Genetic Interactions and Gene-by-Environment Interactions in Evolution

by

Xinzhu Wei

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in The University of Michigan
2018

Doctoral Committee:

Professor Jianzhi Zhang, Chair
Professor Alexey Kondrashov
Professor Trisha Wittkopp
Professor Sebastian Zoellner

Xinzhu Wei

ORCID iD: 0000-0001-8184-7016

xinzhuw@umich.edu

Dedication

To Felix, and in memory of Isaac

Acknowledgements

This work could not exist without numerous help from others. My advisor, George Zhang, is the most helpful person to my thesis work. He is rigorous, diligent, knowledgeable, and above all curious and open-minded. I appreciate the time and effort he spent on me, and his patience and tolerance. It is lucky to be a student of him. I also learnt a lot from my committee. I'd like to thank Alexey Kondrashov, who I also regard as a mentor, for all the fun discussions we had, for the interesting papers he shared with me, and for commenting on my manuscripts. I thank Trisha Wittkopp for her advice, help, and encouragements. Trisha is a real role model, not only for her high-quality science, but also for her art of being a scientist. I thank Sebastian Zoellner for adding his pop-gen and stats perspective. The discussion chapter of thesis would be much shorter without his requirement.

I'd also like to thank pals from my former labs and Zhang lab; it is a pleasure to have people who really focus on research around. I hold similar gratitude to some people in the Wittkopp lab and some other folks in my department. I want to thank my labmate Wei-Chin Ho for the discussions we had. I also want to thank Yifan Dai, for his trust, I have learnt a lot from collaborating with him. Dr. Andrea Hodgins-Davis unintentionally played an important role when she discussed a J Bloom et al. paper in our journal club. I cannot imagine how otherwise, the Bloom et al 2013 and Bloom et al 2015 data would be used in my study. I also thank many

other faculties and stuffs in my department, Deborah, Tom, Diarmaid, Cindy, Carol, Gail, John, and many others, who either discussed science with me, or helped with other problem.

My most sincere acknowledgements also go to all my advisors in undergraduate time, in particular, Felix Li Jin, David Waxman, and Martin Lascoux for their guidance, training, and help. It is unimaginably lucky to have met these amazing scientists at the initial stage of my career.

I thank my friends here and there for the joyful time we had hanging out together, and travelling. I thank Roman Gayduk for lots of discussions and supports. With his influence, I became more diligent during the last two years.

Lastly, I thank four of my loving grandparents, and my mom and dad for nature and nurture. This thesis work also owes special acknowledgements to my mom, for awakening my passion. One morning in my second year of graduate school, she asked me if everything was all right. After assuring her, she pointed out that I had not mentioned anything exciting that I had read nor had I mentioned any good idea I had conceived for more than a year. That moment I decided to stop drowning myself in the self-doubting and dropping out thoughts, and settled back to reading and thinking.

I thank everyone who enlightens my life!

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

The phenotypic effect of a mutation depends on both genetic interactions (G×G) and gene-by-environment interactions (G×E). G×G and G×E can distort the additive relationship between genotypes and phenotypes and complicate biological and biomedical studies. Understanding the patterns and mechanisms of these interactions is important for predicting evolutionary trajectories, designing plant and animal breeding strategies, detecting "missing heritability", and guiding "personalized medicine". In this thesis, I study how G×G and G×E affect mutational effects, including developing new methods and new models. Recent advancements in high-throughput DNA sequencing and high-throughput phenotyping provide powerful tools to study the relationships among genotypes, phenotypes, and the environment at unprecedented scales. Therefore, I take advantage of several published large datasets in my study, each containing hundreds to thousands of different genotypes of model organisms and their corresponding phenotypes in tens of environments. In Chapter 2, I report some general patterns of G×E and demonstrate the importance of considering potential environmental variations in mapping quantitative trait loci. In Chapter 3, I report how the environment affects diminishing returns epistasis and propose a modular life model to explain the patterns of diminishing returns. In Chapter 4, I propose and demonstrate that genetic dominance is a special case of diminishing returns epistasis. In Chapter 5, I report how and why the relationship between growth rate (r) and carrying capacity (K) in density-dependent population growth varies

across environments. In Chapter 6, I demonstrate the existence of an intermediate optimal mating distance for hybrid performance in three model organisms. Overall, I find that large genomic and phenomic data are useful resources to address classical genetic questions, such as the origin of dominance (Chapter 4), the relationship between *r* and *K* (Chapter 5), and presence of an optimal mating distance (Chapter 6). The environment is a key player in the phenotypic effects of mutations, but it is also a high-dimension complex system that is hard to quantify. In this thesis, I define environment quality (*Q*) as the average fitness of many different genotypes measured in the environment. I demonstrate that *Q* is useful in studying how the environment affects additive (Chapter 3), interactive (Chapters 3 and 4), and pleiotropic mutational effects (Chapter 5). Many classical theories and models were developed based on observations made in a single environment, and they are often insufficient to explain across-environment observations. Studying across-environment effects provides valuable information for testing old models and for designing new models when old models fail. I conclude that studying G×G and G×E shed light on underlying biological mechanisms.

Chapter 1

Introduction

"I have studied these things."

— **Isaac Newton**

Mutation and environment are the two fundamental components in evolution, which together determine phenotype. Recent advancements in high throughput DNA sequencing and high throughput phenotyping provide powerful tools to study the relationship between genotypes, phenotypes, and environments at unprecedented scales. I take advantage of several published large datasets of model organisms, each of which includes hundreds to thousands of different genotypes and their corresponding phenotypes, to study how genetic interactions and G×E interactions affect mutational effects and address some classical genetic questions with my new observations.

Fitness related phenotypes, natural polymorphisms, interactions, and environments are key components in my study. Therefore, I introduce the definition and measurements of fitness, the effect of mutations, the relationship between polymorphisms and mutations, genetic interactions, gene by environment interactions, and QTL mapping. I also introduce the

relationship between interactions and phenomena in biology, the relationship between interactions and evolution, as well as the challenges in detecting interactions, which help understand the questions my thesis projects try to tackle.

1.1 Fitness, one compound trait, many meanings

Fitness is, by all means, a fundamental property of all life forms. However, the meaning of fitness differs largely in the level to which it is applied, while the level could be individual, population, species, or a timescale (THODAY 1953). Within species, fitness is a quantitative representation of natural and sexual selection in evolutionary studies. Despite used and discussed all the time, it is still an ambiguous compound trait, which could be measured at different levels and by different approaches. Some canonical measurements preferred over the others depending on the situation and species, but none of them works for all situations.

In theoretical studies, fitness is measured either by an absolute value as measuring the genotype itself (i.e. absolute fitness, usually notated as W) or by a relative value in comparison to all the existing genotypes in a population (i.e. relative fitness, usually noted as w). While W measures the proportional change in the abundance of a genotype over a generation, w measures the change of the genotype frequency over a generation, measuring the reproductive quality of the genotype as in competition to the entire population. In theoretical work, the W is commonly normalized by the highest fitness genotype to get w (CROW AND KIMURA 1970). This is because that w is a more direct measurement of selection. It is also more relevant to competition and finite population.

Theoretical work involving fitness can be summarized into two main directions; one is to apply fitness in population genetic models to predict evolution, such as the rate of adaptation,

and the other is to predict fitness from genotypes and/or phenotypes. In population genetic models, *w* is proved useful, such as in Wright-Fisher model and Moran model (MORAN 1958; WEI *et al.* 2015). In a few cases, *W* can be more straightforward, such as when modeling with branching process (METZ *et al.* 1995). Theoretical population genetic modeling with a simplified trichotomy fitness distribution (lethal: $w = 0$, neutral: $w = 1$, beneficial: $w > 1$) was used in Appendix B of this thesis to study the relationship between robustness and evolvability.

The relationships between fitness and genotypes can be visualized by a fitness landscape, in which similar genotypes locate closer to each other and the height of the landscape represents the fitness value (WRIGHT 1932). The smoothness and the ruggedness of a landscape in an environment are associated with the robustness and evolvability of the genotypes (WEI AND ZHANG 2017b). This notion is used and discussed in Appendix B.

In empirical studies, fitness is measured or estimated with or without competition. With competition, the frequency change of a genotype, an allele, or an inheritable trait is associated with fitness. This frequency measurement is often used in experimental evolution, and the resulting fitness is *w* (MARÉE *et al.* 2000). In Chapter 3, I reanalyzed frequency based fitness measured in a lab environment (KRYAZHIMSKIY *et al.* 2014). Allele frequency change over seasons or over time has been documented in some species in the wild as an indicator of selection (BARRETT AND HOEKSTRA 2011; BERGLAND *et al.* 2014). However, associating selection with allele frequency change can be quite complicated, especially in a natural environment when replications are not available. Fictitious selection may occur due to genetic drift (ZHAO *et al.* 2013), frequency-dependent selection, epistasis, recombination, hitchhiking, and clonal interference. At genotype level and without competition, *W* is the more appropriate measurement. *W* is a combination of viability, mating success, fecundity, and so on, all of these

attributes can give some genotypes better ability to reproduce and to survive (ORR 2009). There are some canonical proxies of *W*, such as the reproductive rates of animals, seed numbers of plants, and maximum growth rates of microbes. Such proxies of *W* are used in my Chapter 2, 3, 5, and 6. Using *W* instead of *w* has the benefit of direct comparison of mutational effect across environments. Moreover, because selection and competition can complicate things, the transformation from proxies of *W* to proxies of *w* can introduce error.

