THE UNIVERSITY OF CHICAGO

USING *DROSOPHILA* NATURAL VARIATION TO STUDY THE ROLE OF POSITIVE SELECTION IN CIS-REGULATORY EVOLUTION AND THE GENETIC BASIS OF A COMPLEX DISEASE TRAIT

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY
BIN HE

CHICAGO, ILLINOIS
DECEMBER 2012

To my parents, Zhao Meijuan and He Jianyin, for bringing me up and encouraging my curiosity as a child.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I remember when I was rotating in Chung-I's lab in my first year, he once said (I paraphrased) "A scientist is like an athlete, who in the end is fighting alone." I agree with him, but I also believe that, just like athletes, we rely on and benefit from great coaches and wonderful teammates for our success. Looking back at my six-year journey of Ph.D., I could not feel more strongly about this. Let me begin by thanking my advisor, Marty Kreitman. He is first and foremost an incredible scientist. His sharp thinking, broad interests and deep insight never seize to enlighten me. He cultivated my interests in natural variation, which I now take as my own, striving to understand how its pattern is shaped by forces of natural selection and how it links to phenotypic variability. His high standard on the originality of scientific work inevitably influenced my judgement for what is good science. I am also grateful to him because he treated me more as a colleague than as a student. Even from the very beginning. he would spend most time with me chatting about interesting ideas and findings in science, but only rarely would he give specific instructions on my projects; rather, he placed full trust in me to define my own questions and to solve them. He would only come in at the end to ensure that the work is up to his standards, and if not, suggest how to improve it. That off-handed style of mentoring was not easy for me at first, but in the end it helped me to grow from a good student to an independent researcher, which I'm now proud of being.

I would like to thank the other four members of my thesis committee, including Misha Ludwig, Ilya Ruvinsky, Kevin White and Sebastian Maerkl. In particular, Misha's passion and original thinking on how enhancers originated and evolved keep inspiring me. I hope I would come back to this fascinating problem in my future work. I also owe nearly all my fly work skills to his patient teaching. I thank Ilya and Kevin for overseeing my progress, patiently listening to the early stage results in committee meetings and giving useful suggestions. I want to thank Sebastian, who, despite being on a different continent, keeps a close connection with me. I worked in his lab in EPFL, Lausanne during 2008 to 2009. Those two visits were among the most memorable time in my phd life, thanks to both the wonderful research experiences and the splendid natural beauty I was able to enjoy there. I especially appreciate Sebastian for showing me how to troubleshoot an experiment, or “make things work”, as he would say. By observing him, I learned that a combination of instinct-guided guesses and positive attitude towards failure are the key to success.

Other than my thesis committee, I also had the luck to work with many faculty members both in and outside the University of Chicago. Dick Hudson and Guy Sella, both of whom I considered the best population geneticists of the day, have generously spent time with me discussing my project. I also want to thank Graeme Bell, the co-PI on the insulin project, for bringing in not only his expertise on human disease genetics, but also his leadership and passion in science. While working on the insulin project, I collaborated with Dan Nicolae in the Stats department in UChicago, Scott Selleck at Penn State University and Trisha Wittkopp at University of Michigan, all of whom have taught me a lot. Trisha and her lab have been especially kind in hosting me for two research trips in Ann Arbor to get the pyro-sequencing work done for the second project.

Although I mentioned faculty members first, it is from my fellow co-workers – graduate students and post-docs in the department – that I learned the most during my Ph.D. I utterly enjoyed the vibrant environment in the Ecology and Evolution department and the closely-related Human Genetics department. I want to specially acknowledge past and current members of the Kreitman and Reinitz lab, for creating an intriguing and collaborative environment; I want to thank Kevin Bullaughey, Joshua Rest, Qiyan Mao, Antoine Barrire, Yong Zhang, Han Liang, J.J. Emerson, Xiang Zhou, Si Tang and indeed many many others whom I cannot list in full for their enlightening discussion. I also want to thank my friends outside the academia, including my Chinese friends in several departments, friends I made through learning French and our basketball team that used to meet every week. All those activities made my life so much richer! Finally, I want to thank my wife, He Bian, for her constant support and perhaps most importantly, being someone whom I can share anything with and indeed with whom I share most of my views upon life and the world, which I take as an invaluable asset in my life.

ABSTRACT

The goal of this thesis is to understand two central questions in evolutionary genetics: (1) What evolutionary forces shape the pattern of genetic variation? and (2) how does genetic variation result in phenotypic differences? These two questions are intimately connected as the result of the latter both fuels and limits the possibility for future adaptation, while natural selection acting on phenotypic variation determines the frequency of mutations.

In the first part, I examined the role of positive selection in cis-regulatory evolution. In comparison with the coding regions, where the importance of positive selection in shaping natural variation patterns has been established by both theoretical and empirical work, the role of natural selection in cis-regulatory regions has been more controversial. On one hand, genome-wide scans of noncoding DNA pointed to strong signals of positive selection, particularly within 5' and 3' UTR regions, where regulatory elements are enriched. On the other, empirical observations of a fast turnover (lineage specific gain and loss) of transcription factor binding sites (TFBS) contrasts with striking functional conservation of other regulatory sequences, which has prompted many researchers to propose neutral evolution under functional constraint. However, a rigorous population genetics approach has not been applied to formally evaluate these and alternative hypotheses. In this study I specifically tested the alternative hypothesis of natural selection driving the turnover of TFBS, using *Drosophila* enhancers as an example. By combining a population genetic approach with a high-quality dataset of TFBS and a state-of-the-art microfluidics technology, I found that the patterns of divergence and polymorphism are not consistent with the neutral hypotheses. Instead they strongly suggested the action of positive selection both in the gain of new binding sites and also in their loss. Consistent with this finding is a nuanced, two-timescale view of regulatory evolution. Frequent and subtle changes in function can occur on a short timescale and drive adaptive changes, while constraints fundamental to developmental processes and genetic network interactions act as a centripetal force and assure functional stability of regulatory components and interactions across a longer timescale. This view is also supported by empirical findings of subtle yet significant differences in the expression patterns driven by orthologous enhancers, whose functions were previously considered unchanged.

The second part of my thesis explores a novel approach of using *Drosophila* natural variation to study the genetic architecture of human complex diseases. The question of identifying the polygenic basis for common human disorders have gained increasing attention, due both to the advances in technology that made genome wide association studies (GWAS) in human possible, and the rising incidence of common diseases that increasingly burden our societies. Hampering this effort, however, is the inability to resolve more basic questions about the types of mutations producing complex traits, their mechanism of action (and interaction), their frequencies in population and their magnitudes of effects. To overcome some of the limitations faced by human studies, such as a low mapping resolution and difficulty in performing functional analysis, we developed a fly model approach, in which we first constructed a model for a Mendelian disease trait, which was subsequently turned into a genetically complex trait by crossing the mutant line into a diverse genetic background (178 inbred lines derived from a wild *Drosophila melanogaster* population). Employing both traditional GWAS approaches and a novel extreme selection scheme, the aim was to identify both common and rare variants underlying the continuously variable disease trait, and to dissect their genetic and molecular effects. The fast decay of LD combined with complete genome sequences enabled us to narrow down the association peak to a 400bp block containing an insertion/deletion (indel) polymorphism in the intron region of the gene *sfl*. Experimental analysis established the functional link between *sfl* and the human mutant proinsulin induced neuro-degeneration phenotype. RNAi analysis of additional genes in the same pathway strongly suggested a previously unknown link between Heparan Sulfate Proteoglycan (HSPG) and cellular responses to misfolded proteins. Finally, by performing allele-specific expression analysis, we revealed the potential mechanism of the intronic variation, suggesting that changes in expression level of *sfl* may be the cause for phenotypic variation.

The two studies highlighted the use of *Drosophila* as a model for understanding both the

evolutionary forces shaping patterns of genetic variation, and resolving the genetic basis for complex traits. In the future, I expect to use knowledge gained through the second part to construct models for investigating how natural selection operates on polygenic traits, which include most life traits such as height and weight, a challenging question in evolutionary biology now.

1

INTRODUCTION

> "...whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning **endless forms most beautiful and most wonderful** have been, and are being, evolved."
>
> Darwin, Charles (1859). On The Origin of Species.
> Chapter XIV, p. 503.

When looked at closely enough, every individual is unique, either by its genetic makeup or the environment in which it has developed and inhabits; In the meantime, however, astoundingly different life forms may possess great similarities at the molecular level, e.g. proteins whose function have hardly changed since the species had a common ancestor (Halder *et al.*, 1995). Underlying the amazing diversities or the surprising similarities are biological processes including mutation and recombination, which generate genetic diversity; environmental factors that makes even genetically identical twins differ, and natural selection, which can either keep gene functions unchanged, or in other cases, promote changes in a very short timescale as species adapt to changing environments. The ultimate goal of this thesis is to understand how natural selection shapes the pattern of genetic variation, and how genetic variation underlies phenotypic variability, which, in turn, serves as the source for future adaptation.

1.1 Questions and Approaches

This thesis consists of two parts: the first part (Chapter 2: Does Positive Selection Drive Transcription Factor Binding Sites Turnover? A Test in *Drosophila* Enhancers) explores the evolutionary forces that shape the pattern of natural variation, particularly focusing on cis-regulatory regions; the second part (Chapter 3: Natural Variation in *Drosophila* Modifies a Human Misfolded Protein Induced Eye Degeneration; 4: GWAS in Drosophila Synthetic

Population Resource (DSPR); 5: Extreme Selection to Identify Common and Rare Variants Influencing a Complex Trait) investigates the genetic architecture of a complex trait in a natural population.

While the two questions may seem distinct, I'd like to argue that they are indeed inherently connected. On one hand, it is common observation that the spectrum of phenotypic variation in natural populations can differ significantly from that of lab-generated mutations, for the reason that natural variation has been subject to various evolutionary forces – most importantly natural selection, which alters the frequency of individual and combinations (haplotype) of mutations. Consequently, understanding how evolutionary forces shape the pattern of natural variation provides the essential backdrop for solving the genetic architecture of a complex trait. On the other hand, knowing what genetic variation is available to influence a trait of adaptive or disease value is also critical for studying how evolution happens, because the former is the raw material that natural selection can operate on, and, importantly, also limits the ways that evolution can proceed by selection. This is because many mutations or combinations are simply not available or only present in very low frequencies either because the mutational input is infrequent, or more likely, they are too deleterious to persist. Therefore, I believe juxtaposing these two questions will benefit our understanding of both.

These two questions not only matter in evolutionary biology, but have also gained increasing importance in disease research. As we know, human common disorders such as diabetes have risen steadily in occurrences, affecting an increasing number of people in today's world. As it turned out, nearly all human common disorders have a complex genetic basis, precluding a simple diagnosis and specific treatment. A lack of understanding of how genetic variation affect the disease traits and how their patterns are shaped by historical forces greatly hamper the development of accurate diagnosis and effective drugs. While recent advances in Genome-Wide Association Studies (GWAS) have produced a great leap forward in terms of identifying novel genetic loci for a number of common disorders, their roles in the biology of disease remains to be worked out. An even more challenging question involves identifying epistatic interaction among loci, which is clearly at play (Carlborg and Haley, 2004; Cho *et al.*, 2012). Resolving these basic questions in human studies have proved extremely challenging, in part due to our large genome size, structure of haplotype blocks and difficulty in performing experimental validation and functional analysis.

Drosophila melanogaster and its related species have served as a indispensible model for studying the evolutionary forces in natural populations and the genetic basis of quantitative traits (McDonald and Kreitman, 1991; Smith and Eyre-Walker, 2002; Andolfatto, 2005; Bergland *et al.*, 2008; Dworkin *et al.*, 2009; Sella *et al.*, 2009; Miles *et al.*, 2011; Wang *et al.*, 2006; Mackay *et al.*, 2012). In this thesis I am going to exclusively utilize the *Drosophila* model to answer both questions, although implications for evolutionary processes in general or human biology more specifically are drawn wherever possible. The first question will be investigated in the context of transcription factor binding sites turnover, centered on a controversy involving a strictly neutral vs. adaptive evolutionary interpretation. For the second question, I will describe several novel approaches for studying the genetic architecture of complex traits, taking advantage of the abundant natural variation in *D. melanogaster* and the the availability of powerful genetic and molecular toolkits, which I show bear important implications for human complex disease studies.

Many efforts have been devoted to address both questions in evolutionary and quantitative genetics; some of them serve as valuable models for my work (Wittkopp *et al.*, 2004; Moses *et al.*, 2006; Burton *et al.*, 2007; Gibson, 2009; King *et al.*, 2012; Mackay *et al.*, 2012). However, two features may distinguish my work from previous efforts: 1) in terms of subjects, I focused primarily on cis-regulatory sequences; 2) in terms of methods, I strive to assign molecular and phenotypic effects to sequence variants whenever possible, in order to analyze them separately to gain further insight into the evolutionary dynamics. Rather than treating all variants indiscriminately, grouping them by their functional effects allowed me to relate the evolutionary dynamics to a biological function. I have endeavored to make both features clear as this thesis unfolds.

1.2 Evolution of Transcription Factor Binding Sites in *Drosophila* Enhancers

To investigate the first question, I chose to focus on cis-regulatory regions, more specifically, transcription factor binding sites, or TFBS. These short DNA motifs, generally 5-15 bp in length, serve as the recognition sites for sequence specific transcription factors (to be distinguished from general transcription factors). Together, they control the target gene expression in a temporally and spatially specific manner. TFBS usually form heterogeneous clusters – together with the intervening and flanking sequences whose functions are often unclear, they form a structure called an “enhancer”, a general feature in most higher eukaryotic genomes.

There are several reasons why transcription factor binding sites are suitable for studying the first question, i.e. how natural selection shapes the pattern of genetic variation. First and foremost, these short DNA motifs exhibit a dynamic pattern of change during evolution. One study estimated that 32-40% human functional binding sites are not functional in rodents (Dermitzakis and Clark, 2002). A similar analysis in 12 Drosophila species also revealed a high rate of binding site loss and gain, a process I will refer to as binding site turnover (or turnover for short) across the phylogeny (Kim *et al.*, 2009). This is in stark contrast to what we know of coding sequence evolution, where 80% of the mouse genes have a 1-to-1 ortholog in human (Mouse Genome Sequencing Consortium *et al.*, 2002), and the level of protein sequence conservation estimated in 1196 coding sequences (CDS) averaged 85% (ranging from 35% to 100%, Makałowski *et al.*, 1996).

Paradoxically, the functional output of enhancers, that is, gene expression patterns, appear to be highly constrained in at least a few developmental genes studied (Ludwig *et al.*, 1998; Ho *et al.*, 2009). A best known example of this is the stripe-2 enhancer for even-skipped in *Drosophila melanogaster*. It has been shown that orthologous enhancers from other *Drosophila* species as well as a group of distantly related sepsid flies, when assayed using a reporter gene in the *D. melanogaster* trans-background, drive nearly identical expression patterns (Ludwig *et al.*, 1998; Hare *et al.*, 2008).

This striking level of functional conservation contrasts sharply with the absence of apparent sequence similarity – the enhancer sequences between *D. melanogaster* and the Sepsid flies have diverged beyond the ability of conventional alignment techniques (Hare *et al.*, 2008). This, and similar observations, have prompted many researchers to suggest that the evolution of TFBS is dominated by genetic drift, and that the realized gain and loss of TFBS should cause no change to enhancer function (Hare *et al.*, 2008; Lusk and Eisen, 2010). However, this is difficult to reconcile with the finding that lab-generated mutations that disrupt combinations or individual binding sites in enhancer constructs were shown to significantly alter the expression pattern (Stanojevic *et al.*, 1991; Small *et al.*, 1992; Arnosti *et al.*, 1996). Some have hence put forward a compensatory neutral evolution model (Durrett and Schmidt, 2008) to explain the paradox. This model is consistent with the finding that whereas the sequences may diverge beyond being alignable, the types of TFBS (i.e. for which transcription factors) and the number of instances of each often remains similar, and in some cases are sufficient for specifying the expression pattern (Crocker *et al.*, 2008; Guss *et al.*, 2001). Therefore, it may seem that the lack of apparent sequence conservation masks the deep level conservation of the composition of TFBS. This phenomenon nevertheless begs the questions of how functionally critical binding sites are lost and/or new ones gained under the assumption that no functional changes were allowed in the enhancers.

One possibility, which I will argue for in this thesis, is that they have indeed caused significant, albeit small, changes to the expression pattern, and those changes are likely to underlie adaptive processes (Crocker *et al.*, 2010). In the second chapter, I will describe evidence supporting this view in the context of testing for positive selection acting on TFBS turnover.

Importance and Challenges in Studying Cis-Regulatory Evolution

Apart from the intriguing evolutionary dynamics of TFBS, cis-regulatory evolution in general is an important and yet less well understood aspect of evolution. As early as the 1970s (Britten and Davidson, 1971), cis-regulatory evolution has been proposed to be a likely source for evolutionary novelties. Later, the discordance between how little protein sequences differed vs. how much human is morphologically and behaviorally different from other primates led King and Wilson to make the famous prediction that most of the human specific evolution lies in regulatory changes (King and Wilson, 1975). While these initial claims remain speculative to this day, the importance of cis-evolution is now well supported by a number of empirical studies, where the genetic changes underlying within or between-species differences were mapped to regulatory rather than coding regions (Gompel *et al.*, 2005; Guss *et al.*, 2001; Frankel *et al.*, 2011; McGregor *et al.*, 2007; Jeong *et al.*, 2006; Prud'homme *et al.*, 2006). In some of these cases, the adaptive values of the morphological change is obvious and thus offers a direct link to adaptation (Chan *et al.*, 2010; Shapiro *et al.*, 2004; Jones *et al.*, 2012). Based on these observations, it has been argued that cis-regulatory changes have greater potential to be utilized by natural selection during adaptation than coding changes. Because enhancers and other regulatory elements are often modular, cis-regulatory changes are less likely to produce pleiotropic effects, a major factor hindering adaptive fixation.

Despite the importance and evidence of cis-regulatory evolution underlying adaptation, a direct proof of natural selection, especially positive selection, acting on regulatory changes is surprisingly rare. This is partly due to the difficulty of identifying enhancers and TFBS in large numbers with high confidence. Unlike for coding regions, there lacks a set of rules for predicting enhancers and defining their boundaries; the counterpart of the genetic code, which allows the interpretation of nucleotide changes in protein coding regions, is also not available for similar interpretations of changes in TFBS. Until recently, studies of TFBS evolution have been limited either to a small and highly specific set of binding sites, relying on pre-existing knowledge about the system (Bachtrog, 2008), or a large dataset mainly based on bioinformatic prediction, which suffers a high false positive rate (Kim *et al.*, 2009). In this thesis, I will take advantage of a curated set of high-confidence TFBS drawn from the literature and recently published polymorphism data for *D. melanogaster* and *D. simulans* to test whether neutral processes or positive selection drives TFBS turnover. Moreover, I applied a state-of-the-art microfluidic technology to help classify sequence variants within TFBS according to their predicted effects on binding affinity, a fundamental property of TFBS that might determine different evolutionary patterns. By analyzing them separately, I was able to gain a more detailed picture of how these short DNA motifs evolve and by implication, how selection might have utilized TFBS variation to modulate enhancer output.

1.3 The Genetic Architecture of Complex Traits

By genetic architecture of a trait, I mean the collection of genetic variation in a population that affect the trait. Specifically it concerns such properties as (1) which genes harbor causal variants, (2) the distribution of their effect sizes and allele frequencies, and (3) the molecular or developmental mechanism through which they act on the trait. Traditional genetic studies, such as forward mutational screens for a particular phenotype, generally reveal genes or genetic pathways with large, direct impact on the trait. Interestingly, however, these genes tend not to be the ones that harbor variation for the same trait in nature populations, likely because mutations in them have strong deleterious effects. By contrast, studies of the genetic architecture of a trait, by focusing on genetic variation in a population that underlie phenotypic variability, often reveal previously unexpected genes or pathways that likely exert their influences through indirect means (Morris *et al.*, 2012; Collins *et al.*, 2012; Sanders *et al.*, 2008).

Studies of the genetic architecture of a trait in a natural population may serve several purposes. In a medical setting, it can help predict the risk of disease or its age of onset, and provide novel targets for potential drug development; In an evolutionary framework, knowing what genetic variation is available for a trait is ultimately more important than knowing what genes can affect the trait, because only the former can be drawn upon by natural selection. For a variety of reasons, major effect genes may not be variable in the population.

The effort to decipher the genetic architecture for complex disease traits has become increasingly important in medical genetics. It has long been known that genetic background can have profound effects on an individual's risk or severity of disease. Even among Mendelian diseases, genetic modifiers are far from rare, influencing various aspects including the age of onset and reaction to drugs (Badano and Katsanis, 2002). The problem is more dire for human common disorders for two reasons. First, the genetic architecture for common disorders tend to involve a large number of genetic variants with very small individual effect size, which means they are poorly understood in most cases (although, see Wright *et al.*, 2010). Gene-by-gene and gene-by-environment interactions, both of which are difficult to study, are also thought to be common, further hampering development of novel theraputics. Second, the importance in understanding and treating human common disorders is more apparent with a rise in their incidence especially in developed countries, perhaps as a combined result of advances in diagnosis / awareness and lifestyle changes that predispose a larger proportion of the population to these diseases. This has posed a great health and economic burdens on modern societies. For example, in the US alone the number of people diagnosed with Type 2 Diabetes (T2D) has reached 25.8M (including 7M undiagnosed patients) – or 8% of the population (US, Jan 2011, Centers for Disease Control, 2011). Another 79M people are considered prediabetic based on fasting glucose and A1C levels. The medical cost for T2D treatment in 2007 has been estimated to be $218 billion (same source as above), accounting for 10.9% of the total medical expenditures (Kaiser Family Foundation, 2007); the latter consumes 16.2% of GDP. Similar to T2D, other common disorders such as cardiovascular diseases, obesity and cancer have all been rising in numbers. Due to the complex genetic architecture and strong gene-by-environment interactions, progress in drug development that is specific to a disease mechanism has been painfully slow despite huge investments.

Faced with these challenges, the human genetics community has responded with a coordinated effort applying genome-wide association studies (GWAS) to a list of human common diseases, with the goal of identifying genetic variants associated with disease phenotypes. The methodology of association studies is not new, but its genome-wide application is only recently made possible by advances in microarray and massively parallel sequencing technology, which dramatically brought down the cost of genotyping hundreds to tens of thousands of individuals at over a million positions in the genome. In the past five years, hundreds of human GWAS results have been published, leading to at least 2,000 robustly associated loci for a list of diseases and quantitative traits (Visscher *et al.*, 2012).

The progresses achieved through human GWAS is undoubtedly huge, yet significant limitations also exist. First and foremost, GWAS relies on linkage between the genotyped markers and the untyped causal variant; while the block structure of linkage disequilibrium (LD) in human initially aids association studies by reducing the total number of markers needed for genotyping, it eventually limits mapping resolution. As a large portion of the genome sequence reside in a small number of large LD blocks spanning 50kb or more (Gabriel *et al.*, 2002), association peaks often encompass tens of genes, a problem that cannot be completely solved by simply adding more individuals to the study. Not surprisingly, finding the causal mutations in human complex diseases is rarely achieved, thus impeding the identification of the molecular mechanisms. Another limitation stems from the difficulty in performing experiments with human subjects. As an alternative, cell lines and mouse models are employed when potential candidate genes are identified. These approaches, however, limit the speed and scale at which functional studies can be carried out.

To overcome some of these difficulties, we explored a novel approach, in which we combined a transgenic fly model expressing a misfolded, disease causing protein (human proinsulin) with the natural variation in the Drosophila melanogaster population and the powerful genetic and molecular toolkits available for the species. By crossing the transgenic line to a panel of fully sequenced inbred lines that are derived from a wild population, this Mendelian disease phenotype expresses as a continuously varying trait with an unambiguously complex genetic architecture. We then used this model system to fine map the underlying loci and to study the detailed mechanism of genetic variation affecting the trait.

2

DOES POSITIVE SELECTION DRIVE TRANSCRIPTION FACTOR BINDING SITES TURNOVER? A TEST IN *DROSOPHILA* ENHANCERS*

2.1 Abstract

Transcription factor binding sites (TFBS) turnover (i.e. lineage specific gain and loss) is a well-documented phenomenon in eukaryote cis-regulatory modules (CRM). The wide spread of the phenomenon and the appearance of conserved expression patterns for diverged orthologous CRM led to the standing view that the observed gain and loss of TFBS were functionally and selectively neutral. To the contrary, genome-wide population genetics analyses have unequivocally identified signatures of positive selection acting in noncoding regions in general, and particularly in 5' and 3' untranscribed regions of genes. To specifically test the neutral vs. selection hypotheses for the TFBS turnover process, I analyzed natural variation patterns within and between two closely related *Drosophila* species. I found that the patterns of divergence and polymorphism for two types of mutations – those inferred to increase or decrease the binding affinity respectively– were not compatible with a neutral hypothesis. Instead, multiple lines of evidence suggested that positive selection have contributed to gain as well as loss of TFBS in the two lineages, with purifying selection maintaining existing TFBS in the population. Spacer sequences also showed signatures of negative and positive selection. We propose a model of CRM evolution to reconcile the finding of frequent adaptive changes with constraints on long-term evolution.

. *A version of this chapter has been published in *PLoS Genet*, Vol. 7, No. 4. (28 April 2011), e1002053. The original publication is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

2.2 Introduction

Gene expression in eukaryotes is generally controlled by transcriptional enhancers, also called cis-regulatory modules (CRM), which are short regions in the genome consisting of a cluster of transcription factor binding sites (TFBS) spaced by intervening sequences (spacers). Individual TFBS have been shown repeatedly to be required for CRM function, yet surprisingly they evolve rapidly and are frequently gained and lost in evolution, attributes that have been demonstrated for a large number of CRM and transcription factors (Balhoff and Wray, 2005; Dermitzakis and Clark, 2002; Kim *et al.*, 2009; Moses *et al.*, 2006; Schmidt *et al.*, 2010). These observations pose a challenge to understanding the forces driving the process, especially in cases where CRM function has been preserved despite sequence and structural divergence (Gregor *et al.*, 2008; Hare *et al.*, 2008; Ludwig *et al.*, 1998).

The gain or loss of a TFBS is unlikely to be functionally irrelevant, as repeatedly shown in TFBS knockout experiments (Arnosti *et al.*, 1996; Shimell *et al.*, 2000; Swanson *et al.*, 2010), and also demonstrated for the evolved differences between two species by a chimeric enhancer study (Ludwig *et al.*, 2000). One possibility for reconciling conservation of CRM function with rapid TFBS turnover is to assume that each loss of a TFBS is precisely balanced by the simultaneous gain of a cognate TFBS elsewhere in the CRM, a process we will call compensatory evolution (Ludwig and Kreitman, 1995). The idea draws on a model first proposed by Kimura (Kimura, 1985), where he considers a pair of tightly linked mutant genes that are individually deleterious but in combination restore wildtype function. As applied to TFBS, the gain of a novel site on an allele carrying a mutation that decreases the quality of an existing binding site can offset the mutants fitness cost, creating a selectively neutral double-mutant allele. Binding site turnover - fixation of the double mutant allele - is achieved entirely by genetic drift, thus preserving both CRM function and population fitness. Recently, a theoretical model of this compensatory turnover process was developed to ask about the feasibility of compensatory evolution for TFBS (Durrett and Schmidt, 2008). With plausible assumptions about the mutation rate, population size and selection coefficient on the individual mutations, a completely neutral model cannot achieve a high enough level of turnover to explain *Drosophila* CRM evolution (as exemplified by *eve* stripe 2 enhancer), whereas a model that assumes the double mutant to be more fit than the wildtype does.

This theoretical finding raises the prospects for positive selection being an important driving force of TFBS gain and loss. Instances of directional selection have been documented in cases where a novel regulatory regime is favored (Ihmels *et al.*, 2005). Functional evolution of a transcription factor (TF) can also drive adaptive co-evolution of its TFBS (Kuo *et al.*, 2010; McGregor *et al.*, 2001; Shaw *et al.*, 2002). Broad-scale studies in noncoding regions and promoters of genes have identified signatures of both selective constraint and positive selection in fruitfly and human (Andolfatto, 2005, 2008; Haddrill *et al.*, 2008; Kohn *et al.*, 2004; Torgerson *et al.*, 2009). However, only a small number of population genetics studies have been carried out to specifically test this hypothesis with TFBS or CRM, and because they focus on a single TF or CRM, they have low statistical power to distinguish between neutrality and selection (Ludwig and Kreitman, 1995). The generality of the conclusions reached in these studies is also not established (Bachtrog, 2008; Macdonald and Long, 2005).

Several different approaches have been designed to detect and quantify selection in the system. One of them has been to consider the genome-wide ensemble of TFBS as evolving at mutation-selection balance, with the fitness of each instance of TFBS being strictly determined by its binding energy (Doniger and Fay, 2007; Kim *et al.*, 2009; Mustonen and Lässig, 2005). This approach proves useful in studying the strength of selective constraints on functional TFBS. However, the assumption of a unidirectional fitness function, i.e. selection always favors affinity-increasing mutations and against affinity-decreasing ones, could be violated if the loss of a TFBS were favored or gain (or strengthening) of a TFBS is deleterious. Another approach calculates the sum of mutational effects in TFBS on binding affinity and compares it to the expectation under a no-selection model (Moses, 2009). A higher than expected sum could imply selective removal of affinity-decreasing mutations and therefore the action of purifying selection. Applying this approach to two of the CRM also included in this study, the author provided evidence for purifying selection acting to preserve the functional TFBS in the anterior *Bicoid*-dependent *hunchback* enhancer and the *even-skipped* stripe 2 enhancer. This test can also be used to detect positive selection, although its power is limited due to the mixed signal with purifying selection, which is expected to be dominant in most cases.

In this study, patterns of polymorphism and divergence are investigated in a pair of closely related *Drosophila* species, *D. melanogaster* (*mel*) and *D. simulans* (*sim*). The short evolutionary distance between the two species ensures unambiguous alignment for noncoding sequences and also allows one to capture the potentially rapid dynamics of TFBS gain and loss. A notable challenge in studying TFBS turnover is assembling a high quality set of TFBS that are precisely defined and contain few false positives. Large numbers of potential TFBS can be identified by methods involving genome-wide scans, such as computational prediction or ChIP, but these approaches generally include a large fraction of false positives, thus reducing their attractiveness for investigating the mechanisms of binding site turnover (see Discussion). Instead, we chose to investigate a curated set of high-confidence TFBS identified by DNaseI footprint in well-studied *D. melanogaster* CRM. Short footprint regions usually contain only a single TFBS motif, which, in most cases, could be perfectly aligned with the other species to allow identification of single nucleotide differences within and between the species. Each of these differences, in turn, was evaluated for the predicted magnitude and direction of effect on TF binding energy. The neutral and selection models generate distinguishable predictions in both divergence to polymorphism ratios and in the site frequency spectra. Analysis of these patterns reveal evidence for purifying selection against affinity-decreasing mutations segregating in the population, while multiple lines of evidence indicate positive selection for both gains and losses of TFBS. These empirical findings challenge the prevailing view of neutral compensatory turnover, and have important implications for understanding CRM functional evolution. In the course of the analysis, we also identified and modeled a potential ascertainment that can impact population genetics studies of genomic features that have been identified only in a reference sequence such as TFBS.

2.3 Results

Our analysis focuses on single nucleotide polymorphism (SNP) and divergence in 645 experimentally identified TFBS for 30 transcription factors in 118 autosomal CRM (Table S.1), all annotated in REDfly (Gallo *et al.*, 2010). These 645 TFBS represent the complete set for which we could obtain unambiguous alignment of both within- and between-species sequences without insertion or deletion. We used position weight matrices (PWM) both to identify TFBS within footprints and to predict the magnitude of binding energy differences among variant alleles. Our bioinformatic and experimental validations showed that the PWM used in this study provide reliable and unbiased estimates for the direction of binding affinity change in both *mel* and *sim* (Materials and Methods).

Single nucleotide changes within or between *mel* and *sim* were polarized with outgroup sequences from *D. sechellia, D. yakuba* and *D. erecta* using PAML (Materials and Methods). Each derived mutation, therefore, could be categorized with respect to species lineage and to direction of binding affinity change.

2.3.1 Lineage specific gain and loss of TFBS as a general pattern

Binding sites for an individual TF or a single CRM usually had too few counts of single nucleotide polymorphism or fixed differences to allow informative statistical analysis. Furthermore, the breadth of the turnover phenomenon across almost all investigated TF and CRM suggests a common underlying evolutionary mechanism (Bradley *et al.*, 2010; Hare *et al.*, 2008; Ludwig *et al.*, 1998; McGregor *et al.*, 2001; Moses *et al.*, 2006). We therefore considered pooling observations from across TFs and CRM. To see if the evolutionary rates in different TFs binding sites are sufficiently uniform, we measured sequence divergence between *mel* and *sim* for the 30 TF. After accounting for sample sizes, no significant departure from the average rate is detected by a binomial test (Figure 2.1). Moreover, the pooling approach should be conservative in deriving a general pattern with respect to among TF variations.

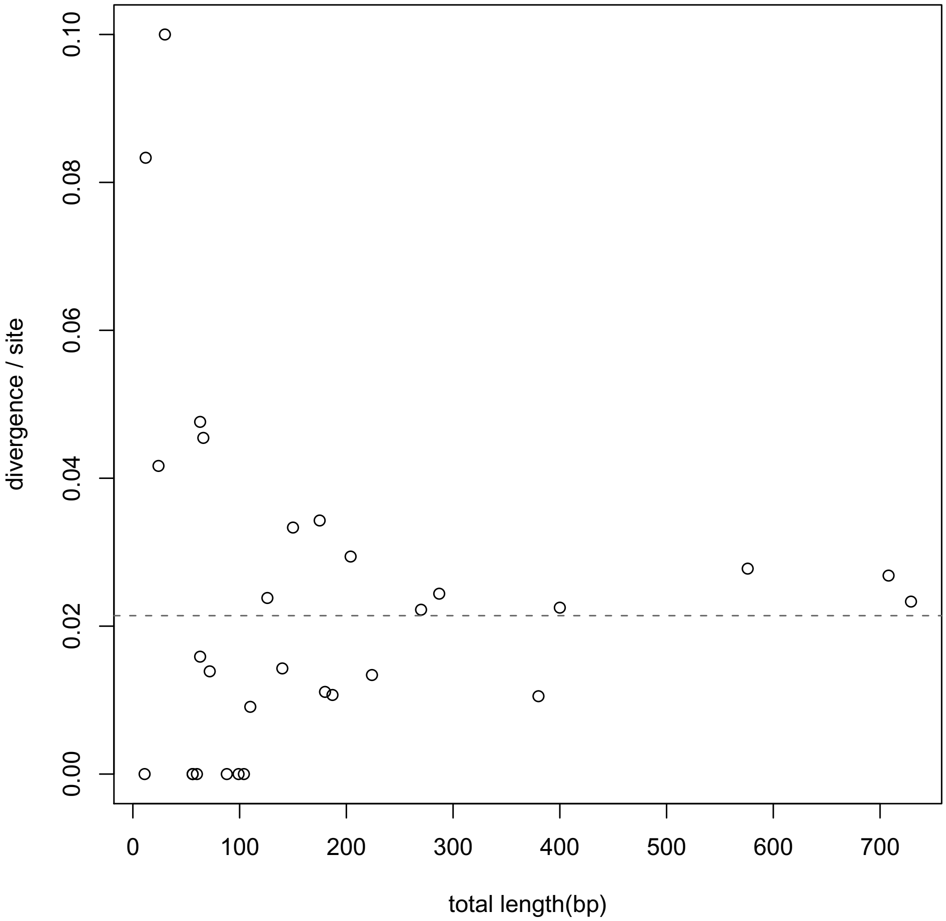


Figure 2.1: **TFBS divergence for 30 TF.** TFBS divergence for 30 TFs is plotted as a function of the total number of nucleotides assigned as binding sites to that TF. A maximum likelihood estimate of the mean divergence is marked by the dashed line. Individual binomial tests find no evidence for heterogeneity in divergence rates among the 30 TFs (0.05 significance level, with Bonferroni correction for multiple testing).

We then estimated percent loss and gain of TFBS on the *mel* and *sim* lineages. For each of the 645 footprint TFBS, a PWM score $S(k)$ was calculated for each occurrence (k) in the alignment of *mel*, *sim* and the inferred *mel-sim* ancestor, by taking the log2 ratio of the probability of a sequence under the functional motif distribution vs. that under the genomic background distribution (Material and Methods). Using $S = 0$ as a cutoff, approximately 2% of all footprint sites were found to be present in *mel* only and may represent *mel* specific gains; and about 2.5% were present in the inferred ancestor (and *mel*) but lost in *sim.* A set of empirical cutoffs were determined for each TF based on the range of PWM scores among its footprint sites, which produced similar results (Table S.2). Consistent with the sequence divergence patterns, gain and loss of TFBS appear to be a general pattern across TF and CRM. A total turnover rate of 4.5% between *mel* and *sim* is slightly higher than a previous finding of 5% for the TF Zeste in four *Drosophila* lineages (*D. mel, D.sim, D.yak, D.ere*) (Moses *et al.*, 2006).

We observed approximately equal numbers of gains vs. losses in our dataset, although the distribution of these events is asymmetric on the two lineages (16 losses, 0 gain along the *sim* lineage vs. 12 gains, 0 losses along the *mel* lineage). This is not unexpected, given that all footprint TFBS were identified as being present in *mel* and the dataset doesn't include *sim*-specific TFBS. We predicted that identification of TFBS by computational methods would produce a more even pattern of gains and losses in both lineages. We tested this prediction for three TF (Hb,Bcd,Kr) using a stringent cutoff procedure and for each TF we found a similar total number of predicted binding sites in the two lineages (Figure S.1). We thus rejected the (unlikely) possibility that there has been a large-scale evolutionary gain of TFBS in *mel* and loss in *sim.*

2.3.2 Classify mutations by direction of affinity change and deal with ascertainment bias

Gain and loss of TFBS may be subject to distinct evolutionary forces. To investigate them separately, we assigned each mutation within a footprint TFBS in *mel* or *sim* to either affinity-increasing or affinity-decreasing group based on PWM score difference between the ancestral and the derived mutation (Materials and Methods). Bioinformatic and experimental investigation showed that this PWM-based procedure for inferring the direction of binding affinity change is reliable when PWM predicted magnitude of change is not too small (Materials and Methods, Figure S.2,S.3). We established a threshold corresponding to a PWM score difference of one, i.e. at least two-fold change in the likelihood ratio between a motif or background distribution, in order to minimize the chance for mis-assignment. Varying this threshold between zero and two do not affect the results qualitatively.

