cause underlying background-dependant phenotypic changes in the wing. The *sd-vg* complex is crucial in regulating the gene expression changes that drive formation of the wing, and previous work has demonstrated the phenotypic consequences of knocking down *sd* gene expression levels. As such, in the current chapter the RNA sequencing data from Chandler et al. (2017) is used to study the joint effect of genetic background and the magnitude of perturbation effects on gene expression and phenotypic expressivity (Figure 1).

3.2 METHODS

Fly rearing and handling

For detailed information on fly rearing and introgression of alleles, see Chandler et al. (2017). Briefly, the two wildtype strains used for this study were ORE and SAM, which are both maintained as inbred lines. Both strains have been marked with a *white* mutation (w^-) that causes a loss of normal eye pigmentation, and was used to facilitate introgression of several mutant alleles caused by the insertion of a transgene (with a mini-w^+ rescue construct). These alleles are almost all regulatory and not expected to alter protein sequence and function, but instead reduce expression of the native transcripts of either *sd* or *vg*.

Wing size quantification

For each mutant and wildtype fly, a single wing was dissected from a minimum of 5 individuals per genotype per sex for each of 2 replicates. There was a total of at least 10 observations per genotype. Wings were imaged using an Olympus DP30BW camera mounted on an Olympus BW51 microscope using DP controller image capture software (ver. 1.43u). Measures of wing area can be confounded by variation in body size, and some of the weaker hypomorphic alleles cause loss of bristles at the wing margins without a change in wing area. Both factors can make modest changes in wing size due to a mutation challenging to quantify. As such, a semi-quantitative ordinal scale (1-10), which has been previously shown to be linearly correlated with wing size (Halder et al. 1998), was used to measure severity of phenotypic effects on adult wings. The advantage of the semi-quantitative measure is that the effects of weak perturbations on wing morphology can be identified in a clear manner. While considerable effort was made to make the differences between ordinal values correspond to those of direct measures on wing size (see Chari and Dworkin (2013) for details), as with any ordinal scaled measure, such comparability is at best only approximate. In Chandler et al. (2017), all wings were both measured quantitatively (wing area) and using the semi-quantitative scale. It was shown that besides the weakest perturbations, which were captured better by the semi-quantitative scale, both methods provided very similar inferences. As such, it is expected that the only meaningful difference (and the reason the semi-quantitative measure is used) is to better account for the background dependence for the weakest alleles (i.e. those with the smallest severity of perturbation). The semi-quantitative values were regressed onto the interaction between maternal allele and F1 background, and the marginal mean for the interaction term was used as a continuous measure for perturbation when modelling the effect of perturbation and background on gene expression (this regression was done by Katie Pelletier).

Wing Imaginal Disc RNA sequencing and Expression Data

Imaginal wing discs were dissected from mature 3rd instar larvae. Each RNA sequencing sample represents the combined imaginal disc tissue of approximately 30 individuals. Reads were trimmed using bbduk. Transcript-level counts were estimated using Salmon (ver. 1.4.0). Salmon is a mapper which is a pseudo-aligner and can use a decoy to prevent spurious alignment to similar genomic sequences, and quantifies transcripts based on mapping rather than alignment. Reads were mapped onto version 6.38 of the *Drosophila melanogaster* transcriptome. Transcript-level counts were collapsed onto gene-level counts using tximport (ver. 1.20.0) using the UCSC dm6 ensemble (ver. 3.12.0). In total, 13,701 genes were reported. The function filterByExpr() from the edgeR (ver. 3.34.1) library was used to filter out genes without a minimum total count of at least 15 reads. After filtering, a value of 1 was added to the expression vector for each gene to avoid complete separation in the cases where a gene was expressed in one background but not at all in the other. . Without adding 1 to every sample, the fit models for these genes would produce nonsensical estimates, and adding a consistent value to all samples avoids this problem. The normalizationFactors() function from DESeq2 (ver. 1.32.0) was used to estimate gene by sample normalization factors, where counts are divided by sample-specific size factors which are the median ratio of gene counts relative to the geometric mean per gene. The normalization factors were then log transformed used as the offsets in the linear models.

Principal component analysis of gene expression

To detect the major sources of gene expression variation in the transcriptome, a principal component analysis (PCA) was conducted using DESeq2 and RNAseqQC (ver. 0.1.4). Genes without a minimum count of at least 5 were removed. A variance stabilizing transformation was applied to the remaining set of genes, normalizing with respect to library size. This analysis was also done with a regularized log transformation, but the results were similar (not shown), suggesting that the size factors in the sample do not vary widely. The 500 most variable genes were plotted and shown here, but plotting the 5000 most variable genes (not shown) led to the same results.

Modelling genetic background and perturbation effects

All statistical analyses were conducted in R (ver. 4.1.3). Using glmmTMB (ver. 1.1.4), models were fit for each gene, h (gene specific subscripts not included). A generalized linear model with a negative binomial distribution where for the i^{th} sample was fit as below:

$$y_i = u_i + \epsilon_i$$

Where,

$$\log(u_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \hat{\beta}_3 x_{i,1} x_{i,2}$$

and the random effect of lane,

$$\beta_0 \sim N(0, \hat{\sigma}_l^2)$$

Where y_i is expression of gene h for the i^{th} sample, β_0 is the model intercept, $\beta_1$is the coefficient for genetic background (SAM or ORE), β_2 is the coefficient (slope) for semi-quantitative measure of the perturbation effect on wing morphology, β_3 is the interaction term for background and perturbation. $\sigma_l^2$is the variance for the random effect for lane of sequencing and ϵ_i is the residual (unfit) variation. In glmmTMB we used the "nb2" (negative binomial 2) quadratic parameterization (Brooks et al. 2017).

The primary focus was on the magnitude of effects within and between backgrounds. To estimate this magnitude, for each gene model coefficients and the estimated slope (amount of change in gene expression per "unit" perturbation) for each of the two backgrounds was extracted, along with associated 95% confidence intervals. However, I also used type II analysis of variance using the Anova function from the car package (v3.1.0), and a false-discovery rate adjusted p-values (Benjamini & Hochberg 1995) at a cut-off of 0.1

to help in filtering genes for additional examination, and to assess whether the contribution of genetic background, perturbation, or their interaction was "significant". Genes were considered to have a perturbation effect (i.e. a general increase or decrease in expression related to severity of perturbation) if 1) there was a significant perturbation term, an insignificant background term, and an insignificant interaction term, or 2) there was a significant perturbation term, a significant background term, and an insignificant interaction term. Genes were considered to have an interaction effect if 1) there was an insignificant perturbation term, an insignificant background term, and a significant interaction term, or 2) an insignificant perturbation term, a significant background term, and a significant interaction term. Genes were considered to have a perturbation and interaction effect if 1) there was a significant perturbation effect, an insignificant background effect, and a significant interaction effect, or 2) there was a significant perturbation effect, a significant background effect, and significant interaction term (Table 1).

Vector Correlations

To elucidate potential mechanisms underlying the observed GBE in wings between SAM and ORE, the difference in the rates of change of gene expression by unit perturbation for genes from key biological processes were compared using vector correlations. GO terms from FlyBase (ver. FB2022_05) were used to group genes according to the following biological processes, chosen for a variety in number of genes and relevance to wing patterning: regulation of intracellular mRNA organization (28 genes), apoptotic signalling pathway (47 genes), regulation of cellular response to growth factor stimulus (61 genes), positive regulation of cell cycle (67 genes), hippo pathway regulators (78 genes), regulation of cell population proliferation (151 genes), regulation of cell differentiation (179 genes), and regulation of cell death (208 genes). For each group, the absolute correlation between the vector of slopes was calculated and compared to the 95% highest absolute correlations from 1000 permutations of correlations between two vectors of random genes of the same length as that group. In each of these cases, the genes from the GO term groups were removed from the pool from which the random gene vectors were selected. As a positive control, this analysis was also conducted with the genes which had a significant perturbation term (1124 genes), as these genes are expected to have a high correlation in their slopes between the two wild type backgrounds. The densities of the slopes of each of these backgrounds are shown in Figure S3.

Selecting genes of interest for follow-up

One of the goals of this study is to identify genes for which to follow up with functional genetic work. First, all genes with an FDR adjusted p-value of less than 0.1 from the interaction only and perturbation plus interaction categories were considered. A plot of gene expression by perturbation for each of these genes was used to visually inspect the relationship, magnitude of effect, and level of noise. Additionally, the FlyBase entry for each gene was considered, where it was assessed whether expression of the gene was expected at the time and location of sampling (with the caveat that the FlyBase entries might have been sourced from only one genetic background and as such may not be accurate for all genes in both backgrounds used in this study). The FlyBase entries also contain phenotypes associated with mutations in genes, which were inspected, but with the caveat that a wing phenotype not being listed does not mean that there is no wing phenotype, it might mean that this was not the phenotype of interest of the research that the entry was pulled from. Genes that showed narrow confidence bands and, in most cases, noted wing shape or size phenotypes, were selected as genes of interest.

3.3 RESULTS

Global transcription among samples correlates highly with wildtype genetic background and less so with the strength of allelic perturbation

In a PCA of the 500 most variable genes across the sample, samples group distinctly by wildtype genetic background (either SAM, ORE, or F1 hybrid of the two background) along PC1, which accounts for 36.9% of variation in gene expression (Figure 2). Within PC1, there is also subtle grouping by perturbation within the genetic background groups (from top to bottom). PC1 explains the largest proportion of the total variation in gene expression of the 500 most variable genes (Figure 3). Along PC3 (8.07%) and PC5 (4.98%) (accounting for a total of 13.05% of the total variation) there is subtle grouping by severity of perturbation, in particular for the most severe perturbations to wing development (in yellow).