The empirical study of fitness is becoming a fruitful field thanks to the advance of technology. However, it is still challenging due to issues with detection power and the obscurity relationship between *w* and different fitness proxies. Currently, detecting fitness by allele frequency change is constrained by the sample size, the number of replicates, duration, and frequency of sampling. Directly measured fitness proxies could be more complicated because one proxy of fitness cannot represent the entire compound trait. Some fitness proxies are correlated due to pleiotropic effect, but this may not always be true (Chapters 5 and 6). For instance, in Chapter 6, I discuss the pleiotropic effects of mutations on growth rate (*r*) and carrying capacity (*K*). In the past, evolutionary biologists view *r* as their fitness proxy because *r* is associated with the growth per generation and it is a character of a genotype, while ecologists prefer population character *K* as fitness proxy because the population sizes of many species in nature are often at or close to the saturation point (MACARTHUR AND WILSON 2016). Both *r* and *K* are important characters of density-dependent growth, on which selection could act. Because fitness is a quantitative representation of natural and sexual selection, using one of them instead of two to measure fitness may result in a biased result, especially when there is a tradeoff between these two fitness proxies.

1.2 The effect of mutation

Mutation is the permanent alteration of inheritable information, most often happens by alternating nucleotide sequences of a genome. It is also the ultimate source of evolution. Mutation could be large-scale, such as change of the ploidy level, change of the copy number of a chromosomal region (i.e. deletion, application, and loss of heterozygosity), or rearrangement of the chromosomes (i.e. translocation, and inversions). Mutation could be small-scale, such as short insertions and deletions (indels), and substitutions (TAJIMA 1989). Besides, there is a special class of mutations caused by transposable elements (LOEWE AND HILL 2010).

The segregating difference (i.e. polymorphisms) within species are usually small-scale mutations, which is what I primarily work on in this thesis, although different substitutions between species are also compared in Appendix A. Depending on where a mutation happens, the effect of the mutation on phenotype could be quite different. Mutations in coding regions pass down the information to mRNAs via transcription and post-transcription modification, and to proteins via translation (WATSON AND CRICK 1953; CRICK 1958). Mutations in noncoding regions can affect expression profile (KHALIL *et al.* 2009). The mutational effect could also be on many other phenotypes, such as chromatin, metabolites, cells, development, physiology, morphology, and behavior (HOULE *et al.* 2010). These phenotypic effects may or may not change fitness. Only germline mutation can stably pass down the information to the next generation (LIAW *et al.* 1997), although somatic mutations may also affect phenotype and fitness (GROUP 2010). The focus of this thesis is the on the effects of germline mutations.

The effect of mutation is also not necessarily on the mean of the phenotype (FORSBERG *et al.* 2015). For example, some mutations that do not affect the mean expression level can change the noise of expression (RASER AND O'SHEA 2004; KÆRN *et al.* 2005). Mutation could also affect the mean and variance of phenotypes of different genotypes, perhaps primarily through genetic

interaction (MCGUIGAN AND SGRO 2009; YADAV *et al.* 2016). Moreover, a mutation without phenotypic effect can still influence the mutational robustness and mutational evolvability of a genotype (WEI AND ZHANG 2017b). Due to genetic epistasis, a mutation may open and close the possibility for other mutation to have an effect (GOOD *et al.* 2017). The effect of a mutation is often context dependent, determined by both the genotype and the environment (LOEWE AND HILL 2010; WEI AND ZHANG 2017a), which I will introduce later.

The distribution of mutational effects on fitness (DME) is a useful measurement in evolution and population genetics. It is either directly studied by mutation accumulation (CHARLESWORTH *et al.* 2004; LOEWE AND HILL 2010) or inferred using population genetics models and DNA sequences (LOEWE AND CHARLESWORTH 2006; KEIGHTLEY AND EYRE-WALKER 2010). While the direct estimation of DME is appropriate for large effect mutations, indirect approach is useful for inferring DME for mutations with small effects (KEIGHTLEY AND EYRE-WALKER 2010). The distributions for direct and indirect DME measurements are different. The observed DME among de novel single-step beneficial mutation follows an exponential distribution (KASSEN AND BATAILLON 2006), while the analytic DME used for indirect inference is usually lognormal or gamma (LOEWE AND HILL 2010).

1.3 Genetic polymorphisms resulting from mutation, selection, drift, and demographic history

Polymorphism, in particular, genetic polymorphism, is important for conservation and biodiversity because it is required for a population to evolve in response to environmental change and it is associated with population fitness via inbreeding depression (REED AND FRANKHAM 2003). Genetic polymorphism refers to the occurrence of two or more alleles of at one locus in

the same population (CAVALLI-SFORZA AND BODMER 1971). Usually, the genetic variants and the common alleles within a population are genetic polymorphisms.

Most of the polymorphisms are selectively neutral or mildly deleterious undergoing weak purifying selection according to Motoo Kimura's neutral theory of molecular evolution because those mutations with under positive selection would sweep to fixation together with nearby linked variants relatively fast (KIMURA 1968). Kimura's theory reconciles the longtime confusion about how to maintain a high level of natural polymorphisms without balancing selection and the penalty of genetic load (BAMSHAD AND WOODING 2003).

The fundamental source of genetic polymorphism is random mutations, but the exact amount of genetic variation carried by a population depends on selection, drift, recombination, migration, as well as the size and demographic history of the population (HUDSON 2002) and mode of mating (BUSTAMANTE *et al.* 2002). The level of polymorphisms can be predicted based on modes of selection and demographic history (NEVO 1978; CHARLESWORTH *et al.* 1997). Because all the evolutionary processes affect polymorphism, it provides valuable information to infer selection, recombination, migration, and time of a demographic event.

The most frequently used genetic polymorphism in quantitative and population genetics is the single-nucleotide polymorphism, often abbreviated as SNP. SNP is the most abundant form of human genetic variation, which is also the most useful source for mapping complex traits (COLLINS *et al.* 1997), for studying haplotype (DALY *et al.* 2001), recombination map (MCVEAN *et al.* 2004; MYERS *et al.* 2005), and demographic history (GUTENKUNST *et al.* 2009). This is because of the nature of mutation, segregation, linkage, and recombination. For example, when mutation rate is low, all existing copies of a SNP in a population relate to each other and

coalesce to the most recent common ancestor (MRCA). The frequencies and patterns of SNPs then reflect the coalescence history as well as the mutational history (ROSENBERG AND NORDBORG 2002). This kind of analyses can answer questions like the human "mitochondria Eve" (VIGILANT *et al.* 1991). Selection can act in a population only if genetic polymorphism exists. How different modes of selection and recombination affect polymorphisms and site frequency spectrum are reviewed by Bamshad and Wooding (BAMSHAD AND WOODING 2003). Because different modes of selection have different effects on neutral and non-neutral genetic variation, SNP map is used to infer natural selection and candidate genes in human (AKEY *et al.* 2002).

Due to the existence of genetic variation, different individuals can have different molecular, cellular, and organismal level phenotypes. Because a linked region in a chromosome passes down to the next generation entirely unless recombination breaks the linkage, minor alleles (SNPs) can represent other small- or large-scale mutations in its nearby region. Therefore, SNP is also the most useful source for mapping complex traits (COLLINS *et al.* 1997). The techniques for genome-wide association studies (GWAS) using common SNPs have progressed a lot over the past decade. It has been shown that SNPs could explain a large proportion of the heritability for human height (YANG *et al.* 2010), body mass index (YANG *et al.* 2015), intelligence (DAVIES *et al.* 2011), and other complex traits (SPEED *et al.* 2017). Enrichment test for GWAS SNPs reveals the nature of the genetic architecture of complex traits (SCHORK *et al.* 2013). It has been shown that disease-associated variations are enriched in regulatory regions (MAURANO *et al.* 2012). Combining multiple pieces of evidence, such as combining GWAS SNPs and tissue-specific expression together, may help identify disease causal genes (LIU *et al.* 2012; LONSDALE *et al.* 2013).

SNPs are also useful in linkage analysis such as QTL mapping (LYNCH AND WALSH 1998). QTL mapping and GWAS are complementary to each other because one suffers from linked genome but benefits with balanced allele frequency and the other benefits from unlinked individuals with allele frequency out of control. Therefore, they are philosophically similar but they use different data and work for slightly different purposes. In chapter 2 to 5 in this thesis, I took advantage of large datasets generated for QTL mapping in yeast to study the patterns of mutational effects for natural polymorphisms (BLOOM *et al.* 2013; BLOOM *et al.* 2015; HALLIN *et al.* 2016).

1.4 Quantitative traits and QTL mapping

The concept of quantitative traits was proposed in the early 1900s to resolve the conflict between Mendelian theory for dichotomy traits and observations of continuous variation for most traits in nature (CASTLE 1903; PATERSON *et al.* 1988; BATESON AND MENDEL 2013). This is a simple yet extremely important conceptual achievement in modern genetics. It defended the principles of heredity and opened a new era of genetic study, which later became the subject quantitative genetics. It also fostered the post-Darwin era of evolutionary study, among which are the work lead by William Castle unifying Mendel's law with Darwin's theory of evolution (CASTLE 1903) and the work by Castle's graduate student Sewall Wright in population genetic theories for quantitative traits and natural variation (WRIGHT 1931).

A major challenge in evolution and in biology is to understand the genetic basis of quantitative traits (MACKAY *et al.* 2009) and to explain heritability, the fraction of phenotypic variation due to genetic variation (KEMPTHORNE 1957). A most common approach to study the genetic basis of quantitative traits is QTL mapping. QTL mapping refers to the statistical practice

of identifying genetic loci that contribute to variation in a quantitative trait through an experimental cross (BROMAN AND SEN 2009). Although genetic mapping was pioneered about a century ago (EAST 1916; SAX 1923), the first modern sense QTL mapping study was conducted in 1989 by Lander and Botstein using restriction fragment length polymorphisms (RFLPs) (LANDER AND BOTSTEIN 1989), it is a breakthrough in terms of the DNA markers used and the statistical approach developed. Soon after that, people realized the multiple testing problems of QTL mapping and developed a series of statistical methods to correct multiple testing or to calculate the confidence interval (JANSEN 1993; VISSCHER *et al.* 1996). This multiple testing problem is discussed in Chapter 2.