We employed two approaches to investigate evolutionary forces acting on affinity increasing and decreasing changes. One approach is based on contrasting polymorphism and divergence patterns in a McDonald-Kreitman (MK) test framework (McDonald and Kreitman, 1991). Positive selection is expected to inflate substitution relative to polymorphism while negative selection will have the reverse but weaker effect (Sawyer and Hartl, 1992). We used synonymous changes in the target genes for the CRM as a proxy for a neutrally evolving class. Following established practices, we further classified each synonymous change as according to its expected impact on codon bias – No-Change, Preferred-to-Unpreferred, or Unpreferred-to-Preferred – and used the No-Change class as the neutral reference. The second approach investigates the site frequency spectrum of TFBS polymorphism to make inferences about selective pressures acting more recently on binding sites.

The fact that all footprints were identified in *mel* impacts the analysis in two ways. First, gains of TFBS can be observed in *mel* but not losses, while the reverse is true in *sim*. Therefore, even though similar processes are most likely operating in both species, our evolutionary analysis of binding site gain will focus on changes in the *mel* lineage, whereas losses will be restricted to changes in the *sim* lineage.

Second, affinity-decreasing and affinity-increasing mutations have the potential to differ in detectability as a footprint site in *mel*. This arises because mutations in TFBS were sampled conditioned on the TFBS being detected in *mel* and affinity-changing mutations in *mel*, in turn, have the potential to affect the detectability of the TFBS. Depending on whether the derived mutation is affinity-increasing or affinity-decreasing, two distinct biases are introduced in the expected neutral frequency spectrum (Figure S.4). Given that the dataset consists only of TFBS that are detectable by footprinting, we assume that the high-affinity allele will always be detectable. Consider the possible situation in which the low-affinity allele is not detectable as a footprint: if the derived mutation is affinity-decreasing, the probability of detecting the TFBS will change inversely with the mutant allele frequency; conversely, if the derived mutation is affinity-increasing, the probability of detection will increase with the mutant allele frequency. Substitutions may be viewed as a special instance of a segregating mutation and treated similarly.

This effect of ascertainment on neutral expectations for the MK test and the site frequency spectrum can be modeled analytically (Supplementary Text S.1); there is no ascertainment if both alleles are equally detectable as footprints. To incorporate uncertainty in the detectability of the low-affinity allele, the model incorporates a parameter, f, which specifies the probability that the weaker affinity allele will not be detected in the footprint assay. While f is likely to be greater than 0, it is unlikely to be close to 1 because footprint sites are degenerate and span a range of affinities. Under the conservative assumption that the lowest affinity among the footprint sites is the detection limit, we estimate $f = 0.27 \pm 0.20$ for the 30 TF (Supplementary Text S.1), indicating that the majority of TFBS changes will be detectable.

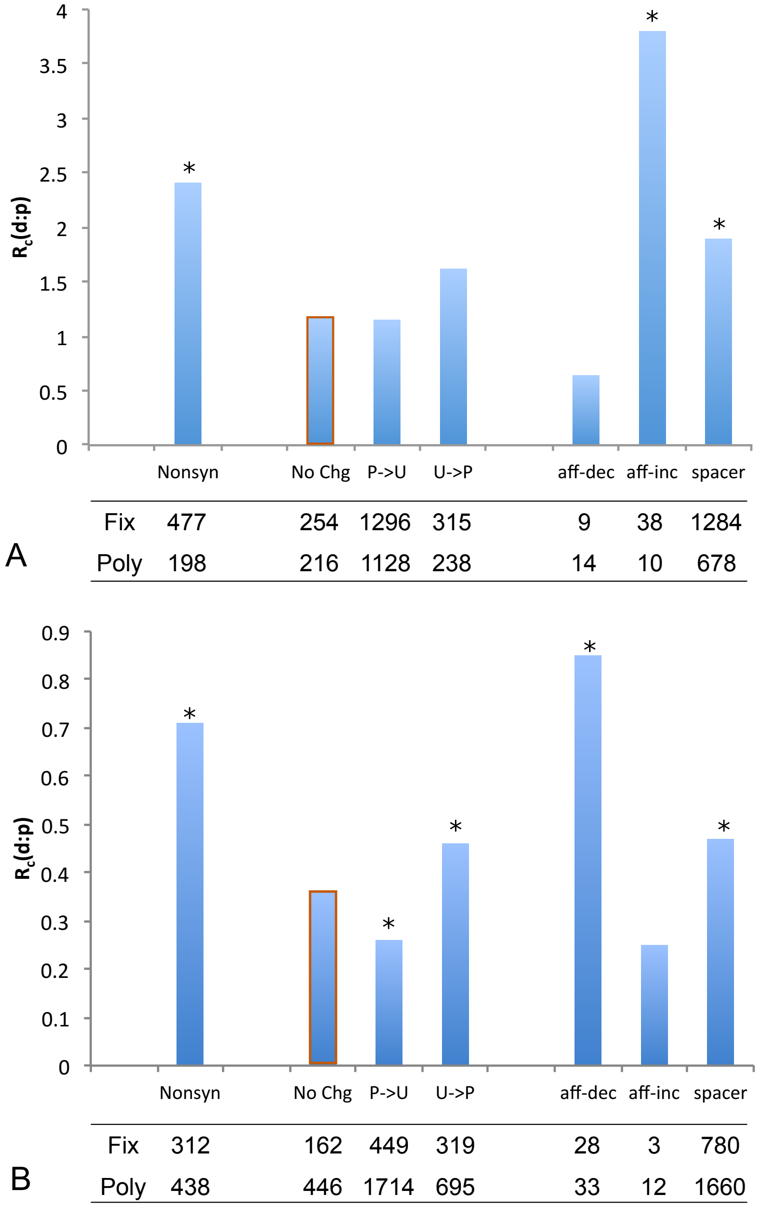
In the following sections, we first present our analysis of polymorphism and divergence in *mel*, focusing on the forces acting to either maintain functional TFBS or to create new ones. We then turn to *sim*, focusing on TFBS loss. Finally, we analyze the spacer sequences between TFBS in both species.

2.3.3 Analysis in mel *suggests positive selection for TFBS gain and purifying selection in maintaining existing TFBS*

For each class of change we summarized the data in the MK table by calculating the ratio, $R(d:p) = \#$ substitution / # polymorphism. The presence of weakly deleterious mutations can mask signatures of positive selection, and if removed can improve the power of the test (Fay *et al.*, 2001). Since most deleterious mutations will be at low frequencies, using 15% as a frequency cutoff has been shown to achieve most of the benefits of a more sophisticated model incorporating the distribution of deleterious effects (Charlesworth and Eyre-Walker, 2008). We applied this cutoff and denote the ratio of substitutions to common polymorphism by $R_c(d:p)$. Under this procedure, $R_c(d:p)$ is significantly higher for nonsynonymous changes than for the synonymous No-Change class (Figure 2.2), consistent with previous findings of positive selection driving amino acid substitutions in *Drosophila* (Smith and Eyre-Walker, 2002).

To delineate the effect of ascertainment from that of selection for the affinity-increasing and affinity-decreasing mutations, we compared the observed $R_c(d:p)$ to the expected neutral ratios under the ascertainment with different f values (Supplementary Text S.1). For affinity-decreasing mutations in *mel*, the difference from the synonymous No-Change class is not statistically significant, even in the absence of ascertainment bias (Figure 2.2A). This seems to suggest only neutral or deleterious mutations are present for this class and therefore no positive selection is involved. The validity of this conclusion can be questioned, however, because any affinity decreasing substitutions in *mel* that led to the loss of a site will not be included in the data while our correction for the ascertainment only accounts for neutral changes but not a potential adaptive excess. Thus, rejection of the neutral model in favor of positive selection is not possible for affinity-decreasing mutations in the *mel* lineage. However, this test is possible for the *sim* lineage (reported in the next section), where the loss of a TFBS is observable.

For affinity-increasing mutations no amount of ascertainment under our model can ac-



A

	Nonsyn	No Chg	P->U	U->P	aff-dec	aff-inc	spacer
Fix	477	254	1296	315	9	38	1284
Poly	198	216	1128	238	14	10	678

B

	Nonsyn	No Chg	P->U	U->P	aff-dec	aff-inc	spacer
Fix	312	162	449	319	28	3	780
Poly	438	446	1714	695	33	12	1660

Figure 2.2: **Substitution-to-polymorphism ratios in *mel* and *sim*.** $R_c(d : p)$ ratios between number of fixed mutations (fix) in each class and number of common polymorphisms (poly; with derived allele frequency > 0.15) for (A) *mel* and (B) *sim*. In *sim*, only TFBS with a predicted ancestral PWM score > 2 are included. Synonymous changes are categorized according to the predicted effect of a mutation on codon preference (P: Preferred codon; U: Unpreferred codon; No Chg: P→P and U→U). Consistent with previous reports, we find evidence for selection on biased codon usage in *sim* but not *mel*. Statistical significance of each class relative to the neutral reference (the No-Change class, outlined in orange) is evaluated by Fishers exact test. Classes that are significant at a 0.05 level (two-sided test) are marked with an asterisk above the bar.

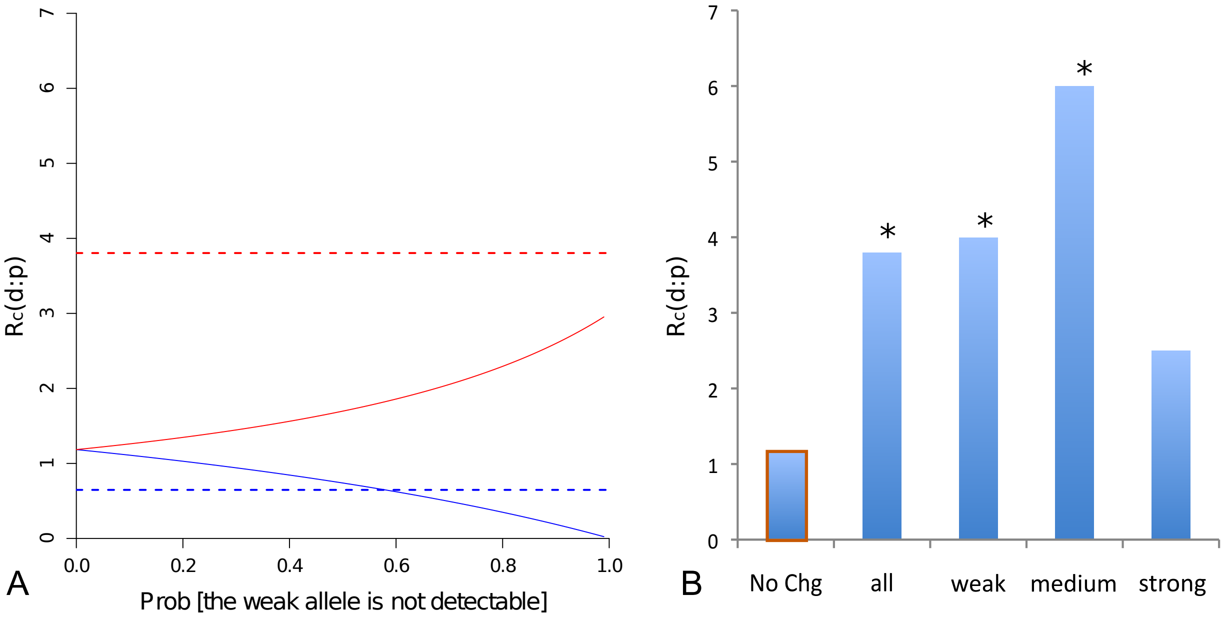


Figure 2.3: **Substitution-to-polymorphism ratios after correction for ascertainment suggests positive selection on affinity-increasing mutations in *mel.*** (A) The expected neutral $R_c(d:p)$ ratio under ascertainment (solid line) as a function of the probability that the weaker allele will not be detectable as a footprint for affinity-decreasing (blue) and affinity-increasing (red) mutations. Dashed lines represent the observed ratios for the two classes respectively. (B) Observed $R_c(d:p)$ for affinity-increasing mutations within TFBS grouped by predicted ancestral PWM score, compared to the No-Change class (orange box). An asterisk above the bar indicates statistical significance at a 0.05 level by Fishers exact test.

count for the observed relative excess of substitutions (Figure 2.3A red). We further reasoned that the ascertainment effect should be weaker or non-existent for TFBS with an ancestrally strong binding affinity, which would be identified with or without the affinity-increasing mutations. We therefore investigated whether the excess of affinity-increasing substitutions differed if TFBS changes were grouped according to the strength of the inferred ancestral binding affinity. We found a consistently larger $R_c(d : p)$ ratio, i.e. an excess of substitutions, across the entire range of inferred ancestral binding affinity classes compared to the No-Change class, including binding sites with the strongest ancestral binding affinity (Figure 2.3B). These results collectively suggested that positive selection has contributed to the fixation of affinity-increasing changes.

To further investigate evolutionary forces acting on the segregating mutations in TFBS in the population, we utilized the site frequency spectrum, for which we generated the neutral expectations for affinity-increasing and affinity-decreasing mutations separately under ascertainment, with $f = 0$ or $f = 1$ (corresponding to no bias or complete bias, respectively). For affinity-decreasing mutations, with the ascertainment expected to shift the frequency spectrum to lower frequency classes (Figure 2.4A, blue vs. grey), the observed spectrum is shifted in that direction but is even more extremely so than the complete bias expectation (Figure 2.4A, orange vs. blue). Since $f = 1$ is clearly an overestimate (compared to our estimate of $f = 0.27 \pm 0.20$), this strongly suggests that forces other than ascertainment must have shaped this pattern. Both a recent selective sweep and population growth can produce an excess of rare variants and one or both mechanisms may be acting in this system, as is suggested by our finding that synonymous changes also show a relative excess of low frequency mutations (Figure S.5B). However, as we compared the site frequency spectrum of the affinity-decreasing mutations to that of synonymous sites (corrected for ascertainment), we found the former is again more significantly shifted than the latter (Figure S.6). Thus we suggest that the observed frequency spectrum is consistent with on-going purifying selection against affinity decrease in functional TFBS. The observed frequency spectrum for

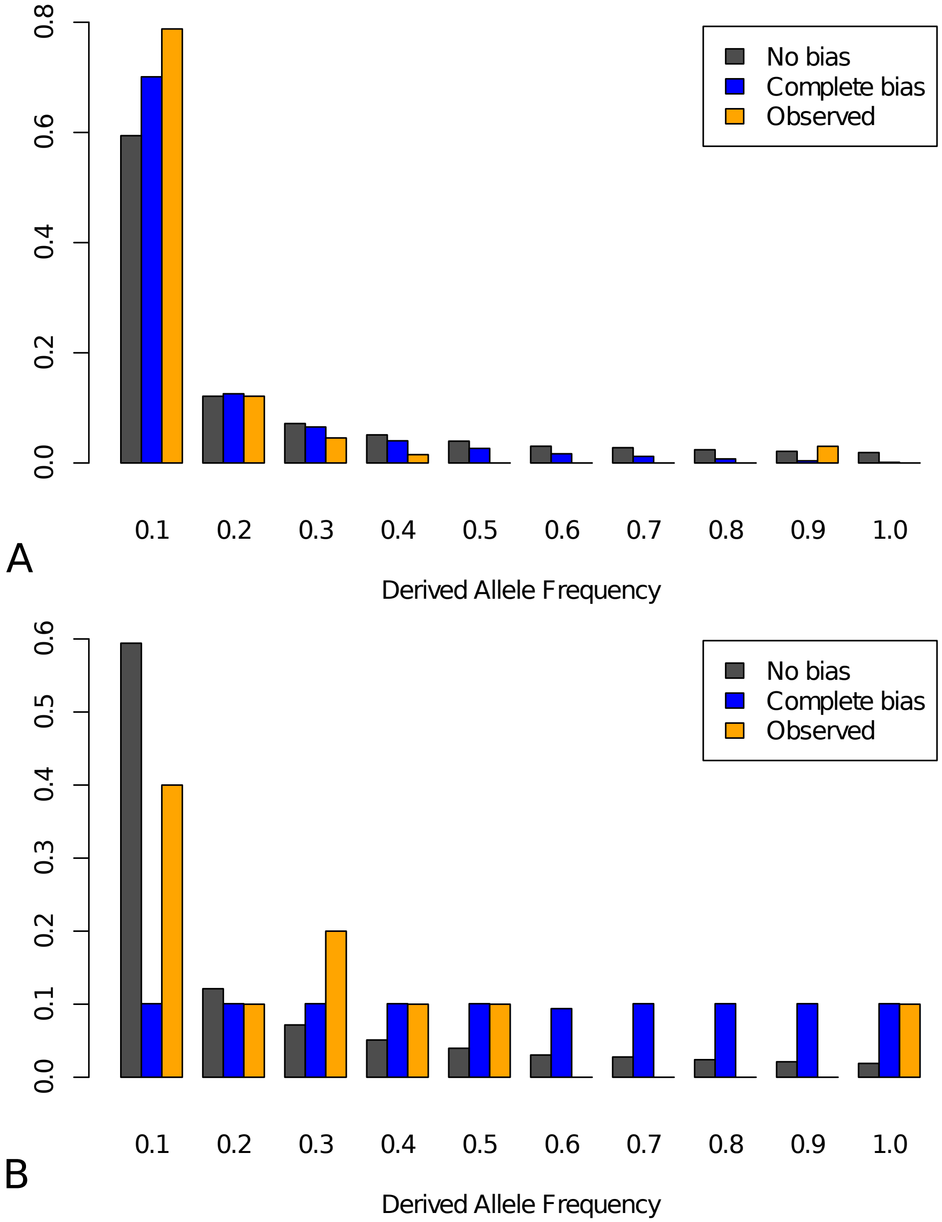


Figure 2.4: **Site frequency spectra in *mel* suggests purifying selection on affinity decreasing mutations.** (A) affinity-decreasing mutations and (B) affinity-increasing mutations. Grey: neutral expectation with no ascertainment ($f = 0$); Blue: neutral expectation under complete ascertainment ($f = 1$); Orange: observed frequency spectrum. The calculations of the expected frequency spectrum under no bias and complete bias are described in supplementary methods. The total number of segregating sites for affinity-decreasing and affinity-increasing mutations is 64 and 15, respectively.

affinity-increasing mutations lies between the two expectations and the differences are not significant from either one, a possible consequence of the small sample size (15 observed affinity-increasing polymorphisms) (Figure 2.4B). Thus, while positive selection is indicated on the basis of the MK test, inference cannot be made about on-going selection for affinity-increasing mutations.

2.3.4 Analysis in sim *suggests loss of TFBS may be adaptive*

Patterns of polymorphism and divergence in *sim* are not influenced by the ascertainment because the identification of TFBS in *mel* is independent of the effect of mutations fixed or segregating in *sim*. However, the inclusion of binding sites gained in *mel* may confound the analysis as their orthologous sequences in *sim* may have evolved under less or different kinds of selective constraints. We thus restricted the analysis to footprint TFBS predicted to be present in the *mel-sim* common ancestor, where we found a significant excess of substitutions for the affinity-decreasing mutations compared to the synonymous No-Change class (Figure 2.2B, Fisher's Exact Test $P = 0.003$). Statistical significance of this pattern is robust to the cutoff for excluding binding sites gained in *mel* (Table S.3). A relative excess of substitutions might also be a consequence of factors other than selection, such as systematic differences in the genealogical histories of CRM vs. synonymous sites. However, these factors seem unlikely to be the cause of this type of departure from neutrality in these two species (Kohn *et al.*, 2004). Therefore we consider positive selection a more plausible explanation.

We also compared the ratio between affinity-decreasing and affinity-increasing mutations in polymorphism to the expected ratio of the two classes in the mutational input, i.e. the probability for a new mutation to be one of the two classes (Materials and Methods). Briefly, the expected ratio was obtained by considering all possible mutations in each of the 645 footprint TFBS and their predicted effects on binding affinity the same way as we did before. Assuming polymorphism for both classes were neutral, we expected similar ratios, whereas the observed results showed a significant deficit of affinity-decreasing polymorphism relative to affinity-increasing polymorphism (Table 2.1), which may suggest that among new mutations, affinity-decreasing ones are more likely to be deleterious, a result consistent with our finding based on frequency spectrum in *mel*. A similar approach has been applied before, using the sum of ΔS (individual mutation's effect on binding affinity predicted by PWM) within a CRM instead of counts of mutations in binary classes (Moses, 2009). There the author also found evidence for purifying selection against affinity-decreasing mutations. The finding of both on-going purifying selection and potentially positive selection acting is not dissimilar to patterns found in nonsynonymous changes (Smith and Eyre-Walker, 2002). We reserve for the Discussion section the attempt to reconcile the adaptive loss of TFBS, as observed between the two species, with on-going purifying selection against affinity-decreasing new mutations.

Table 2.1: **Mutational probability of affinity increase and affinity decrease.**

Affinity-Class	**Mutational Probability**	**Observed**[a]	**Expected**	**Chisq p-value**[b]
Increase	0.105	12	4.7	
Decrease	0.895	33	40.3	0.002

[a] number of segregating mutations of each class among the six *sim* lines.
[b] chi-square test p-value is based on 10,000 simulations.

2.3.5 "Spacer" sequences might contain large numbers of unidentified functional elements

In both *mel* and *sim* we found a significant excess of substitutions in spacer sequences, indicative of positive selection in these intervals (Figure 2.2). Also, the frequency spectrum for this class is strongly shifted towards lower frequencies (Figure S.5E, Tajima's D = -1.09), indicative of on-going purifying selection. The implication of these results is that spacer sequences might contain many unidentified functional elements, for example, TFBS for known or uncharacterized transcription factors, or perhaps other structural features not yet understood.

To summarize, analysis of TFBS changes in *mel* indicates on-going purifying selection against affinity-decreasing polymorphism in the population, and positive selection for affinity-increasing substitutions. In *sim*, the analysis of affinity-decreasing changes indicates a significant, and potentially adaptive excess of substitutions that contributes to binding site loss. Spacer sequences between footprint TFBS in these well-characterized CRM also exhibit patterns of polymorphism and divergence consistent with both functional constraint and adaptive evolution.

2.4 Discussion

Natural selection, both positive and negative, has been shown to act throughout noncoding regions of the Drosophila genome (Andolfatto, 2005; Haddrill *et al.*, 2008), albeit with varying intensities (Kohn *et al.*, 2004). Against this backdrop of ubiquitous selection in noncoding DNA, should it be surprising to find signatures of positive selection in *Drosophila* TFBS? We think not. More surprising perhaps is the incompatibility of this finding with the model of neutral compensatory binding site turnover, a simple and appealing mechanism that allows for both rapid binding site turnover and functional stasis of CRM activity. But as explained below, there are good reasons to doubt whether a strictly neutral compensatory process can actually generate rapid TFBS turnover in *Drosophila*, even with its favorably large population size. Positive selection, in contrast, can drive arbitrarily fast rates of binding site turnover; the question is whether it can also allow for functional stasis of CRM activity. Below, we first discuss the strengths and limits of our analysis and then we describe properties of gene regulatory networks that can promote adaptive binding site turnover and yet also constrain the function of CRM.

Our population genetics analysis identified three major forces in TFBS evolution. First, we found functional TFBS were selectively maintained in the population by purifying selection, as revealed by a frequency spectrum skewed towards rare variants for affinity-decreasing polymorphism in *mel* and a significantly reduced proportion of affinity-decreasing polymorphism compared to mutational input in *sim*. These results are consistent with previous findings of selective constraints on functional TFBS. Mustonen and Lässig estimated that the average selection coefficient to maintain TFBS in bacteria and yeast genomes are on the order of $2N_es = 10$ (Mustonen and Lässig, 2005; Mustonen *et al.*, 2008), and a similar estimate has been obtained for *Drosophila* (Kim *et al.*, 2009). The substitution rate with $S = 10$ is expected to be less than 0.05% of the neutral rate in a population with a size as large as *Drosophila* (Equation B6.4.1, (Charlesworth and Charlesworth, 2010)). This means TFBS loss is unlikely to happen through fixation of deleterious mutations (0.2 losses expected for 645 footprint TFBS vs. 16 inferred in *sim*). We can think of only three mechanisms by which TFBS loss can occur at an appreciable rate: (1) there is loss of constraint; (2) a pair of tightly linked compensatory mutations creates an effectively neutral allele; or 3) positive selection drives the loss of TFBS.

Our second finding – a significant excess of substitutions compared to the neutral class for affinity-decreasing mutations in *sim* – is consistent only with positive selection for TFBS loss. Occasional adaptive loss of a TFBS is not inconsistent with more ubiquitous selection to maintain binding sites (Mustonen and Lässig, 2005), and has been suggested to account for the evolution of fermentation pathways in yeast (Ihmels *et al.*, 2005).

Our third finding is positive selection contributing to the gain of TFBS, as revealed by a significant excess of substitutions for affinity-increasing mutations in *mel*. Collectively, the three findings indicate that natural selection is extensively involved in the maintenance, gain, and loss of TFBS. This conclusion challenges the prevailing view of a neutral TFBS turnover process (Kim *et al.*, 2009; Ludwig and Kreitman, 1995).

We think that a selectionist interpretation of the turnover process is plausible for several reasons. First, the assumption of CRM functional stasis, which is the main argument for the neutral (i.e., compensatory) view, is not well supported experimentally. Reporter transgene assays, in particular, are limited in their quantitative resolution, and yet even in these studies, repeatable differences were found between orthologous CRM (Hare *et al.*, 2008). A functional rescue experiment is potentially more sensitive than a reporter transgene assay. As applied to the *Drosophila even-skipped* stripe 2 enhancer, it demonstrated clear functional differences between CRM that were previously believed to have the same spatial pattern of expression (Ludwig *et al.*, 2005).

Second, compensatory neutral evolution cannot account for the patterns of variation observed in this study. According to this model, affinity-decreasing mutations should in general be deleterious but occasionally become "effectively" neutral when a second compensatory mutation occurs in the CRM of the mutant allele. A mixture of deleterious and compensatory mutations, even if the latter is common, may bring patterns of polymorphism and divergence close to a neutral scenario, but cannot produce a signature of positive selection as observed for both classes of mutations in our analysis. In addition, analytical modeling of the compensatory evolution of TFBS finds that the waiting time for a turnover event is long if complete neutrality of the compensating mutations is assumed (Durrett and Schmidt, 2008). To shorten the waiting time to be compatible with the *Drosophila* TFBS turnover rate, the parameterization of the model requires that the double mutant allele have higher fitness than the non-mutant allele, making it a directional selection model. This supercompensatory scenario could produce signatures of positive selection both for binding site gain and loss, the latter occurring because the fixation of a deleterious mutation in an existing TFBS will have the appearance of being positively selected as it hitchhikes to fixation on the selectively favored allele. However, this scenario is biologically unrealistic, as it requires the second mutation (the gain of a TFBS) to be positively selected only on the background of the first mutation.

As an alternative, consider the following model of positive selection on CRM structure/function. We propose that for CRM with large numbers of interacting partners, the network of cis- and trans-factors will inevitably be constantly evolving — due to both direct selective pressures imposed on the CRM or indirect effects caused by adaptations in other components of the network. For example, egg length variations between and within *Drosophila* species have been studied as potentially adaptive traits; if egg length evolves, genes such as *eve* whose expression pattern need to scale with the embryo may need to change its CRM to adapt to the new context (Lott *et al.*, 2007). This constant flux of change, we propose, imposes continual selection pressure for CRM function within the network to co-evolve and change. This "moving target" hypothesis finds support in an analytical study, which shows that fluctuating selection may be common in *Drosophila*, with changes in the sign of selection coefficient occurring at nearly the rate of neutral evolution (Mustonen and Lässig, 2007). Adaptive substitutions could therefore occur before selection switches its sign again, since positively selected mutations fix at rates much higher than the neutral mutation rate.

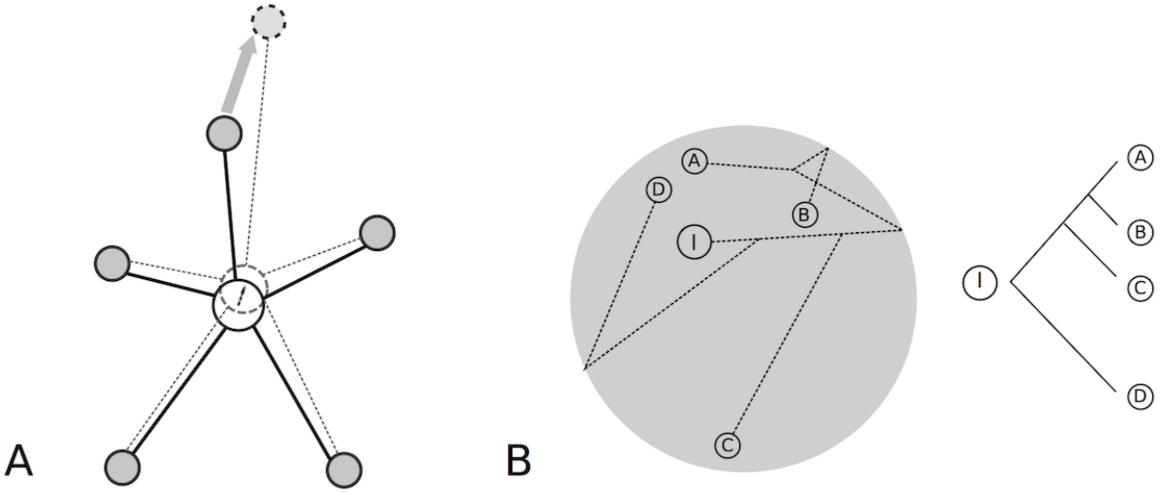


Figure 2.5: **Models of CRM evolution with changes in fitness optimum.** (A) The central node represents the CRM of interest and is connected to many interacting partners. With increasing number of connecting partners, we expect the CRM function to change more frequently in small steps but at the same time to be more constrained in function space. (B) A hypothetical evolutionary trajectory in CRM function space. Small changes in a system under global constraints result in non-linear functional evolution with time. The circle represents permissible space within which CRM function can change without causing strong pleiotropic effects. Depicted on the right is the species phylogeny. Starting from I, the ancestor of the existing species, the CRM function moves in the constrained region and generates a non-clock like evolution pattern in the extant species—species A and D are most distantly related phylogenetically but most similar functionally.

At the same time, the high connectivity in the regulatory network implies pleiotropic effects while the essentiality of genes controlled by the network may call for accurate regulation, both suggesting that the net change in CRM function will be highly constrained (Figure 2.5A). Under this conceptual model, functionally significant change will be possible on short evolutionary timescales, but will remain within constrained bounds over longer timescales. This feature of the model would account for adaptive gain and loss of TFBS in CRM, and could explain the strongly non-linear relationship between function and sequence evolution as exemplified by the *Drosophila eve* stripe 2 enhancer (Hare *et al.*, 2008; Ludwig *et al.*, 1998). Moreover, it provides an explanation for the finding of a non-clocklike evolutionary pattern: sequences from *D. pseudoobscura* rescues a *mel eve* stripe 2 enhancer deficiency almost as well as the native *mel* enhancer and substantially better than ones from much more closely related species (Ludwig *et al.*, 2005, Figure 2.5B).

In conclusion, our findings provide empirical evidence for positive natural selection acting in CRM and TFBS evolution. We suggest that CRM are not as functionally static as commonly believed, but rather may experience frequent adaptation through binding site turnover, even though there may be constraints on net change over longer evolutionary time.

2.5 Materials and Methods

CRM annotation and sequence alignments

REDfly (Gallo *et al.*, 2010) is a database of manually curated CRM and TFBS obtained from the literature from which we chose 118 non-overlapping autosomal CRM for investigation (Table S.1). They regulate 81 target genes and contain binding sites for 82 TF. The 118 CRM range in size from 65bp to 4.3kb (median = 515bp) and contain between 1 to 64 DNase I footprint sites (median = 4). From the set of 82 TF, we identified a subset of 30 with more than 10 footprint sites represented in the dataset and with carefully constructed Position Weight Matrices (Down *et al.*, 2007). In each footprint region plus five flanking bases on each end, we applied the appropriate position weight matrix to identify the highest scoring match as the core motif for the TFBS (referred to as TFBS in the text). We only included those TFBS for which the alignment between *mel* and *sim* sequences contain no insertions or deletions (including both fixed or polymorphic sites). As a result, a total of 645 TFBS for these 30 TF were included for analysis.

For each of the 118 CRM (coordinates in dm3 of *D. melanogaster* reference genome listed in Table S.1), we downloaded pre-aligned MAF blocks from UCSC genome browser for *D. melanogaster* (*mel*), *D. simulans* (*sim*), *D. sechellia* (*sec*), and two outgroup species, *D. yakuba* (*yak*) and *D. erecta* (*ere*). *D. sechellia* is a sister species to *D. simulans* and is included to compensate for the low sequence completeness in the reference *sim* genome. We then used the baseml module in PAML 4.4c (Yang, 2007) to reconstruct the ancestral sequences from the alignments. Following analysis involving polarized changes were done either using a single ancestral sequence for *mel* and *sim* determined by the most probable ancestral state (A,C,G or T) at each position, or summing over the posterior probabilities of all four possible states (full Bayesian approach). The two methods produced essentially the same results and therefore we only presented results using the most probable ancestral state. A maximum parsimony method was also investigated and was found to produce consistent results.

For polymorphism analysis, alignments for the same 118 CRM regions were obtained of a population sample of 162 *D. melanogaster* lines (http://www.hgsc.bcm.tmc.edu/projects/dgrp/) and six *D. simulans* lines (http://www.dpgp.org/). We also compiled the genome sequences of 150 coding regions corresponding to the target genes of the CRM listed in REDfly, for the purpose of compiling synonymous and nonsynonymous changes. For these data, we used codeml module in PAML 4.4c to reconstruct the ancestral sequence states following otherwise the same procedure as described above for CRM regions.

Position Weight Matrix (PWM)

PWM for 30 TF (Antp, Deaf1, Dfd, Kr, Mad, Trl, Ubx, Abd-A, Ap, Bcd, Br-Z1, Br-Z2, Br-Z3, Brk, Cad, Dl, En, Eve, Hb, Kni, Ovo, Pan, Prd, Slbo, Tin, Tll, Twi, Vvl, Z, Zen) were obtained from (Down *et al.*, 2007). This set represents all the TF for which Down *et al.* identified a single best motif for the REDfly footprint sites. For comparison, we also constructed five PWM (Hb, Bcd, Kr, Prd, Twi) from SELEX (Systematic Evolution of Ligands by EXponential enrichment) data (kindly provided by Mark Biggin). We ran MEME (Bailey *et al.*, 2006) with parameters "-evt 0.01 -dna -nmotifs 3 -minw A -maxw B -nostatus -mod zoops -revcomp text" on different selection rounds of the SELEX data.

Use PWM to predict mutation effect on binding affinity

Consider a mutation at the i^{th} position in a binding site motif involving a change from nucleotide j to k (j, k take values 1-4, corresponding to the nucleotides ACGT). We calculated $S[i,k] - S[i,j]$, where S is the PWM matrix of size $L \times 4$. According to previous theories, the PWM score is proportional to the physical discrimination energy of the protein to the sequence and therefore the above calculation may be used to infer the direction and magnitude of binding energy change due to a mutation (Berg and von Hippel, 1987).

To evaluate the accuracy of the PWM-based inference, we experimentally measured the binding energy change of observed mutations in Hb binding sites, using a state-of-the-art microfluidics device that has high sensitivity for relatively weak molecular interactions (MIT-OMI). The experiments were performed as described in Maerkl and Quake (2007). Sixty-four oligonucleotides were synthesized to test 25 SNP in Hb footprint sites and their combination in cases of multiple SNPs in a single TFBS between *mel* and *sim*. Data were analyzed in GenePix 6.0, R, and Prism 5.0. We found that the PWM we used correctly predicted the direction of change in 21/25 cases (Figure S.2). Three of the four disagreements had a predicted PWM score change ΔS close to or smaller than one, which indicates that PWM may not be accurate when its predicted binding energy differences are small. To minimize the chance of misassigning the direction of binding energy change to a mutation, we set a threshold corresponding to a PWM score difference of one, and classified mutations within (smaller in absolute value) that bound as uncertain. The conclusions are robust to the set-point of the threshold (for example, Table S.3). We also compared the PWM derived by Down *et al.* to the five PWM derived from SELEX data: 97% (33/34) of mutations in the TFBS were consistently classified after excluding nine mutations with small predicted effects by either PWM (Figure S.3).

Rate of gain and loss of TFBS in *mel* and *sim*

To examine the extent of binding sites gain and loss between the two species, we calculated PWM scores $S[a_{ij}]$ for each of the 645 footprint TFBS (i from 1 to 645) in orthologous sequences in *mel*, *sim* or the inferred *mel-sim* ancestor (j from 1 to 3), using patser v3e (by Gerald Z. Hertz, 2002). To determine whether a sequence is a binding site or not, we established two sets of cutoffs for PWM scores. First, we used PWM score $S > 0$, corresponding to the sequence being more likely from a binding site distribution than from a background distribution. For the second we used a set of TF-specific cutoff values chosen by first ranking all footprint sites of a TF by their PWM scores in descending order and then taking the 80% quantile value. The two cutoff set produced similar results (Table S.2).

Construct *sim*-PWM from orthologous sequences to the *mel* footprint sites

To test whether the *mel*-derived PWM might be over-optimized so that they would favor *mel* over *sim* sequences independent of the binding affinity differences, we ran MEME on both *mel* footprint sites for three TF (Hb, Bcd, Trl) and their *sim* orthologous sequences with the same parameters. The two set of orthologous PWM were then applied to score the observed variations in the TFBS of the three TF for comparison (Figure S.7).

Mutational probability for affinity-increasing and affinity-decreasing mutations

We attempted to estimate the probability for a random new mutation to be affinity-increasing (P_{inc}) or affinity-decreasing (P_{dec}) by examining all possible mutations that can occur on the inferred ancestral sequence of mel and sim for the 645 footprint TFBS. At the i^{th} site in a TFBS for TF x, the probabilities are calculated as:

$$
\begin{aligned}
P_{inc} &= \sum_{k \neq j} M_{j \to k} \mathbf{1}_{[1,+\infty)}\{S_x[i,k] - S_x[i,j]\}, \quad j,k \in \{A,C,G,T\} && (2.1) \\
P_{dec} &= \sum_{k \neq j} M_{j \to k} \mathbf{1}_{(-\infty,-1]}\{S_x[i,k] - S_x[i,j]\}, \quad j,k \in \{A,C,G,T\} && (2.2) \\
\mathbf{1}_A\{x\} &= \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} && (2.3)
\end{aligned}
$$

where j is the original nucleotide and k varies among the three possible mutations. S_x is the position weight matrix for TF x of size $L \times 4$. These values were then summed across all 645 TFBS and divided by the total number of nucleotides involved. Mutation matrix M is derived from polymorphism of the 4-fold degenerate sites of 9,628 genes in *D. simulans* (Lu *et al.*, 2008).