The correlation in change in gene expression in response to perturbation between genes involved with positive regulation of the cell cycle, hippo pathway regulators, regulation of cell population proliferation, and regulation of cell death is very low across backgrounds

To determine possible biological mechanisms underlying the different wing phenotypes in response to the allelic series between the SAM and ORE backgrounds, for each gene the model slope (representing the of change in gene expression in response to perturbation) was calculated for each background. Genes were grouped according to biological GO terms such that for each background, a vector of gene slopes was assigned for each GO term group. The GO term groups were chosen as to elucidate possible biological mechanisms for the change in phenotype. A low vector correlation suggests that the genes within this group are responding differently between the two backgrounds.

The absolute correlation between these vectors was calculated for each GO term group. As a comparison distribution, 1000 permutations of the vector correlations were calculated with slope vectors for random genes of matched length to the respective GO term group. The top 95% of absolute correlations was used to create the comparison null distribution (Figure 4). The perturbation significant genes were also included as a positive control because it is expected that this group will have a high vector correlation, since the slopes should be in the same direction and not cross (otherwise they would then also have a significant interaction term and would not be included in the perturbation only group). The perturbation significant genes do indeed have a high vector correlation relative to the biological GO term groups, suggesting that the vector correlations are identifying slope correlations as intended.

All eight of the assessed groups have a lower absolute value of correlation of vector of slopes between SAM and ORE relative to the positive control. The lowest of which are the positive regulation of cell cycle group, the Hippo pathway regulators, the regulation of cell population proliferation group, and the regulation of cell death group (Figure 4).

Because low correlations would be guaranteed if the slopes of the genes were flat (ie. A value of 0), scatter plots of the ORE slope vs. the SAM slope for each assessed group were plotted to assess the degree to which zero value slopes are present (Figure 5 and alternatively shown in Figure S3). Overall, the slopes from many genes across the eight groups are of a low value, so this should temper the results from (Figure 4) to an extent, but the low correlations are not driven primarily by zero value slopes. "Outlier" genes for which the ORE slope > 0.1 and SAM slope < 0.1 or ORE slope < - 0.1 and SAM slope < 0.1 have been labeled on these plots (Figure S2), and these are the genes for which the magnitude of the change between backgrounds is largest and so might represent interesting follow-up genes.

Model estimates across SAM and ORE backgrounds reveals set of candidate genes for functional validation

Using a combination of the slopes, FDR-adjusted p-values from the linear models of gene expression, as well as FlyBase GO terms and temporal and spatial gene expression information plus associated phenotypes, effect slopes (Figure 6) and plots of gene expression by perturbation, twelve genes were identified as candidates for future functional work (Table 2, Figure 7). These are genes whose expression shows either an interaction effect or main effect of perturbation plus interaction effect.

The ORE slope and SAM slope for all genes are plotted in Figure 6, with the genes of interest from the interaction only and perturbation and interaction groups labeled. *Vajk2* has a noted wing phenotype and narrow confidence bands (Hevia et al. 2017), and is also a binding partner with *apolpp*, for both of these genes, RNAi expression has been noted to lead to a blistered wing phenotype and smaller wing disc (Panáková et al. 2005), respectively. *Apolpp* is a direct binding partner and is necessary for signalling of critical wing development genes such as *wg* and *hh*. *Orct2* RNAi expression also leads to a smaller wing phenotype (Herranz et al. 2006) with the note that this is due to cell size and not cell number, and *Orct2* is a target of the insulin receptor pathway which is a critical pathway in determining wing size. *P5CS* directly interacts with a downstream target of *Myc* which is implicated in cell size control with a noted smaller wing phenotype (Johnston et al. 1999). *Ranshi* is a direct binding partner of *Mad*, which regulates expression of BMP response in wing development to modulate wing size through increasing cell size, and *Mad* mutants show an altered wing shape (Dworkin and Gibson 2006) and decreased wing size through non-cell autonomous apoptosis of wing disc cells (Umemori et al. 2009).

The vector correlation results, and the genes of interest listed here implicate differences in cell size and cell density in the wing disc across backgrounds as mechanisms to follow-up on.

3.4 DISCUSSION

Global gene expression is influenced to a higher degree by wildtype genetic background than by genotype

RNA seq samples from the *sd/vg* allelic series grouped distinctly along PC1 by wildtype genetic background, accounting for 36.9% of variation in gene expression (Figure 2, Figure 3). This finding was surprising given that as the phenotypes of the wings from the most severe perturbations were similar in both backgrounds, the expectation was that the severely perturbed gene expression profiles might group together and that this would account for most of the variation in gene expression profiles. However, separation according to perturbation severity is only somewhat apparent in PC3, and slightly more so in PC5. This finding suggests that variation in global gene expression in the wing imaginal disc largely reflects the genetic background of origin rather than the severity of perturbation state.

Another explanation for this finding could be that there are a couple of genes for which variation greatly affects phenotype, and although the variation in gene expression of only these genes is sufficient to cause a highly deformed wing, this variation is masked in the PCA by the variation of other genes which wing shape may be robust to. The magnitude of the strong alleles could be such that it swamps out the effect of any modifier genes, as has been shown before, where the strongest alleles are also the least sensitive to genetic background ((Chandler et al. 2017, Daley 2019, Henderson 2021). This robustness is apparent when looking at the phenotypes (Figure 1), but the gene expression profiles of these samples show much more variation.

In studies of developmental biology, researchers will compare the phenotype of a knockout of gene function to a co-isogenic "wild type". This is usually examined in a single wildtype background, to elucidate the gene's function, and then these findings are generalized to be broadly representative of the species (and as such, across many wild type backgrounds). Furthermore, the magnitude of gene expression differences observed between the mutant and wild type are used, in part to infer aspects of the causal relationship between genotype and phenotype. The finding that global gene expression does not group primarily by perturbation but by wildtype genetic background calls to question the generalizability of this classic approach, which assumes that phenotypic changes are mediated by gene expression changes caused by mutant alleles. Though variation in gene expression can indeed lead to phenotypic variation, it is apparent that there is much variation in gene expression between wildtype organisms of different genetic backgrounds, including for many key developmental genes. That is in some genetic contexts, perturbing key gene function (with an associated reduction in gene expression) can have a severe impact on phenotype, suggesting a clear causal relationship. Yet in other contexts (such as across two or more distinct wild type backgrounds) variation in the expression of such genes can vary to a much larger degree with little impact on phenotype (Dworkin et al. 2009). This study adds to the growing body of evidence that genetic background is not trivial, and it is well worth controlling for wildtype genetic background in developmental biology studies. For example, had only the SAM background been used in the current study, one might conclude that *Orct2* does not at all play a role in the altered phenotype due to the perturbation of *sd* or *vg* (Figure 7B), as the expression of this gene remains stable through the allelic series. However, had one conducted this study in only ORE, one might conclude that *Orct2* was indeed important as in this background expression drops through the allelic series.

An interesting question to arise from this finding is that of when gene expression variation does not cause variation in phenotype. The wildtype wings of both the SAM and ORE backgrounds look nearly identical, yet gene expression from these individuals does not group together clearly along the first few PCs. Similarly, several of the mutant genotypes (i.e. *sd*^{E3} in ORE and the *vg*^{1} allele in both backgrounds) have very severely reduced wing sizes, presumably through the same mechanism. Despite this, the ordination plot of the PCA suggests that most of the variation in gene expression does not relate to the effect of perturbation, nor do those genotypic-background combinations seem to be in proximity to one another in this space. An explanation for why the phenotype is robust to variation in gene expression in wildtype individuals is that this variation is still within the range of what is tolerable in the organism. This is supported by findings that variation in gene expression among wildtypes is high (Cowley et al. 2009; Vu et al. 2015). Although less variation in gene expression is due to the mutant alleles, it might be that there is enough variation in key modifier genes to cause the phenotypic outcomes and variation throughout peripheral genes is tolerated well enough (Dworkin et al. 2009; Mathieson 2021).

Differences in the regulation of cell death and regulation of cell population proliferation are candidate mechanisms underlying phenotypic differences in the wings of SAM and ORE individuals.

To understand how patterns of gene expression change with genetic background and perturbation effect in pathways rather than single genes, vectors of the slopes calculated from the linear models from genes grouped by GO biological terms were assessed for their correlation between the SAM and ORE backgrounds (Figure 4). The absolute value of these correlations was compared to the 95% highest absolute correlations from 1000 permutations of vectors of random gene slopes of lengths matching the corresponding GO term groups. This analysis revealed that there is relatively low correlation between the change in gene expression in response to perturbation strength between SAM and ORE for genes involved in the regulation of cell death, regulation of cell population proliferation, *hippo* signalling genes, and genes involved with positive regulation of the cell cycle. The low correlation in the expression of these genes across perturbation suggests that they are, as a group, responding very differently to genetic perturbation in both backgrounds. The ORE and SAM slopes were also plotted for each gene in each pathway to assess the degree to which slopes are equal to zero, as zero value slopes will show low correlation

(Figure 5, Figure S3). Though the slopes are lower than those in the perturbation significant genes group, they are non-zero. However, the low slopes do represent a lower magnitude of effect, and while this does not necessarily mean the genes have no effect it is worth noting, and the conclusions drawn from this result need functional validation. Genes with a higher magnitude of effect (where ORE slope < - 0.1 and SAM slope < 0.1 or ORE slope > 0.1 and SAM slope < 0.1) have been labeled in Figure S2.