One purpose of QTL mapping is to identify the genetic cause of phenotypic variation. However, the large (20 centimorgans level) confidence interval has been a huge problem for many years (GEORGES *et al.* 1995; VAN LAERE *et al.* 2003; GODDARD AND HAYES 2009). It is not until very recently, with the availability of large-scale phenotyping and genome-wide panels of SNPs, and genetic editing, causal identification becomes possible. For example, Sadhu et al used CRISPR (clustered, regularly interspaced, short palindromic repeats) to build mapping panels with targeted recombination events in nematodes, and successfully identified causal genes and variants (SADHU *et al.* 2016). Having a larger panel of individuals and higher recombination density can also map to causal sites in yeast (SHE AND JAROSZ 2018). Despite these successful attempts for identifying causal mutations for simple organisms, mapping to causal sites for large genomes is still challenging and costly.

Another purpose of QTL mapping is to detect and estimate the effect of QTLs on heritable traits. These traits could be phenotypic or molecular. Expression QTL (eQTL) studies at transcript level and at proteome level have generated a lot of insights about cis-regulation and

trans-regulation which help advance the understanding of gene expression regulation and gene expression evolution (CHICK *et al.* 2016; ALBERT *et al.* 2017). For gene expression, the total variance explained by a single eQTL is generally higher than non-expression QTLs (BREM AND KRUGLYAK 2005), and the regulation is much simpler comparing to organismal level phenotype. Many of the QTLs for organismal level traits have very small effects, and only QTLs with relatively large effect can reach statistical significance (MACKAY *et al.* 2009). QTL mapping for organismal phenotype helps understand complex trait, heritability, the genomic architecture of complex traits, patterns of polymorphisms, and evolution (BLOOM *et al.* 2013; JERISON *et al.* 2017; WEI AND ZHANG 2017a).

1.5 The challenge in detecting genetic interactions

Allele by allele interactions (MENDEL 1996) and gene-by-gene interactions (or epistasis) (BATESON 2013) are the two most commonly studied types of genetic interactions. Here we discuss these genetic interactions for fitness. Allele by allele interactions could create complete dominance, incomplete dominance, codominance, overdominance, and recessive of the wildtype allele. Gene by gene interactions is relatively simple in haploid. There are four types of it: positive epistasis (synergistic), negative epistasis (antagonistic), sign epistasis, and reciprocal sign epistasis (PHILLIPS 2008). Gene by gene interactions in diploids can be an order of magnitude more complicated because it involves both allelic interactions and gene-by-gene interactions. In diploids, if the genotypic values cannot be predicted from the single locus additive and dominance effects, there is epistasis (MACKAY 2015). Higher order epistasis involving more than two genes is usually out of our current detection power (TAYLOR AND EHRENREICH 2015).

Most of the large-scale studies of genetic interactions use gene deletions (WILKIE 1994; COSTANZO *et al.* 2016). These studies tested the effects of single mutation and double mutations empirically and provided valuable information about genetic interactions. However, gene deletions and null mutations usually have large effects, representing only a small subset of all mutations. The existence of interactions among gene deletion does not equal to the existence of genetic interactions in nature because many of the deletion pairs are so deleterious that they never exist in the same genome.

QTL mapping has proved itself a powerful tool to study the additive effects of natural polymorphisms, but less so for interactive effects. Moreover, because of the existence of unknown genetic interactions, the accuracy and the power of QTL mapping are affected (PHILLIPS 2008). Because interaction effect is usually smaller than the main additive effect, detecting genetic interactions in natural polymorphisms by QTL mapping is still difficult. This detection power problem constrained our ability to understand the distributions of interactive effects in nature, as well as how these interactions affect adaptation and evolution. The canonical way of QTL mapping involves using an additive model, which means no interactions between two alleles of the same gene and between genes. Dominance or gene-by-gene interactions are ignored or only be considered after taking additive effects into account. However, canonical does not necessarily mean correct, and the majority of the mutational effects may not be additive. For example, a dominant null model can perform equally well as an additive null model (HUANG AND MACKAY 2016). Mapping gene-by-gene interaction is even harder because n polymorphic sites would require n^2 number of tests. Current ways for mapping interactive QTLs either only test QTLs with significant additive effect or reduce the number of markers in pairwise testing; the observed interactive QTLs could explain only a small fraction of the total phenotypic variance

(BLOOM *et al.* 2015). Because of the detection power limitations and the null assumption of additivity, whether the interactive effect is general and how much of the total variance is affected by those interactions may not be fully reflected in these QTL mapping studies. Because detecting significant allelic interactions and gene-by-gene interactions are difficult using QTL mapping approach, we could only compare the general trend of interactions across environments. Such comparisons do not have to require significant and can help understand the amount of interactions among genetic polymorphisms and provide information about how environment effects change the prevalence of genetic interactions and how genetic interactions affect adaptation. In chapter 4 and 5, I took advantage of this approach in studying genetic interactions.

1.6 Heterosis and genetic interactions

Heterosis, or hybrid vigor, refers to the phenomenon that a hybrid is superior to both of its typically inbred parents in any biological quality (e.g., biomass, growth rate, and resistance to pathogens). Darwin was the first to report the observation of heterosis (Darwin 1876); Schull (Shull 1908) and East (East 1908) rediscovered it in 1908.

Heterosis has important relevance to many aspects of our lives. It was first applied to crop breeding by Shull (SHULL 1908), and it is soon widely applied in plant and animal breeding. It is estimated that heterosis increases maize yields by at least 15% (LIPPMAN AND ZAMIR 2007a). Today, 95% of maize acreage in U.S. and 65% worldwide is planted with hybrids (SWANSON-WAGNER *et al.* 2006; HOCHHOLDINGER AND HOECKER 2007; LIPPMAN AND ZAMIR 2007a). Heterosis also affects the pathogenesis of many eukaryotic pathogens. For example, fungal meningitis and encephalitis, especially as a secondary infection for AIDS patients, are often caused by the yeast *Cryptococcus neoformans*. *C. neoformans* has three serotypes: A, D, and AD.

AD is a hybrid of A and D. Most AD isolates exhibit hybrid vigor, and are resistant to the antifungal drug FK506, whereas A and D are not (LI *et al.* 2012). In addition, heterosis occurs to humans. For instance, marital distance, the geographic distance between the birth places of a couple, positively impacts the height of their kids (KOZIEL *et al.* 2011). An analysis of 35,000 human individuals from 35 different population samples showed a highly significant association between height and genome-wide heterozygosity (MCQUILLAN *et al.* 2012). Moreover, higher levels of genetic heterozygosity tend to occur in the outbred group and are associated with lower blood pressure (BP) and total/LDL cholesterol (CAMPBELL *et al.* 2007). Study the basis of heterosis can help optimize hybrid performance, control pathogenesis, and understand human diseases.

Without genetic interactions, the hybrid of two homozygous parents should follow an additive model such that its performance is the average of two parents. Positive genetic interactions can contribute to heterosis. Dominance and overdominance were proposed to explain heterozygote advantage (LIPPMAN AND ZAMIR 2007a). The dominance model posits that each inbred parent contains deleterious alleles at several loci whereas in hybrids these deleterious alleles are complemented by the dominant wild-type alleles from the other parent. Note that this model only requires that the superior allele at a locus is more dominant over the inferior allele at the locus; no complete dominance is required. The overdominance model posits that allelic interactions at a single heterozygous locus result in a synergistic effect on vigor that surpasses both homozygous parents. Positive epistasis from the combination of the two or more parental genes also contributes to heterosis besides overdominance. It is unclear which process is the leading one for creating heterosis.

Studies of mechanisms of heterosis usually require mapping for genetic interactions (LIPPMAN AND ZAMIR 2007b) and are affected by detection power. Whether dominance, overdominance, or positive epistasis, contribute to heterosis most is still unanswerable right now. Despite known for more than 140 years (Darwin 1876) and practiced for at least thousands of years (for example, mules were mentioned in Homer's *Iliad*, 800 BC Greece (EDWARDS 1890; LEIGHTON 1967)), the major cause of simple process may remain mysterious for an indefinite time, until we have better detection power for genetic interactions of natural polymorphism. Understanding the mechanisms of heterosis will greatly improve many related applications in agriculture, conservation biology, pathogen control, and human health.

1.7 Speciation and genetic interactions

Speciation process is a fundamental problem in biology and in evolution. Speciation could be driven by ecological speciation or it could be driven by genetic speciation. Polyploidization (RIESEBERG AND WILLIS 2007), hybridization (MALLET 2007), and transposition (DOBZHANSKY AND DOBZHANSKY 1937; MASLY *et al.* 2006) are all potential causes of genetic speciation. Some of these genome-recreating events can instantly prevent mating with the original population. On the contrary, ecological speciation is more of an accumulation process where reproductive isolation is a gradually evolved feature, presumably due to divergent selection (SCHLUTER 2001).