Generalized McDonald Kreitman (MK) test and site frequency spectrum analysis

For the generalized MK test, we counted the number of fixed and segregating sites for different functional categories in both *mel* and *sim* lineages. In *sim*, we required at least two of the six alleles to be non-missing for a site to be included in the analysis. For coding regions, synonymous sites were further classified into No-Change, Preferred-to-Unpreferred and Unpreferred-to-Preferred, following (Haddrill *et al.*, 2008). Polymorphism and divergence sites in both coding and CRM regions were counted using perl scripts adapted from Polymorphorama (Peter Andolfatto, Doris Bachtrog, 2009).

Following the suggestion of (Fay *et al.*, 2001), we considered only common polymorphism (derived allele frequency $> 15\%$) in the generalized MK test to alleviate the problem caused by negatively selected mutations in detecting positive selection. For each mutation category, we compared the substitution-to-polymorphism ratio to the synonymous No-Change class using Fisher's Exact Test. Two-sided p-values are reported.

Site frequency spectrum (*mel* only): Next-generation sequencing data produce variable coverage. To estimate the site frequency spectrum, for each variable site (TFBS, coding and spacers) with a coverage greater than or equal to 150 (maximum is 162) we randomly chose 150 and combined the counts for each frequency class (from 1/150 to 149/150).

3

A DROSOPHILA MODEL TO INVESTIGATE THE GENETIC BASIS OF A COMPLEX DISEASE TRAIT IN RESPONSE TO EXPRESSION OF A HUMAN MISFOLDED PROTEIN

3.1 Abstract

Genetic background – defined as mutations spread across the genome other than in the major gene – can significantly impact the expressivity of a Mendelian disease mutation; in complex disease, mutations across the genome and from different effect-size groups together determine the individual risk of disease. In both cases, identifying the genetic basis of the disease trait variability is crucial for predicting and treating the disease. Application of genome-wide association studies (GWAS) to human common diseases have yielded thousands of associated loci; however, limited mapping resolution and difficulty in performing experiments leave many basic questions unanswered: what types of variants underlie disease variability? What are their mechanisms of action and interaction? Non-coding variants are a special category that are frequently implicated in GWAS but are difficult to identify and can not be associated with a molecular mechanism. Here we propose a novel approach to the genetic investigation of a complex disease trait, featuring high resolution and experimental tractability that allow many challenging questions to be answered. The approach uses natural genetic variation in *Drosophila* to screen for modifying loci in a sensitized disease background, which we created by expressing a mutant (disease-causing) form of human proinsulin in the developing eye imaginal disc, causing neuro-degeneration in the eye that mimics the beta cell death in human patients. Crossing this transgenic line to a panel of 178 inbred lines of *D. melanogaster* resulted in a continuous distribution of the disease phenotype. GWAS in 154 sequenced lines identified multiple loci, including a strongly associated region (400bp) located within the intron of the gene *sulfateless* (*sfl*). RNAi knock-down of *sfl* enhanced the eye phenotype in a

mutant-proinsulin-dependent manner. Two more genes in the Heparan Sulfate Proteoglycan (HSPG) pathway were also validated as modifying the phenotype, strongly suggesting a previously unknown link between HSPG and cell response to misfolded protein. Finally, using pyro-sequencing, we found evidence of allele-specific expression associated with the *sfl* intronic variants, suggesting that the mechanism of the non-coding variants may be through altering the transcription of the gene.

3.2 Introduction

The genetic background in which a disease mutation acts can have a strong impact on both the risk and severity of disease. Even for Mendelian diseases that are originally characterized as monogenic, further studies almost invariably reveal additional layers of genetic complexity, with variants in other genes acting to suppress or enhance the biological activity of the primary mutation. Examples include phenylketonuria (PKU) and cystic fibrosis (CF), both of which are classical Mendelian diseases that have proved to involve multiple loci (Badano and Katsanis, 2002). In cases of common disorders (such as adult onset diabetes) or quantitative traits (such as height), genetic factors are dispersed among multiple loci spread across the genome, and most have small to modest effect size. Strong gene-by-environment interaction is ubiquitous, and when investigated, gene-by-gene interaction (epistasis) is also an important contributor to overall variation. Together, these factors make the genetic underpinning of common diseases and other complex traits recondite to classical genetic analysis.

Revealing the detailed mechanisms by which multiple genes spread across the genome (henceforth called the ”genetic background”) influence a trait or disease is of key importance and presents a grand challenge. Questions of interest include:

- What genes and pathways harbor natural variants that influence the trait?
- What are the molecular or developmental mechanisms through which they act?

- What are their population frequencies, effect sizes, and contributions to fitness?

While advances in genotyping methods, especially resequencing technologies, have made large-scale genome-wide association studies in human commonplace, various limitations leave the most fundamental questions unanswered. For example, because of the block linkage structure in human, the resolution of association mapping can rarely pinpoint the causal gene. This combined with the fact that most GWAS in human still rely on tag SNPs means that identifying the causal variants is not within reach by the genome-wide approach, but requires case-by-case detailed analysis. As a result of this, the second problem mentioned above, i.e. the molecular and developmental mechanisms of the variants identified in GWAS, are generally not known.

Past experience has shown that model organisms can provide critical knowledge to our understanding of human biology, owing to the deep conservation of fundamental cellular and developmental processes. Here we describe a novel application of a model organism approach to study the genetic architecture of a complex human disease. Our approach leverages the power of a *Drosophila* model and the presence of abundant natural variation. We constructed a fly model by creating a transgene of a diabetes-causing, human mutant proinsulin gene that could be expressed tissue-specifically in the eye imaginal discs and other developing tissues. The misfolded proinsulin protein causes neonatal diabetes in human, most likely by inducing beta cell death in human patients (Støy *et al.*, 2007); it results in tissue degeneration, such as a reduced and deformed eye, in our fly model.

To investigate whether there is natural genetic variation for the extent of eye degeneration induced by mutant proinsulin expression, we crossed the tester line bearing the mutant proinsulin (henceforth called hINS^{C96Y}, with 178 genetically diverse lines from the *Drosophila* Genetics Reference Panel (DGRP), which are inbred lines derived from wild caught flies from a single natural population (Mackay *et al.*, 2012). The resulting F1 offspring exhibited extensive and highly heritable variation in the extent of eye degeneration. We observed a nearly continuous distribution of eye degeneration phenotypes among the lines, suggesting a non-Mendelian, polygenic basis. The analysis (ANOVA) revealed a strong genetic component, indicating that as much as 60% of the total phenotypic variations is genetic.

Completely sequenced genomes are available for 168 of the 192 lines. This allowed us to perform genome-wide association with more than 2 million SNPs in 154 lines, for which we had both phenotype and genotype information. Using a mixed linear model method (Yang *et al.*, 2010), we estimated that 31% (s.e. 17%) of the total variance between individuals can be explained by common autosomal variants ($>$ 5% minor allele frequency; the X chromosome is not variable in our crossing design), which explains more than 50% of the broad sense heritability ($H^2 = 57 \pm 3\%$). The fast decaying linkage disequilibrium in *Drosophila melanogaster* populations allowed us to map a quantitative trait locus (QTL) to a 10kb region in intron 3 of the gene *sulfateless*, which encodes a bifunctional enzyme in the Heparan Sulfate Proteoglycan (HSPG) biosynthesis pathway. Follow-up experiments excluded a causal contribution by another gene that is in this intron of *sfl*, and established that knocking down *sfl* enhances the hINS^{C96Y} induced eye degeneration phenotype. We also found through cDNA sequencing of random heterozygotes of different combinations of *sfl* alleles that the intronic variants are associated with altered gene expression. In the same experiment, we also discovered strong allelic heterogeneity.

Our study demonstrates the utility of an approach that creates a fly model of a human Mendelian disease to reveal abundant quantitative trait variations for the severity of disease. This approach is informative not only in revealing molecular and cellular mechanisms of disease but also in understanding general properties and genetic architecture of human complex disease traits.

3.3 Results

3.3.1 Continuum of phenotypic variation in crosses with 178 DGRP lines

We have characterized our fly model of the hINS^{C96Y}proinsulin expression elsewhere (Park *et al.*, 2012). Briefly, the C96Y mutation (seventh amino acid in the A chain of mature insulin) in the human preproinsulin is known to cause human permanent neonatal diabetes mellitus (PNDM), presumably due to disruption of one of the two interchain disulfide bonds, which prevents the correct folding of the mature insulin (Støy *et al.*, 2007). Studies of a mouse model (called Akita) that carries the same C96Y change through spontaneous mutation, show that the mutant protein dominantly reduces the production and secretion of mature insulin protein. In addition, the intracellular fraction of proinsulin forms a complex with BiP, a molecular chaperons localized to the endoplasmic reticulum (ER) where protein folding occurs; accumulation of misfolded protein in the ER induces cellular stress (Wang *et al.*, 1999). Expression of the hINS^{C96Y} in the fly developing eye imaginal disc causes visible disruption of ommatidial hexagonal packing during eye development and eventually leads to a reduced eye area, rough surface and development of black lesion spots (Figure S.8),.

The *Drosophila* Genetic Reference Panel, or DGRP, is a collection of 192 inbred lines derived from wild caught flies from a single population in Raleigh, NC (Mackay *et al.*, 2012). In the present study, we crossed the transgenic fly line (w; P{GMR-GAL4}, P{UAS-hINS^{C96Y}}/CyO) as the maternal parent into the genetic background of 178 inbred lines from DGRP available at the time. Among several phenotypes observed, including rough eye, reduced total area, distortion of the oval shape and black lesion spots, we chose the total eye area as the phenotype because it is relatively easy to quantify and thus amenable to direct comparisons between individuals. We quantified ten male progeny from each hINS^{C96Y} x DGRP cross, where we observed a continuously varying phenotype distribution, ranging from 13% to 86% of wild type fly eye area (Figure 3.1). In addition, ANOVA revealed that nearly 60% of the variance is between genotypes, suggesting a large genetic component. Males were chosen for measurement and analysis because they generally exhibited a more severe phenotype than females. However, we also measured F1 females for a subset of 38 lines and found a strong correlation between the two sexes (Figure S.9).

We tested and excluded several trivial explanations for the observation. First, natural variability in eye size was investigated among a subset of the inbred lines, which we found to be small and uncorrelated with the eye size in the corresponding cross with hINS^{C96Y} (Fig S.10, $r^2 = 0.146$, $p = 0.526$). Next, we evaluated expression level variation for the hINS^{C96Y} transgene among the inbred lines. Using Western blots to determine the protein level of either GAL4, the transactivator of hINS^{C96Y}, and also an EGFP reporter gene that is co-transcribed with hINS^{C96Y}, we found both levels to have the same range across different genetic backgrounds, and not correlated with the eye degeneration phenotype. Therefore, we concluded that the observed variation in eye area

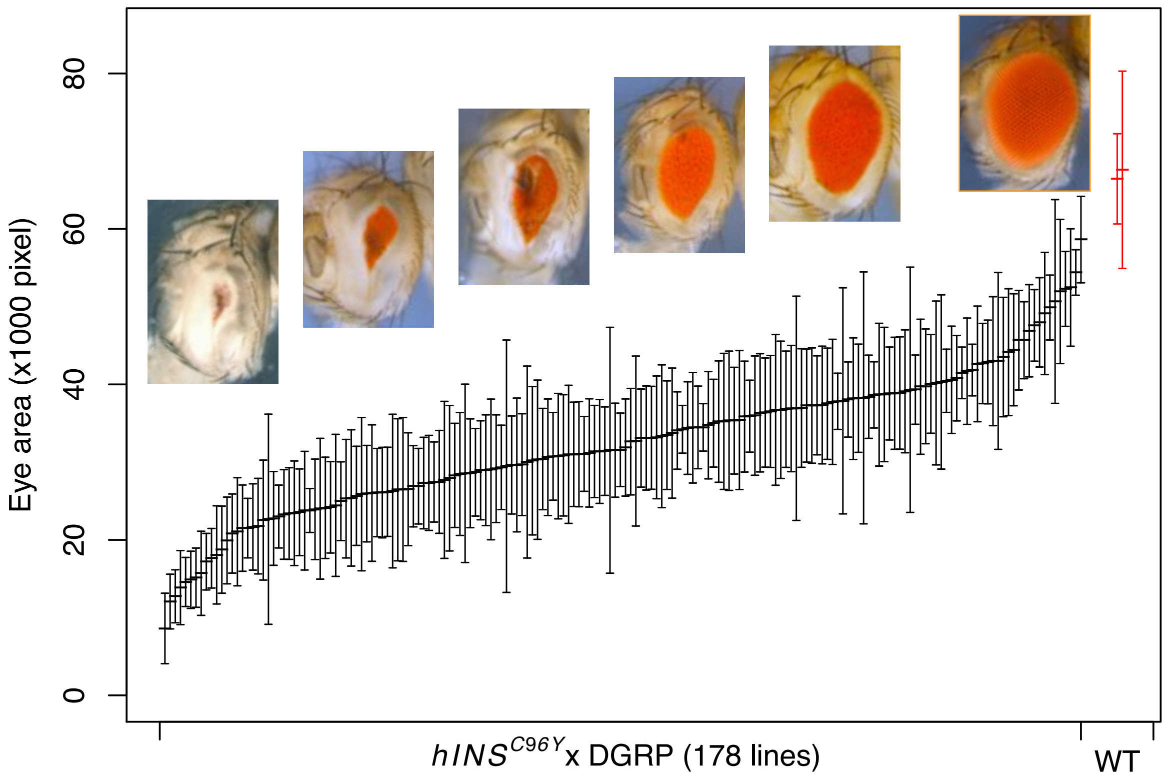


Figure 3.1: **Distribution of eye area in hINS^{C96Y} x DGRP (178) crosses.** Mean ± 1 s.d., sorted by the mean, are shown for 178 crosses as well as two non-transgenic wild type lines (red). Representative pictures of eyes from across the range of the distribution are shown. The rightmost picture shows a non-transgenic wild type fly eye as control.

is caused by differences in cellular response to the expression of hINS^{C96Y} and not to variation in hINS^{C96Y}expression itself. In the genome wide association study detailed in the section below, we also investigated single nucleotide polymorphism (SNP) surrounding the *glass* locus, which encodes the transcription factor expressed in the eye imaginal disc that activates GAL4 expression and found no evidence for association with the eye degeneration phenotype.

3.3.2 Genome-wide genotype-phenotype association

168 of the 192 DGRP lines have been fully sequenced, among which 154 lines were also phenotyped in our study. There is little population structure among the DGRP lines (Mackay *et al.*, 2012), which if present can confound genome-wide analysis. We confirmed this by performing a principal component analysis on 900K autosomal SNPs (obtained by pruning a total of 2 million based on pairwise linkage disequilibrium) (Fig S.11A). Further dividing the 154 lines into three groups by their quantitative phenotype values revealed no correlation between phenotypic severity and the top ten principal components (Fig S.11B, only the first principal component is shown).

We used mean eye area as a quantitative trait to perform single marker regression for 2 million autosomal SNPs. We restricted the analysis to bi-allelic sites for which the minor allele is present in at least 4 of the 168 lines. Because of the direction of the cross, all F1 males inherited their X-chromosome from the hINS^{C96Y} tester line. As a result, the X-chromosome could serve as a useful negative control. For example, population structure in the DGRP sample, if it existed, could induce a correlation between variation in DGRP X-chromosomes and the disease phenotype when no real correlation is expected. We examined quantile-quantile (QQ) plots for autosomal and X-chromosomal SNPs, to find that only the former and not the latter displayed an excess of small p-value SNPs, indicating that the relative excess of small p-value SNPs on autosomes is unlikely to be a consequence of any "hidden" structure in the DGRP sample (Figure 3.2A,B). A Manhattan plot revealed a strong peak on chromosome 3L (Figure 3.2C), with the most significant SNP reaching the conservative genome wide significance threshold (raw $p = 2.4 \times 10^{-8}$, Bonferroni corrected $p = 0.0502$).

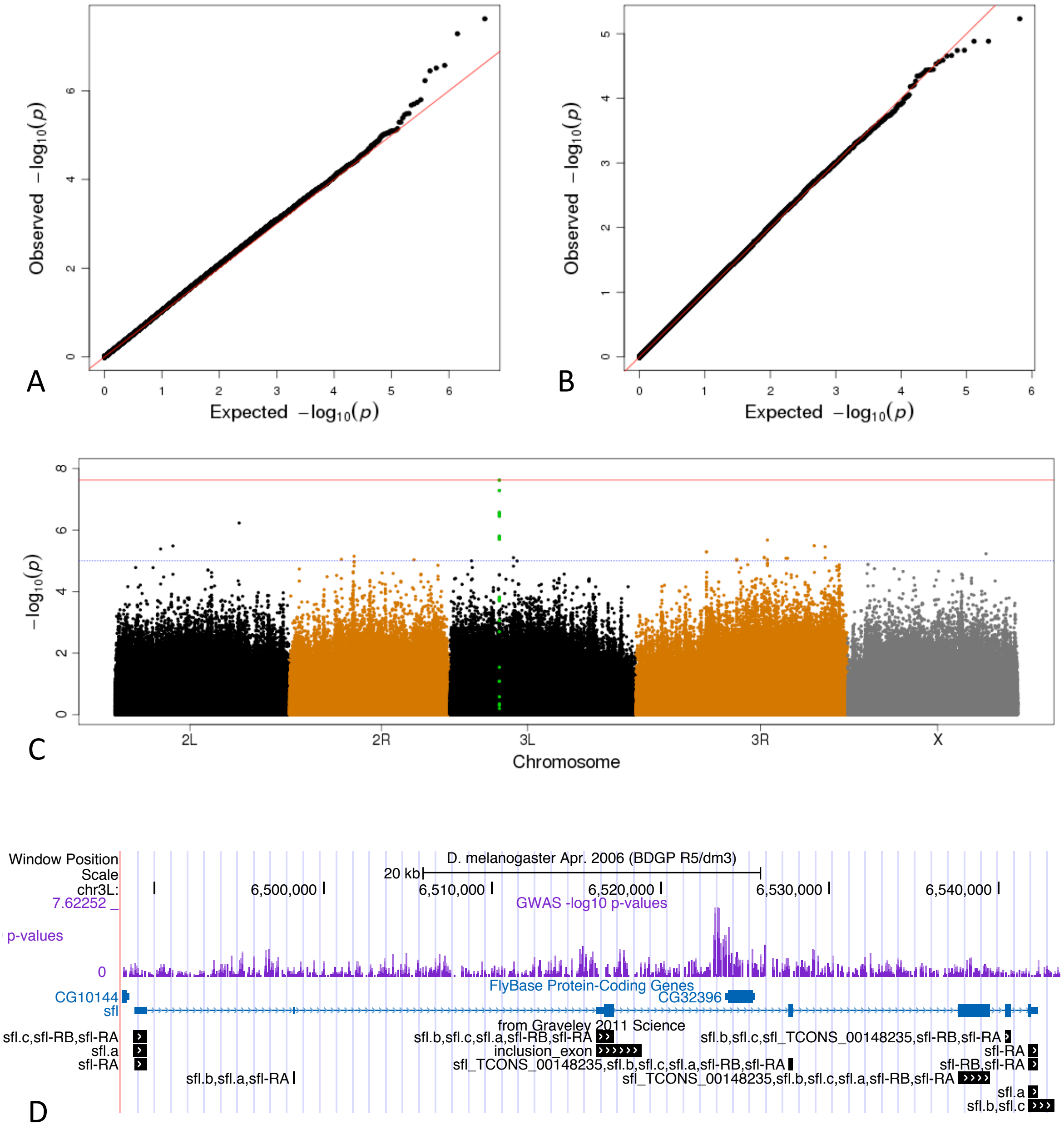


Figure 3.2: **Genome-wide scan identifies candidate locus associated with the hINS^{C96Y} induced phenotype.** Quantile-Quantile (QQ) plot reveals an excess of small p-values on autosomes (A) but not on the X chromosome (B), which is not variable in the mapping population due to cross design. (C) Manhattan plot shows a strong peak (green) on chromosome 3L. The blue and red horizontal lines indicate the suggestive ($p < 10^{-5}$) and the genome-wide threshold (Bonferroni corrected $p < 0.05$), respectively. (D) UCSC browser view of the *sfl* locus containing the association peak. The intron containing the peak also contains a nested gene CG32396. The black track shows RNA-seq supported spliced exons (data from Graveley *et al.*, 2011).

Additional signals from GWAS

Following the nominating threshold suggested by Mackay *et al.* (2012), we identified 30 SNPs passing the arbitrary cutoff of $p < 10^{-5}$ (Table S.4, Figure S.12). We used two methods to assess the false discovery rate at this threshold (see Materials and Methods for details). A permutation test suggested a very modest enrichment in the observed data for SNPs below the threshold (average of 2,000 permutation: 21 SNPs $< 10^{-5}, median = 19, observed = 30$, at 85th percentile). In subsequent analyses, we focused on the peak on chromosome 3L, leaving the secondary candidates to more rigorous testing with additional samples.

Proportion of variance explained by common SNPs

A mixed-linear model (MLM) can be used to estimate the proportion of the variance explained by common SNPs (Yang *et al.*, 2010; Zhou and Stephens, 2012). This method does not identify individual SNPs, and therefore does not suffer from the multiple testing burden. Results for human height and other traits suggest that the proportion of heritability explained by common SNP is much higher than the top GWAS candidates alone, suggesting a large number of unidentified causal SNPs below the GWAS detection threshold (Yang *et al.*, 2011). We applied the method to our data using the GCTA software (Yang *et al.*, 2011). As a reference point, we estimated the proportion of variance between genotypes to be 57% (s.e. 3%). Because this estimate may include genotype-specific environmental errors (such as vial differences), it should be treated with caution as an estimate of the broad sense heritability (see Falconer, 1981, p115) Next, we fit a repeatability model to individual measurements, and estimated that 31% (s.e. 17%) of the variances between individuals can be explained by common variants (minor allele frequency, or MAF$>$ 5%). Compared to the ANOVA estimate, this suggests that more than half of the heritable variation between individuals may be accounted for by combining the effects of all common variants. A corresponding model fit to the mean phenotype estimated that 52% (s.e. 28%) of the variance in mean eye area can be explained by common variants. The estimates of the proportion of variance have a large standard error because of the modest sample size (154 lines). More accurate estimates are expected with a larger sample.

3.3.3 Identifying and validating sfl *as a major effect locus*

To identify the gene underlying the peak on chromosome 3L, we focused in on the region and found the association signal to be sharply confined to the third intron of the gene *sfl* (Figure 3.2D). Because there is also a nested gene (CG32396) lying next to the peak in that intron, we devised two tests to distinguish whether either gene or both could be associated with the phenotype. First, we reasoned that the real causal gene should be expressed in the eye imaginal disc and adult eye tissue. We found CG32396 to have a highly testis-specific expression pattern in adults (from flyAtlas and modEncode), with very low expression in the eye (Fig S.14A). As for the larval stage, we prepared cDNA libraries from dissected eye imaginal discs from third instar wandering larvae. qRT-PCR analysis in this sample failed to identify the expression of CG32396 (data not shown). In comparison, *sfl* expression is consistently detected in the eye imaginal disc samples (data not shown); data from FlyAtlas (Chintapalli *et al.*, 2007) also indicated strong eye and brain expression of the gene in adults (Figure S.14B). Second, we reasoned that RNAi against the causal gene should have an hINS^{C96Y} dependent effect. We found that in the absence of hINS^{C96Y} expression, knocking down neither *sfl* or CG32396 in the eye imaginal disc had any effect on the mean eye area. In contrast, RNAi against *sfl*, but not CG32396, significantly decreased the mean eye area when hINS^{C96Y} was expressed (Figure 3.3). These results strongly suggest *sfl* to be a causal gene underlying the association peak, while CG32396 is not.

3.3.4 Involvement of Heparan Sulfate Proteoglycan (HSPG) in modifying the hINS^{C96Y} induced eye degeneration

The gene *sulfateless (sfl)* encodes a bifunctional enzyme in the HSPG biosynthesis pathway. HSPG is an important component of the cell surface and extracellular matrix (Kirkpatrick and Selleck, 2007). It is best known for its role in development, acting as signaling molecule co-receptors to modulate signaling events at cell surfaces (Häcker *et al.*, 2005). Although a connection between compromised HSPG function and the cellular responses to misfolded protein has not been firmly established, a previous study has shown that mutants in HSPG biosynthesis pathway had impacts on mitochondria density near post-synaptic membranes, and might have resulted in enhanced endocytosis, both of which suggested a potential link to the unfolded protein response (Ren *et al.*,

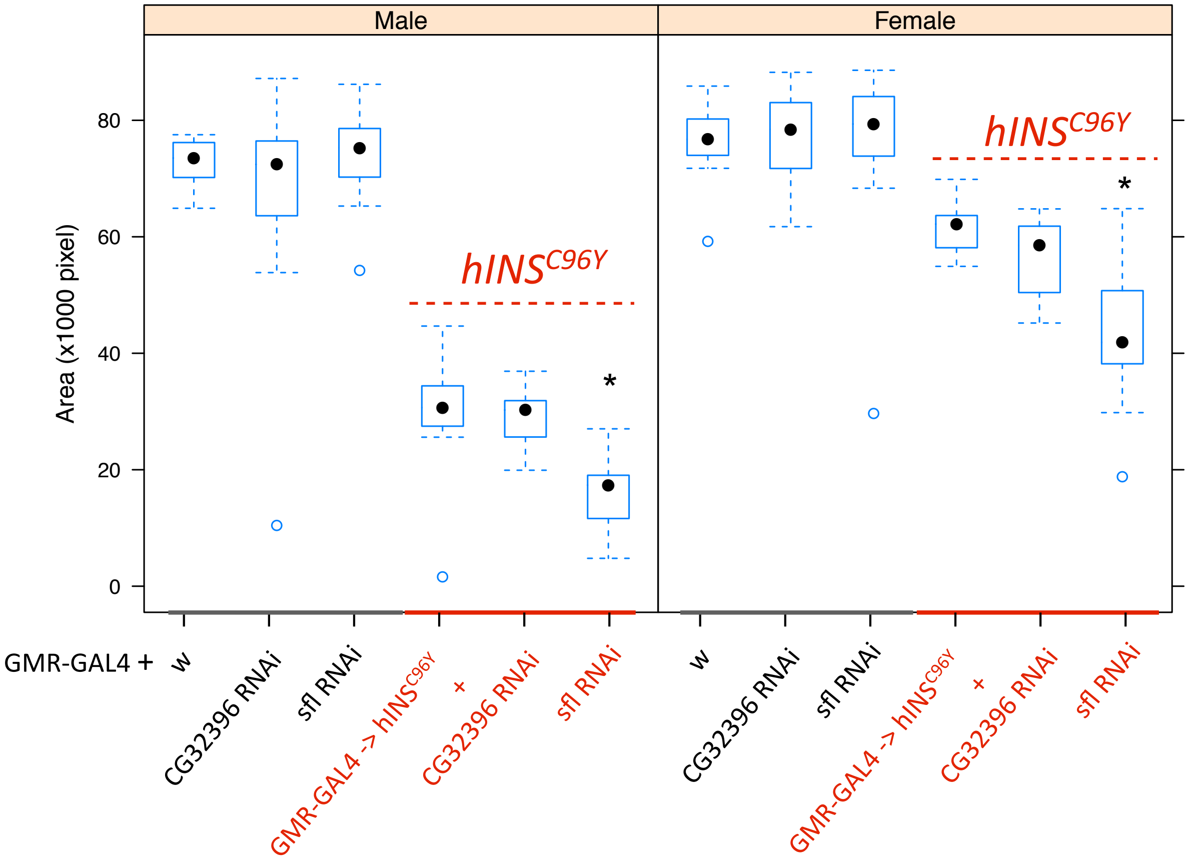


Figure 3.3: **RNAi confirms *sfl*, but excludes CG32396 as the causal gene.** The effect of knocking down either CG32396 or *sfl* was tested in the absence (x GMR-GAL4) or presence (x {GMR-GAL4, UAS-hINS^{C96Y}}) of hINS^{C96Y}. Compared to the control crosses (first and third columns in both sexes), significant difference in mean eye area was observed only with RNAi against *sfl* and only in the presence of hINS^{C96Y} (asterisks above a box plot indicate significant differences at 0.05 level determined by a student's t-test). In box plots, the median (black dot), interquartile (box) and 1.5 times the interquartile range (whiskers) are indicated; data points outside the range are represented by circles.

2009).

To test if the observed effect of *sfl* RNAi is due to the disrupted function of HSPG, we examined RNAi and/or mutant lines against two additional components in the pathway: *ttv* and *botv.* Both genes function upstream of *sfl* (Lin, 2004) (all three have clear homologs in human and mouse). As shown in Figure 3.4, disruption of both genes show a hINS^{C96Y} dependent effect in the same direction as *sfl* RNAi. An exception is *ttv*^{681}, which was characterized as a null mutant of the gene (Bellaiche *et al.*, 1998). It is likely that with the mutant being tested in heterozygotes, its effect may not be visible unless both copies of the genes were mutated. Another possibility is that *sotv*, a sister gene that also functions as a copolymerase as *ttv* does, might partially compensate for the loss of one copy of *ttv* (Lin, 2004). Other than the *ttv* mutant, the experiment provided consistent support for the involvement of HSPG dysfunction in enhancing the misfolded insulin induced neuro-degeneration.

3.3.5 Intronic SNP/Indel modulate sfl *expression level*

Both to validate and to discover untyped SNPs under the association peak, we re-sequenced a 3kb region containing the peak as well as the nested gene CG32396, in 19 of the 154 lines plus the transgenic hINS^{C96Y} line. Using Sanger resequencing, we discovered that the "SNP" achieving the lowest p-values genome-wide is in fact an 18bp / 4bp length polymorphism (relative to the *D. simulans* orthologous sequence) (Figure 3.5A). We also found three other indels in this region, with sizes ranging from 4bp to 30 bp and the minor alleles (deletion in all three cases) being present only once or twice in the sample. In comparison, the 18/4bp polymorphism is present at 50% frequency in the DGRP sample, which makes its evolutionary history of particular interest. In light of the discovery of mislabeled and undiscovered indels, we will use the term "Single Feature Polymorphism" (SFP) when referring to variants in the *sfl* locus. However, in order to be consistent, we will continue to use the term "SNP" when referring to variants genome-wide, acknowledging that a portion of them could be mislabeled indels.

The 18/4bp polymorphism accounted for nearly 20% of the total variance between mean eye area in the 154 lines (likely an overestimate due to the winner's curse, Garner 2007). To test if additional signals exist independent of the 18/4bp polymorphism, we performed a conditional

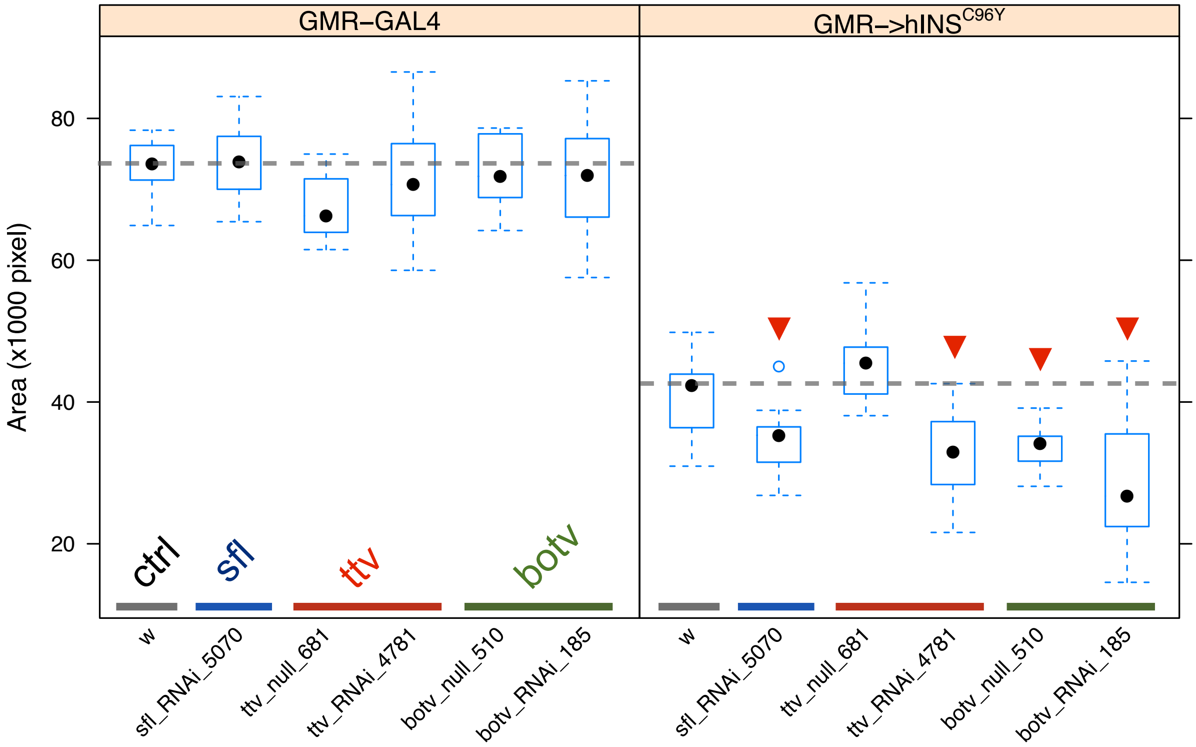


Figure 3.4: **RNAi and mutant analysis for HSPG biosynthesis pathway genes.** The experiments are the same as in Figure 3.3. Left panel shows the effect of RNAi or mutant alleles of the genes in the absence of hINS^{C96Y} expression; right panel shows the effect when hINS^{C96Y} is expressed in the eye imaginal disc with GMR-GAL4. The statistical significance of differences from the control cross (w) were determined by a two-sided student's t test. Those that are significant at 0.05 level are marked with a red arrowhead.

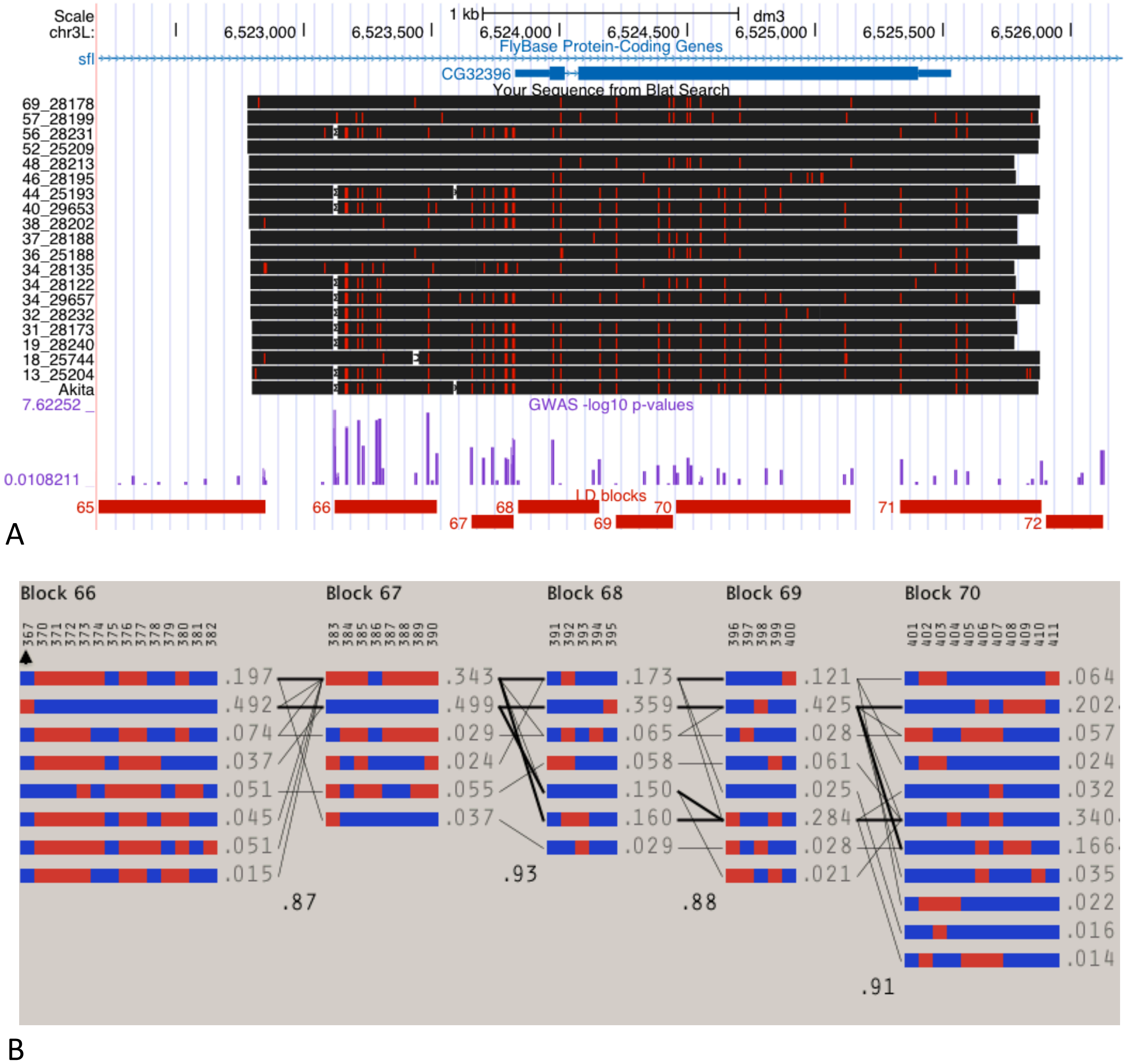


Figure 3.5: **Sanger resequencing of a 3kb region under the peak and the linkage patterns therein.** (A) Alignment of 19 DGRP lines' sequences ordered by their mean eye area. The hINS^{C96Y} transgenic line (the last sequence) carries the 4bp allele on both chromosomes. Red ticks and white spaces indicate SNPs and deletions relative to the reference sequence. The purple track shows the -log10 of GWAS p-values. Red bars at the bottom indicate linkage blocks as determined by Haploview (4.02) using the solid spine method with default settings (D'$>$ 0.8). (B) Detailed haplotype block structures. Each numbered column represents a polymorphic site, with the alleles colored as blue or red; each row represents a haplotype with frequency $>$ 0.01. An arrowhead marks the 18/4bp indel polymorphism (4bp marked as red). Finally, the number between any two blocks represents the multi-allelic D' (maD'), which quantifies the linkage between adjacent blocks. Because the maD's drops below 0.6 for 65/66 and is below 0.7 for 71/70, block 65, 71 and more distant ones are not included here.

analysis with the 18/4bp polymorphism as a covariate. Testing all other SFPs in the *sfl* locus and also across the entire genome, we failed to find any significant ones after accounting for multiple testing, suggesting that, either the 18/4bp polymorphism and its linked variants are the only ones underlying the association peak, or there exist additional variants that are below the statistical power threshold of this study (Figure S.15).