Nevertheless, this finding suggests that variability in cell number (through regulation of cell death, regulation of cell population proliferation, or regulation of the cell cycle) or cell size (through the action of *hippo* signalling genes) across both backgrounds might influence their tolerance to genetic perturbation. The genes labeled in Figure S2, which have a higher magnitude of effect compared to the others in each pathway, might be logical starting places for functional analyses.

Vajk2, Ortc2, ranshi, apolpp, and P5CS have been identified as priority genes of interest for functional testing

One of the goals of this study was to identify prospective candidate genes implicated in the GBE observed across the allelic series. To do this, gene wise linear models were run where change in gene expression was modeled as a function of genetic background, perturbation, and the interaction between genetic background and perturbation. FDR-adjusted p-values were considered significant at $\alpha = 0.1$, and the slope representing the rate of change of gene expression by perturbation strength was calculated for all genes (Figure 6). Using model significance, these slopes, inspection of the plot of gene expression by perturbation, information about known temporal and spatial gene expression, protein-protein interactions, and known phenotypes associated with altered expression or gene mutants, candidate genes were selected (Table 2). Interestingly, and in accordance with the results from the vector correlation analysis (Figure 4), the candidate genes reflect changes in cell size and density as possible mechanisms underlying the observed joint effect of background and perturbation on gene expression (Table 2, Figure 7).

First, *Vajk2* has been previously implicated with wing development, expression of *Vajk2* RNAi leads to smaller and blistered wings (Hevia et al. 2017). The gene expression by perturbation plot for this gene has narrow confidence bands, and it can be observed that while wildtype individuals of both backgrounds have similar levels of gene expression, *Vajk2* expression in ORE wings drops off at a faster rate after the first mutant in the allelic series while expression in SAM remains constant (Figure 7A). This mimics the observed phenotypic changes of the wings, where SAM *sd*^{1} wings appear similar to wildtype wings, but ORE^{sd1} and ORE^{vg2a3} wings have marked phenotypic changes.

The protein produced by *Apolpp* interacts directly with that of *Vajk2*, and *Apolpp* is a gene in which gene expression is affected by the interaction between background and perturbation, where wildtype expression level is similar between ORE and SAM and expression remains steady in SAM, but there appears to potentially be a "threshold" response (which would have to be investigated further) where in ORE expression increases only at the level of the *vg*^{1} allele (Figure 7E). *Apolpp* is also a binding partner of other genes involved with wing development, such as *hh* and *wg* (Panáková et al 2005). Flies with a mutation in *apolpp* have been noted to have smaller wing discs than wildtype flies, though it is unclear whether this is due to a smaller cell size or fewer cells.

Another candidate where expression in ORE decreases faster than that of SAM (per unit change of perturbation), and where model confidence bands are narrow (Figure 7B), is *Orct2*. Interestingly, expression of *Orct2* RNAi has been shown to lead to smaller wings because of smaller cell size, but independent of cell number (Herranz et al. 2006). Furthermore, *Orct2* is a transcriptional target of the insulin receptor pathway, which is critical for regulating wing size during development and alterations to cell size (Shingleton et al. 2005).

Ranshi was also identified as a gene of interest, where a decrease in expression in SAM is observed with an increase in expression in ORE (Figure 7D). A direct binding partner of *ranshi* is *Mad*, mutants of which have been associated with altered wing shape (Dworking and Gibson 2006) and smaller wings due to a reduction in cell density (Umemori et al. 2016).

The implication of *Ranshi*, *Vajk2, Orct2,* and *apolpp* suggests that investigating the background dependence and effect of perturbation strength on genes involved with cell size regulation may elucidate a mechanism underlying the observed GBE in this model system. That is, the regulation of cell size and number, and their joint influences on wing size may mediate the extent of genetic background effects.

Decreased cell size, but not cell number, in wing discs leading to smaller wings is also associated with *P5CS*, which interacts directly with *CTPsyn*, which is a downstream target of *Myc*. The gene expression by perturbation plot of *P5CS,* (significant B:P interaction), indicates an increase in expression in ORE with a slight decrease in expression in SAM (Figure 7C). *P5CS* is associated with d*Myc* (also known as *diminutive*, *dm*) expression, which regulates cell size. Expression of *Myc* RNAi has been associated with smaller wing discs due to reduced cell size (Johnston et al. 1999). A wing phenotype has not yet been noted for *P5CS*, though it is not clear if this has been investigated yet.

The findings from the current study suggest that perhaps in the ORE background cell size is affected as a downstream consequence of the *vg/sd* allelic series, but this is not the case in the SAM background (or it is but to a lesser degree). This could potentially be because cells in the wings of ORE flies are larger to begin with, and as such are impacted by the influence of genes such as *PC5S* or *Orct2* more-so than wings of SAM flies. To examine this relationship, one could use many more genetic backgrounds, such as strains of the DGRP, and measure initial cell sizes and wing phenotypes in response to an allelic series of *sd*. For example, it could be observed that DGRP with a larger cell size are tolerant to perturbation. To assess this correlation, the cell sizes of the smaller DGRP could be increased through temperature manipulations, and if there is a connection then these manipulated lines should show similar response to perturbation as the wildtype lines with the naturally larger cells.

The role of cell size and density in shaping the Drosophila melanogaster wing

The results of this study have pointed towards changes in cell size and number as possible mechanisms underlying the background and perturbation-specific wing shape changes. Change in the size of the *Drosophila melanogaster* wing along latitudinal clines has been associated with changes in cell size (James et al. 1995), which is correlated with temperature effects. Flies reared in lower temperatures grow larger body sizes and wing areas due to an increase in cell size but not number (Partridge et al. 1994). Changes in cell proliferation due to the impact of *sd* and *vg* mutant alleles has been shown to not strongly influence context-dependence of mutational expressivity of the adult wing in this system

(Chandler et al. 2017), so perhaps cell size would be an interesting avenue for future investigation.

It is clear from the various gene expression by perturbation plots presented here, as well as the cell proliferation results from Chandler et al. (2017), that gene expression effects are not necessarily correlated directly with the developmental processes underlying phenotypic expressivity, so I do not think it is the case that a single developmental mechanism or single mediator gene of large effect will explain the background effects observed on phenotype. Combining several levels of biological data, such as gene expression data, sequence binding predictions, and mapping and modifier datasets can lead to a holistic understanding of the causes of GBE as well as identify strong candidate genes (Chandler 2014). It is likely that utilizing multiple datasets and integrating information from many levels of biological organization (such as proteomic or metabolomic), as well as the current approach, will show the complexity of the underpinnings of GBE and reveal it to be a system-wide phenomenon that evades simple explanations (except in the rare and notable instances of large-effect modifier genes).

Conclusions, limitations, and future directions

This chapter has shown the importance of wildtype genetic background and its joint influence with magnitude of genetic perturbation on global gene expression. Most surprisingly, this study has revealed that a large source of variation in gene expression among all samples, including those with the most perturbed wings, is wildtype genetic background. This finding shows that variation in gene expression due to wildtype genetic background is plentiful, yet most of this variation is buffered in wildtype organisms. Gene expression variation underlying trait expression is complex, and gene expression changes do not necessarily correlate to phenotypic changes. Further, this study has implicated changes in cell size and density as possible mechanisms modulating genetic background effects. The genes of interest identified in this study have been shown to interact with either cell size or density in some way, and their RNAi expression across more backgrounds should be a priority future experiment.

This study is not without limitations. The two wildtype genetic backgrounds used here, while illuminating, represent only a fraction of the genetic variation due to genetic background. The inclusion of others (such as DGRP strains) would help to identify trends that only two backgrounds cannot capture. Additionally, the number of samples used in this study were not sufficient to estimate any non-linear effects with sufficient precision. The way that the models are set up, the effect of stronger alleles may mask that of the weaker ones, making it appear as if gene expression levels are not correlated to the magnitude of phenotypic effects. Though this might be the case, models for each allele, where gene expression is modeled as a function of the genotype (wildtype or allele) and background (SAM or ORE) will be needed to understand how individual alleles respond to genetic background and influence gene expression.

3.5 TABLES AND FIGURES

Effect of Interest	Coefficients Used	Rationale
Main effect of perturbation only	1. *P* < 0.1 & *B* > 0.1 & *P:B* > 0.1 2. *P* < 0.1 & *B* < 0.1 & *P:B* > 0.1	Interaction effects need to be ruled out, so in both cases a significant interaction term is excluded. Background effects are not of interest here so they may be significant or not – both cases are included.
Main effect of interaction only	*1.* *P* > 0.1 & *B* > 0.1 & *P:B* < 0.1 *2.* *P* > 0.1 & *B* < 0.1 & *P:B* < 0.1	Perturbation effects need to be ruled out, so in both cases a significant perturbation term is excluded. Background effects are not of interest here so they may be significant or not – both cases are included.