The most common classification of modes of ecological speciation is sympatric, parapatric, and allopatric, categorizing how divergence occurs (BUTLIN *et al.* 2008). Allopatric speciation, which involves geographical isolation is believed the usual mode, can happen simply from the neutral accumulation of genetic incompatibilities. Parapatric speciation was first

described by Fisher in general terms (FISHER 1958) and then by Murray (MURRAY 1972), Bush (BUSH 1975), and Endler (ENDLER 1977) more specifically. The difference between parapatric and sympatric speciation is that in parapatric speciation isolation is incomplete and gene flow between the two populations is allowed (SLATKIN 1982). Sympatric speciation, on the other hand, requires disruptive natural or sexual selection that favors two distinct phenotypes (KONDRASHOV AND KONDRASHOV 1999). According to Darwin, heterogeneous environment with resource competition can lead to disruptive selection and sympatric speciation (DARWIN 1968). J. Maynard Smith later proposed four genetic mechanisms of sympatric speciation: habitat selection, pleiotropic genes, modifying genes, and assortative mating genes (SMITH 1966).

The argument about sympatric speciation used to be old and long-lived because it cannot be easily settled by observations (SMITH 1966), but the situation changed a lot recently. Recent advancements in experimental evolution allow researchers to study different modes of speciation in a forward way. For example, Castillo et al conducted an experimental test for allopatric speciation (CASTILLO *et al.* 2015). In another study, sympatric "speciation" was shown for lambda phage by experimental evolution (MEYER *et al.* 2016), demonstrating the power of experimental evolution in answering questions about speciation. Moreover, because whole genome sequencing is getting cheaper and more sensitive, monitoring contemporary parapatric speciation process by the change of allele frequency becomes possible (EGAN *et al.* 2015).

Another branch of speciation studies focus on identifying the "speciation genes" which either occur at the initiation process of speciation or later as the two species diverge. Although different modes of ecological speciation can all initiate speciation, all speciation events eventually require some genetic changes to keep the two isolated species maintaining isolated genetically by pre- and/or post-mating isolation. One major type of post-mating isolation is due

to Bateson-Dobzhansky-Muller (BDM) incompatibility (BATESON 1909; DOBZHANSKY AND DOBZHANSKY 1937; MULLER 1942), which is a type of negative genetic interactions that involves at least two mutations each occurs in one of the two species. BDM incompatibility is an intellectual advancement because it resolved and bypassed a major problem in speciation, how something extremely deleterious could be allowed by natural selection (ORR 1996). Mapping of incompatible gene pairs usually involves a lot of crosses and experimental validation, and very few studies successfully identified the causal gene pairs (TING *et al.* 1998; COYNE AND ORR 2004; LEE *et al.* 2008). Because BDM incompatibility is a type of genetic interactions, the number of new interactions that exist only in the hybrid but not in the parents increases with divergent time at a speed equal to or faster than quadratic. This process is also called "snowball" effect. The "snowballs" of the number of incompatible genes is proven by two genetic mapping studies using interspecific crossing of plant and animal (MATUTE *et al.* 2010; MOYLE AND NAKAZATO 2010).

Because incompatibility may occur even within species (CORBETT-DETIG *et al.* 2013; SOHAIL *et al.* 2017), studying intraspecific genetic incompatibility may shed light on speciation process. Unlike studying BDM incompatibility by interspecific crosses where the species chosen cannot be too divergent, studying genetic incompatibility by intraspecific cross suits every species. Moreover, hardly are the genes initiating speciation the ones detected, but this downside might be compensated by studying patterns of genetic interactions segregating within species. Intraspecific incompatibility could be maintained by nonrandom mating. Because random mating is unlikely happening in nature (BUSS AND BARNES 1986; MORIN *et al.* 1994; JIANG *et al.* 2013), it would be nice to have a theoretical and empirical study that connects the incompatibility accumulated within species due to non-random mating with modes of ecological speciation.

Speciation is not directly studied in this thesis, but by analyzing intraspecific cross, I demonstrated the existence and the "snowball" of genetic incompatibilities within species in chapter 6.

1.8 Genetic interactions and adaptation

Genetic interactions affect the evolutionary trajectory and the fixation probability of new mutations. Here we discuss the two types of interactions: allelic interactions (MENDEL 1996) and gene-by-gene interactions separately. Chapter 6 used both types of genetic interactions to make inference about hybrid performance.

For allelic interactions, mutations that are beneficial or deleterious have very different fate. While beneficial mutations benefit from being visible to positive selection immediately if it is dominant, a deleterious mutation is more easily purged out by purifying selection. A recessive mutation behaves like a neutral mutation until it by chance creates a homozygous, thus a recessive beneficial mutation is less likely reaching a high frequency and a recessive deleterious mutation is less likely purged out by selection. Haldane first showed this biased fixation toward dominant beneficial allele, and this phenomenon is later termed "Haldane's sieve" (HALDANE 1927; HALDANE 1930). Because of allelic interactions, the fixation of mutations follows more complex trend making the rate and pattern of diploid adaptation different from haploid adaptation (PAQUIN AND ADAMS 1983). Allelic interaction is studied in Chapter 4.

For gene-by-gene interactions, positive epistasis and negative epistasis affect the fate of new mutations a lot (HANSEN 2013). Moreover, because of such dependency, one fixation may open and close some adaptive trajectories due to epistasis. Because of epistasis, different trajectories of mutations are not equally probable. Researchers often use this extra information to

narrow down to a few possible evolutionary trajectories from empirically measured mutational landscape (PALMER *et al.* 2015; STARR AND THORNTON 2016), this is a kind of "reverse approach" in studying the relationship between epistasis and adaptation. Some of the recent works in experimental evolution field also discovered how genetic interactions affects adaptive trajectories. For example in a recent used the 60,000 generations long-term experimental evolution of E.coli, and sequenced the stocks at 500 generation interval (GOOD *et al.* 2017). In this study, they found that the appearance time of beneficial mutations are different for mutations in many adaptive genes, demonstrating that epistasis affects adaptive trajectories. Studying epistasis and adaptation by experimental evolution is a "forward approach". The forward and the reverse approach have different benefits and compensate each other, one explores more possibilities of mutations but does not know the clear evolutionary path, and the other explores only the random mutations happen during the experimental evolution but know the adaptive trajectory for certain.

One overwhelmed pattern in adaptive trajectories of experimental evolution is the diminishing returns of fitness with the number of beneficial mutations (TENAILLON *et al.* 2016). One of the underlying reasons is the widespread diminishing returns epistasis (KRYAZHIMSKIY *et al.* 2014) among adaptive mutations (WÜNSCHE *et al.* 2017). The diminishing returns epistasis, which is a special case of negative epistasis, is general, and I will further introduce this part in Chapter 3.

1.9 Gene by environment interactions (G×E) in evolution

G×E refers to the observation that the same mutation has different phenotypic effects on a trait in different environments (OTTMAN 1996). As early as the first QTL mapping study,

multiple environments were included, although each environment is dealt separately in mapping (LANDER AND BOTSTEIN 1989). Jansen et al quickly noticed this lack of accounting for G×E and developed the first approach that accommodates mapping for multiple QTL as well as G×E (JANSEN *et al.* 1995). G×E is believed to be ubiquitous among all organisms and has long been studied in domestic animals and plants, genetic model organisms, and humans (WEI AND ZHANG 2017a).

G×E exists and is studied for different traits. At cellular trait level, G×E is often discussed under cis-regulatory expression and trans-regulatory expression framework. For example, the G×E for expression has been studied in yeast, where trans-regulating mutations from distant linkage are found to be more environment dependent (SMITH AND KRUGLYAK 2008). G×E is also often studied at phenotype level. For example, it is found to have important effects on human psychiatry disease (DUNCAN AND KELLER 2011). In chapter 2, I studied G×E at growth rate level (WEI AND ZHANG 2017a).

G×E could be studied at genotype level or at mutational level. G×E at genotype level is often studied in the wild. Numerous studies have discussed the G×E responses to climate change, habitat change, or change of other environmental factors (AGRAWAL 2001; GIENAPP *et al.* 2008; VALLADARES *et al.* 2014). At mutational level, G×E studies can be generally divided into two types on the basis of the approach used: forward genetics and reverse genetics. In forward genetics, genes or QTLs that show significantly different phenotypic effects in different environments are identified via linkage or association mapping. In reverse genetics, a mutant carrying a known mutation such as a gene deletion or a point mutation is compared with the wild-type for the trait of interest under two environments, and G×E is detected when the mutational effect on the trait differs significantly in the two environments.

The effect of G×E for a mutation could be divided into two types, antagonistic G×E, and concordant G×E. Antagonistic G×E refers to a mutation that increases the trait value in one environment but decreases the trait value in another; concordant G×E refers to a mutation that affects the trait to the same direction but at a different magnitude in two environments (WEI AND ZHANG 2017a). Although concordant G×E is more general than antagonistic G×E (OSTROWSKI *et al.* 2005; GERKE *et al.* 2010; DILLON *et al.* 2016; WEI AND ZHANG 2017a), the extent of antagonism depends on the tested environments and tested genotypes (WEI AND ZHANG 2017a). The existence of G×E especially antagonistic G×E in nature but not in many of the experimental evolution in the lab (TENAILLON *et al.* 2016) may cause a very different spectrum of fixation of mutations as well as the size of the pool of beneficial mutations.

Investigating G×E can help identify the causal pathways of a trait (GAGNEUR *et al.* 2013), dissect genetic tradeoffs (QIAN *et al.* 2012), understand environmental adaptations (OSTROWSKI *et al.* 2005), and reveal a potential cause of "missing heritability" (MANOLIO *et al.* 2009; EICHLER *et al.* 2010). I reported patterns of G×E and measured the effects of G×E on "missing heritability" in my Chapter 2 (WEI AND ZHANG 2017a).

1.10 Thesis overview

In this thesis, I examine different kinds of genetic interactions and G×E that affects the effects of genetic polymorphisms. I use public available genotype and fitness related phenotype data in budding yeast *Saccharomyces cerevisiae*, house mouse *Mus musculus*, plant *Arabidopsis thaliana*, and human *Homo sapiens* to tackle the questions about mutational effects.