A plot of haplotype structure surrounding the association peak containing the 18/4bp SFP in *sfl* (Haploview v4.2) reveals a linkage block only 400bp in length (block 66 in Figure 3.5, chr3L:6523119-6523518). There are two major haplotypes, which we refer to as the 18bp or 4bp alleles, each represented by two nearly equal-sized groups among the 178 DGRP lines. Because all coding variants in *sfl* are well outside this 400bp linkage block, and therefore cannot be responsible for the association peak, we hypothesized that the intronic SFPs are the causal variant(s) and that they influence the eye phenotype by altering the expression level of *sfl*. Based on the RNAi result, we expect the 18bp allele to be associated with higher expression level than the 4bp allele (mean eye area of 4bp allele = 27590 > 36240 for 18bp allele, consistent with RNAi knockdown of *sfl* resulting in smaller eyes).

To test this hypothesis, we crossed randomly selected pairs of 4bp and 18bp lines to obtain F1 hybrids that carry both alleles. We then used pyro-sequencing to estimate the relative expression in cDNA samples prepared from these heterozygotes. This method allowed us to compare *sfl* expression contributed by the two alleles in the same animal (as measured by the ratio of expression), thereby controlling for both trans-environment differences and experimental noise, resulting in highly reproducible results (Figure S.16). To account for the heterogeneity due to variation in other parts of *sfl* or elsewhere in the genome, we randomly chose 6-8 lines carrying either allele and paired them at random in 15 crosses. The result is summarized in Figure 3.6. Seven crosses showed significantly more expression from the 18bp allele, with a 18bp/4bp ratio between 1.03 to 2.8 (median = 1.15); another six crosses showed no significant differences; two crosses had significantly more expression from the 4bp allele, with a difference of 5 and 6 percent respectively. There is a clear trend towards greater expression associated with the 18bp allele, supporting our hypothesis. However, both the direction of allelic expression difference and its magnitude varies significantly among the 15 crosses, indicating that additional cis-variants in *sfl* or variation in other parts of the genome must contribute to allelic expression differences of *sfl*. This heterogeneity highlights the complexity of the molecular effects of natural (regulatory) variation. We suggest that this is likely a common feature of regulatory variations acting on the expression phenotype.

3.4 Discussion

3.4.1 A fly model for studying the genetic basis of complex disease traits

In this era of inexpensive genome sequencing, which is expected to lead to the maturation of personalized medicine, studying the effect of genetic background on disease risk is clearly an important task. Many common diseases are now recognized as a heterogeneous group of disorders; the lack of accurate genetic classification prevents their effective diagnosis and treatment. Although genetic background is rarely considered as an important factor in predicting Mendelian disorders, it has in fact long been known that genetic modifiers exist in almost every Mendelian disease studied, and can have significant impacts on the expressivity of the primary mutation(s) in the key gene (Badano and Katsanis, 2002). For example, a recent study discovered a region on chromosome 1q21.1 harboring recurrent microdeletions, which, depending on additional variants in other loci, could be associated with a variety of neuropsychiatric phenotypes (Mefford *et al.*, 2008).

In both Mendelian and common complex diseases, however, identifying the causal genes and their mechanisms of action poses great challenges, in part because of the LD structure in human which limits the resolution of association studies – a typical association peak in a well powered study usually encompasses tens of genes, and to identify the causal gene(s) requires either prior knowledge about their functions or additional experiments. For the same reason, causal variants underlying the association are rarely identifiable, greatly hindering efforts in understanding and developing treatment for complex diseases.

The *Drosophila* model offers several advantages. First, with a relatively compact genome, essentially all genome-wide SNP can now be identified in a sample at low cost; several million SNPs have been identified in the resequenced genomes of 192 inbred lines derived from a single population collection, the Drosophila Genetics Reference Panel (DGRP). Many more genomes are currently being sequenced. The vagility of this species and modest population subdivision may allow

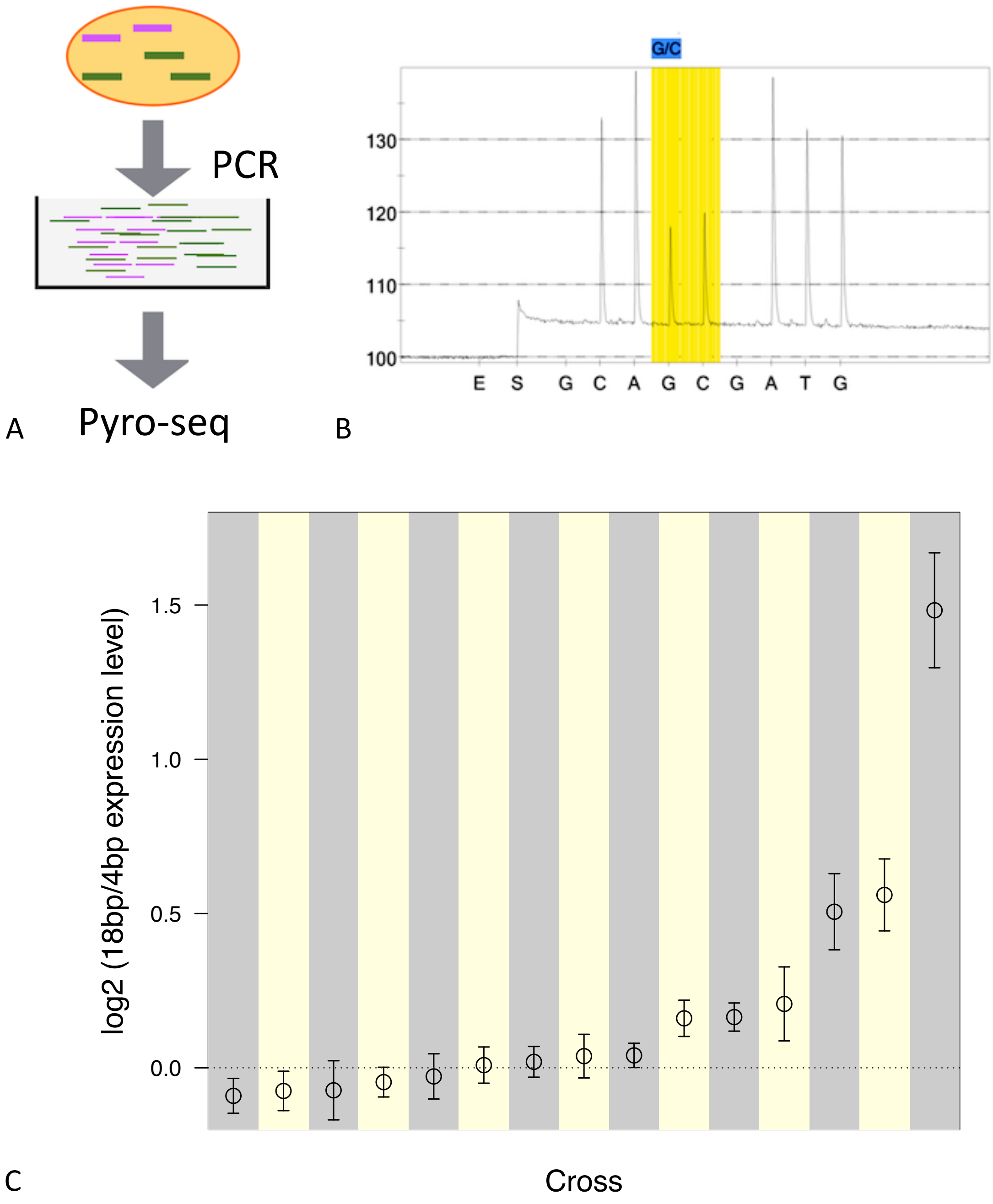


Figure 3.6: **Pyro-sequencing measure of *sfl* allele-specific transcript ratio in 18bp/4bp heterozygotes.** (A) A diagram of the pyro-sequencing approach. (B) An actual pyrogram is shown with the polymorphic site highlighted. Log2 transformed ratio of 18bp/4bp allele expression in 15 crosses between randomly paired 18bp and 4bp lines. Estimates and 95% confidence intervals are plotted. The dotted line corresponds to equal expression from the two alternative alleles.

them to be combined to produce larger samples for investigation. GWAS with complete genome sequences is thus possible in the fly. Second, because linkage disequilibrium typically decays over 10's or 100's of base pairs in Drosophila, association studies of completely resequenced genomes can pinpoint causal variants, as we have shown with *sfl*. Once identified, forward genetics can be immediately brought to bear for validation and biological investigation.

D. melanogaster is also at least 20 times more variable than human in a genome less than one-tenth the size, and is genetically variable for almost every trait ever investigated. Adding to this advantage, wild-derived strains can be made isogenic, which allows repeat measurements of a disease phenotype. Importantly, this variability is penetrant in heterozygotes, such as the one we made between the isogenic lines and our transgenic tester, mimicking its segregation in a natural population. The reduced variance in means of phenotypes compared to individual measurements increases heritability and thus the power to detect a causal association (Mackay *et al.*, 2009). Finally, both forward and reverse genetics can be applied to investigate the biology and pathway genetics of candidate variants.

Beyond what we've shown in this study, this system has several potential applications. First, it is known that a primary mutation(s) in the same gene can manifest in different forms or tissues among different patients (Mefford *et al.*, 2008). Such heterogeneity greatly complicate studies as well as treatment of the disease in human. Our fly model utilizes the binary GAL4-UAS system, which allows us to create a series of models using the same disease mechanism, but directed to different tissues with high specificity. This possibility of constructing and studying multiple related models in parallel can provide insight into the basis of disease heterogeneity. Another prevalent complexity in human diseases is sex difference in disease risk and severity. Our model of hINS^{C96Y} shares this feature: males consistently show more severe phenotypes not only for the eye phenotype but also when hINS^{C96Y}is expressed in the developing notum or wing in our preliminary analysis. Again, the opportunity to discover the genetic basis for such sex differences may be valuable for human disease studies.

3.4.2 Implications and a potential model for human complex diseases

Strictly speaking, this study is a modifier screen for a Mendelian trait, induced by the expression of a misfolded, disease-causing protein; our approach, however, differs from a classical modifier screen in that it assays natural variation instead of lab-induced mutations. Several reasons beyond this difference suggest that our model can have important implications and may be a potential model for human complex disease. First, the disease trait we studied, after crossing to the DGRP panel, is clearly complex in its genetic architecture, as evidenced by the continuously varying phenotype and a high broad sense heritability. The role of the Mendelian mutation is to sensitize the fly to reveal phenotypic effects of background genetic modifiers of disease. The disease trait is, in this manner, transformed into a complex trait that is highly dependent on the genetic background.

Second, the distinction between Mendelian and complex disease is largely historical and has been challenged by some researchers (Badano and Katsanis, 2002), who argue for a continuum between the two. For example, some Mendelian disease can have a strong genetic background dependence (e.g. cystic fibrosis, neuropsychiatric diseases), with additional genetic variants interacting with the major genetic factor(s) to affect the age of onset and severity of the disease. At the same time, common diseases can resemble Mendelian diseases in having a major factor, either genetic or environmental, such as driver mutations in cancer and diet/lifestyle in type-2 diabetes. The common feature shared by complex and Mendelian diseases is that the biological system is pre-stressed by the major factor, which leads to a release of additive genetic variation that is normally cryptic. The decanalization of physiological traits that used to be under stabilizing selection has been proposed as a major reason for the rising incidence of human common disorders in the modern times (Gibson, 2009). From this point of view, common diseases may be viewed as a generalized case of Mendelian diseases.

The commonly made statement – "Complex diseases lack a major genetic factor" – also merits comment. What it really means is that there lacks a COMMON major factor, but not ANY major factor. We propose that the genetic architecture of some complex diseases may consist of two components: the first part a small number of rare variants of relatively large effect size, i.e. the "major" genetic factors; the second part a relatively large number of common variants of modest-to-small effect size. If true, this may explain why human GWAS for most diseases still only account for a small portion of the heritability, even with impressively large sample sizes.

We acknowledge that important differences may exist, possibly even qualitative ones, between the genetic modifiers of Mendelian disorders and the genetic architecture of common diseases. It is conceivable, for example, that variation responding to a strong perturbation such as in a Mendelian disorder may belong to a different class than those responding to some lesser perturbations in a common disease. However, we know of no empirical evidence supporting this idea; in contrast, sensitizing mutants are widely used in *Drosophila* genetics with great success to screen for genetic modifiers of a mutant phenotype. This suggest to us that, the advantage of our fly model in its mapping resolution and experimental tractability, make it an ideal tool for testing the above hypothesis.

Finally, it is often assumed that effects of natural variation on a trait should in general be smaller than lab-induced mutations as large-effect mutations are more likely to be deleterious and therefore eliminated by natural selection. However, what we observed contradicts this expectation. Using total eye area as a quantitative measure of disease severity, the 178 genetic backgrounds range from nearly 13% to almost 86% of the wild-type eye area. In comparison, none of the RNAi or genetic mutant lines caused phenotypes nearly as severe. It is also worth pointing out that all of the natural variation we investigated was introduced in a single copy in heterozygous flies that carry only one copy of the DGRP autosomes. This surprising observation raises questions about the nature of the causal natural variation and their natural fitness effects in the population, which is amenable to future studies in the *Drosophila* model.

3.4.3 Connection between unfolded protein and HSPG function

Unfolded and misfolded proteins underlie a diverse group of diseases, including most neuro-degenerative diseases such as Parkinson's disease and Alzheimer's disease (Bucciantini *et al.*, 2002). Recently, it has been suggested that cell apoptosis might play an important role in type 2 diabetes onset (Rhodes, 2005). While the detailed mechanism causing apoptosis has not been established, one possibility is that the increased demand of insulin production together with the inherently error-prone protein folding process leads to the production of unfolded or misfolded proteins, which may overwhelm the molecular chaperones for maintaining proteostasis.

Our genome-wide association study identified *sfl*, a bifunctional enzyme that modifies the polysaccharide chains in the biosynthesis of Heparan Sulfate Proteoglycan (HSPG). Further RNAi experiments identified two more genes (*ttv* and *botv*) in the same pathway as being involved in modifying the eye-degeneration phenotype, strongly implicating a connection between HSPG function and cellular response to unfolded proteins.

HSPG are abundant components of cell surfaces and extracellular matrices. They consist of a core protein with unbranched disaccharide chains. HSPG are best known for their roles in interacting with signaling molecules and functioning as co-receptor, which make them an integral component in development (Häcker *et al.*, 2005; Kirkpatrick and Selleck, 2007). Although best known for their roles on the cell surface as co-receptors, they also have less well recognized functions in regulating vesicle trafficking: HSPG independently mediates the uptake of triglyceride-rich lipoproteins in mice (Stanford *et al.*, 2009). A study in *Drosophila* showed that mutations in *sfl* or *ttv* led to an activity-dependent increase in endocytosis in the neuro-muscular junctions (Ren *et al.*, 2009).

Furthermore, a link between vesicle trafficking, including both endo- and exocytosis, and cellular response to unfolded protein has been strongly implicated in two yeast studies (Kim *et al.*, 2009; Kimmig *et al.*, 2012). In a genome-wide screen, Kim, Gilbert and colleagues searched for genes that would cause synthetic lethality with a mutant PDI protein that lacks its disulfide isomerase activity. To their surprise, only 10/130 genes they identified belonged to the unfolded protein response pathway, while more than half of the genes were related to vesicle trafficking. The authors hypothesized that yeast cells compensate for the slower exit of the PDI substrates through a decrease in the rate of endocytosis. Further compromising cells' regulation of vesicle trafficking, therefore, could lead to synthetic lethality. A similar finding was made in fission yeast, where Kimmig *et al.* used a chemical to interrupt disulfide bond formation in the cells. As an immediate response, many genes' expression levels were down-regulated by Ire1, a highly conserved protein-folding sensor. Among those genes, 10% were related to trafficking as defined by Gene Ontology (GO) terms.

Another interesting finding in Ren *et al.* (2009) is that mutations in the HSPG biosynthesis pathway seemed to also alter the organization of several internal organelles, including mitochondria, endoplasmic reticulum and Golgi, which provided another potential clue for how disrupting HSPG function may affect cellular response to unfolded proteins.

3.5 Materials and Methods

Fly strains

The {GMR-GAL4, UAS-hINS^{C96Y}} line was generated by crossing the GMR-GAL4 line (Stock #1104, Bloominton Stock Center) with the UAS-hINS^{C96Y} line (Park *et al.*, 2012), and obtaining the recombinant 2nd chromosome, which was immediately balanced over CyO. 178 DGRP lines were obtained from the Bloomington stock center. RNAi lines against *sfl* (GD5070), *ttv* (GD4871), *botv* (GD37186) as well as mutant lines for *ttv* (*ttv*^{681}) and *botv* (*botv*^{510}) were generous gifts from Dr. Scott Selleck (Penn State University).

Eye area measurement

All crosses were reared in 25C. To quantify total eye area, about 20 1-5 day old adult flies were positioned on their side on a glass slide, prepared by applying a thin layer of vacuum grease (Beckman cat 335148) to the surface. Three glass capillaries (76x1.2 mm) were positioned in parallel at a distance about 1.5 adult body lengths apart to create two rows of spaces. Halocarbon oil 700 (Sigma H8898) was then added to fill the space and to hold the cover glass. Eyes were imaged using a Leica M205FA dissection scope and Leica DFC420 camera. A composite picture was taken using the scope's multifocus function to make the entire eye area in focus. Image analysis used an in-house ImageJ macro, which reported the area of the eye in pixels. At least 10 images (independent flies) passing the quality check were collected for each cross.

Principal Component Analysis

The whole-genome SNP dataset for the 154 DGRP lines used for GWAS was downloaded from the DGRP website (http://dgrp.gnets.ncsu.edu/). To detect population structures among these lines, 900K SNPs (after LD pruning using PLINK v1.07, with parameter –indep-pairwise 50 5 0.5) were input into the SmartPCA software (in EIGENSOFT v3.0), and the top 15 principal componentes

(PCs) were calculated (no outlier exclusion). To test whether the presence of a weak population structure could be confounding the association analysis, we tested for correlation between the hINS^{C96Y} phenotype (line mean) and the length of the first five eigenvectors in each DGRP line. In each case the correlation was not significant.

Genome wide association

Mean eye area of 154 DGRP lines crossed to the hINS^{C96Y} line was regressed on each SNP genome-wide with a minor allele frequency $> 5\%$ (PLINK 1.07, quantitative trait mode). Altogether 2,106,077 autosomal SNPs and 324,253 SNPs on the X chromosome were tested. Because the X chromosomes were not variable in the F1 by design, the X-linked SNPs were expected to conform to a null distribution where no association exists. This was tested and confirmed with the quantile-quantile plot.

Because an estimate of the total number of independent SNPs genome-wide doesn't exist for fly, we adopted two arbitrary thresholds to identify candidate SNPs on the autosomes: the first is a Bonferroni corrected threshold at 0.05 level (-log10 p-value > 7.62). This is conservative because it assumes all tests are independent while the number of independent SNPs must be much smaller than the total number tested. The second is a nominating threshold of p-value $< 10^{-5}$, suggested by Mackay and colleagues (Mackay *et al.*, 2012).

For the second threshold, which is less conservative, we used two ways to estimate the false discovery rate (FDR). First, assuming the false positive rate for autosomal and X-linked SNPs are the same, we expected 6.5 autosomal SNPs while observing 29, which gave an FDR of 22%. Second, we randomly shuffled the line labels of the phenotype column 2,000 times, and carried out GWAS on each of the permuted datasets. The resulting number of SNPs passing the threshold in each of the 2,000 trials has a mean of 21 and a median of 19, while the observed number 30 is at the 85th percentile. Both methods suggest a fairly modest enrichment of true positives under the second threshold.

To determine if there are secondary signals in *sfl* or elsewhere in the genome independent of the intronic SFPs in *sfl*, we fit a linear model with the 18/4bp *sfl* polymorphism as a covariate. This analysis was performed either within the *sfl* locus or genome wide, and in each case the p-values were corrected for multiple testing using Bonferroni's method. The lack of significant SNPs in both cases could be due to either a lack of genuine signals or a lack of power due to the limited sample size and highly conservative threshold.

Expression analysis for sfl *and CG32396*

Expression profiles for both genes in the adult tissues were assessed with data from FlyAtlas (Chintapalli *et al.*, 2007) and modENCODE (Roy *et al.*, 2010). To directly assay expression in the eye imaginal discs, we isolated total RNA from 10 pairs of eye imaginal discs in 3rd instar wandering larvae. Briefly, individual larva were sexed and dissected in 1x PBS, and the isolated eye imaginal discs are immediately dissolved in 300ul Trizol (Invitrogen 15596-026). Total RNA was extracted according to the manufacturer's manual. We then made cDNA libraries using (dT)20 primers after DNase I treatment (Invitrogen 18080-051,18068-015). Real time quantitative PCR was performed with primer pairs targeting either *sfl* or CG32396, with RP49 as an endogenous reference (SYBR-Green assay).

RNAi and mutant validation for candidate genes

All RNAi strains were originally from the Vienna *Drosophila* RNAi Center as P-element insertion lines in a co-isogenic w1118 background. For each RNAi strain, we first tested whether it alone had an effect on eye development by crossing it to GMR-GAL4 (stock #1104) and comparing the eye area of the F1 males or females to the control cross between w1118 and GMR-GAL4. In all crosses, GMR-GAL4 line was used as the maternal parent. To test its effect on the hINS^{C96Y} induced eye degeneration phenotype, we crossed the RNAi strain to our hINS^{C96Y} line (used as maternal parent), so that both hINS^{C96Y} and the RNAi constructs are driven by GMR-GAL4. The resulting phenotype was compared to the cross between hINS^{C96Y} (maternal) and w1118. At least 10 individual flies were measured per cross and a t-test was used to determine significance at 0.05 level with multiple testing correction.

For mutant lines, we substituted GMR-GAL4 with w1118 in the first test and used w1118 as a control. The same scheme was used for the second test.

Pyro-sequencing assays

Six 18bp lines and eight 4bp lines were randomly chosen and paired to form 15 crosses (see Figure S.17 for cross design). All crosses were reared at 25C. Three sets of ten 3rd instar wandering larvae were collected from each cross and dissected in 1x PBS to isolate the eye imaginal discs. RNA isolation and cDNA library preparation are the same as described above for qPCR analysis. Genomic DNA were extracted from adult flies from the same cross.

Because the 18bp/4bp polymorphism is in the intron, we identified SNPs in the cDNA that could be used to distinguish the two alleles in each cross. Five such exonic SNPs were identified and their corresponding pyro-sequencing assays covered all 15 crosses. Pyro-sequencing was performed as previously described (Wittkopp *et al.*, 2004). Briefly, each of the three cDNA and one gDNA sample per cross was analyzed in four replicate PCR amplifications and subsequent pyrosequencing to determine relative expression. The ratio in genomic DNA analysis was used to account for amplification bias. The resulting 3 (cDNA biological replicates) x 4 (PCR and pyro technical replicates) = 12 corrected ratios were first log2 transformed and analyzed using ANOVA $y_{ij} = \alpha + L_i + \epsilon_{ij}$, where L_i is a random effect term for the biological replicates ($i = 1, 2, 3$). If the random effect term has a p-value > 0.1 (true for 13 of the 15 crosses), all data were pooled to fit a reduced ANOVA model $y_i = \alpha + \epsilon_i$, from which the estimate and the 95% confidence interval for the ratio of expression (α) was calculated. In the two cases where the random effect term was nominally significant (p< 0.1), a linear mixed-effect model was fit using the lme package in R to obtain an estimate and 95% confidence interval for the same ratio.

4

GWAS IN *DROSOPHILA* SYNTHETIC POPULATION RESOURCE (DSPR)

4.1 Abstract

DSPR represent 1,700 inbred lines of *D. melanogaster* derived from two synthetic, advanced intercross populations, each founded by eight genetically diverse inbred lines. This larger panel of lines allowed us to pursue two goals: (1) to replicate the *sfl* locus identified in DGRP and (2) to identify novel loci taking advantage of the higher statistical power. In the preliminary study, we phenotyped 100 RILs from each of the two synthetic populations (A and B), mainly for the purpose of replicating the *sfl* locus. These 200 lines together nearly matched the range of phenotype spanned by the 178 DGRP lines, suggesting that the genetic architecture observed in DGRP is a common feature rather than special property of that population. Despite the similar range of phenotype and the presence of the *sfl* 18/4bp intronic polymorphism at the same intermediate frequency in DSPR and DGRP, the results from the 200 DSPR lines did not replicate the previous finding: while the 4bp allele is associated with significantly more severe phenotype in DGRP, the association is non-significant in the 100 B population lines while it is marginally significant at 0.05 level in the A population, but in the reverse direction. One possible reason for the failure lies in the linkage structure: the minimum non-recombining region is more than 300kb in the DSPR lines examined, far surpassing the 55kb *sfl* locus; therefore the genotypic variation is organized into long haplotypes consisting of eight distinct founder types, rather than individual variants as are tested in DGRP. Genome-wide QTL scan performed separately for the two synthetic populations revealed a significant peak in addition to several promising ones in the B population, with the former spanning about 10 genes located on chromosome 2L. We are therefore planning to phenotype all the B lines in the hope of refining the identified peak and identifying additional novel loci. In summary, 200 RILs from DSPR failed to replicate the *sfl* variants identified in DGRP, but pointed to a novel locus. The failure of replication, however, provides an exciting opportunity for us to probe into its underlying mechanism, which could provide valuable insight into this notoriously difficult and

wide-spread problem in human studies of complex disease.

4.2 Background

DSPR is a genetic mapping resource developed by Stuart MacDonald, Anthony Long and colleagues (King *et al.*, 2012). It consists of nearly 1,700 recombinant inbred lines (RILs) derived from two multi-parent, advanced intercross populations. The two populations were each founded by eight inbred lines from a worldwide collection. After an initial mixing stage, the populations were maintained as large, random mating cohorts for 50 generations, at which point nearly 1,700 RILs were created. Because DSPR populations are highly recombined, and have a much larger number of RILs as compared to DGRP, it is well suited for discovering loci of smaller effect size. In addition, the frequency of any rare mutation in a founder line will be elevated in the synthetic population to 12.5% initially, which makes DSPR potentially more power to identify low frequency variants in a natural population.

4.3 Results and Discussion

While the best use of DSPR is for *de novo* discoveries of genetic loci associated with our phenotype, the current experiment uses it primarily to ask whether the GWAS result obtained in DGRP – the identification of *sfl* – could be replicated in an independent set of wild-derived lines. The DSPR is suitable for this purpose, because the indel polymorphism in *sfl* and the linked SNP (chr3L:6523484, dm3), two of the most significant variants identified in DGRP, were both present in the 16 DSPR founder lines at very similar minor allele frequencies (close to 50% in both). This experiment is less conservative than one which uses lines from the DGRP to found a synthetic population because differences between these two founder sets in haplotype structures and additional variants either in the *sfl* locus or elsewhere in the genome could result in failure of replication by several means. The synthetic population we will describe in chapter 5 is, in fact, founded with DGRP lines, and therefore may be a more direct test to validate the *sfl* QTL.

There are several reasons for replicating an association result from GWAS. Most importantly, any statistical association, even at the most stringent threshold, is susceptible to being a false positive. In human studies where experimental validation is generally not possible with human subjects, replication of the QTL in an independent population is a necessity. Experimental validation is often straightforward in *Drosophila* with RNAi and other genetic tools, which we did with *sfl* as well as two other genes in the HSPG biosynthesis pathway. Therefore, the replication in DSPR serves a different purpose in our study, namely to ask whether the same variants have the same or different effect in a different set of genomic backgrounds. If they do, the DSPR sample would allow us to estimate the effect size of the variants, which, due to a known statistical effect (Garner, 2007; Sun *et al.*, 2011), can not be estimated without bias in the original sample (DGRP). If they don't, it will offer us an opportunity to study the mechanism of disease heterogeneity in a tractable system, which we hope will bring light to this prevalent and challenging problem in human diseases.

4.3.1 Power to repicate sfl *in another sample*

I first calculated the sample size (the number of RIL) needed to detect the *sfl* intronic variants with 80% power. To account for the winner's curse effect, which likely resulted in an over-estimate of the true effect size in the original DGRP sample, I applied a method described in Sun *et al.* (2011), which suggested an adjusted estimate approximately 50% of the original estimate from DGRP. Then, using G*Power3 (Faul *et al.*, 2007), I calculated the required sample size for replicating this single locus at $\alpha = 0.05$ level in an independent population (by a two-tailed t-test), assuming the same minor allele frequency and effect size in the replication sample. As shown in Figure S.18, an estimate of 150-200 lines for both allele groups is expected to provide 80% statistical power.

4.3.2 Similar Phenotype Range in 200 RILs from DSPR compared to 178 DGRP lines

To perform the study, I randomly chose 100 RILs that carry either the 18bp or the 4bp allele at the indel site. I also attempted to evenly sample from both synthetic populations (labeled as A and B, which are initiated with different founder sets), so as to maximize the amount of haplotype variation in the sample. However, this reduces the overall power to test for an effect of the *sfl* polymorphism, as explained below. I followed the same procedure for crossing to the hINS^{C96Y}line and phenotyping the F1 male progeny, as described in Chapter 3. The resulting phenotypic distribution of 192 F1 lines (8 lines did not yield useful data) closely matches both the range and the nearly continuous distribution of eye phenotypes spanned by the 178 DGRP lines (Figure 4.1). That the DSPR RIL, founded by just 15 independent lines, could nearly recapitulate the full range of phenotypic variation seen in 178 DGRP lines should not come as any surprise (see Falconer, 1981, pp94). However, it does reassures us that the extreme phenotypes observed in DGRP were not due simply to rare variants that are specific to DGRP. Instead, a combination of common variants shared between DGRP and DSPR, as well as a group of low frequency variants that are specific to each but are similar in the total number, are likely to underlie the phenotypic variability.

4.3.3 Failure to replicate: possible reasons and implications

Despite the similarities in phenotypic range, the *sfl* intronic variants, to our surprise, showed an opposite effect in DSPR: the 4bp allele seemed to be associated with larger eyes, a difference marginally significant at 0.05 level in a sample of 200 lines (Figure 4.2). Upon more careful examination, this difference appears to be entirely driven by population A (p-value = 0.0074 for A vs. 0.87 for B, S.20).

An obvious difference between DGRP and DSPR is the population-specific variation they harbor – in the 55Kb *sfl* locus, for example, approximately 60% of the common variation (MAF > 0.05) are shared between the two (Figure S.19); rare variants are more likely to be specific to either population.

A less obvious, yet crucial difference between the two is that DSPR has far more linkage disequilibrium due to a vastly smaller number of recombination events in its history. According to King *et al.* (2012), the average genetic distance between breakpoints in DSPR RILs is 3.0 cM for autosomes. This means large segments – as long as 1.5 - 3.0 Mb – of founder chromosomes are inherited as a unit. To evaluate the LD structure among the 100 RILs from either the A or B population, I color-coded and sorted the 1Mb chromosome segments surrounding the *sfl* locus in these lines by their inferred founder ancestry (Figure 4.3). The result shows that the minimum non-recombining region in either A or B population is approximately 300Kb, which extends well

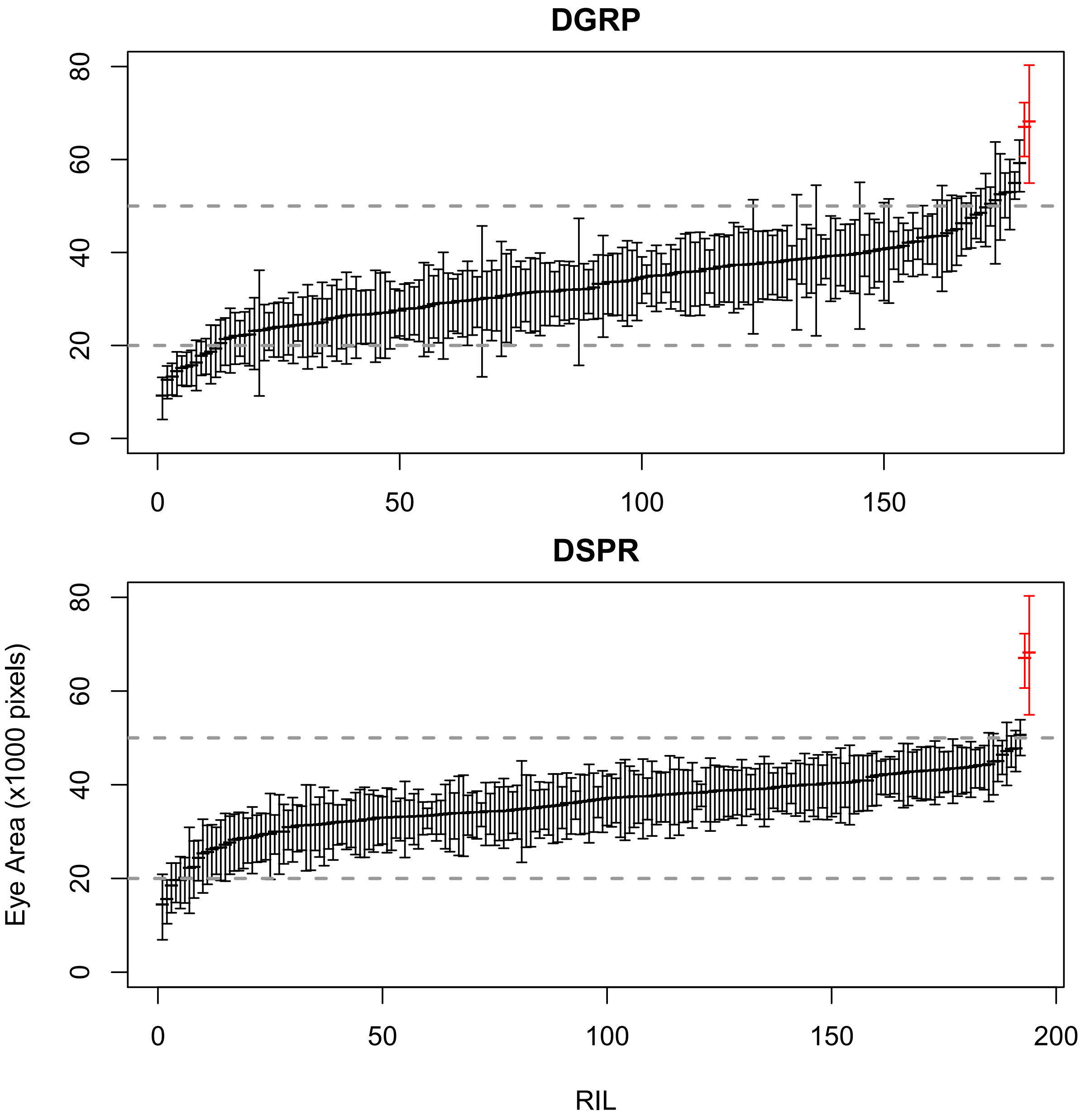


Figure 4.1: **Phenotype distributions in DGRP or DSPR lines crossed with hINS^{C96Y}.** In both panels, the circle and the error bar represent line mean ± 1 s.d. The data points on the right (red) represent wild type fly eye area (the same lines were plotted in both panels). Top panel: 178 DGRP lines; bottom panel: 192 lines from DSPR. The dashed lines mark 20,000 and 50,000 pixel points on the y-axis in both panels for visual comparison.

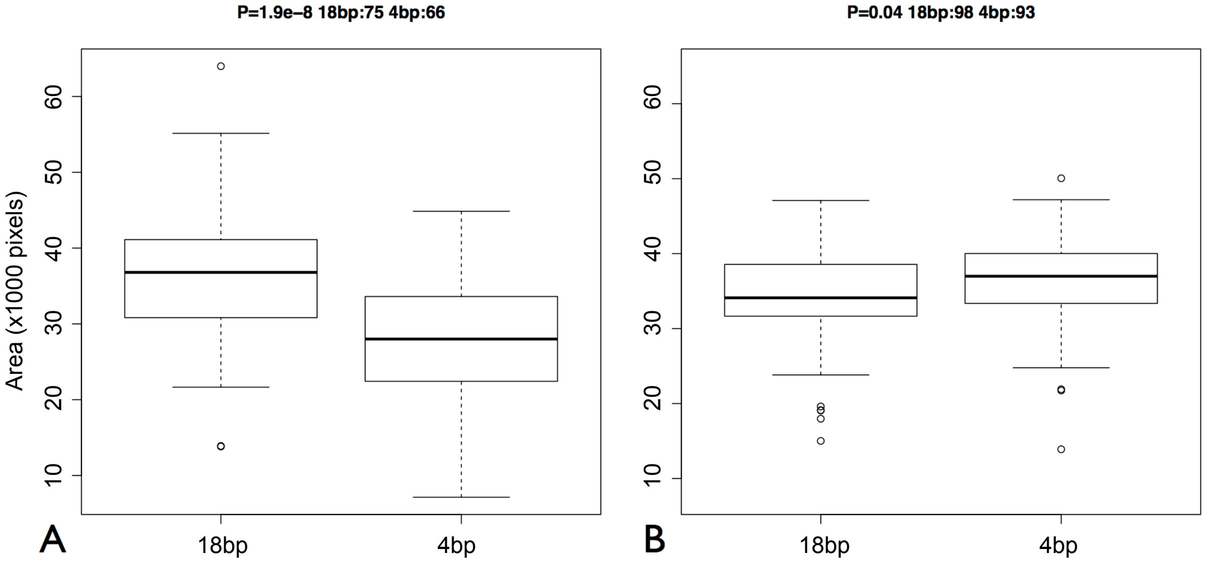


Figure 4.2: **Impact of *sfl* variants in DSPR on eye area in DSPR x hINS^{C96Y}heterozygotes.** Boxplots of the line means in either the 18bp or 4bp allele group were shown for DGRP (A) and DSPR (B) respectively. p-value from a students t-test and the sample size in each allelic group are indicated above the figures. In box plots, the thick line represents the median, the box the interquartile range, the whiskers the 1.5 times the interquartile range and circles for data outside the whiskers' range.