Main effect of perturbation and interaction between background and perturbation	*1.* $P < 0.1$ & $B > 0.1$ & $P:B < 0.1$ *2.* $P < 0.1$ & $B > 0.1$ & $P:B < 0.1$	In both cases there needs to be a significant interaction and perturbation term, however the background term can be either significant or not since it is not of interest.

Table 1. Categorizing gene responses to perturbation based on model coefficients and FDR adjusted p-values. *P* = perturbation, *P:B* = interaction between perturbation and background, *B* = background.

Gene name	Reason for interest	Category
Vajk2	Gene expression plot has very narrow confidence bands suggesting a sharp decrease in ORE expression as perturbation increases with a comparatively slower SAM decrease; *Vajk2* RNAi expression leads to small and blistered wing phenotype (Hevia et al. 2017); is a direct binding partner with *apolpp*	Perturbation and Interaction*
Orct2	Gene by expression plot has narrow confidence band where SAM expression stays consistent but ORE expression decreases steadily; *Orct2* RNAi expression leads to smaller wings where cell size but not number is affected (Herranz et al. 2006); is a transcriptional target of the insulin receptor pathway	Perturbation and Interaction*
P5CS	Gene expression by perturbation plot indicates that in SAM expression is stable while in ORE	Interaction only*

	expression increases, directly interacts with *CTPsyn* which acts downstream of *Myc* and is required for *Myc*-mediated cell size control (*Myc* RNAi expression leads to reduced wing size mediated by smaller wing disc cell size) (Johnston et al. 1999); wing phenotype has not yet been noted for this gene	
ranshi	Gene expression by perturbation plot indicates higher wildtype expression in SAM that decreases with perturbation while expression in ORE slightly increases, *ranshi* is a direct binding partner with *Mad*, which regulates expression of BMP response target genes in wing development and overexpression of *Mad* has been found to lead to altered wing shape and (Dworkin and Gibson 2006) and a reduction in wing size through non-cell autonomous apoptosis in the wing disc (Umemori et al. 2009)	Interaction only*

apolpp	Gene expression by perturbation plot indicates that in ORE expression increases while in SAM it slightly decreases; is a direct binding partner with *Vajk2*, a smaller wing disc phenotype has been noted in *apolpp* mutants (Panáková et al 2005); is also a direct binding partner of many genes involved in wing development including *wg* and *hh* and is required for signalling of these genes (Panáková et al. 2005)	Interaction only*

Table 2. Genes of interest identified by comparing change in gene expression by perturbation strength in the SAM and ORE wildtype genetic backgrounds.

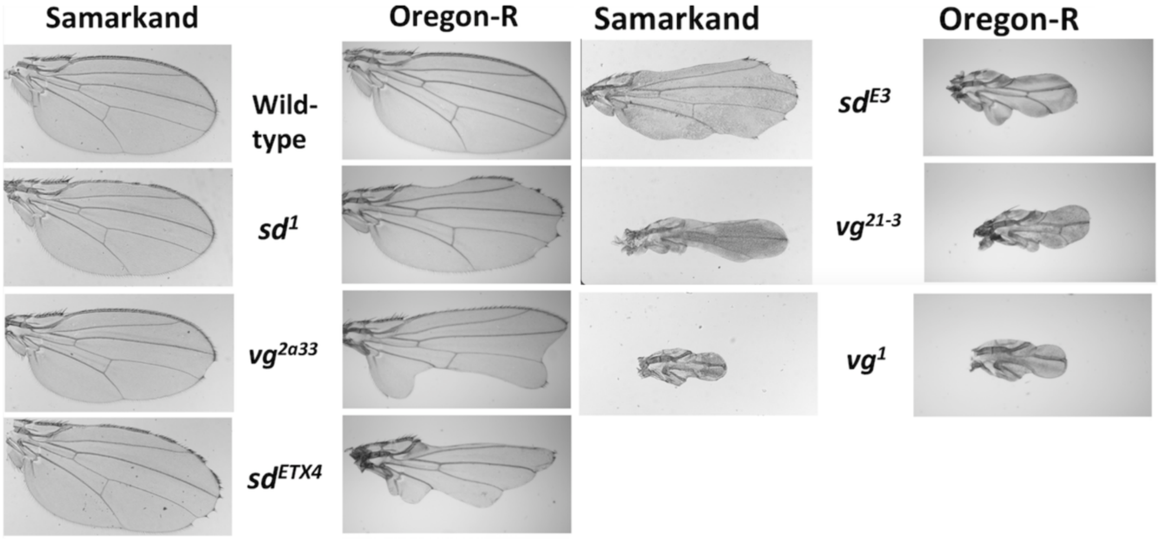

Figure 1. Wing phenotypes associated with each allele from the allelic series in both the SAM and ORE backgrounds. Adapted from (Chandler et al. 2017)

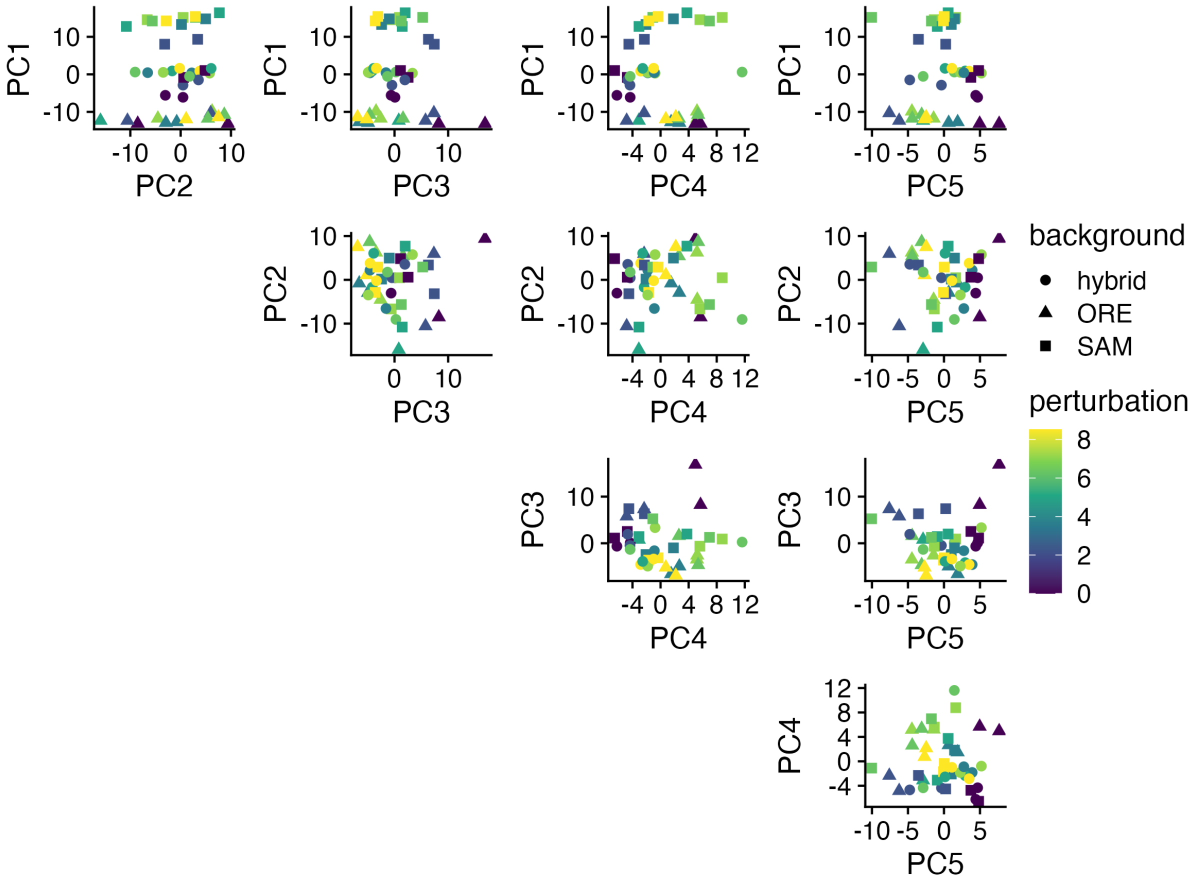

Figure 2. PCA of gene expression of 500 most variable genes suggests that global transcription correlates highly with genetic background. Plots of pairwise comparisons of principal components 1 through 5. Samples separate by wildtype genetic background and not by perturbation along PC1. Some variation from perturbation separates along PC3 and PC5, but not as distinctly as separation by genetic background. Counts are variance stabilized and represent the 500 most variable genes. Darker (blue) points represent the weakest perturbations, lighter

(yellow) represent the strongest, and shapes represent each genetic background and the hybrid background.