In Chapter 2, I addressed the question of how the environment affects the mutations by conducting G×E QTL mapping in budding yeast *Saccharomyces cerevisiae* for 1081 pairs of

environments. I reported many general patterns G×E, such as, how likely a QTL has a different effect in a different environment, how likely a G×E is antagonistic versus concordant, where are the genomic location of G×E sites, and how much G×E causes "missing heritability". Because I found that the mutational effects are often environment dependent, I went on to study the details of such environment dependent effects in later chapters. This chapter is the basis of Chapters 3 to 5.

In Chapter 3, I studied how the environment affects diminishing returns epistasis. Diminishing returns epistasis means that the same advantageous mutation is less beneficial when occurring on a fitter genotype background; it is often found during experimental evolution of microbes and was suggested to be general. In this chapter, I developed a high-throughput approach to study diminishing returns epistasis with population data. I then used this approach to quantify the fraction of diminishing returns epistasis for yeast growth across 47 environments. I found diminishing returns epistasis is general, and the fraction of diminishing returns epistasis increases as *Q* increases. I also calculated the effect size for each polymorphic locus and found that the benefit of a SNP also decreases as *Q* increases. I developed a new model named modular life model which takes both environment contribution and genetic contribution into account. This new model successfully explains all the observed patterns of diminishing returns.

In Chapter 4, I follow the findings of Chapter 3 to study genetic dominance. Theories on the origin of genetic dominance have experienced a century-long debate, but none satisfactorily explained all currently observed patterns of dominance. In this chapter, I propose that dominance is a special case of diminishing returns epistasis because the common observation is that the benefit from gaining a wildtype allele on a homozygous deleterious background is bigger than the benefit from gaining the same wildtype allele on a heterozygous background. I first used

modular life model to predict the known patterns of genetic dominance and the unknown patterns of genetic dominance. According to modular life model, all current observations of dominance are expected. Moreover, it predicts that as *Q* gets higher, the beneficial mutation gets more dominant, resembling the patterns of diminishing returns. Two independent yeast dataset confirmed the predicted pattern from using the modular life model. This observed pattern is opposite to the prediction from the Wright-Kacser-Burns model, the previous leading model.

In Chapter 5, I study the pleiotropic effect of mutations on *r* and *K*, two fitness proxies, to test whether the pleiotropic effect is environment dependent. *r-K* relationship has been a long-standing question in life history ecology. Studying the genetic basis of *r-K* and the mutational relationship of *r-K* by quantitative genetic approach helps understand the *r-K* relationship and predict life history evolution. In this study, I found positive *r-K* pleiotropy is prevalent in low *Q* environment and negative *r-K* pleiotropy is prevalent in high *Q* environment. I also observe the same mutation can change from concordant pleiotropy to antagonistic pleiotropy when the environment changes. This finding is hard to explain by a simple energy tradeoff model. I proposed a new model, which includes the tradeoff of rate and yield of ATP production and the cost of maintenance relative to reproduction. The model predictions match well with the observed patterns.

After studying different types of genetic interactions individually for fitness proxies, I study them together to predict phenotype. Having great application potential, hybrid performance is an important topic in biology. As I introduced before, hybrid performance experience two counteracting process, one is heterosis, and another is genetic incompatibility, both rise from genetic interactions. Because of these counteracting forces, it is believed that the fitness of a genotype is a hump-shaped function of the mating distance, culminating at an intermediate

distance referred to as the optimal mating distance (OMD). I derived the model between genetic distance *D* and a hybrid performance measurement and tested the model using large datasets from the plant *Arabidopsis thaliana*, fungus *Saccharomyces cerevisiae*, and animal *Mus musculus*. I confirmed the existence of OMD in all three species.

While Chapters 2 to 6 focus on the mutational effects due to genetic interactions and gene by environment interactions, two appendices Chapters A and B discussed other mutational effects in evolution. Appendix A discussed the case when selection on one gene and the selection on the other gene occur on the same genomic region, such that the net effect of a beneficial mutation to one gene may not be beneficial due to its deleterious effect on the other gene. In this chapter, I developed a simple method to disentangle individual selection strength for overlapping genes whose coding regions overlapped with each other. Appendix B discussed the relationship between the robustness of a phenotype and the evolvability of a phenotype. In this project, I provide the mathematical proof for the relationship between phenotype robustness (PR) and phenotype evolvability (PE) defined in a random genotype-phenotype map (GPM). I showed that the PR and PE are positively correlated in random GPM, suggesting PR and PE are by default positively correlated.

In Chapter 7, I discussed some of the topics and models proposed in this thesis in a unified way. I also discussed some ideas I conceived while working on this thesis. A couple sections have preliminary results; the majority are still at hypothetical stages. In the end, I discuss the questions that interest me most and my future research goals.

1.11 References

Agrawal, A. A., 2001 Phenotypic plasticity in the interactions and evolution of species. Science 294**:** 321-326.

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome research 12**:** 1805-1814.
Albert, F. W., J. S. Bloom, J. Siegel, L. Day and L. Kruglyak, 2017 Genetics of trans-regulatory variation in gene expression. bioRxiv**:** 208447.
Bamshad, M., and S. P. Wooding, 2003 Signatures of natural selection in the human genome. Nature Reviews Genetics 4**:** 99.
Barrett, R. D., and H. E. Hoekstra, 2011 Molecular spandrels: tests of adaptation at the genetic level. Nature Reviews Genetics 12**:** 767.
Bateson, W., 1909 Heredity and variation in modern lights. Darwin and modern science 85**:** 101.
Bateson, W., 2013 *Mendel's principles of heredity*. Courier Corporation.
Bateson, W., and G. Mendel, 2013 *Mendel's principles of heredity*. Courier Corporation.
Bergland, A. O., E. L. Behrman, K. R. O'Brien, P. S. Schmidt and D. A. Petrov, 2014 Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. PLoS Genetics 10**:** e1004775.
Bloom, J. S., I. M. Ehrenreich, W. T. Loo, T.-L. V. Lite and L. Kruglyak, 2013 Finding the sources of missing heritability in a yeast cross. Nature 494**:** 234.
Bloom, J. S., I. Kotenko, M. J. Sadhu, S. Treusch, F. W. Albert *et al.*, 2015 Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nature communications 6**:** 8712.
Brem, R. B., and L. Kruglyak, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of the National Academy of Sciences of the United States of America 102**:** 1572-1577.
Broman, K. W., and S. Sen, 2009 *A Guide to QTL Mapping with R/qtl*. Springer.
Bush, G. L., 1975 Modes of animal speciation. Annual Review of Ecology and Systematics 6**:** 339-364.
Buss, D. M., and M. Barnes, 1986 Preferences in human mate selection. Journal of personality and social psychology 50**:** 559.
Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan *et al.*, 2002 The cost of inbreeding in Arabidopsis. Nature 416**:** 531.
Butlin, R. K., J. Galindo and J. W. Grahame, 2008 Sympatric, parapatric or allopatric: the most important way to classify speciation? Philosophical Transactions of the Royal Society B: Biological Sciences 363**:** 2997-3007.
Campbell, H., A. D. Carothers, I. Rudan, C. Hayward, Z. Biloglav *et al.*, 2007 Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Hum Mol Genet 16**:** 233-241.
Castillo, D. M., M. K. Burger, C. M. Lively and L. F. Delph, 2015 Experimental evolution: assortative mating and sexual selection, independent of local adaptation, lead to reproductive isolation in the nematode Caenorhabditis remanei. Evolution 69**:** 3141-3155.
Castle, W. E., 1903 *The laws of heredity of Galton and Mendel: and some laws governing race improvement by selection*. Academy.
Cavalli-Sforza, L., and W. Bodmer, 1971 Human population genetics. San Francisco, CA: Free-man.
Charlesworth, B., H. Borthwick, C. Bartolomé and P. Pignatelli, 2004 Estimates of the genomic mutation rate for detrimental alleles in Drosophila melanogaster. Genetics 167**:** 815-826.

Charlesworth, B., M. Nordborg and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genetics Research 70**:** 155-174.
Chick, J. M., S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi *et al.*, 2016 Defining the consequences of genetic variation on a proteome-wide scale. Nature 534**:** 500.
Collins, F. S., M. S. Guyer and A. Chakravarti, 1997 Variations on a theme: cataloging human DNA sequence variation. Science 278**:** 1580-1581.
Corbett-Detig, R. B., J. Zhou, A. G. Clark, D. L. Hartl and J. F. Ayroles, 2013 Genetic incompatibilities are widespread within species. Nature 504**:** 135-+.
Costanzo, M., B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons *et al.*, 2016 A global genetic interaction network maps a wiring diagram of cellular function. Science 353**:** aaf1420.
Coyne, J. A., and H. A. Orr, 2004 Speciation. Sunderland, MA, pp. Sinauer Associates, Inc.
Crick, F. H., 1958 On protein synthesis, pp. 8 in *Symp Soc Exp Biol*.
Crow, J. F., and M. Kimura, 1970 An introduction to population genetics theory. An introduction to population genetics theory.
Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander, 2001 High-resolution haplotype structure in the human genome. Nature genetics 29**:** 229.
Darwin, C., 1876 *The Effects of Cross- and Self-fertilisation in the Vegetable Kingdom*. John Murray.
Darwin, C., 1968 On the origin of species by means of natural selection. 1859. London: Murray Google Scholar.
Davies, G., A. Tenesa, A. Payton, J. Yang, S. E. Harris *et al.*, 2011 Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Molecular psychiatry 16**:** 996.
Dillon, M. M., N. P. Rouillard, B. Van Dam, R. Gallet and V. S. Cooper, 2016 Diverse phenotypic and genetic responses to short-term selection in evolving Escherichia coli populations. Evolution 70**:** 586-599.
Dobzhansky, T., and T. G. Dobzhansky, 1937 *Genetics and the Origin of Species*. Columbia university press.
Duncan, L. E., and M. C. Keller, 2011 A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. American Journal of Psychiatry 168**:** 1041-1049.
East, E. M., 1908 Inbreeding in corn. Conn. Agric. Exp. Sta. Rpt. 1907**:** 419-428.
East, E. M., 1916 Studies on size inheritance in Nicotiana. Genetics 1**:** 164-176.
Edwards, M. W., 1890 *Iliad*. Wiley Online Library.
Egan, S. P., G. J. Ragland, L. Assour, T. H. Powell, G. R. Hood *et al.*, 2015 Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. Ecology letters 18**:** 817-825.
Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal *et al.*, 2010 Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics 11**:** 446-450.
Endler, J. A., 1977 *Geographic variation, speciation, and clines*. Princeton University Press.
Fisher, R. A., 1958 *The genetic theory of natural selection*. Dover.
Forsberg, S. K., M. E. Andreatta, X.-Y. Huang, J. Danku, D. E. Salt *et al.*, 2015 The multi-allelic genetic architecture of a variance-heterogeneity locus for molybdenum concentration in