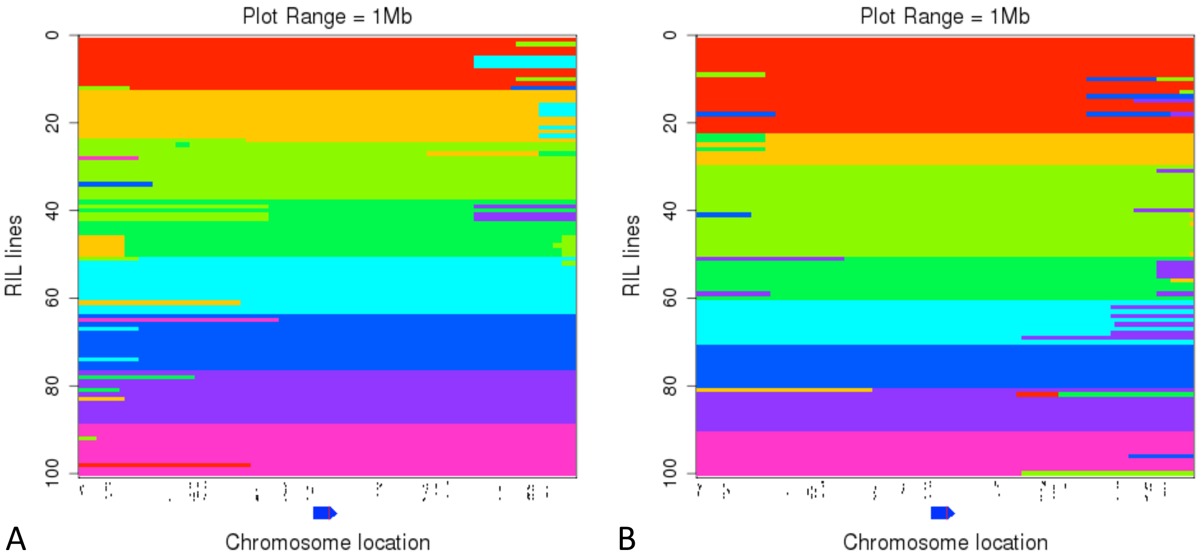


Figure 4.3: **Haplotype structures in DSPR samples around *sfl*.** In both panels, each of the 100 rows is a recombinant inbred line (RIL) either in population A (A) or population B (B). The colors identify the eight different founder chromosomes from which that region of the chromosome is descended from. The region in the plot is centered on *sfl*, which is represented by the blue shape at the bottom. The black vertical marks indicate the actual sequenced tags, from which the ancestry of each segment for each RIL is inferred (King *et al.*, 2012).

beyond the 55Kb *sfl* locus. This has two implications for the analysis: first, it suggests that we should test for association between the phenotype and the eight founder haplotypes rather than individual sequence variants; second, because the A and B populations only share one of their eight founders and almost certainly have different haplotype alleles, they should be analyzed separately rather than combined. It is therefore not surprising to learn that the opposite-direction effect of the *sfl* variants is driven by just one of the two populations (Figure S.20).

4.3.4 Genome-wide QTL scan points to a potential locus in population B

All DSPR RILs are genotyped at 10,275 high quality SNP markers across the genome, allowing the founder ancestry of each segments to be inferred (King *et al.*, 2012). Taking advantage of this, I performed a standard QTL mapping in population A and B separately, using the eight additive probabilities of founder chromosomes inferred from the sequenced tags. As shown in Figure 4.4, one locus on chromosome 2L was identified to be significant at 0.05 level based on 1,000 permutation test. A conservative confidence interval of the linked locus, defined by a drop of 2 LOD score from the peak (Lander and Botstein, 1989), contains about 10 genes (Table 4.1). Among these genes, Ge-1 is a particularly interesting candidate: its gene product serves as a central component of the P bodies, which are cytoplasmic foci involved in mRNA degradation, nonsense-mediated mRNA decay (NMD), translational repression, and RNA-mediated gene silencing (Kulkarni *et al.*, 2010). Intriguingly, following exposure of cells to oxidative stress, Ge-1-containing P-bodies were found adjacent to a particular type of stress granules (Yu *et al.*, 2005). These results suggest that variants that impact Ge-1 function could contribute to cellular response to stress, as in the case of our hINS^{C96Y} model. We are now planning to test this candidate using RNAi and other genetic tools.

Several other intervals are close to the genome wide significance threshold. Given the polygenic basis as evidenced by the continuum of phenotype variation among the 200 RILs, we are certain of additional loci awaiting discoveries. The sample size in this pilot analysis is certainly not adequate to find most loci, and it did not take full advantage of the DSPR, which contains more than 800 RILs for population B alone. Thus we plan to phenotype an additional set of B lines to identify additional loci and also to refine the resolution of the existing one on chromosome 2L.

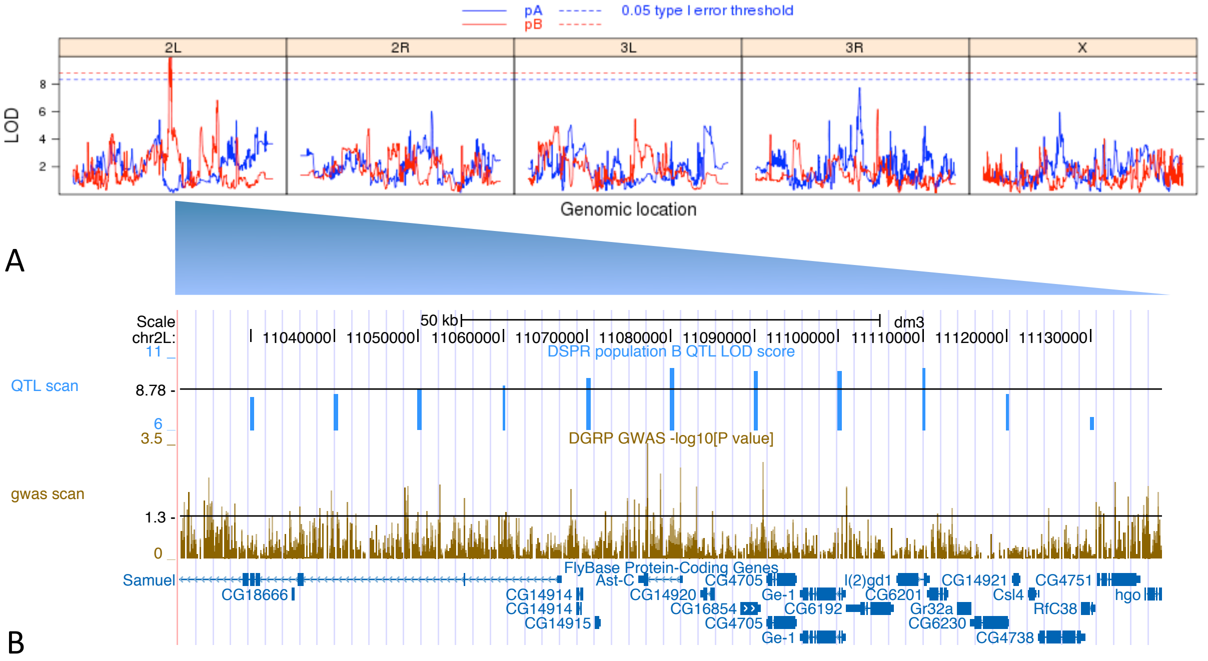


Figure 4.4: **Genome wide interval mapping in DSPR.** (A) Genome scans for QTL in the two populations (blue: pop. A; red: pop. B). Independent tests for association were performed at a grid of genomic locations at 10kb intervals, using the eight additive probabilities corresponding to the eight ancestor haplotypes. The dashed line indicates the 5% type I error threshold obtained from 1000 permutation tests. (B) A zoom-in view of the significant peak in population B on chromosome 2L. In comparison, the log10 p-values from GWAS in DGRP is plotted below for the same region. The black horizontal line indicates a 0.05 level for type I error without multiple-test correction.

Table 4.1: **DSPR Population B candidate gene annotation**

Gene Symbol	**Gene Name**	**Molecular Functionl**	**Expression**
Sameul	Protein coding	Unknown; regulation of gene silencing/growth	Moderate to high expression, not in eye.
Ast-C	Allatostatin	Neuropeptide hormone; Myoinhibitory hormone	High in brain and midgut, not in eye.
CG16854	Protein coding	Catalytic activity	In brain and testis, not in eye
CG4705	Protein coding	Unknown	Moderate expression in brain and eye
Ge-1	Protein coding	found in P bodies, depletion of it causes loss of P bodies	Moderate expression in brain and eye
Reps/CG6192	Protein coding	Unknown; predicted to bind calcium ion	High expression in brain, moderate expression in eye
l(2)gd1	Lethal (2) giant disc 1	Phospholipid binding;	No flyatlas data available
CG6201	Protein coding	carbohydrate metabolic process; Glycoside hydrolase	Moderate to high expression in brain, no info in eye
Gr32a	Protein coding	Taste receptor activity, involved in male courtship	Testis specific
CG6230	Protein coding	ATPase activity	Moderate to high expression in brain and eye

4.4 Materials and Methods

RILs from DSPR

200 RILs (Table S.5) were chosen based on their inferred allelic state at the *sfl* intronic indel (100 each for the 18bp and 4bp allele). The strategy for selecting lines was to sample all 16 founder haplotypes represented at and around that indel site as evenly as possible (inferred from sequenced tags using an HMM algorithm, King *et al.* 2012). As part of this strategy, the A and B populations were equally represented, with 100 RILs from each, in the sample of 200. This strategy assured the widest representation of founder variation but at the expense of statistical power to test for replicating the *sfl* QTL, which is best carried out in the A and B subsamples separately.

Calculating sample size required for replication

To account for the winner's curse effect, which predicts an upward-bias in the estimate of effect size, I ran the br2 program (Sun *et al.*, 2011) with the following parameters: –alpha 2.5e-08 –qt –B1 500 –B2 100. The program predicts a lower bound for the true effect size as 52% of the original estimate ($\beta/\beta_0 \approx 52.2\%$). To estimate the sample size required to replicate the variant, I used G*Power3 (Faul *et al.*, 2007) to produce a curve for the required sample size as a function of the true effect size of the variant (Figure S.18). Based on the predicted lower bound, 125 RILs are expected to achieve 80% power.

Fly cross and phenotyping

All crosses and phenotyping were carried out exactly as described in Chapter 3. Images were processed in the same pipeline as for DGRP. 192 of the 200 RILs passed all quality control to produce useful data for analysis.

QTL mapping

In each of the A and B population, I regressed the mean eye area for each RIL (based on > 10 individual measurements per RIL) on the eight additive probabilities for founder haplotypes inferred at a grid of genomic locations at 10kb intervals. An early version of the mapping software in R was kindly provided by Dr. Elizabeth King. The up-to-date version is now available through the website flyrils.org. The 0.05 type I error rate threshold was established for each of the A and B population through 1,000 permutation tests, in which the RIL IDs for the phenotype column were randomly shuffled in each permutation.

5

EXTREME SELECTION TO IDENTIFY COMMON AND RARE VARIANTS INFLUENCING A COMPLEX TRAIT

5.1 Abstract

Part of identifying the genetic architecture of complex disease is to understand the relative contribution from common vs. rare variants to the phenotypic variability. The widely used approach of genome-wide association studies (GWAS), however, are inherently biased towards common variants; rare variants of moderate effect-size are generally not targeted and have been suggested to underlie the "missing heritability" in common disease and quantitative trait studies (Manolio *et al.*, 2009). Once instances from both categories are identified, we would also like to know whether the two classes of variants differ in aspects other than their population frequency. To make inroads into both questions, we designed a novel strategy to improve the power for detecting rare variants. By creating a synthetic population initiated by a limited number of founders – eight in this study – any rare variant carried by one founder becomes elevated in its frequency, in this case to a minor allele frequency (MAF) of 12.5%. To balance the gain in statistical power with the amount of variation examined, we propose using a 8x8 matrix utilizing 64 DGRP inbred lines, from which 16 eight-founder synthetic populations will be created. After letting each synthetic population randomly mate and reproduce for a number of generations, the resulting advanced intercross population is used for a one-step, extreme selection experiment. The basic idea is to select for the phenotypic extremes in the genetically admixed population and use sequencing to estimate the allele frequencies in both tails of the distribution. Variants associated with the phenotype are expected to exhibit large differences in their frequencies, which can be detected by a statistical test such as the Fisher's Exact Test. Simulation suggests that this strategy achieves 80% power for a variant that starts as a singleton among the eight founders and has a moderate effect size of $d/\sigma = 0.5$. In comparison, GWAS would need more than 5,000 lines to detect a variant of the same effect size and has a population frequency of 5%. Simulation also suggests that this approach has high accuracy and reasonable resolution. Future studies will focus on (1) using simulation to explore

distinct signatures in the allele frequency data that can be used to distinguish causal variants from different frequency classes; (2) to test and refine the analysis methods by applying them to the pilot experiment data. In summary, we expect to use this novel strategy to obtain both an overall picture of the relative contribution from rare variants to phenotypic variance and also to identify specific instances for genetic and molecular investigation.

5.2 Background

In Chapter 3, I described the results from a genome-wide association study to identify genetic variants that modify the disease phenotype. However, a clear discordance exists between the continuous phenotype distribution, which undoubtedly suggests a complex genetic architecture, and the single large-effect candidate locus (*sfl*) identified in the study. The reason underlying this discrepancy is likely to be a practical one: GWAS is highly conservative in calling significant associations to assure a low false positive rate. Furthermore, while the method is unbiased with respect to the location and types of variants, it is not unbiased when it comes to the population frequency and effect size of the variants. As a result, GWAS is most powerful for identifying intermediate frequency, moderate- to large-effect variants, but not their complements. The missing heritability problem, i.e. combining top GWAS candidates only explains a small percentage of the total heritability (Manolio *et al.*, 2009), leads to the hypothesis that a substantial part of the phenotypic variance is attributable to low frequency variants of modest effect size, which are undetectable in GWAS (the actual limit for frequency and effect size depends on the sample size of the study). The approach that I will describe in this chapter is designed specifically to have enhanced power to detect low frequency variants (although, it is still more powerful for common variants). The goal of the study is to both estimate the total contribution from low frequency variants as a proportion of the total, and also identify particular instances, which can be subjected to further characterization with respect to genomic location, molecular mechanism and fitness effect.

5.3 Study Design

In order to enhance the power to detect rare variants, the key of the approach is to increase their frequencies in a synthetic population that is initiated with a limited number of founder lines. Our design incorporates the following features:

- Randomly chose 64 lines from the 192 inbred lines in DGRP;
- Use them to found 16 synthetic populations;
- Each population will be founded by eight DGRP lines, and
- Each DGRP line will be used twice in the 16 populations

To achieve this design, an 8x8 matrix of 64 lines is sampled across the rows and columns to produce 16 synthetic populations, each with 8 founders. This design balances the desire to maximize the total number of genomes to be examined and to achieve a minimum minor allele frequency in consideration of mapping power. The 64 lines are expected to capture the majority of common variants, while rare variants will be present only once in the 64 lines but twice in the 16 synthetic populations. The latter allows a rare mutation detected in one synthetic population to be replicated in another. After an initial mixing stage (a round-robin cross scheme), each synthetic population is maintained in discrete generations for n generations, resulting in a multi-parent, advanced intercross population. In the pilot study, one such population has been constructed, and had reached generation 11 ($n = 11$) at the time of the mapping experiment.

A major feature of this synthetic population is that all segregating variants have a minimum minor allele frequency (MAF) of 1/8 at the time of population initiation. This means any low frequency allele in a founder line (chosen from the 192 DGRP lines) is automatically boosted to at least 12.5% MAF. However, it is important to note that both unintended selection and genetic drift can shift the allele frequencies in the n generations of random mating, which could result in lower MAF for certain variants, possibly rendering them below the detection limit of the study. To minimize the influence of genetic drift, we maintained the population at a size of 1,500 adults or more. Through a simple calculation (Text S.2), we expect genetic drift in such a population to shift allele frequencies by 2% for variants with an initial frequency of 12.5%, or 1/8, during a period of

11 generations; for variants with an initial frequency of 50%, we expect a shift of 3% on average. By contrast, in a population of 100 diploids, we expect an average shift of 7.8% or 11.7% for the two frequency classes, respectively. *De novo* mutations may arise during the course of the study, but they can be safely ignored for our purpose as they hardly reach an appreciable frequency for them to matter in the downstream mapping.

Multiple strategies may be employed for mapping QTLs in this synthetic population. For cost and time efficiency, we employed a strategy that is akin to the bulk segregant analysis in yeast (Segrè *et al.*, 2006; Ehrenreich *et al.*, 2010), which we will refer to as extreme mapping. In brief, we took the synthetic, advanced intercross population (at its 11th generation), and crossed it en masse to the hINS^{C96Y} line. The resulting progeny each carries one set of the lab strain chromosomes and one set of chromosomes from the synthetic population. Then, instead of phenotyping and genotyping a moderate number of offspring, we sorted through a much larger pool of F1 flies ($\sim 4,000$) and extracted 200 flies in each of the two extremes of the phenotypic distribution. We plan to pool and sequence the two sets of 200 flies. With allele frequencies estimated from the sequencing data, statistical tests will be applied at each polymorphic site to determine if the MAF is significantly different between the two extremes, which will be used as evidence of association between that site and the phenotype.

Comment on mapping resolution

In terms of mapping resolution, the strategy proposed here shares features of both conventional QTL mapping and GWAS in an outbred population. The creation of the advanced intercross population allows more time for recombination to break down the linkage between variation across the founder chromosomes, though the number will still be vastly smaller than that in a typical outbred population. The estimated average number of break points per chromosome is 2.75 for the second and third chromosome and 3.7 for the X chromosome. An advantage of this study design is that the longer the population is maintained, the more the initial linkage disequilibrium is reduced, and therefore the higher the resolution. At present, however, the relatively small number of recombination events per chromosome in the experimental populations will inevitably limit the resolution. The goal of this pilot study is to troubleshoot the procedure and optimize the parameters in various stages of the experimental design.

5.4 Simulation

I first used simulation to explore several aspects of the study design, including power of the study to detect a variant of a specific initial frequency and effect size, mapping accuracy – how close is the peak to the causal site and mapping resolution.

The simulation was implemented in several stages. First, a population of 4,000 diploid individuals was created with their genotypes randomly sampled from eight ancestor chromosomes (initial mixing). Without loss of generality, I simulated one chromosome arm instead of the whole genome, which has a physical length of 25Mb and a genetic length of 50cM, values that resemble chromosome 3L of *D. melanogaster.* Next, in every subsequent generation, offspring were generated by randomly choosing two parents from the previous generation. Recombination was implemented in the procedure to generate gametes. After 20 generations, a sample of 4,000 chromosomes were sampled (one chromosome randomly chosen from each diploid) to mimic the process of crossing the synthetic population to the lab strain. Phenotype was assigned to each according to an additive model

$$y = \mu + \sum_{l=1}^{k} \beta_k g_k + \epsilon \tag{5.1}$$

k represents the number of causal loci; β_k is the scaled effect size, defined as d/σ, where d is the difference in the phenotype mean between the two alleles and σ is the standard deviation of the phenotype. The error term, ϵ, is normally distributed with mean of 0 and variance of 1. Following phenotype-based sorting, individuals in the 5% tails on both sides of the distribution were extracted, in which the minor allele frequencies were counted and a Fisher's Exact Test was used to determine the significance of the difference between the two tails. For more details of the simulation procedure, please see Materials and Methods.

5.4.1 Power to detect variants of various MAF and effect sizes

The extreme mapping identifies a SNP marker as associated when the null hypothesis of the allele frequencies in the two tails being equal can be rejected at a given threshold. To evaluate the power of the test, I calculated the expected frequencies in the 5% tails of the phenotypic distribution using the mixed normal distribution, for a series of variants of different effect sizes ($d/\sigma \in [0.1, 1]$). The power of the Fisher's Exact Test was simulated in R assuming a sample size of 200 (the number of Bernoulli trials with the probability equal to the frequency in either tail). Power also depends on the initial frequency of the variant, which was set to one of the three values (0.125, 0.25 and 0.5) corresponding to singleton, doubleton and four copies of the eight ancestral chromosomes, respectively. The singleton class is of most interest to us because it contains the highest proportion of rare variants in the original population. As shown in the Figure 5.1, the test achieves $> 80\%$ power for a singleton with a scaled effect size (d/σ) of 0.4 or more (Figure 5.1 top right panel, solid line). In contrast, GWAS generally has poor power to detect rare variants. For example, it takes $> 5,000$ lines to achieve 80% power for detecting a variant of the same effect size with a minor allele frequency of 5% or less. If the minor allele frequency were 10%, the number of lines required is reduced to about 3,000, which is still a fairly large number. This calculation also depends on unknown parameters such as the number of independent SNPs in either mapping strategy. However, my calculations suggest that the overall conclusion stands. For example, assuming 1,000 independent tests instead of 100 for extreme mapping has very small influences on the power–effect-size curve (Figure 5.1, right panel, solid vs. dashed lines). If I loosen the genome-wide significance threshold from 10^{-8} to 10^{-6} for GWAS – a fairly liberal threshold, it still takes 2,300 lines, instead of 3,000, for the latter case above. Therefore, I showed in this section that the extreme mapping strategy complements GWAS in detecting relatively rare variants, by increasing their minor allele frequencies in a synthetic population initiated by a small number of founder lines.

5.4.2 Mapping accuracy

Next, I examined the accuracy of mapping, i.e. the distance between the peak in the chromosome map of p-values to the assigned causal site. I simulated for eight effect size categories. In each category, I ran 50 replicate simulations, each of which used an independently generated synthetic population from eight DGRP lines. One of the eight lines was seeded with a randomly placed causal mutation (singleton in the eight founders). After phenotype sorting and statistical testing, the SNP with the most significant p-value was identified and its distance to the seeded causal SNP

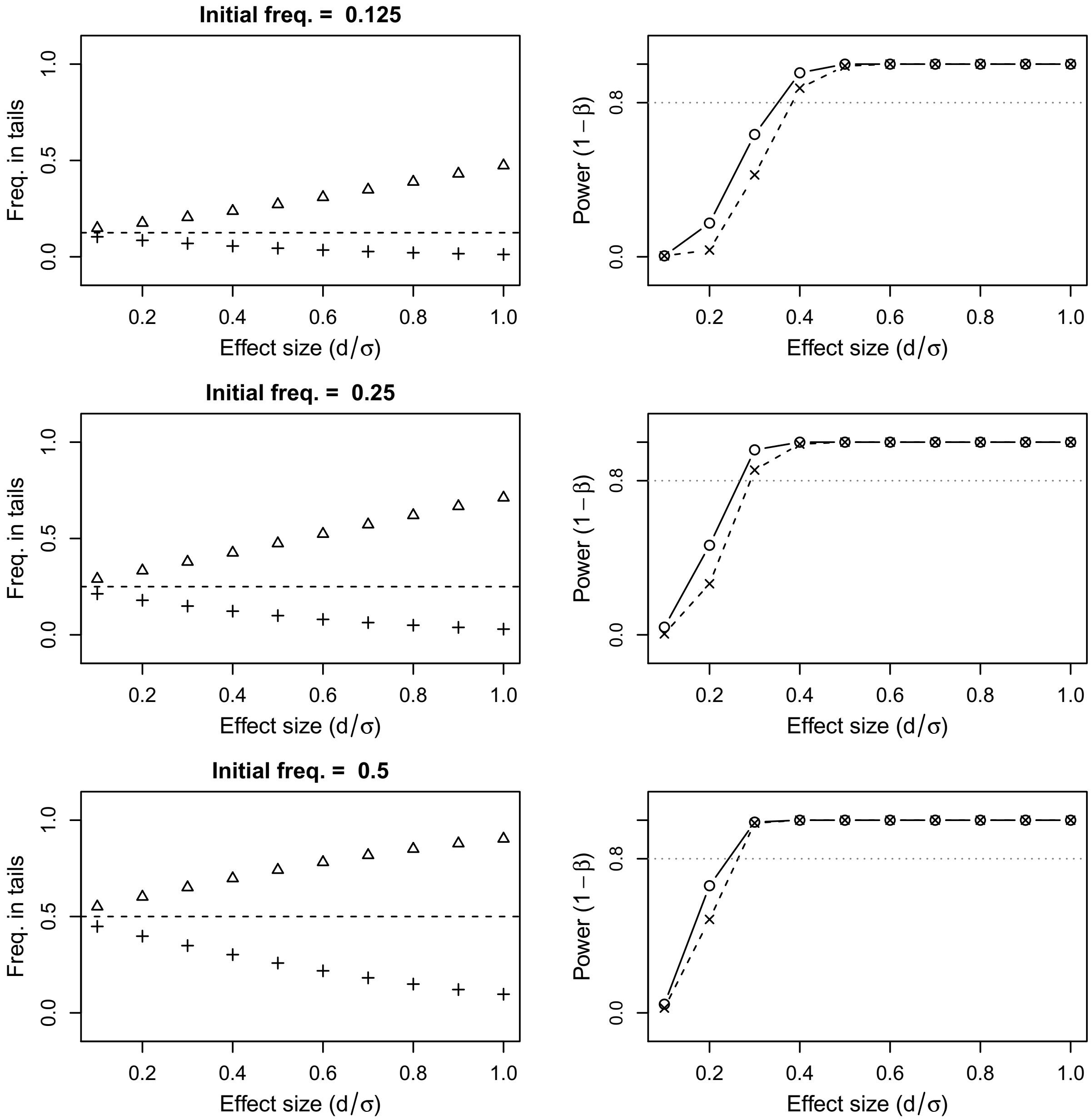


Figure 5.1: **Power of Extreme Mapping.** Each of the three rows represents a different initial frequency of the causal mutation (indicated above each left panel figure and by the dashed horizontal line in each plot). In the left column, the allele frequencies (y-axis) in both 5% tails (triangle: left; plus sign: right; the minor allele is assumed to reduce the phenotypic mean by d) of the phenotype distribution are plotted for 10 different effect sizes (given in units of d/σ). As expected, large effect size variants produce greater differences in their allele frequencies between the two tails. The corresponding right column plots the power of a Fisher's Exact Test comparing the frequencies of the allele in the two tails (200 individuals per tail sampled). The test is performed at 0.05 level with multiple testing corrected by Bonferroni's method, assuming 100 (circle) or 1,000 (cross) independent tests.

was recorded as a measure of the mapping accuracy. As shown in Table 5.1, variants with a scaled effect size of 0.5 or more can be mapped to within 100kb on average, or within 10 kb 50% of the time.

Table 5.1: **Distance to predicted peaks from a causal mutation***

Effect Size	**Distance (kb) to causal site**		
	Mean	Median	80% Quantile
0.3	1,466	335	718
0.5	92	4	230
0.7	68	2	82
0.9	51	5	14
1.1	46	5	20
1.3	17	8	10
1.5	17	0	4
1.7	8	0	1

*Simulated bulk segregant analysis in an artificial population (N=4000) with 8 founders following 20 generations of random mating. We assumed uniform recombination along a single Drosophila chromosome arm (25Mb; 50cM). 200 flies in each tail of the phenotypic distribution were selected for SNP frequency comparison. Each simulation used an independently generated synthetic population derived from 8 Raleigh inbred lines, one of which was seeded with a randomly placed causal mutation. The causal SNP was assigned an effect size in units of d/σ for a normally distributed trait. Shown are results for 50 replicate simulations for each of 8 effect sizes.

5.4.3 Mapping resolution

When two causal variants are spaced closely to each other, sharing of linked polymorphism as well as imperfect mapping accuracy can lead to the merging of the two peaks. To evaluate how far apart two variants can be to still allow the method to reliably distinguish them, I seeded two singleton variants at a distance of either 173kb or 489kb apart on different ancestral chromosomes. I then ran the simulation and located the position of the most significant p-value SNP among the singletons on each of the two ancestral chromosomes that carry the causal variants. This is different from scanning all SNPs as in an actual experiment, as one would not know which of the eight ancestral chromosomes contains a causal variant. Instead, the purpose of this exercise is to explore the phenomenon of “peak merging” as a result of background noise and linkage disequilibrium. Another purpose is to determine whether the difference between two causal singletons vs. one doubleton on the same two ancestral chromosomes are distinguishable. To do so, I ran a second simulation for either distance parameter, in which I seeded a doubleton in the same two chromosomes that carry the causal singletons in the other experiment, and placed the doubleton in approximately the same region (see Figure 5.2). I recorded the peak positions among singletons of either of the two chromosomes, as I did for the first simulation, and compared the result to that in the two-singleton case. As shown in Figure 5.2 right panel, when two causal singletons are 173kb apart, and each has a scaled effect size of 1, the respective peak positions can have substantial overlap, and the difference between it and one causal doubleton cannot be reliably determined. By contrast, when the two singletons were placed 500kb apart, their respective peaks rarely overlap, and the result is clearly distinguishable from that caused by one doubleton on the same ancestral chromosomes.

5.4.4 Pilot experiment and future directions

In a pilot experiment, we created one synthetic population from eight DGRP lines. A round-robin cross scheme was employed initially to enforce heterozygosity, so as to avoid differential fitness among inbred lines causing severe alteration of the allele frequencies. Subsequently, the population was maintained in a 45x45x80 (cm) cage on a 14d cycle (discrete generations) at 25C. Two replicate cages were established following the same protocol and maintained in parallel. At the end of the 11th generation, $2 \times 10 = 20$ bottles in total were set in each cage to collect eggs. From the hatched adults, 400 virgin females per cage were collected and crossed to males from the hINS^{C96Y} line. Therefore, a total of 400 (individuals) x 2 (haploid genomes) x 2 (replicate cages) = 1,600 haploid genomes were sampled.

From the progeny of the cross, nearly 4,000 adult males of the desired genotype were visually sorted based on their eye size, from which 200 flies with the most extreme phenotype (smallest or largest eyes) were collected. An additional sample from both tails were also collected for quantitative measurement of their eye area. Also measured was a random sample of about 60 flies prior to phenotype sorting to estimate the population overall distribution (Figure 5.3). Using these samples, I estimated that our phenotype sorting achieved approximately 20% tail instead of the targeted

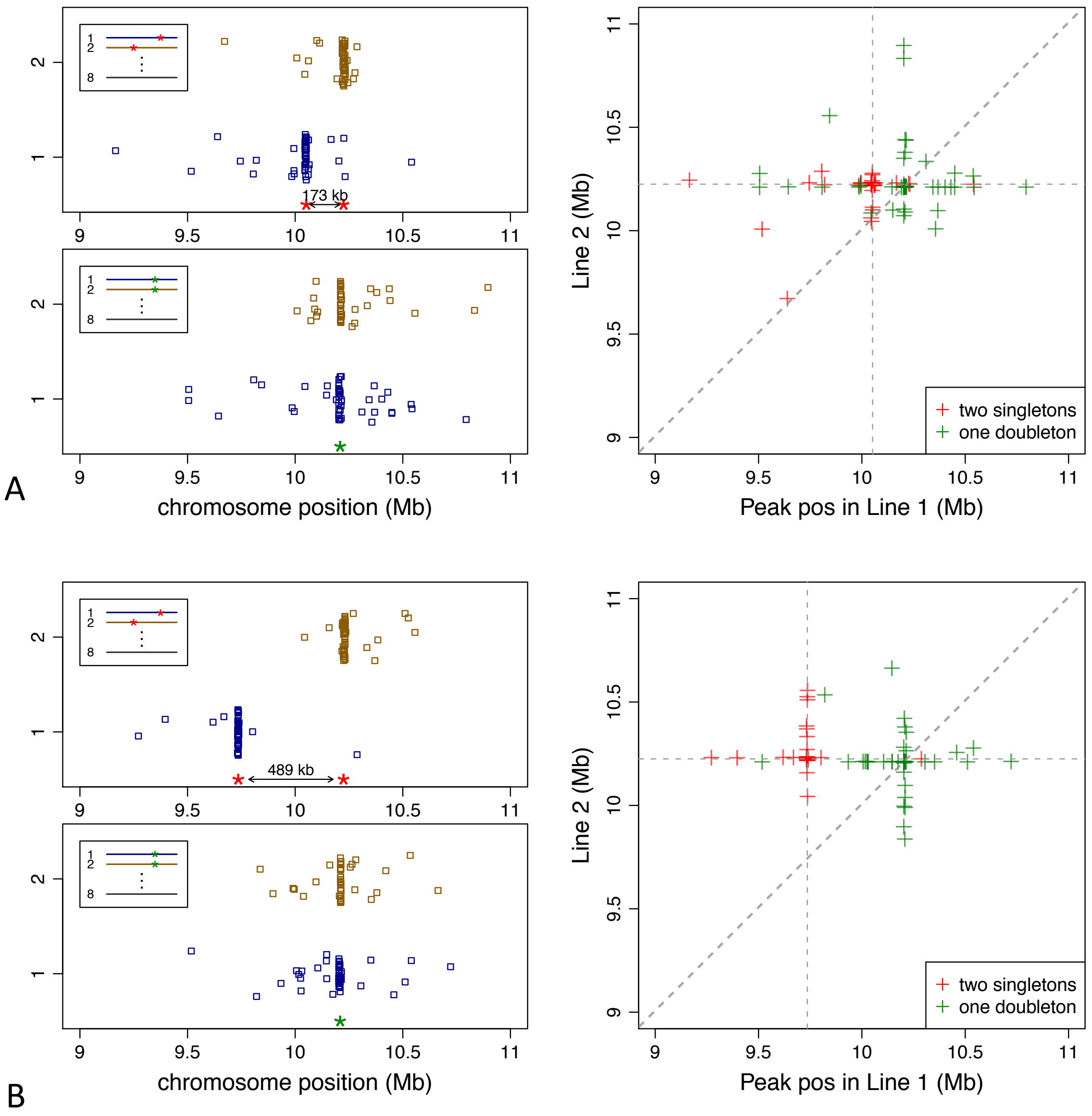


Figure 5.2: **Ability to distinguish two independent causal singletons depends on the distance between them.** In both (A) and (B), the top-left panel depicts two chromosomes each seeded with a singleton at a distance of 173kb (A) or 489kb (B) apart. 50 replicate simulations were run, each with an independently simulated synthetic population. Following extreme selection, the positions of the most significant SNP on each of the two chromosomes were plotted. For the bottom-left panel, the same procedure was applied, except that a doubleton (i.e. a mutation at the same site in two of the eight founder chromosomes) instead of two singletons was seeded. The position of the seeded mutation(s) are indicated by red or green asterisks. The right panel re-plots the top-left panel using a scatter plot, with x and y being the positions of the recorded peaks on each of the two chromosomes for one of the 50 replicate simulations. The horizontal and vertical dashed lines mark the position of the seeded singletons, while the diagonal dashed line is $y = x$.

5%. Further improvement in the sorting scheme is expected to improve this result towards the target. Simulation suggests that even with 20% tails, the experiment is still expected to provide decent power for detecting rare variants with effect size 0.5 or larger (see Figure S.21); we therefore proceed to sequence the selected pools.

In the follow-up of this experiment, a number of questions remain to be answered through a combination of simulation and empirical testing with the pilot data. The first question concerns our ability to attribute the identified peak to a particular variant, and therefore a frequency class, which will be used to determine the relative contribution of rare vs. common variants. In the previous simulation for investigating mapping accuracy, I demonstrated that for a variant with effect size of 0.5 (d/σ), with 50% chance the peak (in the p-value map) will be located within 10kb of the causal variant. However, as the causal variant is not guaranteed to have the most significant p-value, more information is required to determine or infer with confidence its frequency. I believe this is possible because a causal singleton should produce a different signature in the chromosome map of p-values than a doubleton, for example. If we focus on linked singletons, a causal singleton should only influence other singletons belonging to the same founder chromosome, but not singletons on other chromosomes. In contrast, non-singleton variants will influence singletons on multiple founder chromosomes, and to a lesser degree compared with a causal singleton if they have the same effect size. Exactly how this and other signatures may help us distinguish the causal variant from the linked ones will depend on factors such as the density of SNPs and linkage structure, and will be a key problem to be explored through simulation.

Another issue concerns the strategy to sequence the selected pools of flies. An ideal strategy would be to tag each fly before pooling them for sequencing (individual-seq). This way all the linkage information will be preserved, which will allow us to test epistatic interactions between variants. It will also allow us to infer recombination breakpoints, which help reconstruct the haplotype structure and can be used for association testing. However, there are technical challenges associated with the individual-seq strategy, such as multiplexing, an essential step in this strategy. With respect to scale, the current routine is 12-plex, while 96-plex is needed in our approach but is still challenging to perform. With respect to quality, remixing individual fly DNA after barcoding can introduce considerable variability in DNA contribution from each fly and lead to variable coverages along the

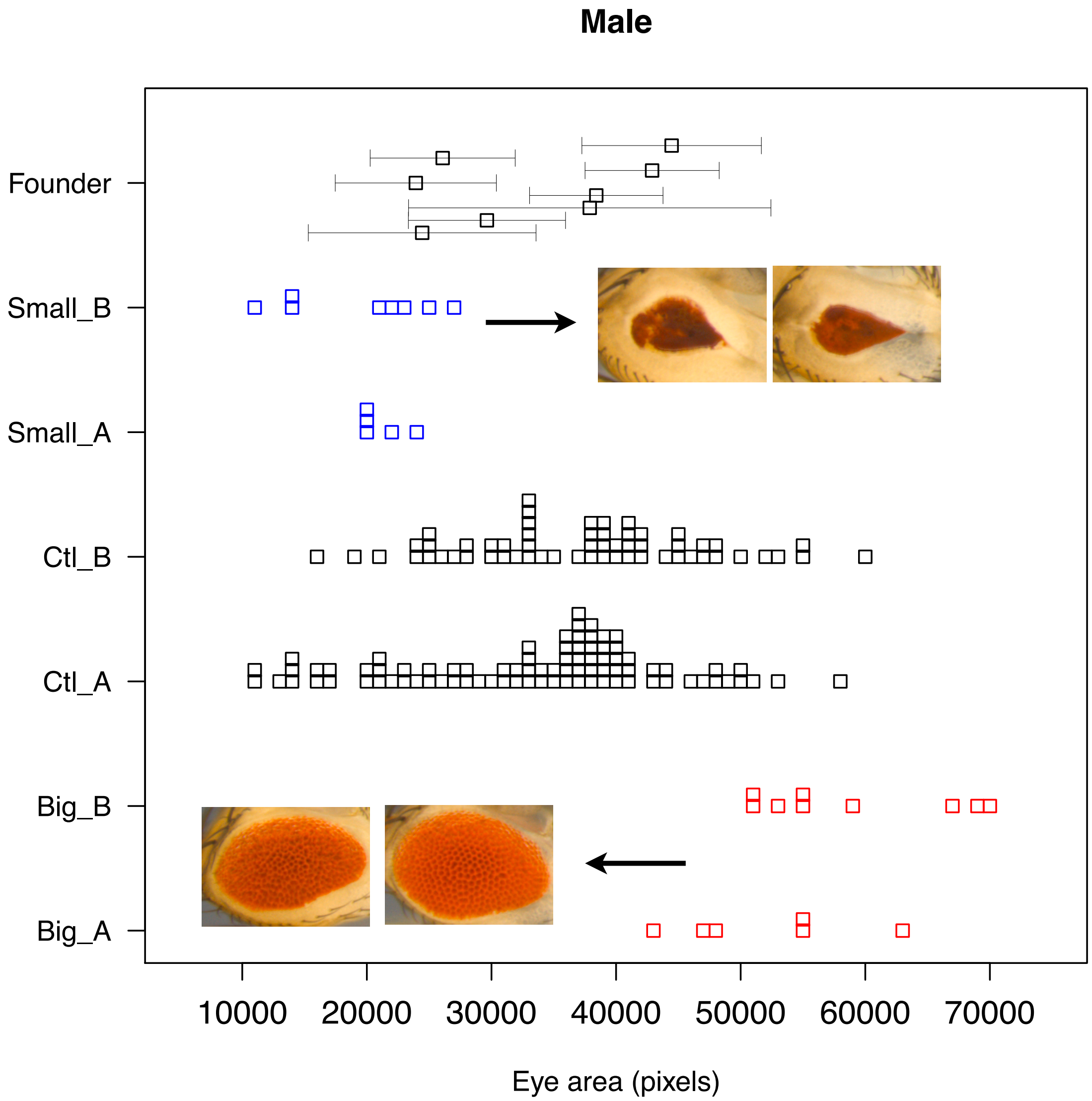


Figure 5.3: **Extreme selection phenotype sorting.** The top row shows the eye area mean ± 1 s.d. (light grey bars) of the eight founder lines. From the second to the last row: sample distribution of eye area from one of the following: small eye (blue), random sample from the cross (black) or large eye (red). Representative pictures of small and large eyes are shown alongside the respective distributions. The realized extreme selection portion as estimated from these samples is approximately 20% ± 3%.)

genome, which will both reduce the sample size and therefore the power of statistical tests, and also increase the variance along the genome. Alternatively, pooling all 200 flies at the DNA extraction step (pool-seq) is more cost-effective and also more likely to yield high-quality data. This strategy, however, inevitably results in the loss of haplotype information. Crucially, the choice between these two strategies will determine the appropriate downstream analyses, and thus we need to balance between the technical challenges and the breadth and ease of the downstream analyses.