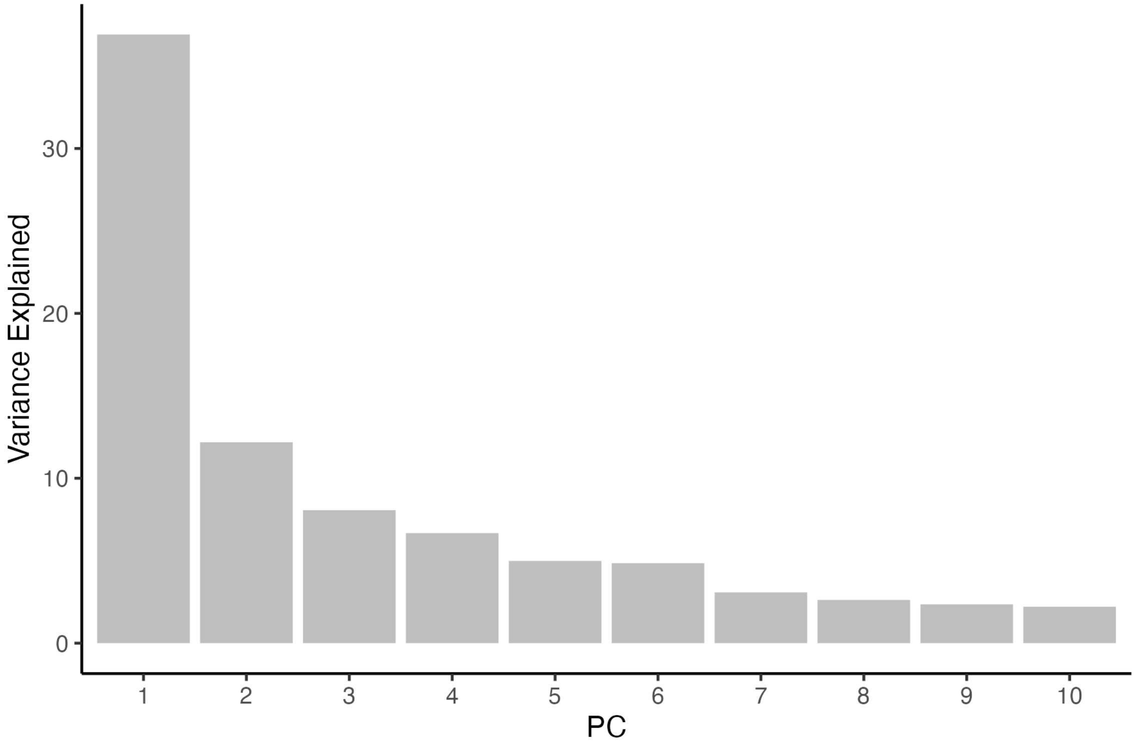

Figure 3. PC1, for which samples group along by genetic background, explains 36.9% of the variation in gene expression. The variance explained by each of the first 10 PCs. PC3 and PC5, for which samples group along by perturbation (to a weak extent) explain 8.07% and 4.99% of the variation in gene expression, respectively.

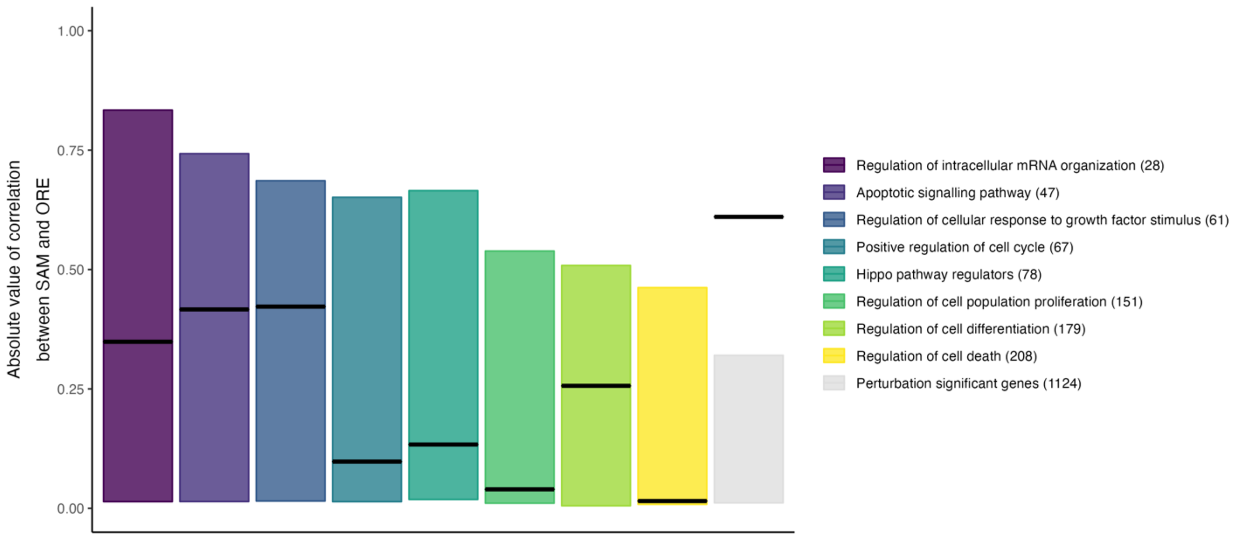

Figure 4. Genes involved with positive regulation of cell cycle, hippo signalling, regulation of cell population proliferation, and regulation of cell death respond differently to background and perturbation effect in SAM and ORE when compared to groups of random genes.

Solid black lines represent the absolute value of the correlation between the vectors of slopes of the genes for each of the indicated gene groups for the SAM and ORE backgrounds. A high absolute correlation indicates that the effect of genetic perturbation on the expression of genes in that group are similar for both backgrounds, with a low correlation indicating little similarity between effects. The coloured boxes represent the distribution of the top 95% absolute values of correlations from groups of 1000 genes, each group matched to the number of genes in the groups listed above. The grey "Perturbation significant genes" category is used as a positive control, and represents genes with significant perturbation values from the linear models at an FDR adjusted alpha of 0.1

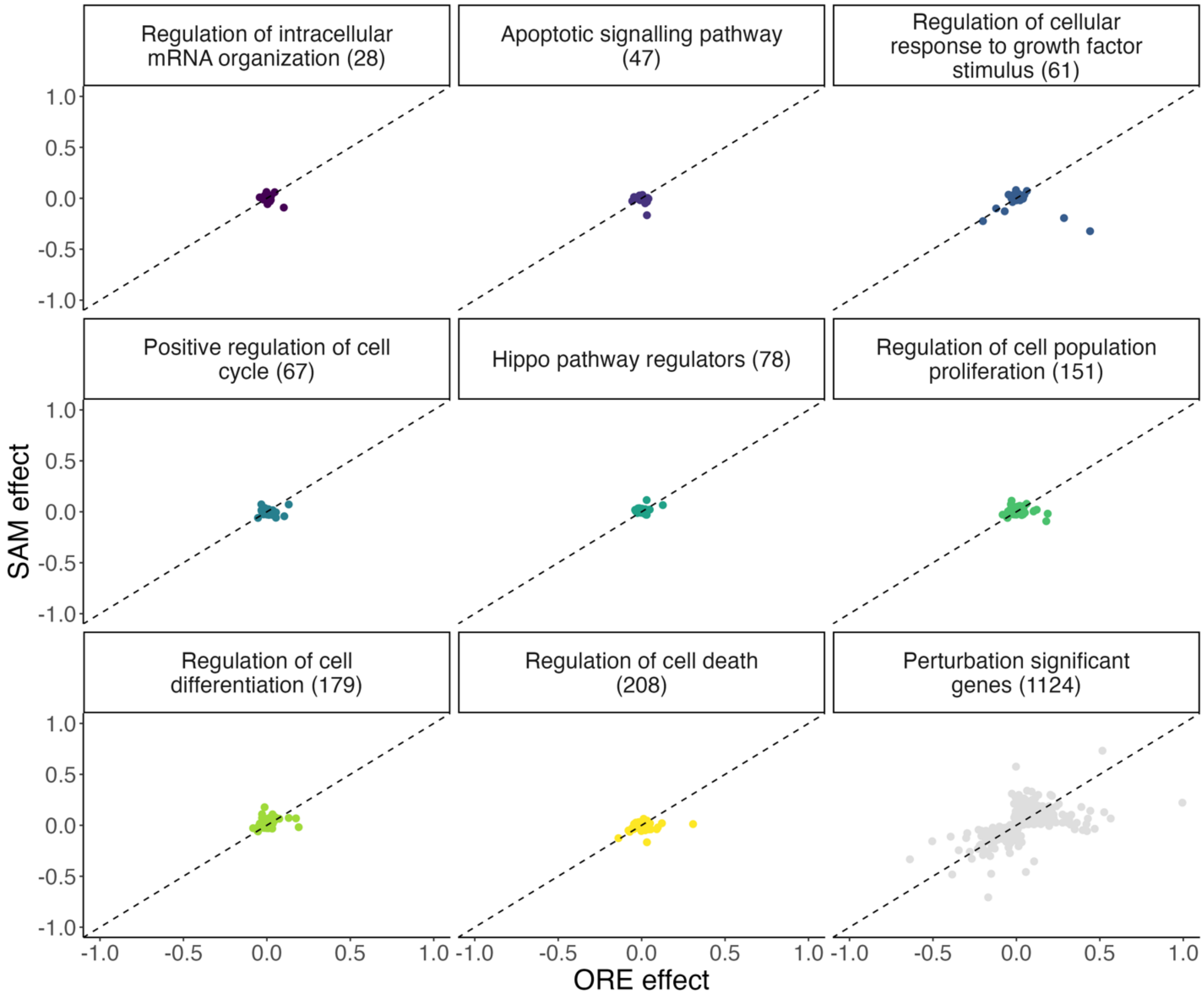

Figure 5. Scatter plot of gene slopes for each of the eight investigated biological pathways (plus the perturbation significant genes). The dashed line represents a correlation of 1 between the effect of change in expression with change in perturbation for a gene (in which case the effect between ORE and SAM will be identical).