leaves acts as a source of unexplained additive genetic variance. PLoS genetics 11**:** e1005648.
Gagneur, J., O. Stegle, C. Zhu, P. Jakob, M. M. Tekkedil *et al.*, 2013 Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. PLoS Genet 9**:** e1003803.
Georges, M., D. Nielsen, M. Mackinnon, A. Mishra, R. Okimoto *et al.*, 1995 Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics 139**:** 907-920.
Gerke, J., K. Lorenz, S. Ramnarine and B. Cohen, 2010 Gene-environment interactions at nucleotide resolution. PLoS Genet 6**:** e1001144.
Gienapp, P., C. Teplitsky, J. Alho, J. Mills and J. Merilä, 2008 Climate change and evolution: disentangling environmental and genetic responses. Molecular ecology 17**:** 167-178.
Goddard, M. E., and B. J. Hayes, 2009 Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics 10**:** 381.
Good, B. H., M. J. McDonald, J. E. Barrick, R. E. Lenski and M. M. Desai, 2017 The dynamics of molecular evolution over 60,000 generations. Nature 551**:** 45.
Group, U. C. S. W., 2010 United States cancer statistics: 1999-2006 incidence and mortality Web-based report. Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute.
Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS genetics 5**:** e1000695.
Haldane, J. B. S., 1927 A mathematical theory of natural and artificial selection, part V: selection and mutation, pp. 838-844 in *Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge University Press.
Haldane, J. B. S., 1930 A mathematical theory of natural and artificial selection.(Part VI, Isolation.), pp. 220-230 in *Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge University Press.
Hallin, J., K. Märtens, A. I. Young, M. Zackrisson, F. Salinas *et al.*, 2016 Powerful decomposition of complex traits in a diploid model. Nature communications 7**:** 13311.
Hansen, T. F., 2013 Why epistasis is important for selection and adaptation. Evolution 67**:** 3501-3511.
Hochholdinger, F., and N. Hoecker, 2007 Towards the molecular basis of heterosis. Trends Plant Sci 12**:** 427-432.
Houle, D., D. R. Govindaraju and S. Omholt, 2010 Phenomics: the next challenge. Nature reviews genetics 11**:** 855.
Huang, W., and T. F. Mackay, 2016 The genetic architecture of quantitative traits cannot be inferred from variance component analysis. PLoS genetics 12**:** e1006421.
Hudson, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18**:** 337-338.
Jansen, R., J. Van Ooijen, P. Stam, C. Lister and C. Dean, 1995 Genotype-by-environment interaction in genetic mapping of multiple quantitative trait loci. Theoretical and Applied Genetics 91**:** 33-37.
Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics 135**:** 205-211.
Jerison, E. R., S. Kryazhimskiy, J. K. Mitchell, J. S. Bloom, L. Kruglyak *et al.*, 2017 Genetic variation in adaptability and pleiotropy in budding yeast. eLife 6.

Jiang, Y., D. I. Bolnick and M. Kirkpatrick, 2013 Assortative mating in animals. The American Naturalist 181**:** E125-E138.
Kærn, M., T. C. Elston, W. J. Blake and J. J. Collins, 2005 Stochasticity in gene expression: from theories to phenotypes. Nature Reviews Genetics 6**:** 451.
Kassen, R., and T. Bataillon, 2006 Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nature genetics 38**:** 484.
Keightley, P. D., and A. Eyre-Walker, 2010 What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philosophical Transactions of the Royal Society B: Biological Sciences 365**:** 1187-1193.
Kempthorne, O., 1957 *An introduction to genetic statistics*. John Wiley And Sons, Inc.; New York.
Khalil, A. M., M. Guttman, M. Huarte, M. Garber, A. Raj *et al.*, 2009 Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proceedings of the National Academy of Sciences 106**:** 11667-11672.
Kimura, M., 1968 Evolutionary rate at the molecular level. Nature 217**:** 624-626.
Kondrashov, A. S., and F. A. Kondrashov, 1999 Interactions among quantitative traits in the course of sympatric speciation. Nature 400**:** 351.
Koziel, S., D. P. Danel and M. Zareba, 2011 Isolation by distance between spouses and its effect on children's growth in height. Am J Phys Anthropol 146**:** 14-19.
Kryazhimskiy, S., D. P. Rice, E. R. Jerison and M. M. Desai, 2014 Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science 344**:** 1519-1522.
Lander, E. S., and D. Botstein, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121**:** 185-199.
Lee, H.-Y., J.-Y. Chou, L. Cheong, N.-H. Chang, S.-Y. Yang *et al.*, 2008 Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. Cell 135**:** 1065-1073.
Leighton, A. C., 1967 The mule as a cultural invention. Technology and Culture 8**:** 45-52.
Li, W., A. F. Averette, M. Desnos-Ollivier, M. Ni, F. Dromer *et al.*, 2012 Genetic Diversity and Genomic Plasticity of Cryptococcus neoformans AD Hybrid Strains. G3 (Bethesda) 2**:** 83-97.
Liaw, D., D. J. Marsh, J. Li, P. L. Dahia, S. I. Wang *et al.*, 1997 Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. Nature genetics 16**:** 64.
Lippman, Z. B., and D. Zamir, 2007a Heterosis: revisiting the magic. Trends Genet 23**:** 60-66.
Lippman, Z. B., and D. Zamir, 2007b Heterosis: revisiting the magic. Trends in genetics 23**:** 60-66.
Liu, C., F. Zhang, T. Li, M. Lu, L. Wang *et al.*, 2012 MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. BMC genomics 13**:** 661.
Loewe, L., and B. Charlesworth, 2006 Inferring the distribution of mutational effects on fitness in Drosophila. Biology Letters 2**:** 426-430.
Loewe, L., and W. G. Hill, 2010 The population genetics of mutations: good, bad and indifferent, pp. The Royal Society.
Lonsdale, J., J. Thomas, M. Salvatore, R. Phillips, E. Lo *et al.*, 2013 The genotype-tissue expression (GTEx) project. Nature genetics 45**:** 580.

Lynch, M., and B. Walsh, 1998 *Genetics and analysis of quantitative traits*. Sinauer Sunderland, MA.
MacArthur, R. H., and E. O. Wilson, 2016 *The theory of island biogeography*. Princeton university press.
Mackay, T. F., 2015 Epistasis for quantitative traits in Drosophila, pp. 47-70 in *Epistasis*. Springer.
Mackay, T. F., E. A. Stone and J. F. Ayroles, 2009 The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10**:** 565-577.
Mallet, J., 2007 Hybrid speciation. Nature 446**:** 279.
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**:** 747-753.
Marée, A. F., W. Keulen, C. A. Boucher and R. J. De Boer, 2000 Estimating relative fitness in viral competition experiments. Journal of virology 74**:** 11067-11072.
Masly, J. P., C. D. Jones, M. A. Noor, J. Locke and H. A. Orr, 2006 Gene transposition as a cause of hybrid sterility in Drosophila. Science 313**:** 1448-1450.
Matute, D. R., I. A. Butler, D. A. Turissini and J. A. Coyne, 2010 A Test of the Snowball Theory for the Rate of Evolution of Hybrid Incompatibilities. Science 329**:** 1518-1521.
Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen *et al.*, 2012 Systematic localization of common disease-associated variation in regulatory DNA. Science**:** 1222794.
McGuigan, K., and C. M. Sgro, 2009 Evolutionary consequences of cryptic genetic variation. Trends in ecology & evolution 24**:** 305-311.
McQuillan, R., N. Eklund, N. Pirastu, M. Kuningas, B. P. McEvoy *et al.*, 2012 Evidence of inbreeding depression on human height. PLoS Genet 8**:** e1002655.
McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science 304**:** 581-584.
Mendel, G., 1996 Experiments in plant hybridization (1865). Verhandlungen des naturforschenden Vereins Brünn.) Available online: www. mendelweb. org/Mendel. html (accessed on 1 January 2013).
Metz, J. A., S. A. Geritz, G. Meszéna, F. J. Jacobs and J. S. Van Heerwaarden, 1995 Adaptive dynamics: a geometrical study of the consequences of nearly faithful reproduction.
Meyer, J. R., D. T. Dobias, S. J. Medina, L. Servilio, A. Gupta *et al.*, 2016 Ecological speciation of bacteriophage lambda in allopatry and sympatry. Science**:** aai8446.
Moran, P. A. P., 1958 Random processes in genetics, pp. 60-71 in *Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge University Press.
Morin, P. A., J. J. Moore, R. Chakraborty, L. Jin, J. Goodall *et al.*, 1994 Kin selection, social structure, gene flow, and the evolution of chimpanzees. Science 265**:** 1193-1201.
Moyle, L. C., and T. Nakazato, 2010 Hybrid Incompatibility "Snowballs" Between Solanum Species. Science 329**:** 1521-1523.
Muller, H., 1942 Isolating mechanisms, evolution, and temperature, pp. 71-125 in *Biol. Symp.*
Murray, J., 1972 Genetic diversity and natural selection. Genetic diversity and natural selection.
Myers, S., L. Bottolo, C. Freeman, G. McVean and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science 310**:** 321-324.
Nevo, E., 1978 Genetic variation in natural populations: patterns and theory. Theoretical population biology 13**:** 121-177.
Orr, H. A., 1996 Dobzhansky, Bateson, and the genetics of speciation. Genetics 144**:** 1331-1335.