5.5 Materials and Methods

Create synthetic population

Eight DGRP lines were used to found the population. A round-robin crossing scheme was implemented in the beginning to enforce heterozygosity and to avoid unintended competition due to differential fitness between inbred lines. The population was then maintained as a large, random mating cohort, on a 14-day discrete cycle. At the time of the pilot experiment, the population is at its 11th generation.

Extreme Selection

Briefly, 10 food bottles were left in each of the two replicate cages for 2d to collect eggs. This procedure was then repeated once to obtain 20 bottles per cage. After the adults had emerged, virgin females were collected and crossed to males from hINS^{C96Y}. Five mating pairs were placed in a vial, and 80 vials in total were set up for each cage, resulting in 2 x 5 x 80 x 2 = 1,600 distinct haploid genomes being sampled in the whole crossing scheme (the actual number is slightly larger because of the additional recombination in the females when producing gametes). All vials were transferred once after 2d, and in another 2d the adults were discarded. Then, from each two vials (same 10 parents) 25 male flies were collected and pooled. Before phenotype sorting, a sample of 60 flies were randomly picked and set aside for estimating the overall population distribution. Then a two-round sorting scheme was used to pick out the phenotypic extremes. Sorting was done manually by a single person under a general-purpose dissection scope, with flies on the CO_2 bed.

Phenotype sorting was done separately for the two replicate cages. Approximately 2,000 flies were examined and sorted in each and 110 were picked in each tail. Ten flies of the 110 were not used for sequencing, but for quantitative measurement of the eye area to estimate the achieved selection proportion.

Power analysis

For a variant of effect size d/σ, the population consists of two subpopulations:

$$X_1, ..., X_{N_1} \sim \text{i.i.d. } \mathbf{N}(\mu, \sigma^2);\ X_{N_1+1}, ..., X_N \sim \text{i.i.d. } \mathbf{N}(\mu - d, \sigma^2) \tag{5.2}$$

A mixed normal distribution is used to calculate the theoretical quantiles at 5% and 95%, which were then used to count the frequencies of the two alleles in the two tails. These expected frequencies were used to calculate the power of a Fisher's Exact Test, achieved through simulation using power.fisher.test() in R. 200 samples were assumed to be drawn for each tail in calculating the power.

Synthetic population simulation

An in-house R script is provided for simulating the synthetic population. The $drift\_rec()$ function was kindly provided by R. R. Hudson. Briefly, an initial diploid population of size N (=4,000) is created by randomly sampling with replacement from eight distinct founder chromosomes. Without loss of generality, the simulation assumes a single chromosome genome, whose genetic length and physical length are matched to the chromosome 3L of *D. melanogaster*; moreover, the first half of the population is assumed to be males and the other half females. To produce each of the N individuals of the next generation, two parents are randomly drawn from each sex. Consistent with the lack of recombination in males in *D. melanogaster*, the male parent directly descends one of its two chromosomes at random to its offspring; in the female parent, recombination is simulated by assuming a Poisson distributed number of events (mean=0.5, corresponding to 50cM) and a uniform distribution of the events along the chromosome.

Simulation for evaluating mapping accuracy

A synthetic population was generated following the above protocol. Genotypes were assigned to the recombinant chromosomes based on the actual sequence variation present in eight randomly chosen DGRP lines (chromosome 3L). One of the eight founder chromosomes was seeded with a single causal variant, which was also a singleton that is only present in that founder out of the eight. Phenotypes were assigned based on an additive model:

$$y_i = \mu + \beta g_i + \epsilon_i \tag{5.3}$$

where β is the scaled effect size, defined as d/σ, where d is the difference in the phenotype mean between the two alleles and σ is the standard deviation of the phenotype. ϵ is a Gaussian distributed error with mean of 0 and variance of 1. Then, the 200 phenotypic extremes on both ends of the distribution were identified, and a genome-wide scan was performed to identify sites with significantly different minor allele frequencies in the two tails, using a Fisher's Exact Test. The SNP with the smallest p-value was identified and its position recorded as the peak location.

To evaluate the mapping accuracy for variants of different effect sizes, eight categories were established (scaled effect size between 0.3 and 1.7). 50 replicate simulations were run for each category, each starting with an independently generated synthetic population. The absolute distance between the peak location and the seeded causal site was used to derive median, mean and 80th percentile estimates.

Simulation for evaluating mapping resolution

The procedure is similar as above. The difference is that instead of one chromosome, two founder chromosomes were each seeded with a singleton causal variant at a distance of either 173kb or 489kb apart. The choice of these two distances were constrained by the existing singletons in the eight DGRP lines that were used as the founder lines of the synthetic population.

To compare the scenario between two causal singletons to that of one causal doubleton, the same synthetic population was used in a second simulation, where a doubleton causal variant was assigned to the same two founder chromosomes. Like the singletons, the doubleton was also picked from available variation from the eight DGRP lines, and was chosen to be between the two causal singletons used above. In both simulations, genome-wide scans were performed for all variants, but only singletons on the two chromosomes carrying the causal variant(s) were evaluated. Peak locations on each of the two chromosomes were recorded and used in the subsequent analyses.

APPENDIX A

SUPPLEMENTARY TEXT

Text S.1 Neutral expectations for the MK test under ascertainment bias

The ascertainment of footprint TFBS exclusively in mel may alter the neutral expectations for MK test and site frequency spectrum in and only in mel. This arises as a result of the mutations in mel being sampled conditioned on the TFBS being detected in the same species. Since affinity-increasing and affinity-decreasing mutations have the potential to change the detectability of the TFBS as a footprint, the conditioned expectation for the neutral pattern is different from the unascertained case. Where there is a fixed or segregating mutation in mel that changes the binding affinity, we assumed that the high-affinity allele is detected with probability 1 while the low affinity one with probability f. Applying these assumptions, we calculate the expected number of fixed mutations in the ascertained sample under neutrality, as well as the number of segregating mutations in each frequency classes for the affinity-increasing and affinity-decreasing mutations. For affinity decreasing mutations, the expected number of segregating mutations at frequency j out of n in the population sample is (Ewens, 2004, equation 9.17)

$$\zeta_j = \frac{\theta}{j}, \text{ where } \theta = 4N\mu \tag{A.1}$$

In the ascertained sample, however, the number of ζ_j^* is expected to be

$$\zeta_{j,dec}^* = \zeta_j \times (1 - \frac{j}{n} \times f) \tag{A.2}$$

where the factor in parenthesis is the probability of sampling the derived allele multiplied by the probability that the TFBS is not detectable, which gives the expected frequency spectrum for affinity-decreasing mutations under ascertainment. For MK table, let D be the actual number of mutations fixed in mel, of which $D * f$ will not be detectable in the ascertained sample. We take the observed values for divergence (D_0) and common polymorphism (P_0, derived allele frequency

> 0.15) for the synonymous No-Change class to estimate the expected neutral substitution-to-polymorphism ratio under ascertainment

$$R^*_{c,dec}(d:p) = \frac{D_0 \times (1-f)}{\sum\limits_{j>a}^{n-1} \zeta_{j0} \times (1 - \frac{j}{n} \times f)}, \text{ where } \frac{a}{n} > 15\% \tag{A.3}$$

For affinity-increasing mutations, the expected number of mutations segregating at frequency j/n in the ascertained sample is

$$\zeta^*_{j,inc} = \zeta_j \times (1 - \frac{n-j}{n} \times f) \tag{A.4}$$

For MK table, we have

$$R^*_{c,inc}(d:p) = \frac{D_0}{\sum\limits_{j>a}^{n-1} \zeta_{j0} \times (1 - \frac{n-j}{n} \times f)}, \text{ where } \frac{a}{n} > 15\% \tag{A.5}$$

Finally, we attempt to obtain an empirical estimate of f. Note that footprint sites for any TF consists of more than a single consensus sequence but spanning a range of affinities. Therefore f must be smaller than 1. To estimate f for a particular TF, we make the conservative assumption that the lowest PWM score among the footprint sites for that TF is the threshold of the detection limit. We then enumerate all possible single nucleotide mutations ($3 * L$, L is the total number of nucleotides belonging to the TFBS for that TF), and among those predicted to be affinity-decreasing we calculated the proportion $\hat{f}$ which would lead the PWM score of the TFBS to drop below the threshold. Averaging across the 30 TF, we estimated $\hat{f} = 0.27 \pm 0.20$

Text S.2 Expected allele frequency shift after n generations of random mating

Let p_0 be the initial frequency of an allele. After one generation of genetic drift, the frequency in the next generation is expected to follow a binomial distribution with a variance

$$V(p) = \frac{p_0(1-p_0)}{2N} \tag{A.6}$$

where N is the population size of diploids. If we treat the process as a random walk, we can calculate the expected change in frequency regardless of direction after n generations. One caveat is that the step size in this random walk, $Z_j = \sqrt{p_j(1-p_j)/2N}$, changes with the frequency. However, if the total change is small in the end, we can treat the step size as fixed as determined by the initial frequency p_0. Using this approximation, we have

$$E[|S_n|] = \sqrt{E[(Z_1 + Z_2 + ... + Z_n)^2]} = \sqrt{nE[Z_j]^2} = \sqrt{\frac{p_0(1-p_0)n}{2N}} \tag{A.7}$$

Note that in the above calculation, for $i \neq j$, $Z_i \perp Z_j$ and therefore $E[Z_i Z_j] = 0$.

Equation A.7 gives the expected frequency shift in n generations. One can also calculate the upper boundary of the shift by replacing $\sqrt{n}$ with n, although the probability of n steps taking the same direction is vanishingly small as n increases.

APPENDIX B

SUPPLEMENTARY TABLES

Table S.1: **CRM studied in this study**

name	symbol	Chromosome	start	end	footprints
CE1000	E74 Promoter	chr3L	17612224	17612895	8
CE1004	h stripe 6	chr3L	8659948	8660496	25
CE1006	Fbp1 enhancer	chr3L	14091229	14091302	2
CE1007	DNA IIA237 enhancer	chr2R	5822784	5823022	2
CE1008	475bp sev enhancer	chrX	10973456	10973933	6
CE1010	dpp 812bp BE/VM enhancer	chr2L	2445768	2446580	30
CE1012	h7 element	chr3L	8658176	8659109	41
CE1013	Stripe 3+7 enhancer	chr2R	5863004	5863516	16
CE1014	1.4kb posterior enhancer region	chr3R	4526528	4527944	9
CE1016	2.7kb Dfd EAE	chr3R	2611055	2613713	15
CE1018	NK-1 promoter	chr3R	17382901	17383486	2
CE1019	dpp proximal promoter	chr2L	2454605	2454756	4
CE1020	DNApol-alpha180 promoter	chr3R	17497407	17497685	4
CE1021	eve promoter	chr2R	5866417	5866922	18
CE1022	tin B enhancer	chr3R	17205669	17206042	3
CE1025	twist ventral activator	chr2R	18933197	18933859	14
CE1026	ChAT 0.3kb 5' region	chr3R	14530829	14531165	2
CE1027	eve MAS	chr2R	5861380	5861582	6
CE1028	ems enhancer element IV	chr3R	9720483	9720788	4
CE1029	iab-2(1.7) enhancer	chr3R	12636228	12637974	11
CE1030	sal enhancer	chr2L	11455638	11456155	18
CE1031	en promoter	chr2R	7415172	7415811	8
CE1032	tin 5' region	chr3R	17203781	17205010	1

Table S.1: (continued)

name	symbol	Chromosome	start	end	footprints
CE1034	Ubx BRE enhancer	chr3R	12526644	12527144	17
CE1035	knirps 5' regulatory region	chr3L	20689622	20690740	39
CE1036	eve mesodermal enhancer	chr2R	5872553	5873441	17
CE1037	SRF-A intervein "c" enhancer	chr2R	20229766	20229892	1
CE1038	omb wing enhancer	chrX	4280499	4281987	1
CE1039		chr3L	14070015	14072441	16
CE1041	dpp promoter	chr2L	2452152	2452471	1
CE1042	Antp P1 promoter	chr3R	2830803	2831280	3
CE1047	E2f promoter	chr3R	17458890	17459274	5
CE1048	tud promoter	chr2R	17070613	17071035	1
CE1049	Antp P2 promoter	chr3R	2758361	2760662	64
CE1050	gsb early enhancer	chr2R	20943848	20944882	19
CE1051	slbo P2	chr2R	20221605	20221785	1
CE1054	otp regulatory region	chr2R	16786018	16787830	9
CE1055	slp1	chr2L	3823118	3824696	9
CE1056	eve stripe 2	chr2R	5865215	5865887	24
CE1057	alpha1 tubulin	chr3R	2912163	2913017	14
CE1058	Ubx promoter	chr3R	12559789	12560360	20
CE1061	Race 533bp enhancer	chr2L	13904731	13905265	3
CE1062	dorsal vessel enhancer	chr2R	20191677	20191999	2
CE1064	Trl promoter	chr3L	14750248	14751686	27
CE1067	so10 enhancer	chr2R	3318590	3319018	8
CE1068	Ubx PRE	chr3R	12589355	12590918	50
CE1069	tin D	chr3R	17209338	17209690	12
CE1073	B3-15 vm1 enhancer	chr2R	20196541	20196901	6
CE1074	tll CD1	chr3R	26676774	26677274	8

Table S.1: (continued)

name	symbol	Chromosome	start	end	footprints
CE1075	ftz-f1 promoter	chr3L	18758302	18758908	3
CE1076	zen promoter	chr3R	2579902	2581525	25
CE1077	rho NEE	chr3L	1461790	1462115	10
CE1078	snail promoter	chr2L	15478269	15480270	12
CE1079	Dll NRE	chr2R	20690304	20691195	3
CE1080	vg boundary	chr2R	8776378	8777133	5
CE1081	vg quadrant enhancer	chr2R	8783677	8784481	8
CE1082	vvl autoreg enhancer	chr3L	6778064	6778580	2
CE1084	ph-p1	chrX	2034243	2034683	6
CE1085	ph-d1	chrX	2019326	2019622	5
CE1086	ph-d2	chrX	2018150	2018483	2
CE1088	scs fragment C	chr3R	7775364	7775488	1
CE1089	hairy stripe 3+4	chr3L	8657462	8658374	15
CE1090	Dpt promoter	chr2R	14752946	14753349	4
CE1094	Deb-A	chr2R	8085631	8086424	2
CE1098	Jra promoter	chr2R	5984349	5984683	1
CE1099	Sgs4	chrX	3143318	3144141	30
CE1101	Act 5C proximal promoter	chrX	5795566	5796038	1
CE1102	sim 2.8 Pe	chr3R	8895635	8898460	9
CE1103	Sgs3 promoter	chr3L	11505175	11505337	3
CE1104	Sgs3 63bp 5'region	chr3L	11504674	11504738	2
CE1105	E74 KpnI RcoRI intron	chr3L	17580603	17580926	3
CE1106	Eip74Eg promoter	chr3L	15650460	15650886	9
CE1107	sc L3/TSM enhancer	chrX	289651	289904	2
CE1108	sc SMC enhancer	chrX	286995	287352	1
CE1109	ko intronic enhancer	chr3L	21080606	21081587	3

Table S.1: (continued)

name	symbol	Chromosome	start	end	footprints
CE1110	ras promoter cBE76	chrX	10638961	10639282	3
CE1112	hsp23 promoter	chr3L	9373335	9373683	9
CE1115	P1 promoter fragment I	chr3R	2827205	2827388	4
CE1116	P1 promoter fragment II	chr3R	2825808	2826148	4
CE1118	P1 promoter fragment IV	chr3R	2823441	2823507	3
CE1121	ftz zebra element	chr3R	2689374	2690045	19
CE1122	ind cis-regulatory	chr3L	15033494	15033709	3
CE1123	Lz LMEE	chrX	9181194	9181476	1
CE1126	copia LTR	chrX	4178496	4183743	1
CE1127	ftz USE	chr3R	2683637	2686199	58
CE1128	Cp15 promoter	chr3L	8721445	8721581	3
CE1129	Cf2 promoter	chr2L	4883103	4883234	1
CE1131	sal 1.1 enhancer	chr2L	11454343	11455436	8
CE1132	kni L2 enhancer	chr3L	20699779	20700471	5
CE1133	ct enhancer	chrX	7424257	7424926	7
CE1134	PCNA promoter	chr2R	16150202	16150749	6
CE1135	bab1 intron1	chr3L	1085706	1086298	4
CE1136	bab2 intron1	chr3L	1175395	1175613	2
CE1138	Adh promoter	chr2L	14614893	14616302	26
CE1140	scs	chr3R	7788458	7789554	5
CE1141	otu promoter	chrX	8383414	8383958	3
CE1144	Sxl intron	chrX	6985658	6986870	2
CE1145	Orb promoter	chr3R	19105085	19106451	1
CE1146	ovo promter	chrX	4957413	4958494	6
CE1147	glass promoter	chr3R	14200149	14200999	2
CE1148	ninaE promoter	chr3R	15713875	15714387	5

Table S.1: (continued)

name	symbol	Chromosome	start	end	footprints
CE1149	en promoter	chr2R	7415997	7416555	14
CE1150	en intron	chr2R	7412745	7414506	15
CE1151	U1 promoter	chr3R	19685239	19685733	1
CE1152	ftz 3' region	chr3R	2691949	2692335	1
CE1155	Actin5C promoter	chrX	5794681	5795112	4
CE1156	Hsp27 promoter	chr3L	9376508	9376729	1
CE1157	ACE	chr3L	8719259	8719581	8
CE1159	tll promoter	chr3R	26677662	26678030	8
CE1160	dpp disk BS1.1	chr2L	2469788	2471532	5
CE1161	dppVRR intron	chr2L	2455937	2457661	23
CE1162	hb promoter	chr3R	4520300	4524677	20
CE1164	Ddc intron promoter	chr2L	19119460	19123023	32
CE1169	117	chr3R	7799721	7799941	2
CE1170	410	chr3L	11552404	11552514	1
CE1172	11B11	chrX	2071273	2071411	2
CE1173	11G4	chr3R	13210846	13211094	2
CE1174	11G5	chr2R	17085621	17085745	2
CE1175	11F6	chrX	12524387	12524538	1
CE1176	116	chr3R	13520975	13521188	2
CE1177	417	chr3R	21067232	21067420	3
CE1178	110	chr3L	5571882	5572102	2
CE1179	407	chr2R	6202914	6203028	3
CE1181	pbx extended	chr3R	12598627	12600114	25
CE1182	DNA pol alpha 73	chr3R	23064563	23065162	3
CE1183	tsh enhancer	chr2L	21852963	21853592	16
CE1184	lab550 enhancer	chr3R	2507112	2507664	3

Table S.1: (continued)

name	symbol	Chromosome	start	end	footprints
CE1185	Antp P1 promoter	chr3R	2824411	2825376	9
CE1186	twi promoter	chr2R	18932628	18933070	3
CE1187	white promoter	chrX	2690577	2691909	9
CE1188	white intron	chrX	2689167	2689286	2
CE1189	Ubx promoter	chr3R	12562550	12563348	4
CE1190	Ubx 3’	chr3R	12481453	12481572	3
CE1191	dpp promoter	chr2L	2450830	2450949	3
CE1192	zeste promoter	chrX	2341986	2342105	2
CE1193	Hsp70 promoter	chr3R	7783054	7784301	13
CE1194	hsp26 promoter	chr3L	9370273	9370988	5
CE1195	hsp83 promoter	chr3L	3193048	3193256	1
CE1196	E(spl) promoter	chr3R	21865315	21866073	2
CE1197	ac promoter	chrX	263167	264014	2
CE1198	l(1)sc promoter	chrX	302783	303755	2
CE1199	yp1 promoter	chrX	9947564	9947891	11
CE1200	Ser wing enhancer	chr3R	22997005	22997817	25
CE1201	ems enhancer	chr3R	9723513	9724763	10
CE1204	ftz neurogenic enhancer	chr3R	2687125	2689005	4
CE1205	Sxl promoter	chrX	6986907	6988310	1

coordinates based on dm3

Table S.2: **Percentage of gain and loss of TFBS**

Category	**Species**	**Cutoff >80%**^{a}	**Cutoff >0**
Loss	*mel*	2	0
	sim	16	16
Gain	*mel*	14	12
	sim	1	0

Table S.3: **MK table in *sim***

Category	**Fix**	**Poly**	**Ratio**	**FET p**
Coding				
Nonsyn	312	438	0.71	9E-09
Syn				
No Chg.	162	446	0.36	–
P→U	449	1714	0.26	0.002
U→P	319	695	0.46	0.04
CRM (aff-dec)^a				
all	38	35	1.09	2E-05
anc.score>0	31	35	0.89	1E-04
anc.score>2	28	33	0.85	3E-03

Note: to evaluate the influence of including *mel* specific binding sites (gained in *mel*, not present in ancestor), three criteria are tested, including all footprint sites, sites with predicted ancestral PWM score > 0 or 2.

Table S.4: **GWAS SNPs below 10^{-5} p-value threshold**

Chrs	Position	Gene Symbol	Site Class	MAF	pval
2L	6018257	H2.0	intronic	0.147651007	4.11E-06
2L	7639113	CG13792	intergenic	0.420560748	3.28E-06
2L	7639113	CG6739	intergenic	0.420560748	3.28E-06
2L	16378839	CG5888	intergenic	0.105960265	5.89E-07
2L	16378839	jhamt	intergenic	0.105960265	5.89E-07
2R	6830823	Spn47C	intergenic	0.048275862	8.91E-06
2R	6830823	luna	intergenic	0.048275862	8.91E-06
2R	8514952	CG17760	intronic	0.079470199	7.11E-06
2R	16411003	CG13422	intergenic	0.173333333	9.29E-06
2R	16411003	CG13426	intergenic	0.173333333	9.29E-06
3L	6523119	sfl	intronic	0.474452555	2.38E-08
3L	6523164	sfl	intronic	0.429530201	1.98E-06
3L	6523166	sfl	intronic	0.429530201	1.82E-06
3L	6523167	sfl	intronic	0.42384106	1.59E-06
3L	6523212	sfl	intronic	0.486666667	3.56E-07
3L	6523285	sfl	intronic	0.482758621	3.08E-07
3L	6523298	sfl	intronic	0.492857143	2.67E-07
3L	6523484	sfl	intronic	0.421052632	5.18E-08
3L	8372543	ImpE1	intronic	0.21192053	7.87E-06
3R	9282003	CG14372	intronic	0.045751634	5.12E-06
3R	9282011	CG14372	intronic	0.045751634	5.12E-06
3R	13265256	CG5873	intronic	0.033783784	8.95E-06
3R	13265256	CG14331	intronic	0.033783784	8.95E-06
3R	13265265	CG5873	intronic	0.034013605	9.52E-06
3R	13265265	CG14331	intronic	0.034013605	9.52E-06
3R	13265268	CG5873	intronic	0.034013605	9.52E-06
3R	13265268	CG14331	intronic	0.034013605	9.52E-06
3R	16891400	AnnIX	intronic	0.452702703	7.66E-06
3R	16891456	AnnIX	intronic	0.445945946	7.97E-06
3R	17323042	C15	intergenic	0.430555556	9.08E-06
3R	17323042	CG7922	intergenic	0.430555556	9.08E-06
3R	17323100	C15	intergenic	0.315068493	2.11E-06
3R	17323100	CG7922	intergenic	0.315068493	2.11E-06
3R	19762489	Gdh	intronic	0.391891892	8.26E-06
3R	19968993	kal-1	intronic	0.358108108	8.25E-06
3R	23486244	CG18437	intergenic	0.486486487	3.24E-06
3R	23486244	Mlc1	intergenic	0.486486487	3.24E-06
3R	24918157	CG11873	intronic	0.489932886	3.48E-06
3R	24920070	CG11873	intronic	0.406896552	8.00E-06
X	18241719	CG6123	intronic	0.033112583	5.88E-06

Table S.5: DSPR RILs used in Extreme Selection

RIL	RIL	RIL	RIL	RIL	RIL
11005	11448	12179	21085	21353	22319
11033	11451	12180	21087	21357	22344
11040	11452	12192	21090	21359	22352
11045	11461	12200	21095	21362	22390
11048	11476	12201	21100	21380	22395
11054	11481	12214	21117	21383	22412
11073	11485	12216	21118	21398	22416
11078	12009	12217	21123	22005	22430
11093	12014	12231	21126	22009	22431
11103	12015	12240	21149	22012	22434
11133	12019	12252	21156	22018	22439
11145	12033	12272	21158	22028	
11155	12037	12273	21159	22038	
11162	12045	12330	21162	22044	
11179	12050	12332	21164	22059	
11188	12051	12337	21176	22071	
11191	12063	12347	21181	22077	
11197	12064	12353	21198	22099	
11200	12070	12359	21199	22107	
11205	12074	12361	21231	22113	
11240	12080	12365	21234	22139	
11256	12087	12372	21239	22141	
11259	12093	12377	21248	22143	
11267	12117	12379	21258	22151	
11289	12122	12382	21266	22205	
11331	12128	21016	21270	22207	
11335	12130	21024	21274	22210	
11344	12132	21026	21277	22225	
11348	12134	21033	21282	22241	
11349	12148	21037	21292	22245	
11358	12150	21038	21293	22250	
11360	12151	21068	21311	22255	
11380	12132	21073	21331	22261	
11381	12167	21074	21342	22268	
11394	12168	21075	21346	22272	
11421	12169	21076	21347	22288	
11422	12171	21078	21348	22302	
11425	12177	21080	21350	22305	

APPENDIX C

SUPPLEMENTARY FIGURES

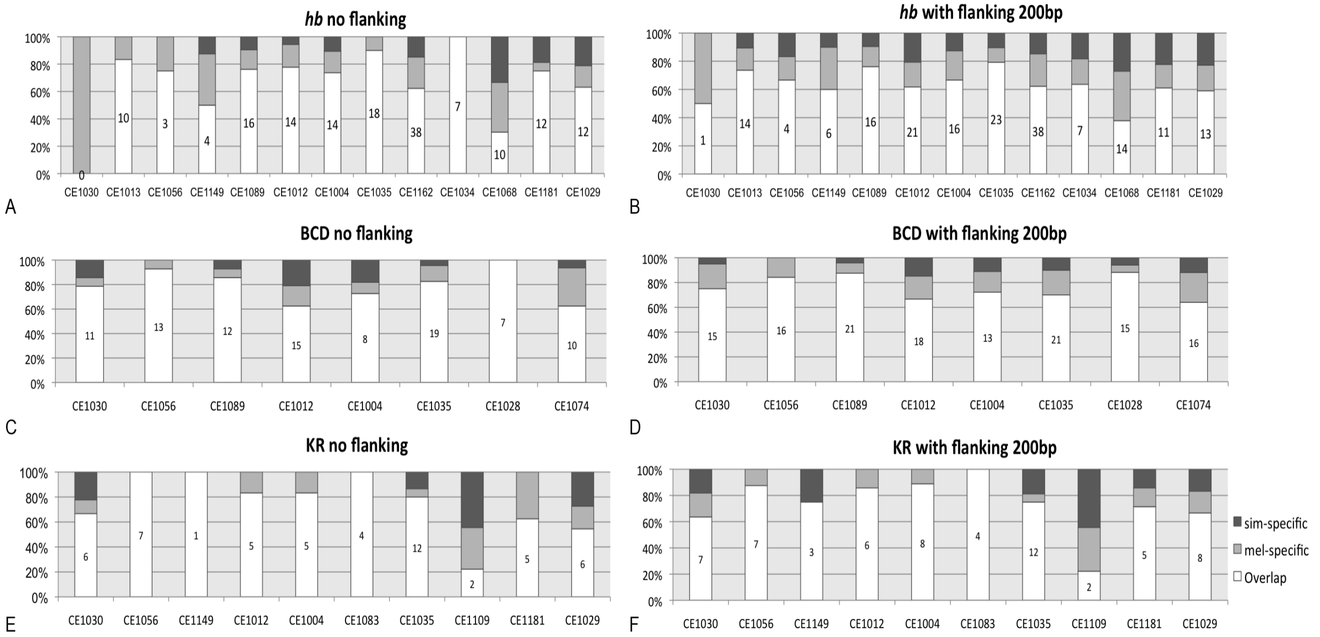


Figure S.1: ***De novo* TFBS prediction in *mel* and *sim*** show potential compensatory sites in *sim* (A), (C) and (E), Proportions of predicted matches to *Hunchback (hb), Bicoid (bcd) or Krüpple (Kr)* PWM that are *mel*-specific (black), *sim*-specific (grey) or shared in both species (white, with numbers indicated) in each HB, BCD or KR regulated enhancer region (defined as regions that contain at least one mel footprint site for the TF). (B), (D) and (F): similar to (A),(C),(E) except that they include 200 bp flanking sequences on each side of an enhancer. Prediction method: briefly, patser v3e (Gerald Hertz, 2002) was used to scan the CRMs in both species. The cutoff for calling TFBS was individually chosen for each TF based on the sensitivity (proportion of footprint TFBS recovered in prediction) and the specificity (additional predicted sites that dont overlap a footprint). For HB, the cutoff used recovered 91.8% of the footprint sites while predicting nearly twice as many (1.94 times) sites. Each TFBS was then aligned in the two species and classified according to whether it was mel-specific, sim-specific, or shared

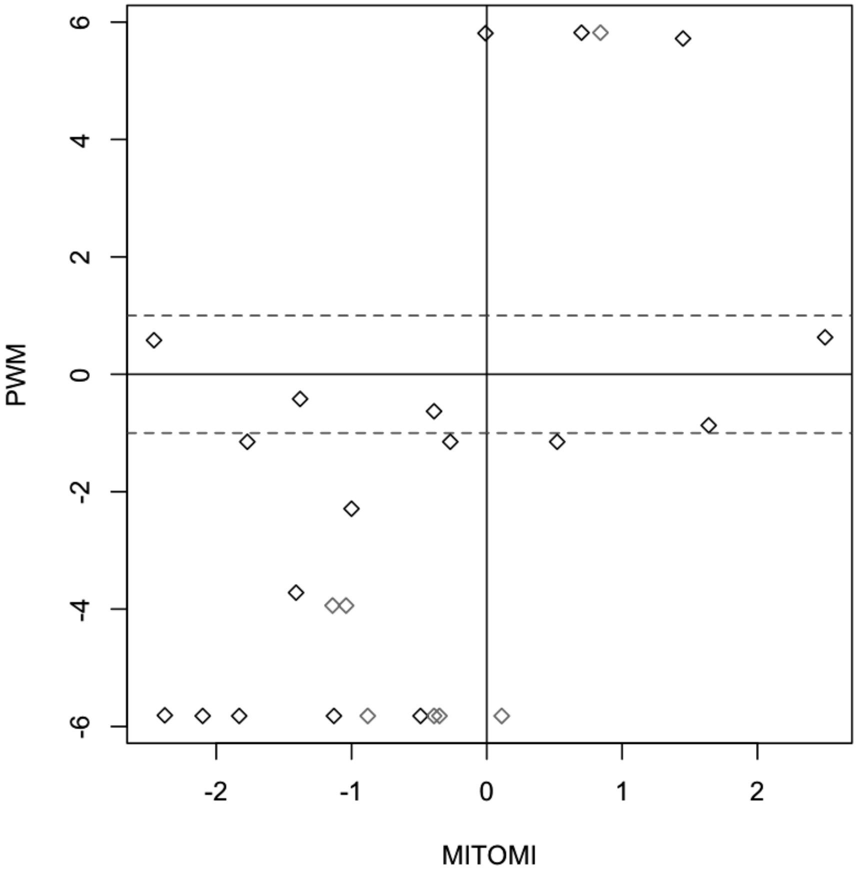


Figure S.2: **Compare PWM prediction and MITOMI measurement for binding affinity change.** Each circle represents an observed nucleotide change between *mel* and *sim* in a HB binding site. MITOMI experiments were performed as described in the methods. Each mutation was measured in two oligonucleotides carrying the original and mutant nucleotide respectively. The two dashed lines indicate the cutoff for PWM scores we applied in the study, in order to reduce mis-assignment.

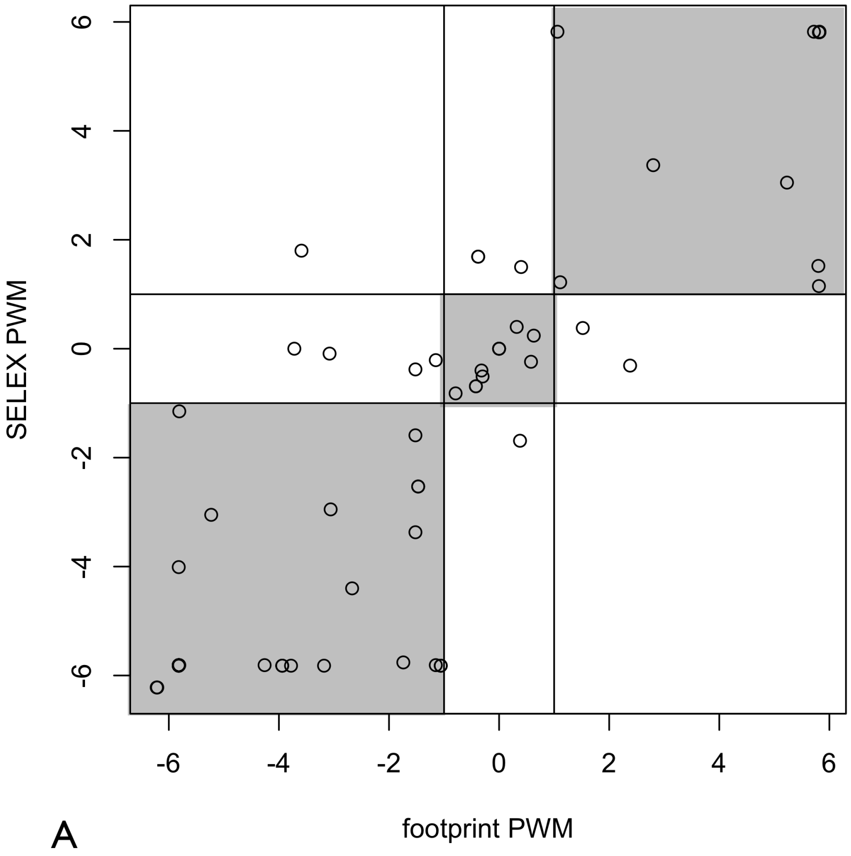


Figure S.3: **Compare PWM based on mel footprints and SELEX PWM.** Each point represents one substitution and its x, y values are the estimates of its effect on binding affinity using the footprint PWM or the SELEX PWM, respectively. 33/34 strong-effect substitutions are consistently assigned by the two sets of PWM into either affinity-increasing or affinity-decreasing categories.

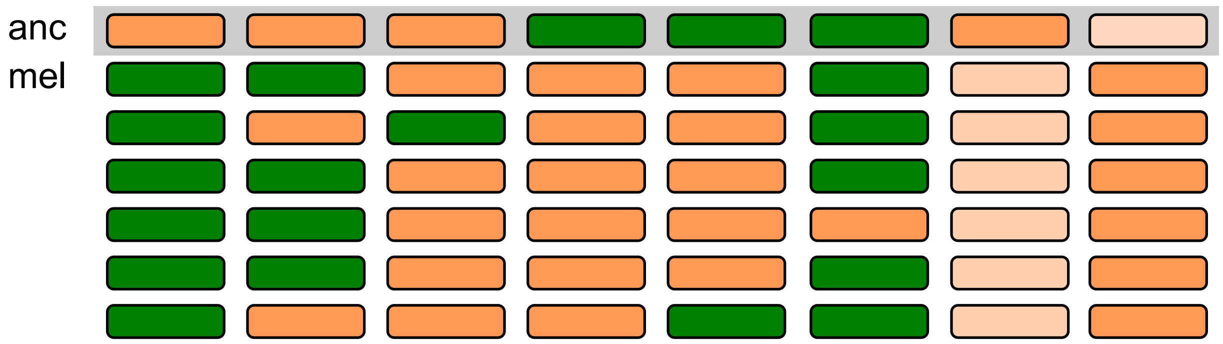


Figure S.4: **Impact of ascertainment on MK table and site frequency spectra in *mel*.** Each box represents a TFBS, where orange indicates relatively strong binding affinity while greens indicates weak affinity. Each column is an alignment of a sample of six *mel* alleles with the inferred ancestral allele. In the first column, a fixed affinity-decreasing mutation in *mel* with a relatively large effect makes the TFBS not detectable as a footprint. In column 2 and 3, the affinity-decreasing mutations are not fixed but segregating, therefore the probability of not detecting the TFBS is proportional to the derived allele frequency (assuming a random *mel* allele is used in the footprint assay). Column 3-6 illustrate the situation for affinity-increasing mutations, where the substitutions are always detectable but the segregating mutations are detected with higher probability when the derived allele frequency is low. The last two columns represent cases where both alleles are detectable. To incorporate the uncertainty in the detectability of the low-affinity allele, we define a parameter f for the probability that the weak allele is not detectable.