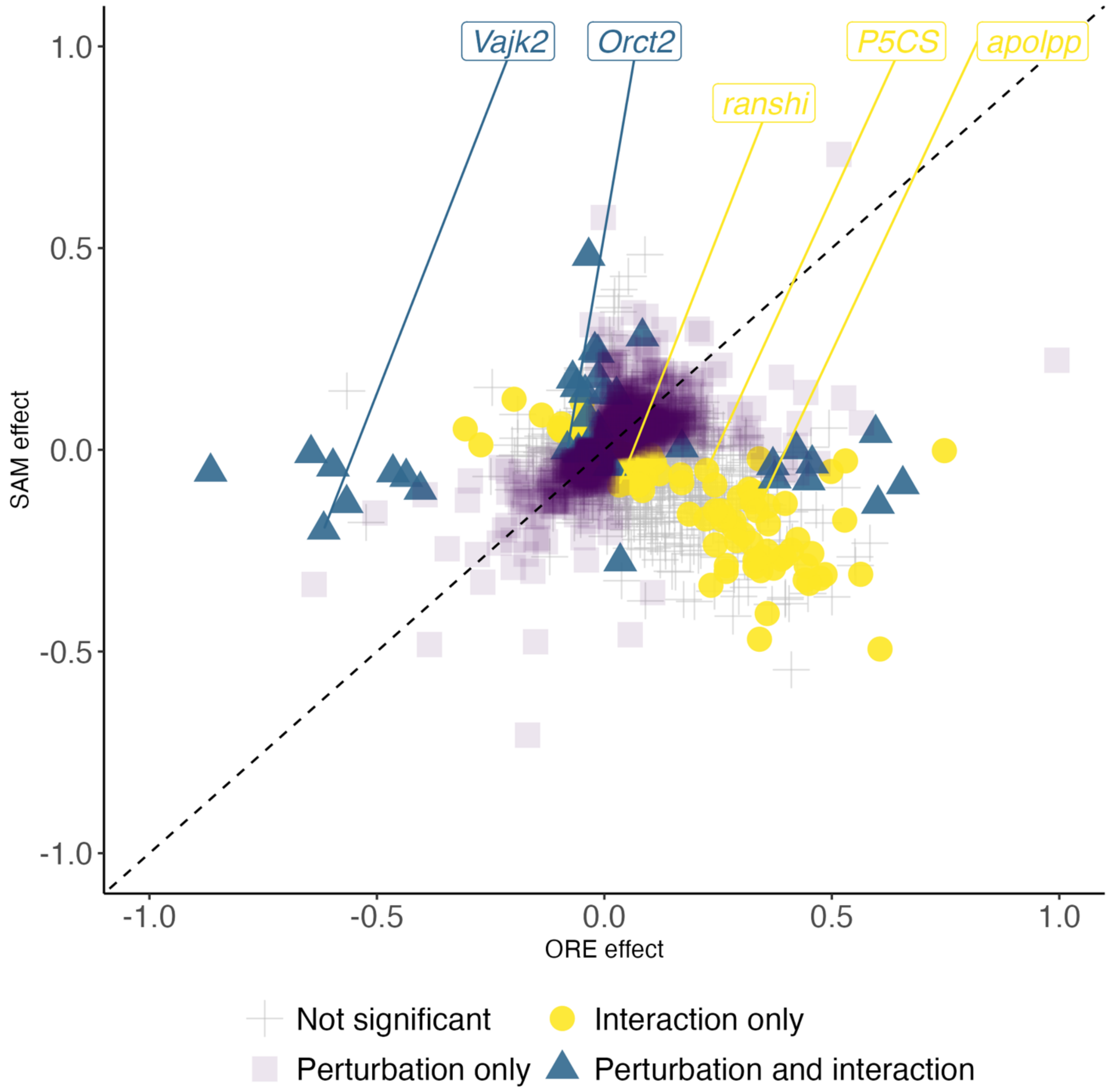

Figure 6. Scatter plot of the model estimates for both backgrounds of each gene. Points are coloured according to the significance of model terms (See table X). Dashed line represents perfect correlation between the SAM and ORE effects. Labels represent genes of interest for future functional analyses, based

on their individual gene expression by perturbation plots (Figure 7) and FlyBase entries.

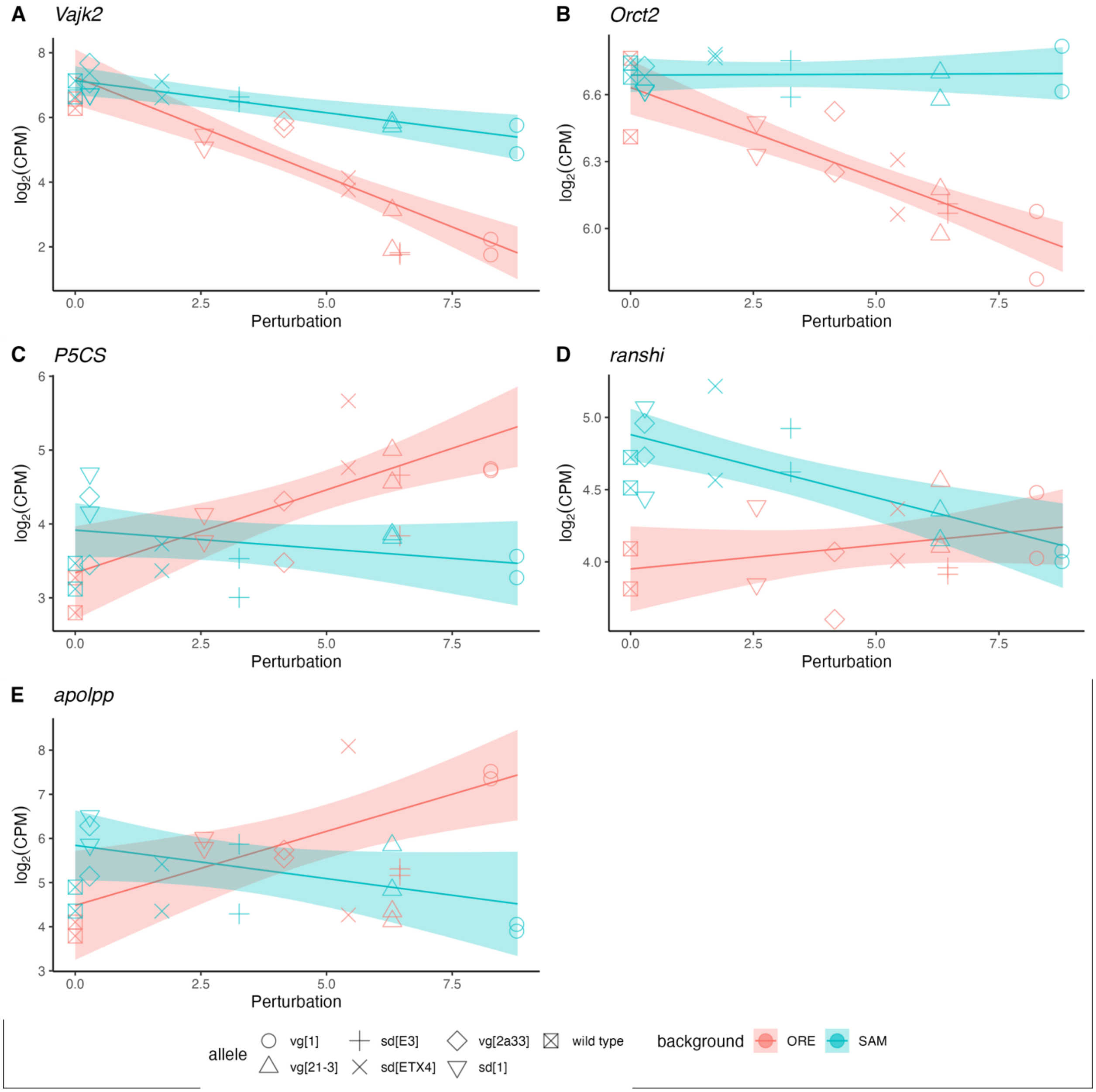

Figure 7. Gene expression by perturbation plots for genes of interest. Each plot has gene expression ($\log_2$(CPM)) on the y-axis and perturbation on the x-axis. Model regression estimates are shown by the solid lines with confidence bands

shaded. Shapes represent estimated gene expression. Colours represent the two wildtype genetic backgrounds for each gene.