Orr, H. A., 2009 Fitness and its role in evolutionary genetics. Nature Reviews Genetics 10**:** 531.
Ostrowski, E. A., D. E. Rozen and R. E. Lenski, 2005 Pleiotropic effects of beneficial mutations in Escherichia coli. Evolution 59**:** 2343-2352.
Ottman, R., 1996 Gene-environment interaction: definitions and study designs. Prev Med 25**:** 764-770.
Palmer, A. C., E. Toprak, M. Baym, S. Kim, A. Veres *et al.*, 2015 Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes. Nature communications 6**:** 7385.
Paquin, C., and J. Adams, 1983 Frequency of fixation of adaptive mutations is higher in evolving diploid than haploid yeast populations. Nature 302**:** 495.
Paterson, A. H., E. S. Lander, J. D. Hewitt, S. Peterson, S. E. Lincoln *et al.*, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335**:** 721.
Phillips, P. C., 2008 Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics 9**:** 855.
Qian, W., D. Ma, C. Xiao, Z. Wang and J. Zhang, 2012 The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. Cell Rep 2**:** 1399-1410.
Raser, J. M., and E. K. O'shea, 2004 Control of stochasticity in eukaryotic gene expression. science 304**:** 1811-1814.
Reed, D. H., and R. Frankham, 2003 Correlation between fitness and genetic diversity. Conservation biology 17**:** 230-237.
Rieseberg, L. H., and J. H. Willis, 2007 Plant speciation. science 317**:** 910-914.
Rosenberg, N. A., and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics 3**:** 380.
Sadhu, M. J., J. S. Bloom, L. Day and L. Kruglyak, 2016 CRISPR-directed mitotic recombination enables genetic mapping without crosses. Science 352**:** 1113-1116.
Sax, K., 1923 The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. Genetics 8**:** 552-560.
Schluter, D., 2001 Ecology and the origin of species. Trends in ecology & evolution 16**:** 372-380.
Schork, A. J., W. K. Thompson, P. Pham, A. Torkamani, J. C. Roddey *et al.*, 2013 All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS genetics 9**:** e1003449.
She, R., and D. F. Jarosz, 2018 Mapping causal variants with single-nucleotide resolution reveals biochemical drivers of phenotypic change. Cell 172**:** 478-490. e415.
Shull, G. H., 1908 The composition of a field of maize. Amer. Breeders Assoc. Rep. 4**:** 296-301.
Slatkin, M., 1982 Pleiotropy and parapatric speciation. Evolution 36**:** 263-270.
Smith, E. N., and L. Kruglyak, 2008 Gene–environment interaction in yeast gene expression. PLoS biology 6**:** e83.
Smith, J. M., 1966 Sympatric speciation. The American Naturalist 100**:** 637-650.
Sohail, M., O. A. Vakhrusheva, J. H. Sul, S. L. Pulit, L. C. Francioli *et al.*, 2017 Negative selection in humans and fruit flies involves synergistic epistasis. Science 356**:** 539-542.
Speed, D., N. Cai, M. R. Johnson, S. Nejentsev, D. J. Balding *et al.*, 2017 Reevaluation of SNP heritability in complex human traits. Nature genetics 49**:** 986.
Starr, T. N., and J. W. Thornton, 2016 Epistasis in protein evolution. Protein Science 25**:** 1204-1218.

Swanson-Wagner, R. A., Y. Jia, R. DeCook, L. A. Borsuk, D. Nettleton *et al.*, 2006 All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. Proc Natl Acad Sci U S A 103**:** 6805-6810.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123**:** 585-595.

Taylor, M. B., and I. M. Ehrenreich, 2015 Higher-order genetic interactions and their contribution to complex traits. Trends in genetics 31**:** 34-40.

Tenaillon, O., J. E. Barrick, N. Ribeck, D. E. Deatherage, J. L. Blanchard *et al.*, 2016 Tempo and mode of genome evolution in a 50,000-generation experiment. Nature 536**:** 165.

Thoday, J. M., 1953 Components of fitness, pp. 113 in *Symp. Soc. Exp. Biol*.

Ting, C.-T., S.-C. Tsaur, M.-L. Wu and C.-I. Wu, 1998 A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 282**:** 1501-1504.

Valladares, F., S. Matesanz, F. Guilhaumon, M. B. Araújo, L. Balaguer *et al.*, 2014 The effects of phenotypic plasticity and local adaptation on forecasts of species range shifts under climate change. Ecology letters 17**:** 1351-1364.

Van Laere, A.-S., M. Nguyen, M. Braunschweig, C. Nezer, C. Collette *et al.*, 2003 A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425**:** 832.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes and A. C. Wilson, 1991 African populations and the evolution of human mitochondrial DNA. Science 253**:** 1503-1507.

Visscher, P. M., R. Thompson and C. S. Haley, 1996 Confidence intervals in QTL mapping by bootstrapping. Genetics 143**:** 1013-1020.

Watson, J. D., and F. H. Crick, 1953 Molecular structure of nucleic acids. Nature 171**:** 737-738.

Wei, X., and J. Zhang, 2017a The genomic architecture of interactions between natural genetic polymorphisms and environments in yeast growth. Genetics 205**:** 925-937.

Wei, X., and J. Zhang, 2017b Why phenotype robustness promotes phenotype evolvability. Genome biology and evolution 9**:** 3509-3515.

Wei, X., L. Zhao, M. Lascoux and D. Waxman, 2015 Population structure and the rate of evolution. J Theor Biol 365**:** 486-495.

Wilkie, A. O., 1994 The molecular basis of genetic dominance. J Med Genet 31**:** 89-98.

Wright, S., 1931 Evolution in Mendelian populations. Genetics 16**:** 97-159.

Wright, S., 1932 *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*. na.

Wünsche, A., D. M. Dinh, R. S. Satterwhite, C. D. Arenas, D. M. Stoebel *et al.*, 2017 Diminishing-returns epistasis decreases adaptability along an evolutionary trajectory. Nature ecology & evolution 1**:** 0061.

Yadav, A., K. Dhole and H. Sinha, 2016 Differential regulation of cryptic genetic variation shapes the genetic interactome underlying complex traits. Genome biology and evolution 8**:** 3559-3573.

Yang, J., A. Bakshi, Z. Zhu, G. Hemani, A. A. Vinkhuyzen *et al.*, 2015 Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nature genetics 47**:** 1114.

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**:** 565.

Zhao, L., M. Lascoux, A. D. Overall and D. Waxman, 2013 The characteristic trajectory of a fixing allele: A consequence of fictitious selection that arises from conditioning. Genetics 195**:** 993-1006.

Chapter 2

Gene by environment interaction: the genomic architecture of interactions between natural genetic polymorphisms and environments in yeast growth

"I was born at the right time and place. I won the Ovarian Lottery."

— **Warren Buffet**

2.1 Abstract

Gene-environment interaction (G×E) refers to the phenomenon that the same mutation has different phenotypic effects in different environments. Although quantitative trait loci (QTLs) exhibiting G×E have been reported, little is known about the general properties of G×E and those of its underlying QTLs. Here we use the genotypes of 1005 segregants from a cross between two *Saccharomyces cerevisiae* strains and the growth rates of these segregants in 47 environments to identify growth rate QTLs (gQTLs) in each environment and QTLs that have different growth effects in each pair of environments (g×eQTLs). The average number of g×eQTLs identified between two environments is 0.58 times the number of unique gQTLs identified in these environments, revealing a high abundance of G×E. Eighty-seven percent of g×eQTLs belong to gQTLs, supporting the practice of identifying g×eQTLs from gQTLs. Most g×eQTLs identified from gQTLs have concordant effects between environments, but as the effect size of a mutation in one environment enlarges, the probability of antagonism in the other

environment increases. Antagonistic g×eQTLs are enriched in dissimilar environments. Relative to gQTLs, g×eQTLs tend to occur at intronic and synonymous sites. The gene ontology distributions of gQTLs and g×eQTLs are significantly different, so are those of antagonistic and concordant g×eQTLs. Simulations based on the yeast data showed that ignoring G×E causes substantial missing heritability. Together, our findings reveal the genomic architecture of G×E in yeast growth and demonstrate the importance of G×E in explaining phenotypic variation and missing heritability.

2.2 Introduction

Gene-environment interaction (G×E) refers to the observation that the same mutation has different phenotypic effects on a trait in different environments (OTTMAN 1996). G×E is believed to be ubiquitous among all organisms and has long been studied in domestic animals and plants, genetic model organisms, and humans. In humans, G×E has been implicated in cancer (THORGEIRSSON *et al.* 2008), inflammatory disorder (CHAMAILLARD *et al.* 2003), immune system diseases (PADYUKOV *et al.* 2004), and mental disorders (RISCH *et al.* 2009; BYRD and MANUCK 2014; LUCK *et al.* 2014). Investigating G×E can help identify the causal pathways of a trait (GAGNEUR *et al.* 2013), dissect genetic tradeoffs (QIAN *et al.* 2012), understand environmental adaptations (OSTROWSKI *et al.* 2005), and reveal a potential cause of "missing heritability" (MANOLIO *et al.* 2009; EICHLER *et al.* 2010).