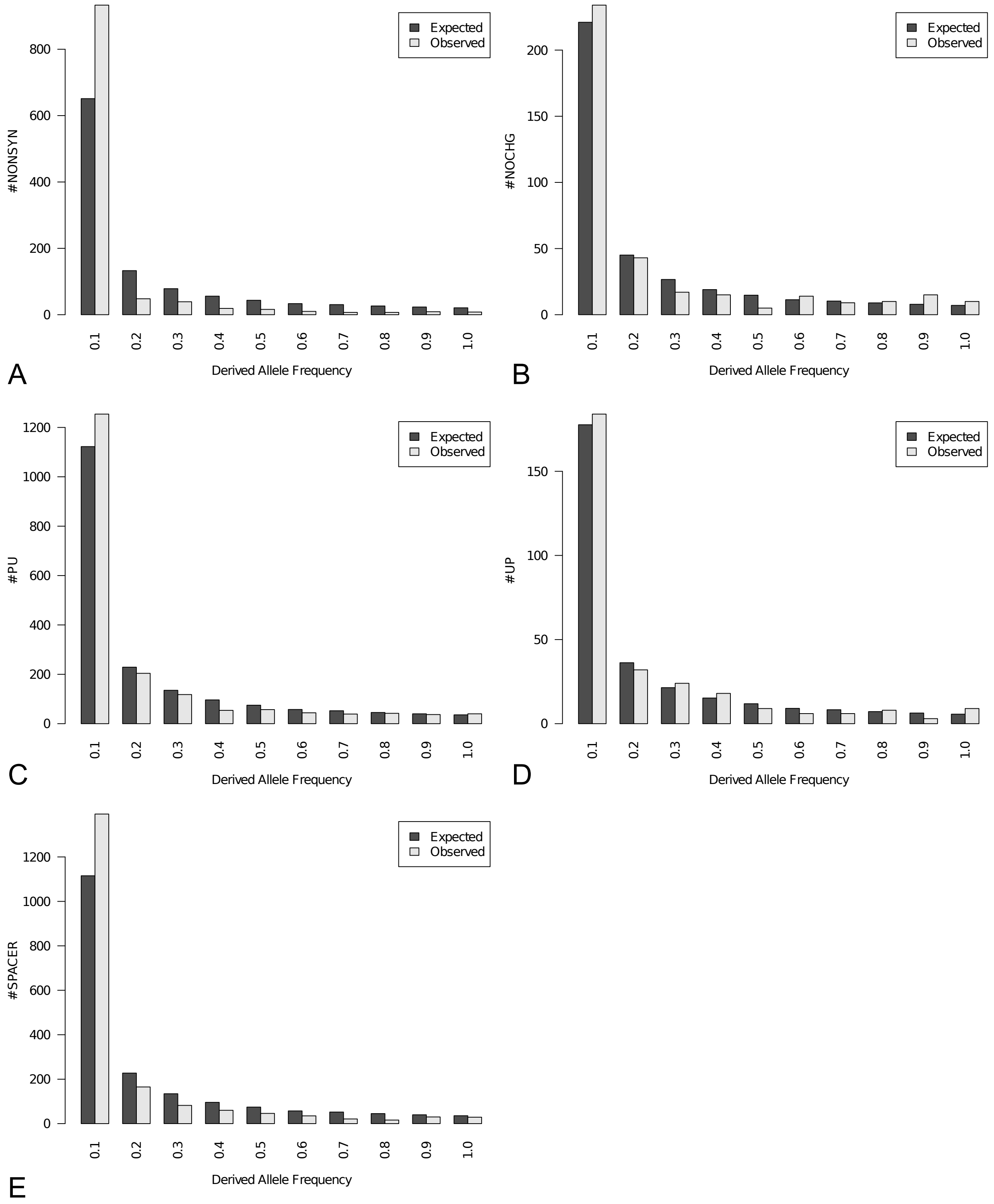


Figure S.5: **Site frequency spectra for different classes of sites** (A) Non-synonymous; (B) Synonymous No-Change (C) Preferred-to-Unpreferred; (D) Unpreferred-to-Preferred; (E) Spacers in CRM. Black: neutral expectation; Gray: observed site frequency spectrum.

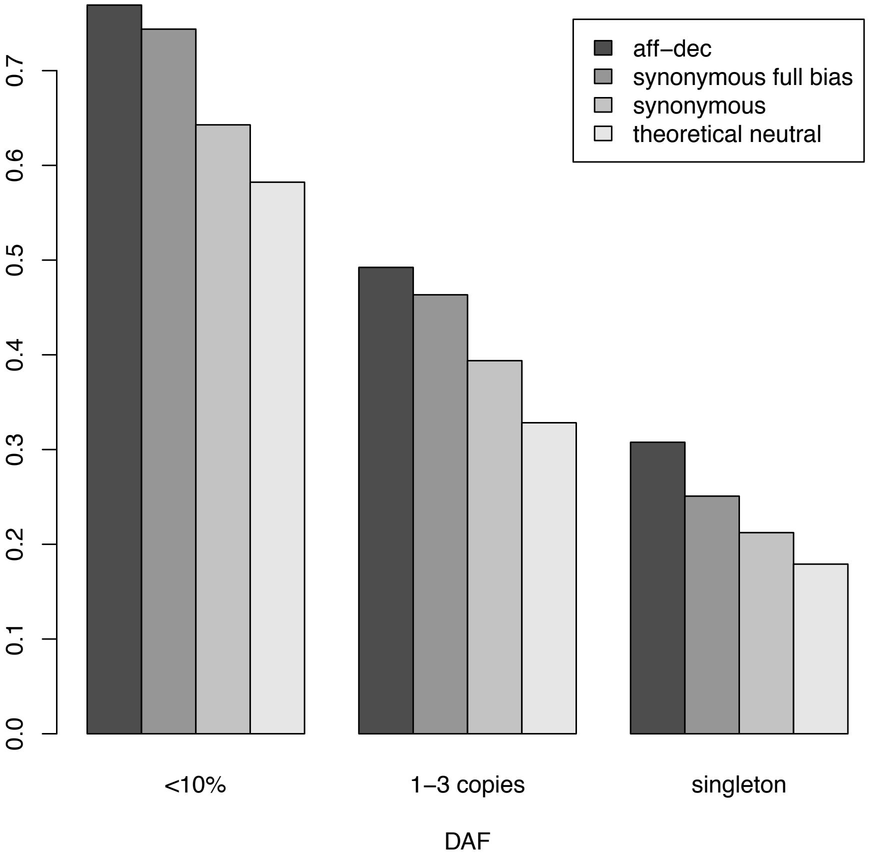


Figure S.6: **Excess of rare variants in affinity decreasing mutations in *mel* suggests purifying selection.** The proportion of low frequency class(es) for affinity-decreasing mutations compared to the theoretical neutral expectation, the observed synonymous sites, or the expected proportion for synonymous sites under ascertainment assuming $f = 1$. DAF: derived allele frequency.

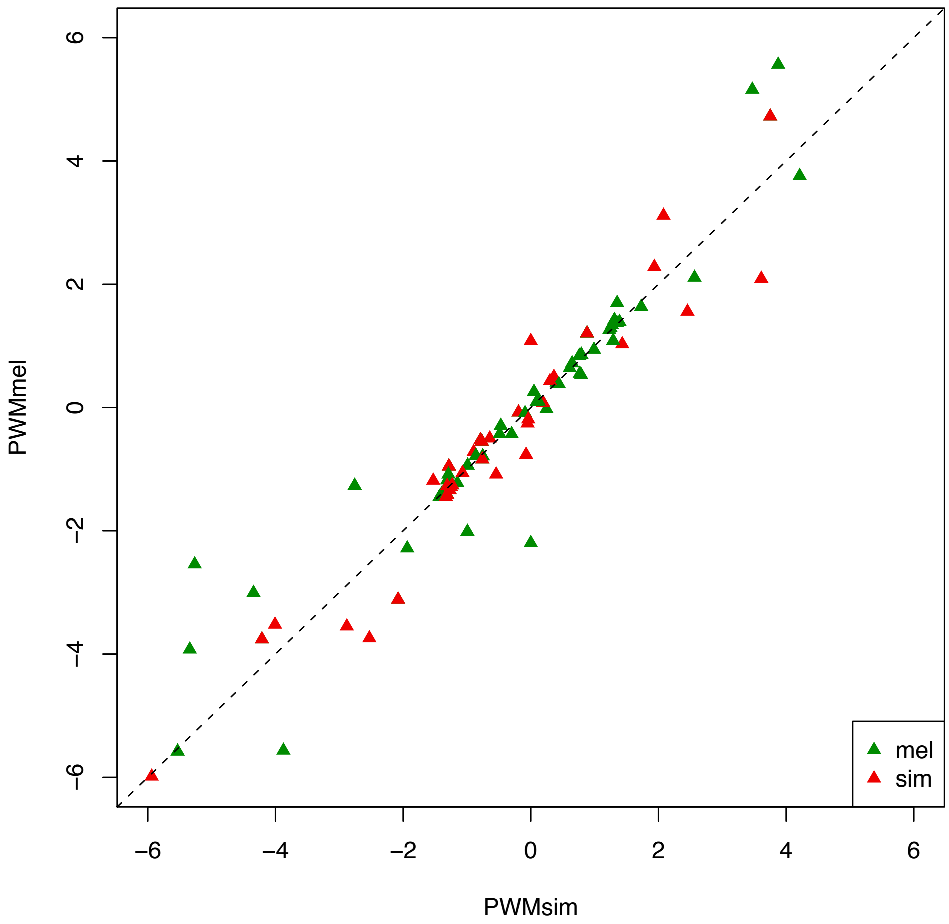


Figure S.7: **PWM derived from *mel* footprints or orthologous sequences in *sim* produce consistent results.** On the scatter plot each point represents a single nucleotide mutation with its x, y values being the estimates of its effect on binding affinity using either the *mel* PWM or the *sim* PWM, respectively. Green and red triangles are mutations occurring on *mel* or sim lineages. From the figure, the PWM have very little biases with respect to scoring mutations from the species where it is derived or the other species.

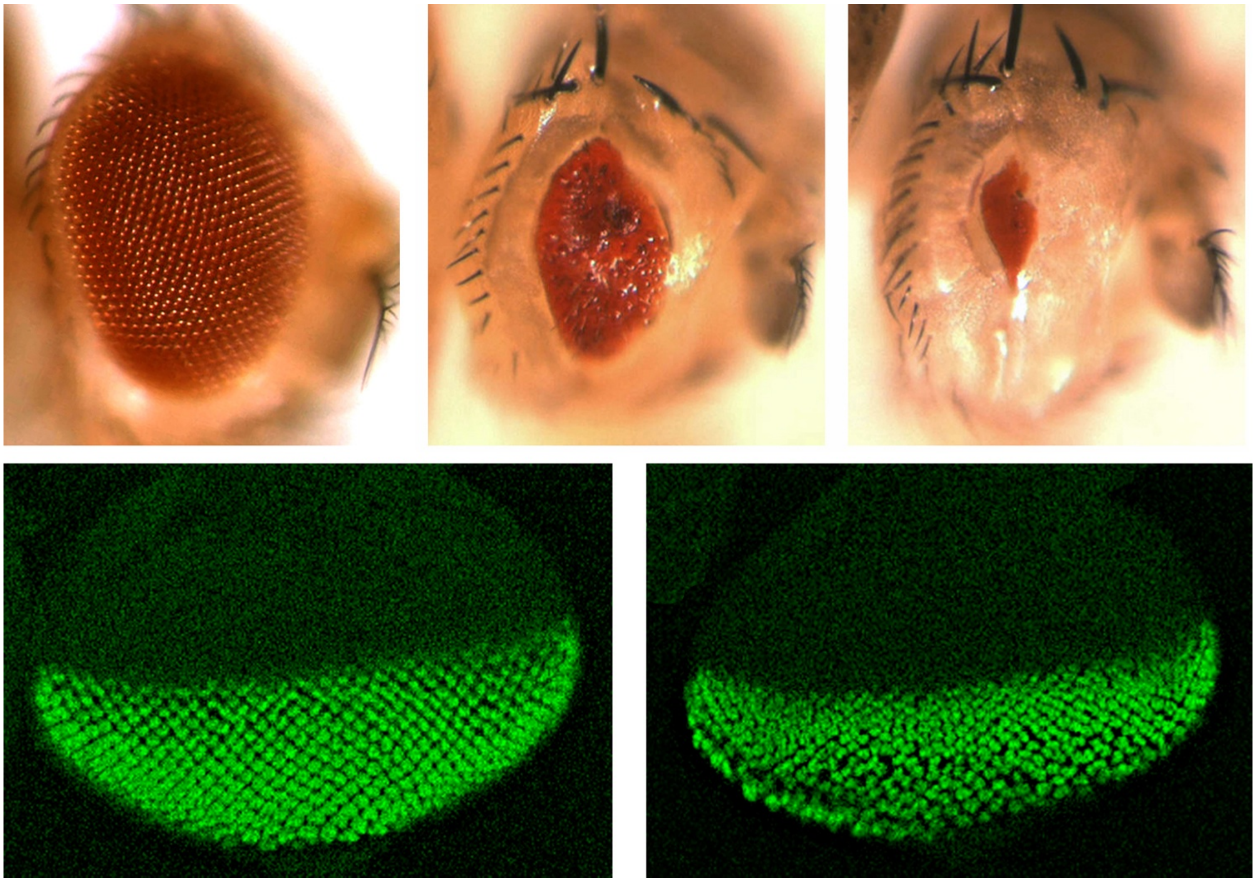


Figure S.8: **Developmental and adult phenotypes of hINS$^{\mathbf{C96Y}}$ expression in the eye imaginal disc.** Top: male, 2-5d adults. Genotype from left to right: hINS^{WT}/CyO (control); hINS^{C96Y}/CyO; hINS^{C96Y}/hINS^{C96Y}. Bottom: eye portion of the eye-antenna imaginal discs from 3rd instar larvae, stained with ELAV. Left: hINS^{WT}/CyO (control); Right: hINS^{C96Y}/CyO.

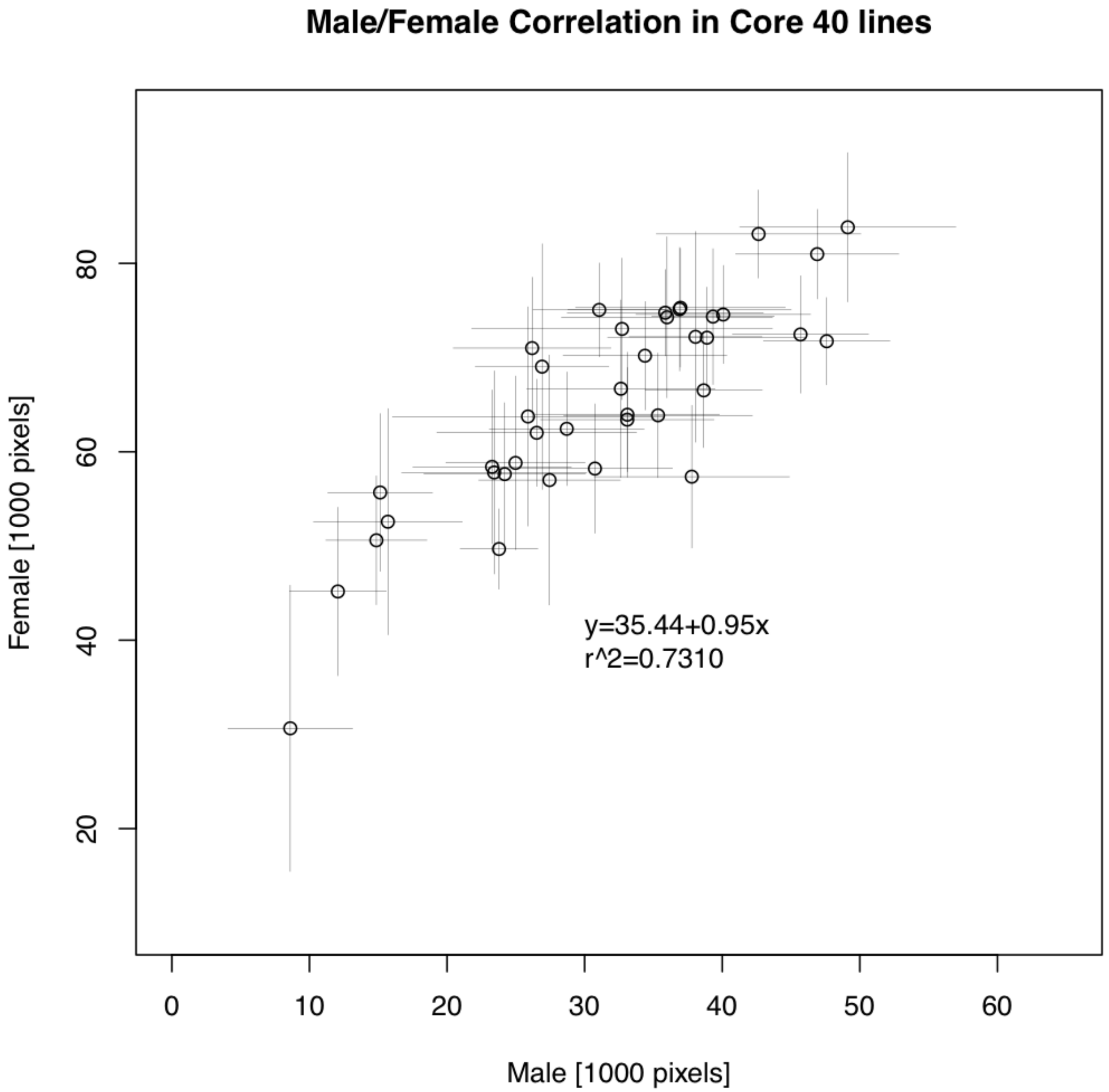


Figure S.9: **Correlations of eye area between F1 males and females within the same cross.** mean ± 1 s.d. are plotted for a subset of 38 lines. The least square linear fit is indicated.

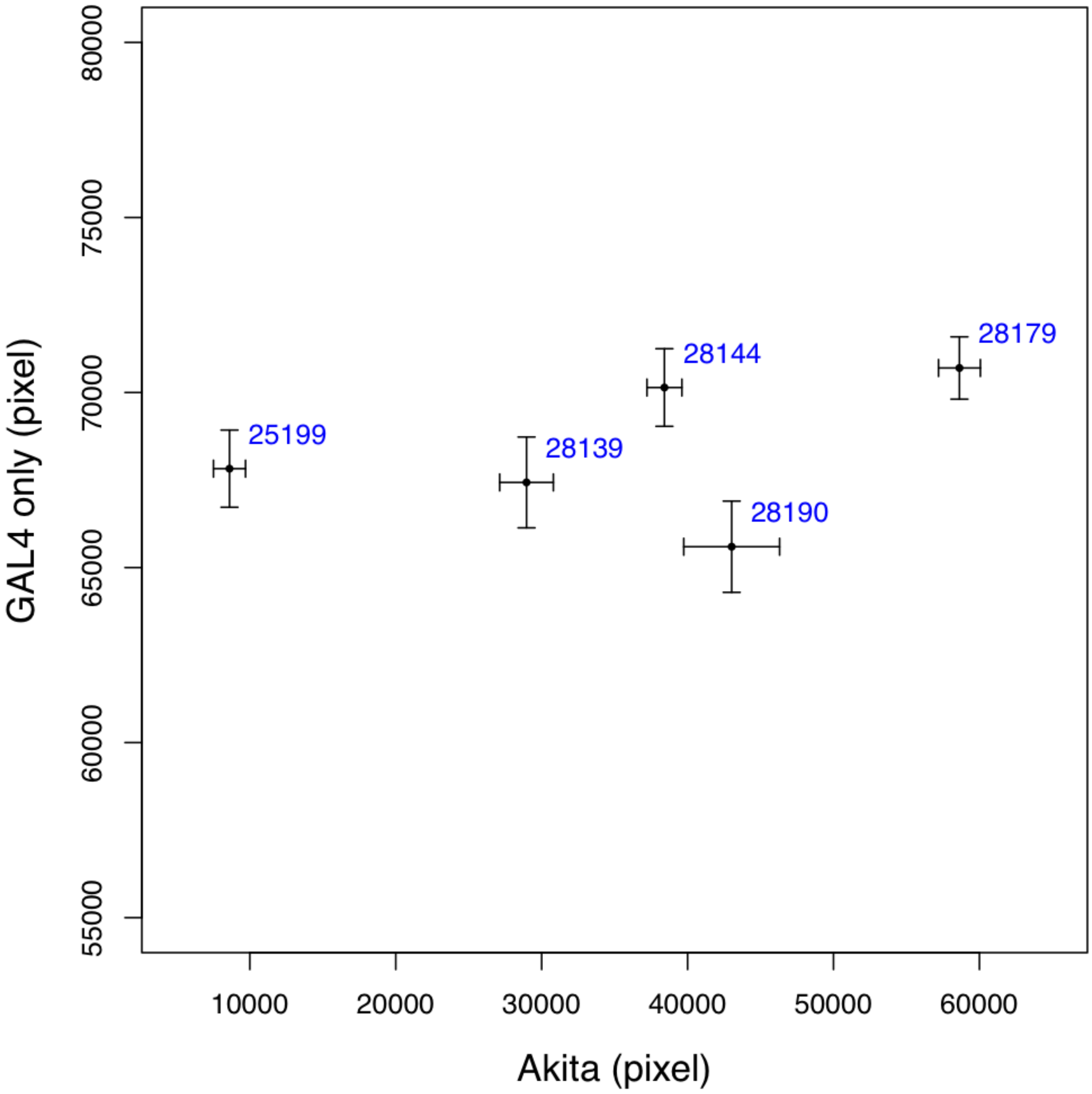


Figure S.10: **Observed variation in eye area in crosses to hINS^{C96Y}not correlated with natural variation among the inbred lines.** Five inbred lines were sampled across the phenotypic distribution of the crosses with hINS^{C96Y}, including the two extremes. They were crossed to a control line (w;GMR-GAL4), whose male progeny were measured for their eye area. No correlation is observed between results from the hINS^{C96Y}cross and the GMR-GAL4 cross.

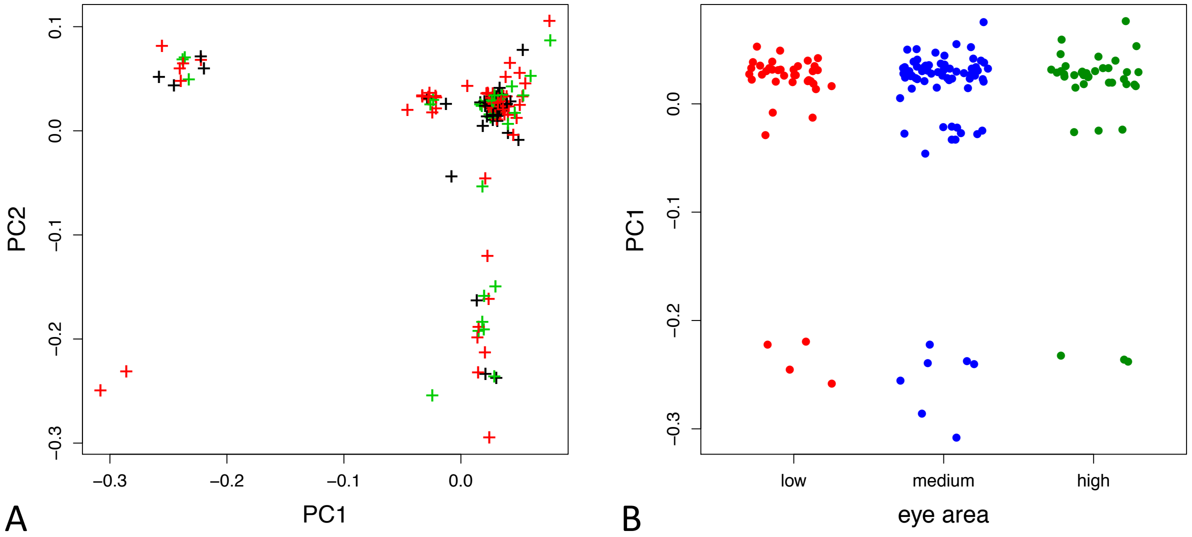


Figure S.11: **Population structure assessed through principal component analysis (PCA) using 900K autosomal SNPs after LD pruning.** (A) 154 DGRP inbred lines projected onto the plane spanned by the first two principal components (PC1, PC2). The points are colored according to the phenotype severity in the hINS^{C96Y} crosses (red: severe, or first 25%; blue: intermediate, 25%-75%; green: mild, 75%-100%, percentiles in eye area distribution from small to large). (B) projection onto PC1 grouped by their phenotype severity showed no correlation between the two.

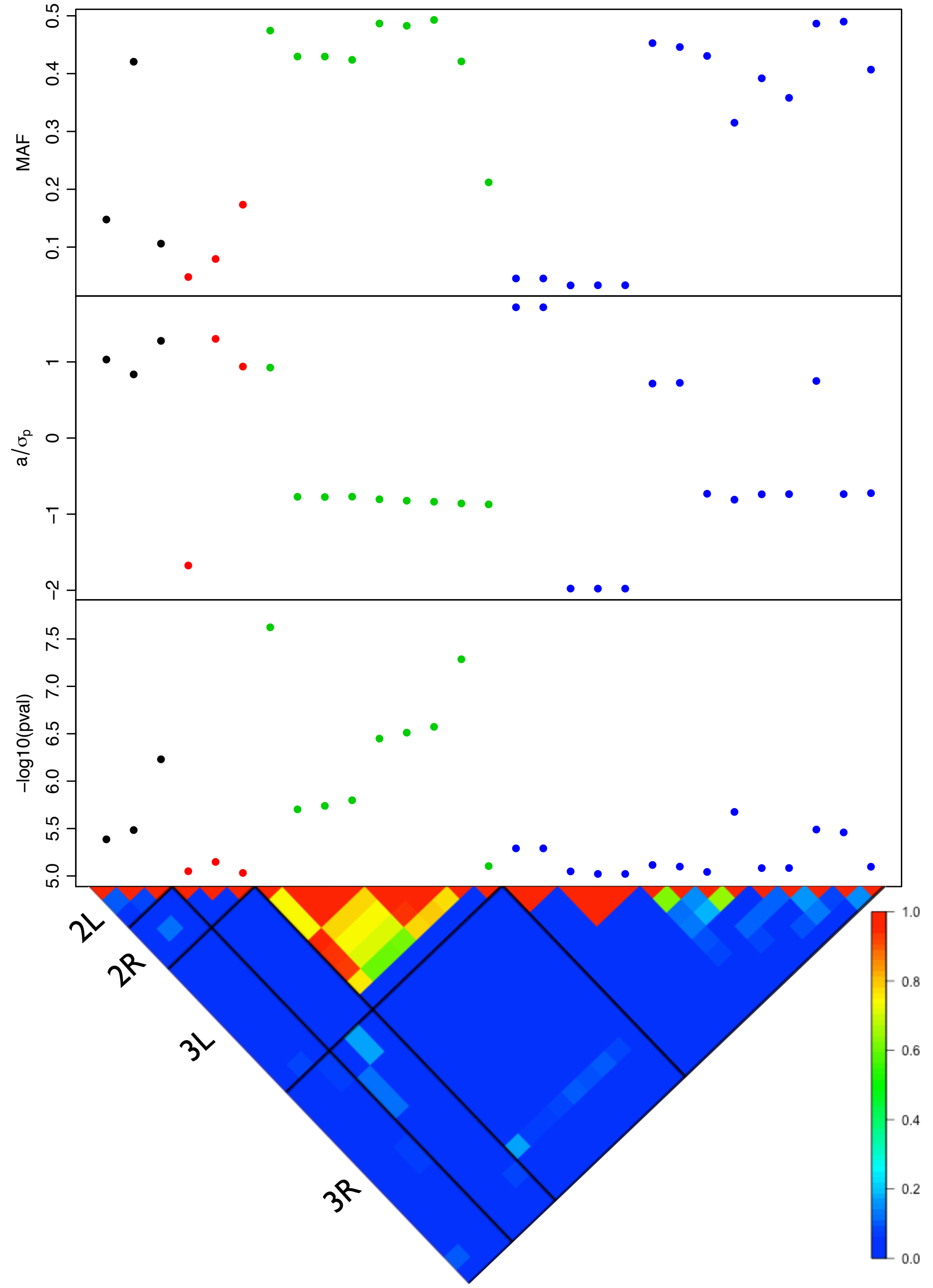


Figure S.12: **All SNPs with p-values below** 10^{-5} **from GWAS.** From top to bottom are minor allele frequencies, effect sizes (d/σ_p), -log10 (p-values) and the bottom triangle represent the linkage map between these SNPs split by chromosomes.

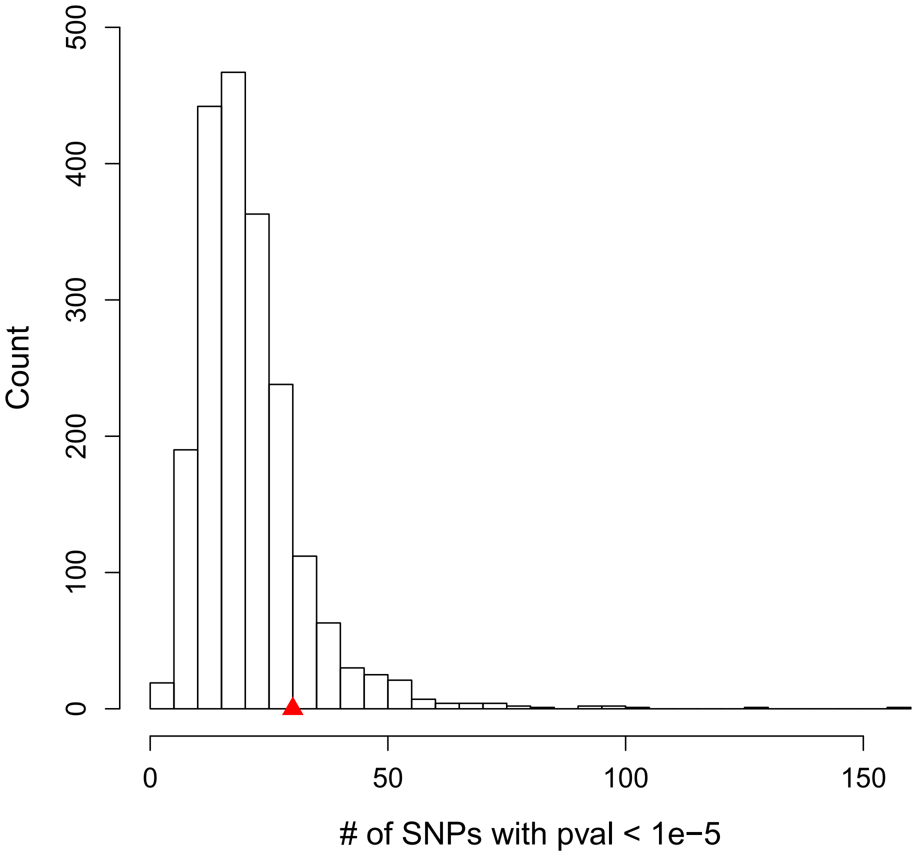


Figure S.13: **Permutation test to assess the FDR of the $p < 10^{-5}$ threshold for GWAS.** Shown is the histogram of the number of SNPs with p-values $< 10^{-5}$ in 2,000 permutation tests. The red triangle indicates the observed number in the real data.

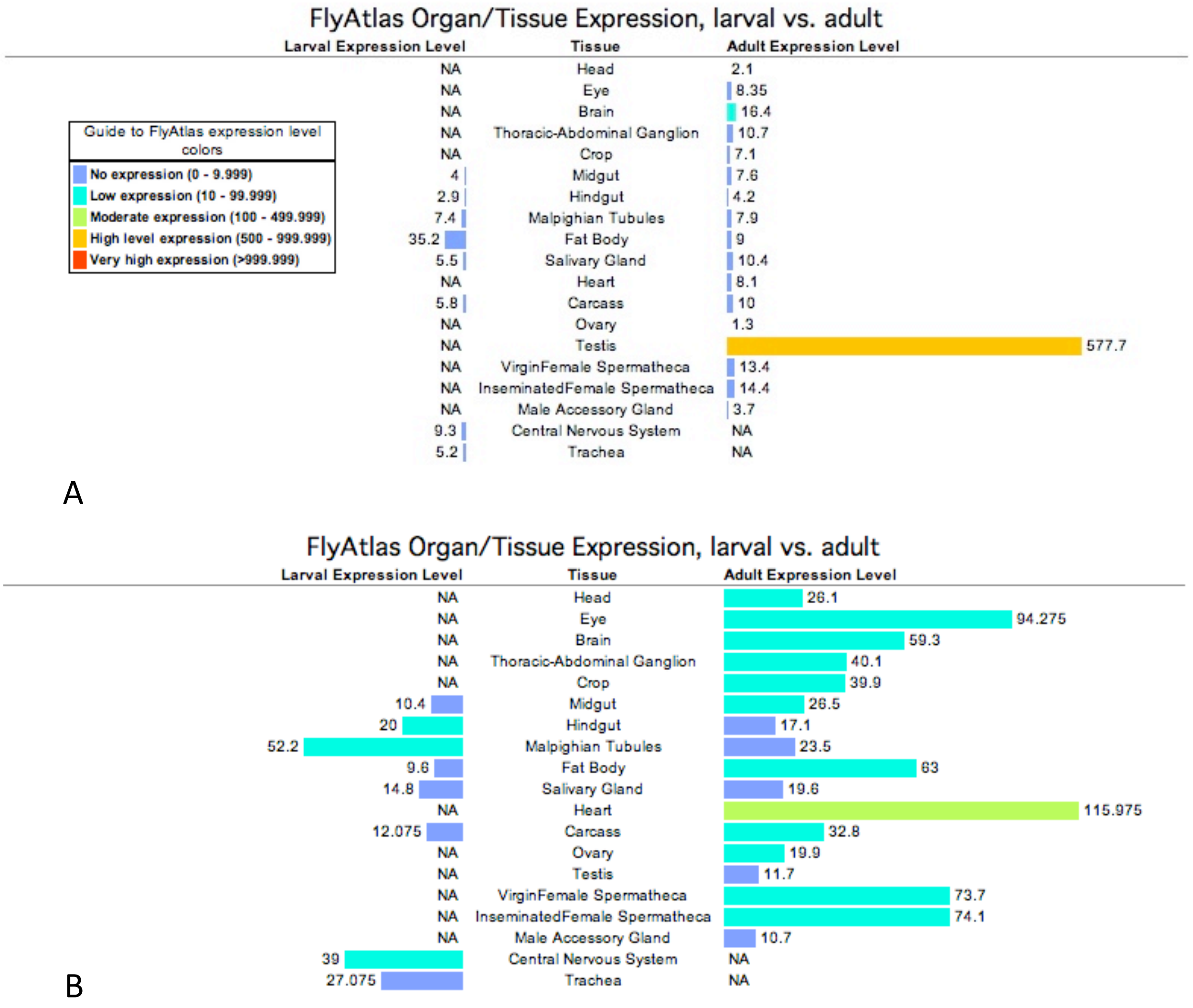


Figure S.14: **FlyAtlas expression report for CG32396 and *sfl*.** (A) CG32396 (B) *sfl*

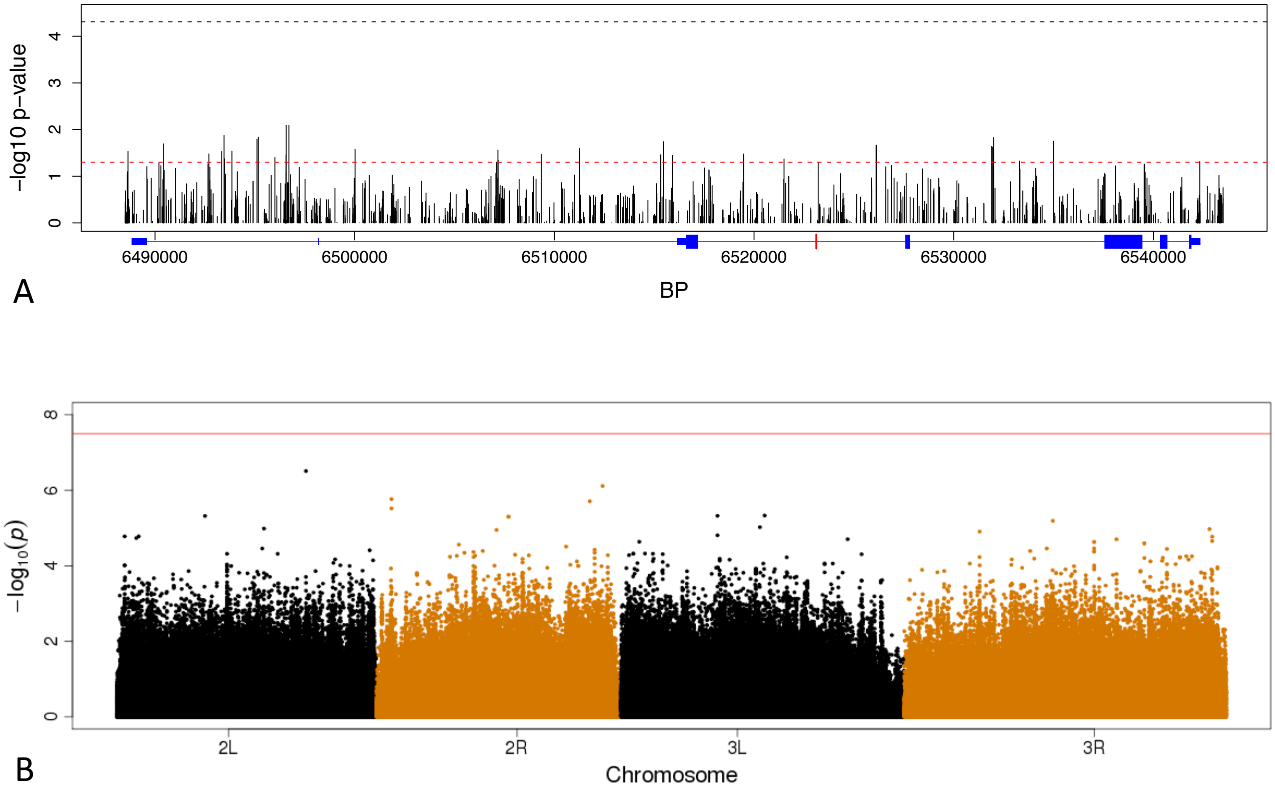


Figure S.15: **Conditional analysis shows no additional SNPs associated with the phenotype of interest.** (A) within the *sfl* locus; (B) all chromosomes. The intronic 18/4bp polymorphism is included in the linear model as a covariate in both cases.The two dotted lines in (A) correspond to a single test 0.05 level (red) or the multiple testing corrected 0.05 level using Bonferroni's method (blue). The red line in (B) represents the Bonferroni corrected 0.05 level.