3.6 SUPPLEMENTAL FIGURES

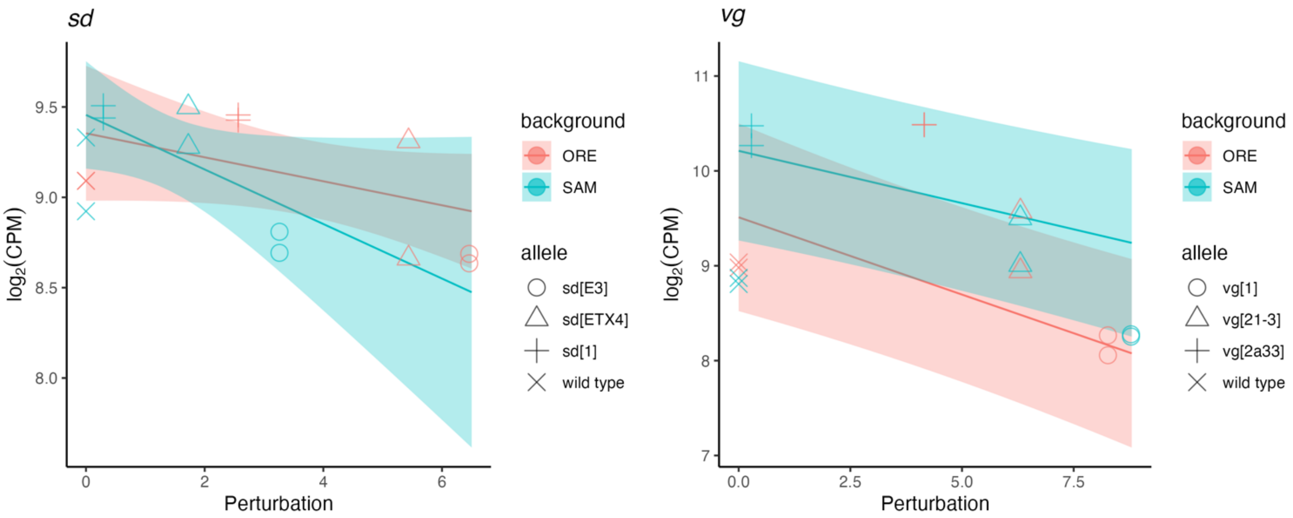

Figure S1. *Scalloped* and *vestigial* expression decrease with increase in perturbation. Gene expression by perturbation plots for *sd* and *vg*. Note that each plot contains expression for only the corresponding alleles rather than the full allelic series.

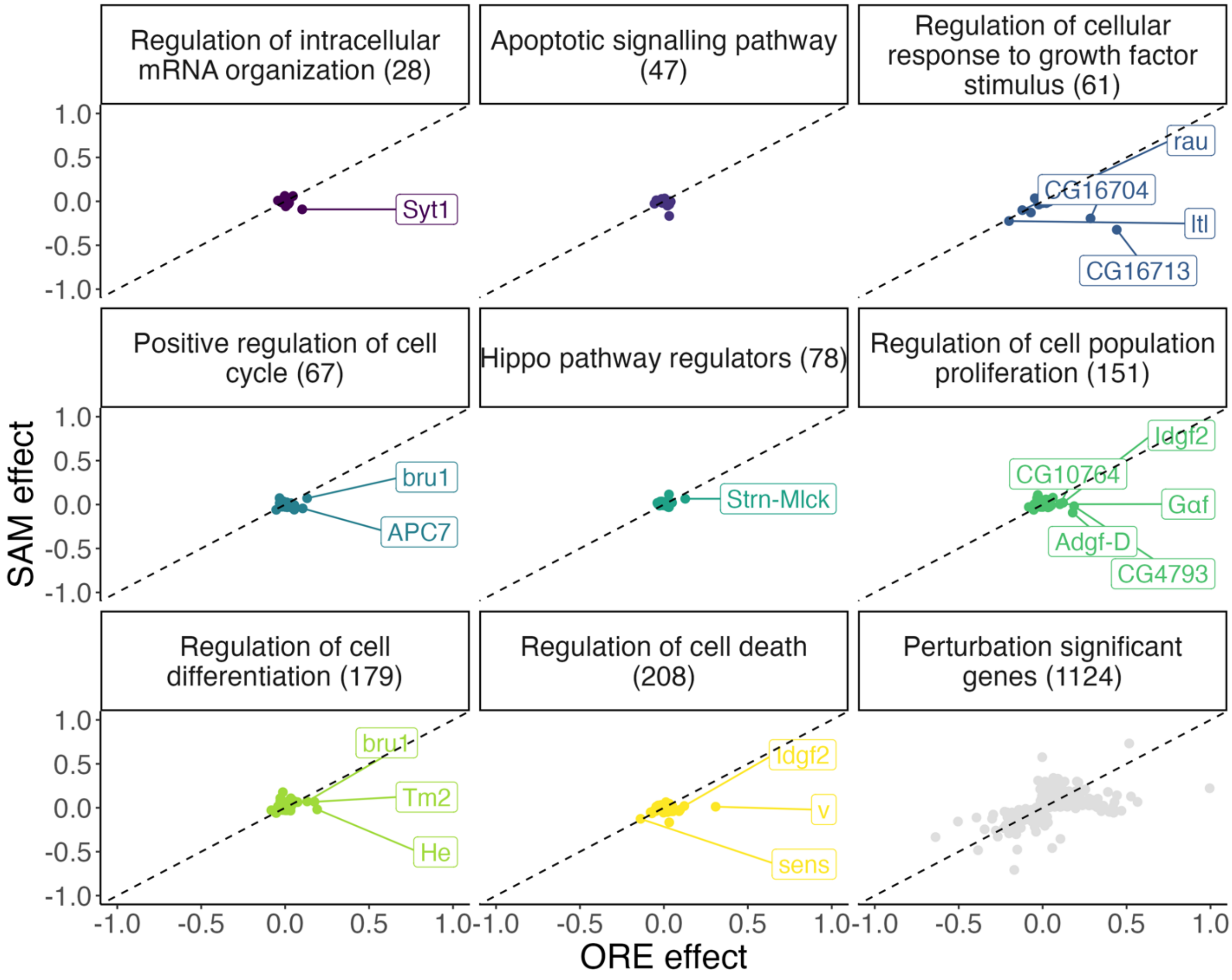

Figure S2. Plots of ORE slopes by SAM slopes for genes from each of the biological pathways investigated with "outlier" genes of interest highlighted.

Genes labeled are genes for which either the ORE slope < 0.1 and SAM slope < 0.1 or ORE slope < -0.1 and SAM slope < 0.1

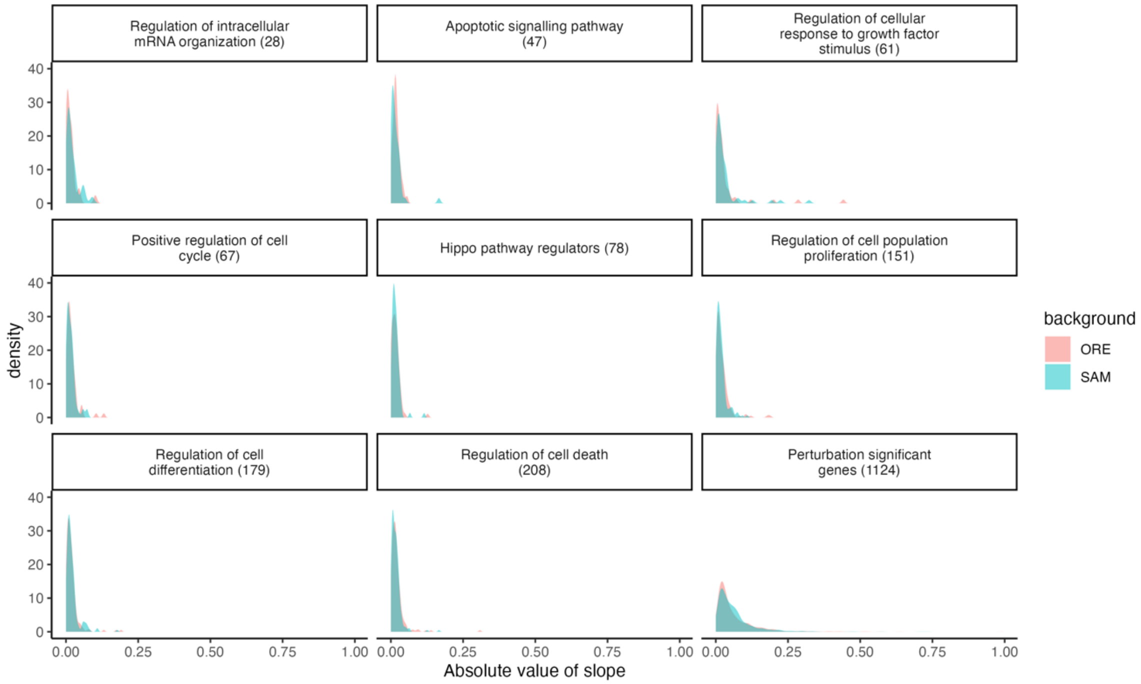

Figure S3. Density plots of slopes for genes from each of the biological pathways assessed.

3.6 REFERENCES

Altenburg, E., and H. J. Muller. 1920. The Genetic Basis of Truncate Wing,—an Inconstant and Modifiable Character in *Drosophila*. Genetics 5:1–59.

Arning, L. 2016. The search for modifier genes in Huntington disease - Multifactorial aspects of a monogenic disorder. Mol Cell Probes 30:404–409.

Bandura, J. L., and B. A. Edgar. 2008. Yorkie and Scalloped: Partners in Growth Activation. Dev Cell 14:315–316.

Bateson, W., and G. Mendel. 2013. Mendel's Principles of Heredity. Courier Corporation.

Brooks, M. E., K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Mächler, and B. M. Bolker. 2017. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. R J 9:378–400.

Campbell, S., M. Inamdar, V. Rodrigues, V. Raghavan, M. Palazzolo, and A. Chovnick. 1992. The scalloped gene encodes a novel, evolutionarily conserved transcription factor required for sensory organ differentiation in *Drosophila*. Genes Dev 6:367–379.

Chandler, C. H., S. Chari, A. Kowalski, L. Choi, D. Tack, M. DeNieu, W. Pitchers, A. Sonnenschein, L. Marvin, K. Hummel, C. Marier, A. Victory, C. Porter, A. Mammel, J. Holms, G. Sivaratnam, and I. Dworkin. 2017. How well do you know your mutation? Complex effects of genetic background on expressivity, complementation, and ordering of allelic effects. PLOS Genet 13:e1007075.