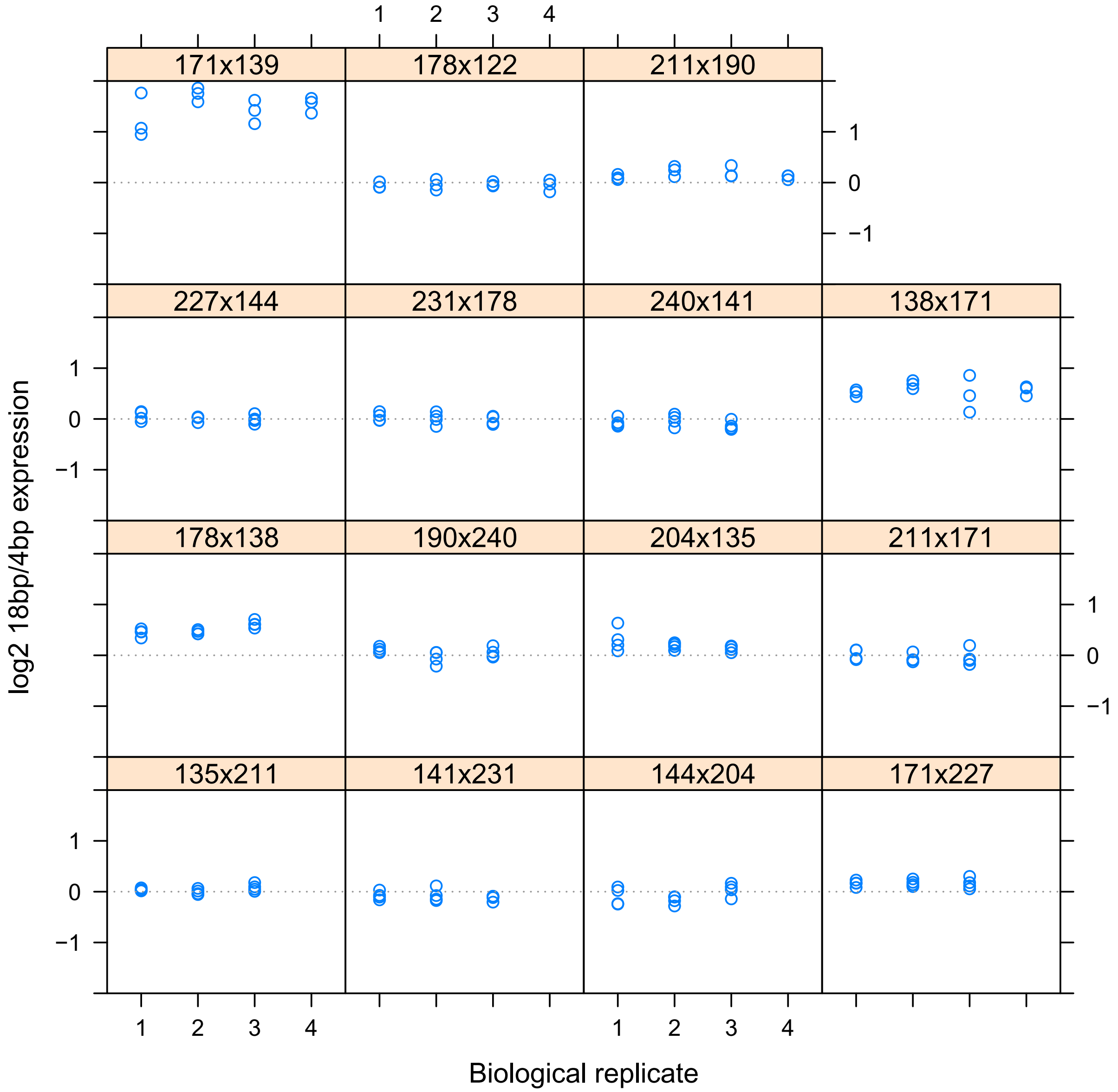
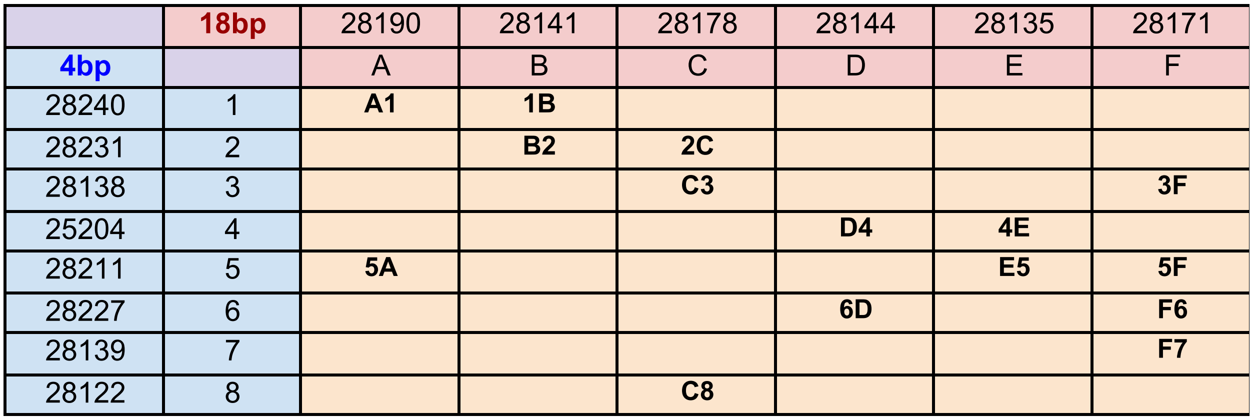


Figure S.16: **log2 transformed ratios between transcript levels associated with 18bp/4bp alleles.** The allele-specific expression ratios were measured in F1 hybrid individuals by pyro-sequencing, either with three biological replicates and four pyro-technical replicates, or four and three, to obtain a total of 12 measurements. In each of the 15 crosses, the technical replicates were plotted in a single column, with different columns representing the biological replicates.



	18bp	28190	28141	28178	28144	28135	28171
4bp		A	B	C	D	E	F
28240	1	**A1**	**1B**				
28231	2		**B2**	**2C**			
28138	3			**C3**			**3F**
25204	4				**D4**	**4E**	
28211	5	**5A**				**E5**	**5F**
28227	6				**6D**		**F6**
28139	7						**F7**
28122	8			**C8**			

Figure S.17: **Cross design for pyro-sequencing** Six 18bp and eight 4bp lines were randomly chosen from the 154 DGRP lines used in GWAS. The bloomington center stock# is listed. In each cell, the order of the letter/number indicate the direction of the cross. For example, A1 indicates that males of 28240 was crossed to virgin females of 28190.

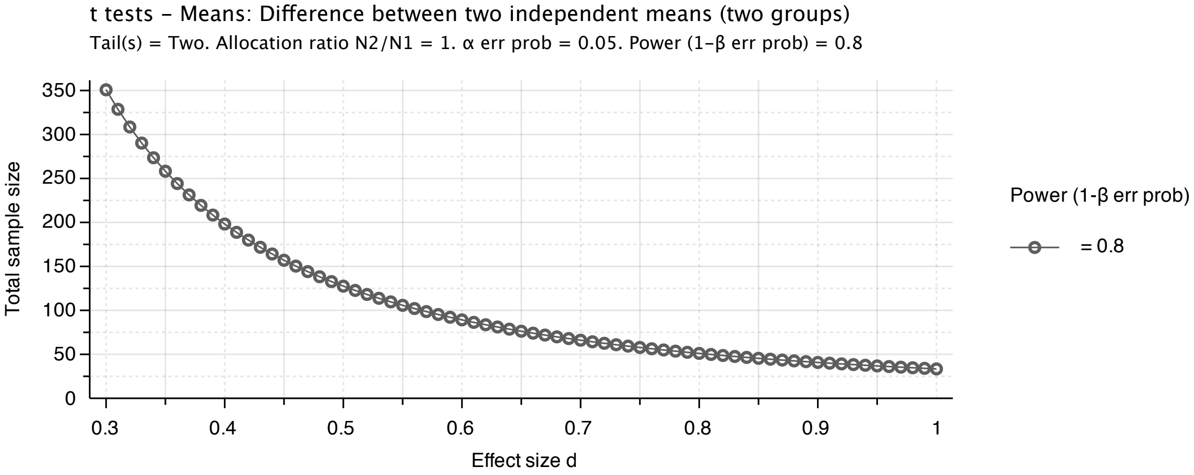


Figure S.18: **Sample size required for replication.** The sample size required to replicate an association with 80% power given an effect size (in units of d/σ) is plotted on the y-axis, calculated by G*Power3 (Faul *et al.*, 2007)

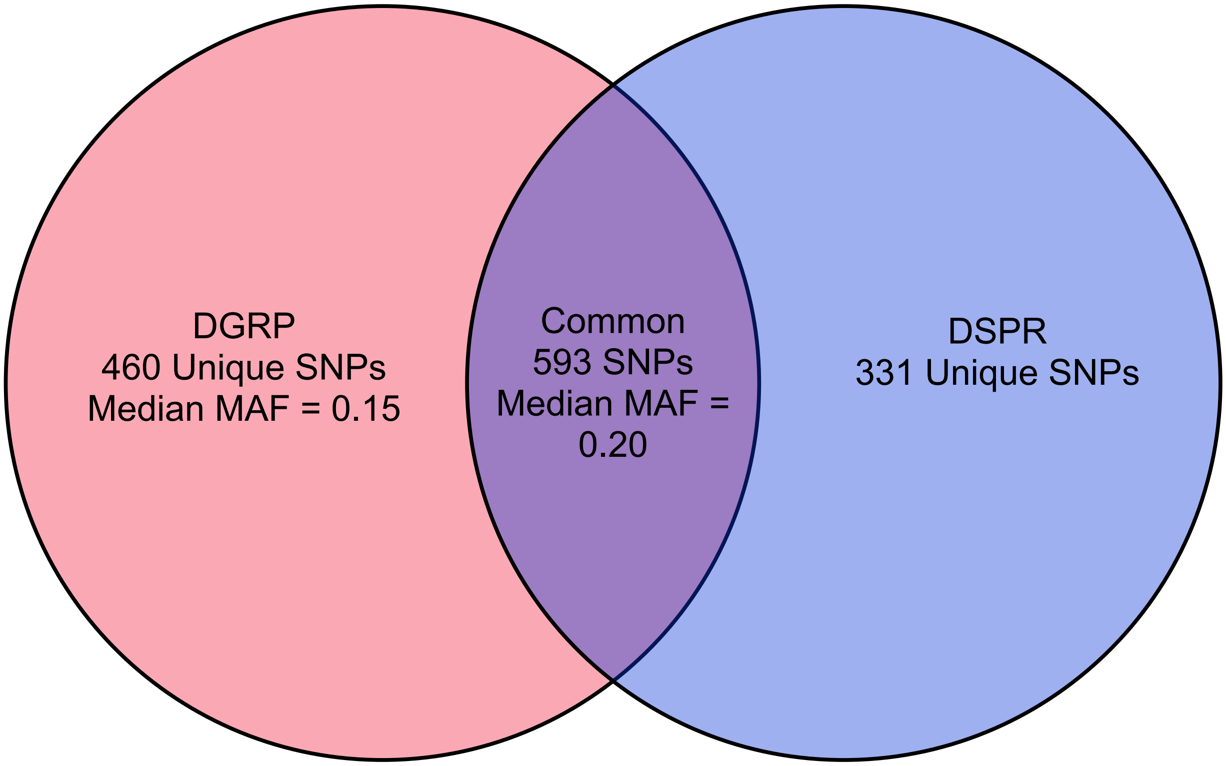


Figure S.19: **Venn diagram for overlap of SNPs in the *sfl* locus between DSPR and DGRP.** In the 55kb *sfl* locus, there are a total of 924 SNPs in DSPR and 3018 SNPs in DGRP (1053 have MAF$>$ 0.05). Focusing on the DGRP SNPs with MAF$>$ 0.05, the pie chart shows the amount of overlap between the two.

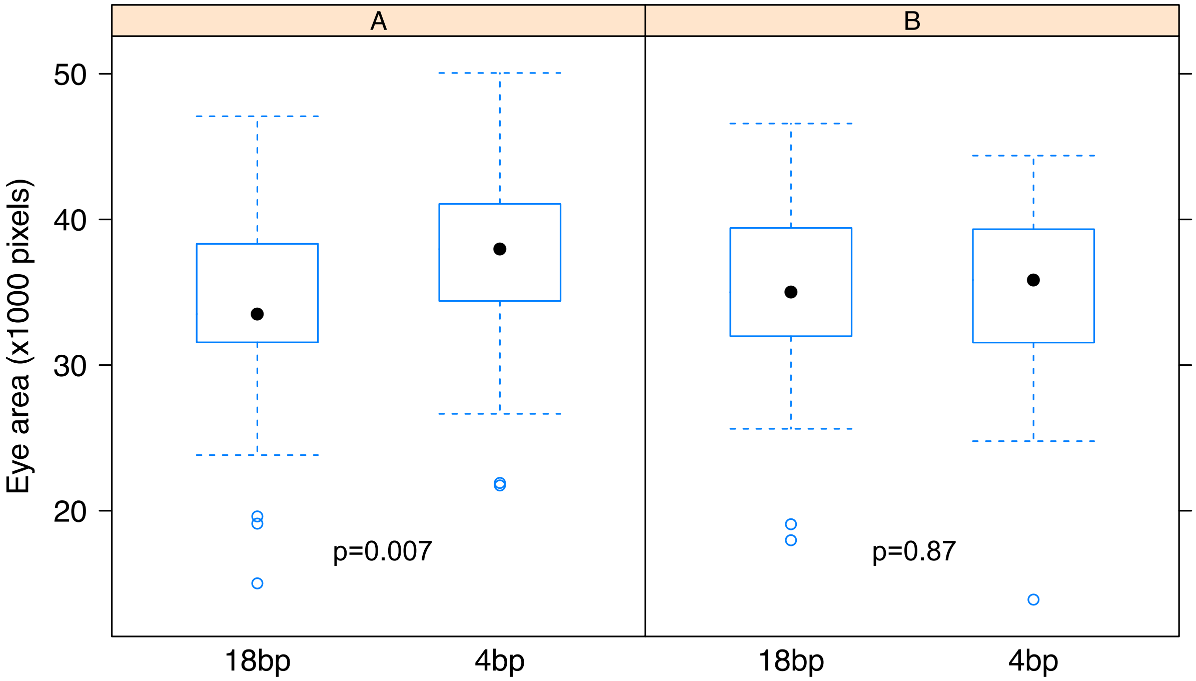


Figure S.20: **The reverse direction of *sfl* intronic variation effect is driven by just one of the two synthetic DSPR populations.** The difference between the 18bp vs. 4bp alleles were plotted and tested separately for population A and population B in DSPR.

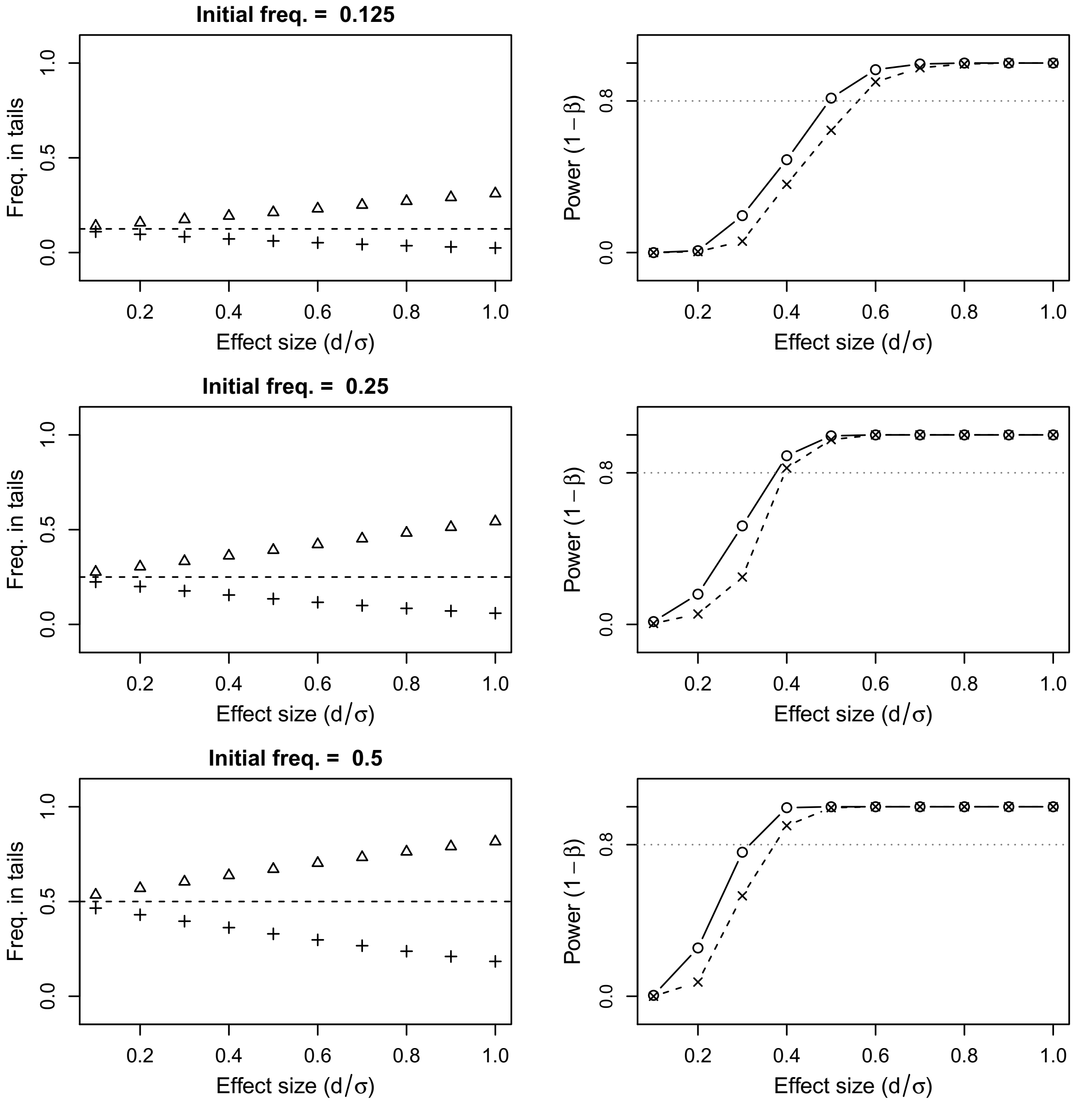


Figure S.21: **Power of extreme mapping when 20% tail is selected instead of 5%.** Same simulation procedure as in 5.1, except that 20% tail instead of 5% was selected.

REFERENCES

Andolfatto, P., 2005 Adaptive evolution of non-coding DNA in Drosophila. Nature *437*(7062): 1149–1152.

Andolfatto, P., 2008 Controlling type-I error of the McDonald-Kreitman test in genomewide scans for selection on noncoding DNA. Genetics *180*(3): 1767–1771.

Arnosti, D. N., S. Barolo, M. Levine, and S. Small, 1996 The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development (Cambridge, England) *122*(1): 205–214.

Bachtrog, D., 2008 Positive Selection at the Binding Sites of the Male-Specific Lethal Complex Involved in Dosage Compensation in Drosophila. Genetics *180*(2): 1123–1129.

Badano, J. L. and N. Katsanis, 2002 Beyond Mendel: an evolving view of human genetic disease transmission. Nat Rev Genet *3*(10): 779–789.

Bailey, T. L., N. Williams, C. Misleh, and W. W. Li, 2006 MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research *34*(suppl 2): W369–W373.

Balhoff, J. P. and G. A. Wray, 2005 Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. Proc Natl Acad Sci U S A *102*(24): 8591–8596.

Bellaiche, Y., I. The, and N. Perrimon, 1998 Tout-velu is a Drosophila homologue of the putative tumour suppressor EXT-1 and is needed for Hh diffusion. Nature *394*(6688): 85–88.

Berg, O. G. and P. H. von Hippel, 1987 Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. Journal of molecular biology *193*(4): 723–750.

Bergland, A. O., A. Genissel, S. V. Nuzhdin, and M. Tatar, 2008 Quantitative trait loci affecting phenotypic plasticity and the allometric relationship of ovariole number and thorax length in Drosophila melanogaster. Genetics *180*(1): 567–582.

Bradley, R. K., X.-Y. Y. Li, C. Trapnell, S. Davidson, L. Pachter et al., 2010 Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS biology *8*(3): e1000343+.

Britten, R. J. and E. H. Davidson, 1971 Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. The Quarterly review of biology *46*(2): 111–138.

Bucciantini, M., E. Giannoni, F. Chiti, F. Baroni, L. Formigli et al., 2002 Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature *416*(6880): 507–511.

Burton, P. R., D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas et al., 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*(7145): 661–678.

Carlborg, O. and C. S. Haley, 2004 Epistasis: too often neglected in complex trait studies? Nat Rev Genet *5*(8): 618–625.

Centers for Disease Control, 2011 National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. Technical report, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA.

Chan, Y. F. F., M. E. Marks, F. C. Jones, G. Villarreal, M. D. Shapiro et al., 2010 Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science (New York, N.Y.) *327*(5963): 302–305.

Charlesworth, B. and D. Charlesworth, 2010 *Elements of Evolutionary Genetics* (1 ed.). Roberts & Company Publishers.

Charlesworth, J. and A. Eyre-Walker, 2008 The McDonald-Kreitman Test and Slightly Deleterious Mutations. Mol Biol Evol *25*(6): 1007–1015.

Chintapalli, V. R., J. Wang, and J. A. T. Dow, 2007 Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet *39*(6): 715–720.

Cho, Y. S., C.-H. Chen, C. Hu, J. Long, R. T. Hee Ong et al., 2012 Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat Genet *44*(1): 67–72.

Collins, A. L., Y. Kim, P. Sklar, International Schizophrenia Consortium, M. C. O'Donovan et al., 2012 Hypothesis-driven candidate genes for schizophrenia compared to genome-wide association results. Psychological medicine *42*(3): 607–616.

Crocker, J., N. Potter, and A. Erives, 2010 Dynamic evolution of precise regulatory encodings creates the clustered site signature of enhancers. Nature Communications *1*(7): 99+.

Crocker, J., Y. Tamori, and A. Erives, 2008 Evolution Acts on Enhancer Organization to Fine-Tune Gradient Threshold Readouts. PLoS Biol *6*(11): e263+.

Dermitzakis, E. T. and A. G. Clark, 2002 Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover. Molecular Biology and Evolution *19*(7): 1114–1121.

Doniger, S. W. and J. C. Fay, 2007 Frequent Gain and Loss of Functional Transcription Factor Binding Sites. PLoS Comput Biol *3*(5): e99+.

Down, T. A., C. M. Bergman, J. Su, and T. J. P. Hubbard, 2007 Large-Scale Discovery of Promoter Motifs in Drosophila melanogaster. PLoS Comput Biol *3*(1): e7+.

Durrett, R. and D. Schmidt, 2008 Waiting for Two Mutations: With Applications to Regulatory Sequence Evolution and the Limits of Darwinian Evolution. Genetics *180*(3): 1501–1509.

Dworkin, I., E. Kennerly, D. Tack, J. Hutchinson, J. Brown et al., 2009 Genomic consequences of background effects on scalloped mutant expressivity in the wing of Drosophila melanogaster. Genetics *181*(3): 1065–1076.

Ehrenreich, I. M., N. Torabi, Y. Jia, J. Kent, S. Martis et al., 2010 Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature *464*(7291): 1039–1042.

Ewens, W. J., 2004 *Mathematical Population Genetics* (2nd ed.). Springer.

Falconer, D. S., 1981 *Introduction to Quantitative Genetics* (2nd ed.). Longman.

Faul, F., E. Erdfelder, A.-G. G. Lang, and A. Buchner, 2007 G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods *39*(2): 175–191.

Fay, J. C., G. J. Wyckoff, and C.-I. Wu, 2001 Positive and Negative Selection on the Human Genome. Genetics *158*(3): 1227–1234.

Frankel, N., D. F. Erezyilmaz, A. P. McGregor, S. Wang, F. Payre et al., 2011 Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. Nature *474*(7353): 598–603.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy et al., 2002 The Structure of Haplotype Blocks in the Human Genome. Science *296*(5576): 2225–2229.

Gallo, S. M., D. T. Gerrard, D. Miner, M. Simich, B. Des Soye et al., 2010 REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. Nucleic Acids Research *39*(suppl 1): D118–D123.

Garner, C., 2007 Upward bias in odds ratio estimates from genome-wide association studies. Genet. Epidemiol. *31*(4): 288–295.

Gibson, G., 2009 Decanalization and the origin of complex disease. Nat Rev Genet *10*(2): 134–140.

Gompel, N., B. Prud'homme, P. J. Wittkopp, V. A. Kassner, and S. B. Carroll, 2005 Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature *433*(7025): 481–487.

Graveley, B. R., A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin et al., 2011 The developmental transcriptome of Drosophila melanogaster. Nature *471*(7339): 473–479.

Gregor, T., A. P. McGregor, and E. F. Wieschaus, 2008 Shape and function of the Bicoid morphogen gradient in dipteran species with different sized embryos. Developmental Biology *316*(2): 350–358.

Guss, K. A., C. E. Nelson, A. Hudson, M. E. Kraus, and S. B. Carroll, 2001 Control of a Genetic Regulatory Network by a Selector Gene. Science *292*(5519): 1164–1167.

Häcker, U., K. Nybakken, and N. Perrimon, 2005 Heparan sulphate proteoglycans: the sweet side of development. Nature reviews. Molecular cell biology *6*(7): 530–541.

Haddrill, P. R., D. Bachtrog, and P. Andolfatto, 2008 Positive and negative selection on noncoding DNA in Drosophila simulans. Molecular biology and evolution *25*(9): 1825–1834.

Halder, G., P. Callaerts, and W. J. Gehring, 1995 Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. Science (New York, N.Y.) *267*(5205): 1788–1792.

Hare, E. E., B. K. Peterson, and M. B. Eisen, 2008 A Careful Look at Binding Site Reorganization in the even-skipped Enhancers of Drosophila and Sepsids. PLoS Genet *4*(11): e1000268+.

Hare, E. E., B. K. Peterson, V. N. Iyer, R. Meier, and M. B. Eisen, 2008 Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS genetics *4*(6): e1000106+.

Ho, M. C., H. Johnsen, S. E. Goetz, B. J. Schiller, E. Bae et al., 2009 Functional evolution of cis-regulatory modules at a homeotic gene in Drosophila. PLoS genetics *5*(11): e1000709+.

Ihmels, J., S. Bergmann, M. Gerami-Nejad, I. Yanai, M. McClellan et al., 2005 Rewiring of the Yeast Transcriptional Network Through the Evolution of Motif Usage. Science *309*(5736): 938–940.

Jeong, S., A. Rokas, and S. B. Carroll, 2006 Regulation of Body Pigmentation by the Abdominal-B Hox Protein and Its Gain and Loss in Drosophila Evolution. Cell *125*(7): 1387–1399.

Jones, F. C., M. G. Grabherr, Y. F. F. Chan, P. Russell, E. Mauceli et al., 2012 The genomic basis of adaptive evolution in threespine sticklebacks. Nature *484*(7392): 55–61.

Kaiser Family Foundation, 2007 Trends in Health Care Costs and Spending. Technical report, The Henry J. Kaiser Family Foundation.

Kim, J., X. He, and S. Sinha, 2009 Evolution of Regulatory Sequences in 12 Drosophila Species. PLoS Genet *5*(1): e1000330+.

Kim, J.-H. H., Y. Zhao, X. Pan, X. He, and H. F. Gilbert, 2009 The unfolded protein response is necessary but not sufficient to compensate for defects in disulfide isomerization. The Journal of biological chemistry *284*(16): 10400–10408.

Kimmig, P., M. Diaz, J. Zheng, C. C. Williams, A. Lang et al., 2012 The unfolded protein response in fission yeast modulates stability of select mRNAs to maintain protein homeostasis. eLife **1**.

Kimura, M., 1985 The role of compensatory neutral mutations in molecular evolution. Journal of Genetics *64*(1): 7–19.

King, E. G., C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Sen et al., 2012 Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome research *22*(8): 1558–1566.

King, M. C. and A. C. Wilson, 1975 Evolution at two levels in humans and chimpanzees. Science (New York, N.Y.) *188*(4184): 107–116.

Kirkpatrick, C. A. and S. B. Selleck, 2007 Heparan sulfate proteoglycans at a glance. Journal of Cell Science *120*(11): 1829–1832.

Kohn, M. H., S. Fang, and C.-I. Wu, 2004 Inference of Positive and Negative Selection on the 5â Regulatory Regions of Drosophila Genes. Molecular Biology and Evolution *21*(2): 374–383.

Kulkarni, M., S. Ozgur, and G. Stoecklin, 2010 On track with P-bodies. Biochemical Society transactions *38*(Pt 1): 242–251.

Kuo, D., K. Licon, S. Bandyopadhyay, R. Chuang, C. Luo et al., 2010 Coevolution within a transcriptional network by compensatory trans and cis mutations. Genome Research *20*(12): 1672–1678.

Lander, E. S. and D. Botstein, 1989 Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. Genetics *121*(1): 185–199.

Lin, X., 2004 Functions of heparan sulfate proteoglycans in cell signaling during development. Development *131*(24): 6009–6021.

Lott, S. E., M. Kreitman, A. Palsson, E. Alekseeva, and M. Z. Ludwig, 2007 Canalization of segmentation and its evolution in Drosophila. Proceedings of the National Academy of Sciences *104*(26): 10926–10931.

Lu, J., Y. Shen, Q. Wu, S. Kumar, B. He et al., 2008 The birth and death of microRNA genes in Drosophila. Nat Genet *40*(3): 351–355.

Ludwig, M. Z., C. Bergman, N. H. Patel, and M. Kreitman, 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature *403*(6769): 564–567.

Ludwig, M. Z. and M. Kreitman, 1995 Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. Molecular biology and evolution *12*(6): 1002–1011.

Ludwig, M. Z., A. Palsson, E. Alekseeva, C. M. Bergman, J. Nathan et al., 2005 Functional Evolution of a cis-Regulatory Module. PLoS Biol *3*(4): e93+.

Ludwig, M. Z., N. H. Patel, and M. Kreitman, 1998 Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development (Cambridge, England) *125*(5): 949–958.

Lusk, R. W. and M. B. Eisen, 2010 Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. PLoS genetics *6*(1): e1000829+.

Macdonald, S. J. and A. D. Long, 2005 Identifying signatures of selection at the enhancer of split neurogenic gene complex in Drosophila. Molecular biology and evolution *22*(3): 607–619.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles et al., 2012 The Drosophila melanogaster Genetic Reference Panel. Nature *482*(7384): 173–178.

Mackay, T. F. C., E. A. Stone, and J. F. Ayroles, 2009 The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics *10*(8): 565–577.

Maerkl, S. J. and S. R. Quake, 2007 A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. Science *315*(5809): 233–237.

Makałowski, W., J. Zhang, and M. S. Boguski, 1996 Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. Genome Research *6*(9): 846–857.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff et al., 2009 Finding the missing heritability of complex diseases. Nature *461*(7265): 747–753.

McDonald, J. H. and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*(6328): 652–654.

McGregor, A. P., V. Orgogozo, I. Delon, J. Zanet, D. G. Srinivasan et al., 2007 Morphological evolution through multiple cis-regulatory mutations at a single gene. Nature *448*(7153): 587–590.

McGregor, A. P., P. J. Shaw, J. M. Hancock, D. Bopp, M. Hediger et al., 2001 Rapid restructuring of bicoid-dependent hunchback promoters within and between Dipteran species: implications for molecular coevolution. Evol Dev *3*(6): 397–407.

Mefford, H. C., A. J. Sharp, C. Baker, A. Itsara, Z. Jiang et al., 2008 Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. The New England journal of medicine *359*(16): 1685–1699.

Miles, C. M., S. E. Lott, C. L. L. Hendriks, M. Z. Ludwig, Manu et al., 2011 Artificial selection on egg size perturbs early pattern formation in Drosophila melanogaster. Evolution; international journal of organic evolution *65*(1): 33–42.

Morris, A. P., B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segrè et al., 2012 Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics *44*(9): 981–990.

Moses, A. M., 2009 Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. BMC evolutionary biology *9*(1): 286+.

Moses, A. M., D. A. Pollard, D. A. Nix, V. N. Iyer, X.-Y. Li et al., 2006 Large-Scale Turnover of Functional Transcription Factor Binding Sites in Drosophila. PLoS Computational Biology *2*(10): e130+.

Mouse Genome Sequencing Consortium, R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers et al., 2002 Initial sequencing and comparative analysis of the mouse genome. Nature *420*(6915): 520–562.

Mustonen, V., J. Kinney, C. G. Callan, and M. Lässig, 2008 Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. Proceedings of the National Academy of Sciences *105*(34): 12376–12381.

Mustonen, V. and M. Lässig, 2005 Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. Proceedings of the National Academy of Sciences of the United States of America *102*(44): 15936–15941.

Mustonen, V. and M. Lässig, 2007 Adaptations to fluctuating selection in Drosophila. Proceedings of the National Academy of Sciences of the United States of America *104*(7): 2277–2282.

Park, S.-Y., M. Z. Ludwig, N. A. Tamarina, C. L. Williams, S. Carl et al., 2012 A Drosophila Model for Misfolded Protein Induced Neurodegeneration. under review.

Prud'homme, B., N. Gompel, A. Rokas, V. A. Kassner, T. M. Williams et al., 2006 Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature *440*(7087): 1050–1053.

Ren, Y., C. A. Kirkpatrick, J. M. Rawson, M. Sun, and S. B. Selleck, 2009 Cell type-specific requirements for heparan sulfate biosynthesis at the Drosophila neuromuscular junction: effects on synapse function, membrane trafficking, and mitochondrial localization. The Journal of neuroscience : the official journal of the Society for Neuroscience *29*(26): 8539–8550.

Rhodes, C. J., 2005 Type 2 Diabetes-a Matter of ß-Cell Life and Death? Science *307*(5708): 380–384.

Roy, S., J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre et al., 2010 Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science *330*(6012): 1787–1797.

Sanders, A. R., J. Duan, D. F. Levinson, J. Shi, D. He et al., 2008 No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics. The American journal of psychiatry *165*(4): 497–506.

Sawyer, S. A. and D. L. Hartl, 1992 Population Genetics of Polymorphism and Divergence. Genetics *132*(4): 1161–1176.

Schmidt, D., M. D. Wilson, B. Ballester, P. C. Schwalie, G. D. Brown et al., 2010 Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science (New York, N.Y.) *328*(5981): 1036–1040.

Segrè, A. V., A. W. Murray, and J.-Y. Y. Leu, 2006 High-resolution mutation mapping reveals parallel experimental evolution in yeast. PLoS biology *4*(8): e256+.

Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto, 2009 Pervasive Natural Selection in the Drosophila Genome? PLoS Genet *5*(6): e1000495+.

Shapiro, M. D., M. E. Marks, C. L. Peichel, B. K. Blackman, K. S. Nereng et al., 2004 Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature *428*(6984): 717–723.

Shaw, P. J., N. S. Wratten, A. P. McGregor, and G. A. Dover, 2002 Coevolution in bicoid-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera. Evolution & development *4*(4): 265–277.

Shimell, M. J., A. J. Peterson, J. Burr, J. A. Simon, and M. B. O'Connor, 2000 Functional analysis of repressor binding sites in the iab-2 regulatory region of the abdominal-A homeotic gene. Developmental biology *218*(1): 38–52.

Small, S., A. Blair, and M. Levine, 1992 Regulation of even-skipped stripe 2 in the Drosophila embryo. The EMBO journal *11*(11): 4047–4057.

Smith, N. G. C. and A. Eyre-Walker, 2002 Adaptive protein evolution in Drosophila. Nature *415*(6875): 1022–1024.

Stanford, K. I., J. R. Bishop, E. M. Foley, J. C. Gonzales, I. R. Niesman et al., 2009 Syndecan-1 is the primary heparan sulfate proteoglycan mediating hepatic clearance of triglyceride-rich lipoproteins in mice. The Journal of clinical investigation *119*(11): 3236–3245.

Stanojevic, D., S. Small, and M. Levine, 1991 Regulation of a Segmentation Stripe by Overlapping Activators and Repressors in the Drosophila Embryo. Science *254*(5036): 1385–1387.

Støy, J., E. L. Edghill, S. E. Flanagan, H. Ye, V. P. Paz et al., 2007 Insulin gene mutations as a cause of permanent neonatal diabetes. Proceedings of the National Academy of Sciences of the United States of America *104*(38): 15040–15044.

Sun, L., A. Dimitromanolakis, L. L. Faye, A. D. Paterson, D. Waggott et al., 2011 BR-squared: a practical solution to the winner's curse in genome-wide scans. Human genetics *129*(5): 545–552.

Swanson, C. I., N. C. Evans, and S. Barolo, 2010 Structural Rules and Complex Regulatory Circuitry Constrain Expression of a Notch- and EGFR-Regulated Eye Enhancer. Dev Cell *18*(3): 359–370.

Torgerson, D. G., A. R. Boyko, R. D. Hernandez, A. Indap, X. Hu et al., 2009 Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS genetics *5*(8): e1000592+.

Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang, 2012 Five years of GWAS discovery. American journal of human genetics *90*(1): 7–24.

Wang, J., T. Takeuchi, S. Tanaka, S. K. Kubo, T. Kayo et al., 1999 A mutation in the insulin 2 gene induces diabetes with severe pancreatic beta-cell dysfunction in the Mody mouse. The Journal of clinical investigation *103*(1): 27–37.

Wang, Y., D. Pot, S. D. Kachman, S. V. Nuzhdin, and L. G. Harshman, 2006 A quantitative trait locus analysis of natural genetic variation for Drosophila melanogaster oxidative stress survival. The Journal of heredity *97*(4): 355–366.

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

Wright, A. F., C. F. Chakarova, M. M. Abd El-Aziz, and S. S. Bhattacharya, 2010 Photoreceptor degeneration: genetic and mechanistic dissection of a complex trait. Nat Rev Genet *11*(4): 273–284.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders et al., 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics *42*(7): 565–569.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: a tool for genome-wide complex trait analysis. American journal of human genetics *88*(1): 76–82.

Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso et al., 2011 Genome partitioning of genetic variation for complex traits using common SNPs. Nature genetics *43*(6): 519–525.

Yang, Z., 2007 PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution *24*(8): 1586–1591.

Yu, J. H. H., W.-H. H. Yang, T. Gulick, K. D. Bloch, and D. B. Bloch, 2005 Ge-1 is a central component of the mammalian cytoplasmic mRNA processing body. RNA (New York, N.Y.) *11*(12): 1795–1802.

Zhou, X. and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nat Genet *44*(7): 821–824.