Chari, S., and I. Dworkin. 2013. The Conditional Nature of Genetic Interactions: The Consequences of Wild-Type Backgrounds on Mutational Interactions in a Genome-Wide Modifier Screen. PLOS Genet 9:e1003661.

Chen, R., L. Shi, J. Hakenberg, B. Naughton, P. Sklar, J. Zhang, H. Zhou, L. Tian, O. Prakash, M. Lemire, P. Sleiman, W.-Y. Cheng, W. Chen, H. Shah, Y. Shen, M.

Fromer, L. Omberg, M. A. Deardorff, E. Zackai, J. R. Bobe, E. Levin, T. J. Hudson, L. Groop, J. Wang, H. Hakonarson, A. Wojcicki, G. A. Diaz, L. Edelmann, E. E. Schadt, and S. H. Friend. 2016. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. Nat Biotechnol 34:531–538.

Cowley, M. J., C. J. Cotsapas, R. B. H. Williams, E. K. F. Chan, J. N. Pulvers, M. Y. Liu, O. J. Luo, D. J. Nott, and P. F. R. Little. 2009. Intra- and inter-individual genetic differences in gene expression. Mamm Genome 20:281–295.

Daley, Caitlyn. 2019. Examining The Predictability of Genetic Background Effects in The *Drosophila* Wing. Master's Thesis. McMaster University. http://hdl.handle.net/11375/25162

de Belle, J. S., and M. Heisenberg. 1996. Expression of *Drosophila* mushroom body mutations in alternative genetic backgrounds: a case study of the *mushroom body miniature gene (mbm)*. PNAS 93:9875–9880.

Delanoue, R., K. Legent, N. Godefroy, D. Flagiello, A. Dutriaux, P. Vaudin, J. L. Becker, and J. Silber. 2004. The *Drosophila* wing differentiation factor vestigial-scalloped is required for cell proliferation and cell survival at the dorso-ventral boundary of the wing imaginal disc. Cell Death Differ 11:110–122.

Di Tella, S., M. Ri. Lo Monaco, M. Petracca, P. Zinzi, M. Solito, C. Piano, P. Calabresi, M. C. Silveri, and A. R. Bentivoglio. 2022. Beyond the CAG triplet number: exploring potential predictors of delayed age of onset in Huntington's disease. J Neurol 269:6634–6640.

Dworkin, I., and G. Gibson. 2006. Epidermal Growth Factor Receptor and Transforming Growth Factor-**β** Signaling Contributes to Variation for Wing Shape in *Drosophila melanogaster*. Genetics 173:1417–1431.

Dworkin, I., E. Kennerly, D. Tack, J. Hutchinson, J. Brown, J. Mahaffey, and G. Gibson. 2009. Genomic consequences of background effects on scalloped mutant expressivity in the wing of *Drosophila melanogaster*. Genetics 181:1065–1076.

Fahed, A. C., M. Wang, J. R. Homburger, A. P. Patel, A. G. Bick, C. L. Neben, C. Lai, D. Brockman, A. Philippakis, P. T. Ellinor, C. A. Cassa, M. Lebo, K. Ng, E. S. Lander, A. Y. Zhou, S. Kathiresan, and A. V. Khera. 2020. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. Nat Commun 11:3635.

Fournier, T., and J. Schacherer. 2017. Genetic backgrounds and hidden trait complexity in natural populations. Curr Opin Genet Dev 47:48–53.

Halder, G., P. Polaczyk, M. E. Kraus, A. Hudson, J. Kim, A. Laughon, and S. Carroll. 1998. The Vestigial and Scalloped proteins act together to directly regulate wing-specific gene expression in *Drosophila*. Genes Dev 12:3900–3909.

Henderson, D. 2021. Can We Predict the Magnitude and Direction of Epistasis from Individual Allelic Effects? Master's Thesis. McMaster University http://hdl.handle.net/11375/26226

Herranz, H., G. Morata, and M. Milán. 2006. calderón encodes an organic cation transporter of the major facilitator superfamily required for cell growth and proliferation of *Drosophila* tissues. Development 133:2617–2625.

Hevia, C. F., A. López-Varea, N. Esteban, and J. F. de Celis. 2017. A Search for Genes Mediating the Growth-Promoting Function of TGF**β** in the *Drosophila melanogaster* Wing Disc. Genetics 206:231–249.

James, A. C., R. B. Azevedo, and L. Partridge. 1995. Cellular basis and developmental timing in a size cline of *Drosophila melanogaster*. Genetics 140:659–666.

Johnson, K. R., Q. Y. Zheng, and K. Noben-Trauth. 2006. Strain background effects and genetic modifiers of hearing in mice. Brain Res 1091:79–88.

Johnston, L. A., D. A. Prober, B. A. Edgar, R. N. Eisenman, and P. Gallant. 1999. *Drosophila myc* Regulates Cellular Growth during Development. Cell 98:779–790.

Little, T. J., and N. Colegrave. 2016. Caging and Uncaging Genetics. PLoS Biol 14:e1002525.

Mackay, T. F. C. 2014. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet 15:22–33.

Mathieson, I. 2021. The omnigenic model and polygenic prediction of complex traits. Am J Hum Genet 108:1558–1563.

Nadeau, J. H. 2001. Modifier genes in mice and humans. Nat Rev Genet 2:165–174.

Noben-Trauth, K., Q. Y. Zheng, K. R. Johnson, and P. M. Nishina. 1997. mdfw: a deafness susceptibility locus that interacts with *deaf waddler (dfw)*. Genomics 44:266–272.

Panáková, D., H. Sprong, E. Marois, C. Thiele, and S. Eaton. 2005. Lipoprotein particles are required for Hedgehog and Wingless signalling. Nature 435:58–65. Nature Publishing Group.

Partridge, L., B. Barrie, K. Fowler, and V. French. 1994. Evolution and development of body size and cell size in *Drosophila melanogaster* in response to temperature. Evolution 48:1269–1276.

Phillips, P. C. 2008. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9:855–867.

Shingleton, A. W., J. Das, L. Vinicius, and D. L. Stern. 2005. The Temporal Requirements for Insulin Signaling During Development in *Drosophila*. PLoS Biol 3:e289.

Simmonds, A. J., X. Liu, K. H. Soanes, H. M. Krause, K. D. Irvine, and J. B. Bell. 1998. Molecular interactions between Vestigial and Scalloped promote wing formation in *Drosophila*. Genes Dev 12:3815–3820.

Sittig, L. J., P. Carbonetto, K. A. Engel, K. S. Krauss, C. M. Barrios-Camacho, and A. A. Palmer. 2016. Genetic Background Limits Generalizability of Genotype-Phenotype Relationships. Neuron 91:1253–1259.

Umemori, M., O. Habara, T. Iwata, K. Maeda, K. Nishinoue, A. Okabe, M. Takemura, K. Takahashi, K. Saigo, R. Ueda, and T. Adachi-Yamada. 2009. RNAi-Mediated Knockdown Showing Impaired Cell Survival in Drosophila Wing Imaginal Disc. Gene Regul Syst Bio 3:GRSB.S2100.

Vu, V., A. J. Verster, M. Schertzberg, T. Chuluunbaatar, M. Spensley, D. Pajkic, G. T. Hart, J. Moffat, and A. G. Fraser. 2015. Natural Variation in Gene Expression Modulates the Severity of Mutant Phenotypes. Cell 162:391–402.

Zlotogora, J. 2003. Penetrance and expressivity in the molecular age. Genet Med 5:347–352.

CHAPTER 4: ONE MIGHT SAY THAT COMPLEX TRAITS ARE INDEED COMPLEX

In the first portion of this thesis, wing shape was used as a model complex trait to understand the relationship between gene expression variation in developing tissue and adult wing shape. The major finding of which was that the variation in gene expression among genes grouped by function has similar effects on the direction of variation in wing shape, even when correlated effects of gene expression were accounted for.

In the second portion, gene expression variation and its relationship to trait expressivity was studied using a model system for studying the joint effect of genetic background and magnitude of allelic perturbation on global gene expression variation. The takeaway was that wildtype genetic background has a profound effect on gene expression variability from across an allelic spectrum ranging from wildtype individuals to those with severe genetic perturbation. Additionally, genes to follow-up on with regards to the effect of variation in cell size and cell shape on background dependence have been suggested. In both chapters, the Hippo signalling pathway stood out as implicated with wing shape variation and context dependence.

Overall, complex traits are complex. I started this work with an interest biological diversity and the relationship between genotype and phenotype, and what I hope to have shown with this work is that understanding complex trait variation requires a holistic understanding of the systems of biological organization that contribute to phenotypic variation. Gene expression variation plays a role in complex trait variation and expressivity, and these effects are modified by the genetic background of an individual and the magnitude of genetic perturbation. As was shown in this thesis, developing ways to understand large, multivariate datasets such as gene expression and multivariate shape are crucial for understanding the effects of gene expression variation on complex trait variation. Vector correlations are a useful method to recover signal from datasets with few replicates that capture subtle effects. The Hippo signalling pathway was identified in both chapters presented here through gene expression variation, and it has been previously implicated for its role in wing shape development using GWAS. These results should point towards this pathway as a priority candidate for future research.