G×E studies can be generally divided into two types on the basis of the approach used: forward genetics and reverse genetics. In forward genetics, genes or quantitative trait loci (QTLs) that show significantly different phenotypic effects in different environments are identified via linkage or association mapping. In reverse genetics, a mutant carrying a known mutation such as a gene deletion or a point mutation is compared with the wild-type for the trait of interest under

two environments, and G×E is detected when the mutational effect on the trait differs significantly in the two environments. For example, Qian and colleagues measured the fitness effects of single gene deletions in yeast for nearly 5000 nonessential genes in six different environments and identified many antagonistic G×E cases where deleting a gene is deleterious in one environment but beneficial in another (QIAN *et al.* 2012). Although such systematic reverse genetic studies can provide a broad picture of G×E, to date they are limited to gene deletions (DUDLEY *et al.* 2005; BROWN *et al.* 2006; HILLENMEYER *et al.* 2008; QIAN *et al.* 2012), which constitute a special group of mutations. In theory, the reverse genetic approach can also be applied to all natural genetic polymorphisms, but studies of this sort are universally small in scale (OSTROWSKI *et al.* 2005; GERKE *et al.* 2010; DILLON *et al.* 2016) and thus do not offer an overview of G×E for natural genetic polymorphisms. By contrast, large forward genetic analysis in principle allows deciphering general properties of G×E for natural genetic variants.

Many recent forward genetic studies of G×E in humans are driven by the idea of personalized medicine and focus on finding candidate genes and environmental factors that interact in influencing disease, drug response, or behavior (CASPI *et al.* 2002; HOOD *et al.* 2004; CASPI *et al.* 2005; KENDLER *et al.* 2012; BYRD and MANUCK 2014; LUCK *et al.* 2014). Although a number of genes have been reported to interact with environmental factors, the reproducibility of these genome-wide association study (GWAS) results tends to be low (HUNTER 2005; DUNCAN and KELLER 2011), and one likely reason is that environmental factors are hard to control in human studies. The power to detect genetic variants that interact with environments is generally lower than the power to detect genetic variants that have effects in one environment. Furthermore, the detection of interaction is affected by how interaction is measured (DUNCAN and KELLER 2011), because the null hypothesis of no interaction may be based on an additivity

or multiplicity assumption. That is, if the phenotypes of two genotypes are A1 and B1 in environment 1 and A2 and B2 in environment 2, respectively, the null hypothesis of no G×E under additivity is A1−B1 = A2−B2, whereas that under multiplicity is A1/B1 = A2/B2. In model organisms such as the mouse *Mus musculus* and fly *Drosophila melanogaster*, recombinant inbred lines established from a cross between two parental lines are typically used to identify G×E QTLs via linkage mapping (FRY *et al.* 1998; UNGERER *et al.* 2003; LI *et al.* 2006; FLINT and MACKAY 2009; GERKE *et al.* 2010; EL-SODA *et al.* 2014; MATSUI and EHRENREICH 2016). Generally speaking, environments are better controlled, detection power is higher, and the detected interactions are more readily verifiable in model organism studies, compared with human studies.

Although the abundance of G×E has been demonstrated in various model organisms, there is no systematic study about the genomic and functional distributions of G×E QTLs. Furthermore, it is unknown whether G×E is mostly antagonistic (i.e., the same allele has opposite phenotypic effects in two environments) or concordant among natural genetic polymorphisms. It is also unclear how much ignoring G×E impacts the identification of QTLs underlying natural phenotypic variations among individuals that cannot possibly have identical environments. Methodologically, some human studies identify G×E by directly testing if genes with known effects in one environment have different effects in another environment (CASPI *et al.* 2003), instead of testing all pairs of genetic variants by GWAS. Although the former approach has been criticized to have publication bias, low statistical power, and high false discovery rates when compared with GWAS (DUNCAN and KELLER 2011), some authors consider it to be more replicable and superior for finding causal genes (MOFFITT *et al.* 2005; UHER 2014). Which of the two methods performs better depends on the probability that an influential mutation in one

environment has a different effect in another environment. It also depends on the probability that a G×E QTL between two environments has detectable effects in at least one of the environments. But neither of these probabilities is currently known. Here we address all these questions using a recently published dataset of the budding yeast *Saccharomyces cerevisiae*, which includes the genome sequences and the growth rates in 47 environments of 1005 haploid segregants produced by the F1 resulting from a cross between strains BY and RM (BLOOM *et al.* 2013). BY is derived from the commonly used laboratory strain S288c, whereas RM is derived from the vineyard strain RM11-1a. The 47 growth environments varied in temperature, pH, carbon source, metal ions, and small molecules (BLOOM *et al.* 2013). The growth rate of each segregant was measured by the mean end-point colony radius on agar plates. Although a more recently published dataset (BLOOM *et al.* 2015) contained 4390 segregants from the same F1, only 21 environments were examined. We thus focused on the earlier data, which include more environments and hence suit better the study of G×E. We analyzed the later data (BLOOM *et al.* 2015) only to verify the key findings from the earlier data. Note that several yeast studies mapped growth rate QTLs in each of an array of environments (CUBILLOS *et al.* 2011; EHRENREICH *et al.* 2012; BLOOM *et al.* 2013; WILKENING *et al.* 2014) or mapped plasticity QTLs across environments (YADAV *et al.* 2016), but these studies either treated growth rates in different environments as different traits or treated growth rate variance among environments as a phenotypic trait. Hence, yeast G×E in growth rate has not been studied.

2.3 Materials and methods

2.3.1 Genotype and phenotype data

We acquired from the Kruglyak lab the genotype data of 1040 segregants from a cross between the BY and RM strains of *S. cerevisiae*, including a total of 28,220 single nucleotide polymorphisms (SNPs) mapped to the reference genome sequence R64-1-1 (BLOOM *et al.* 2013). We similarly obtained the average end-point colony radius of each segregant in each of the 47 environments (BLOOM *et al.* 2013). After requiring each segregant to have both genotype data and phenotype data in at least one environment, we retained 1005 qualified segregants for subsequent analysis. Narrow-sense heritability data were from the supplementary materials of the original publication (BLOOM *et al.* 2013). We also acquired the genotype and phenotype data from a follow-up study (BLOOM *et al.* 2015) where the growth rates of 4390 segregants from the same cross were similarly measured in 21 of the original 47 environments. We downloaded the cDNA sequences, genome annotations, GO terms, and GO domains from Ensembl biomart for reference R64-1-1, and used Matlab scripts for all enrichment tests.

2.3.2 Mapping growth rate QTLs (gQTLs) in an environment

We started the first round of gQTL mapping using the filtered growth rates as the phenotype. The filtered growth rate of a segregant is its colony radius after 48h growth on agar plates averaged between two replicates, followed by a series of data filtering and correction by the original authors (BLOOM *et al.* 2013). Given an environment, for each SNP, we compared the growth rates between the two groups of segregants that carry the alternative alleles, using a *t*-test. We converted *P*-values to *Q*-values (STOREY and TIBSHIRANI 2003). A stringent *Q*-value of 0.005 was used as the cutoff for statistical significance, on the basis of the simulation described below. On each chromosome, we chose the SNP with the lowest *Q*-value. Sometimes, a chromosome carried multiple SNPs with exactly the same minimal *Q*-values; these were

always adjacent SNPs (i.e., with no intervening SNP), and the middle SNP was chosen. We combined all chosen SNPs from all chromosomes to fit the linear model $\mathbf{Y} = \beta_0 + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$, where **Y** is a vector of the growth rates of all segregants, β_0 is the fitted population mean growth rate, **β** is a vector of gQTL effect sizes, **ε** is an error vector, and **X** is a matrix of genotypes (number of segregants × number of gQTLs). If the allele at a SNP is from BY, the corresponding element in **X** is -1; otherwise, it is 1. We estimated **β**, growth rate residuals, and *t*-statistics from regression using the embedded Matlab function LinearModel. A SNP is removed if its contribution in the linear model is not significant at $P = 0.05$ by a *t*-test. We then used all remaining SNPs to fit a linear model and calculated the growth rate residuals.

We started the second round of gQTL mapping using the growth rate residuals as phenotypes, following the procedure described above. We then combined the SNPs identified from the first two cycles to fit a linear model, removed SNPs with insignificant contribution to the linear model, and calculated growth rate residuals using the remaining SNPs. This process was repeated until no more SNP is added in a cycle of gQTL mapping. In all environments, four or fewer cycles were needed. That is, each chromosome has at most three gQTLs identified in an environment.

2.3.3 Mapping growth rate by environment interaction QTLs (g×eQTLs) in each pair of environments

The 47 environments form 1081 pairs. We first used the identified gQTLs to test G×E (class I g×eQTLs). That is, for a given environment pair and a gQTL identified from one or both of these two environments, we used a genotype's growth rate difference between the two environments as its phenotype and then used a *t*-test to compare the phenotypes of the groups of

genotypes with alternative alleles at the gQTL. $P = 0.05$ from a *t*-test was used to determine whether significant G×E is present for the gQTL; simulation results suggested no need for multiple-testing correction here. Given that on average only 10.3 gQTLs were mapped per environment, we assumed that any two gQTLs that are identified from different environments and lie within 7500 nucleotides from each other (corresponding to the average distance spanned by ~4 genes) have the same underlying causal genetic variant. In such cases, we tested the middle SNP between the two gQTLs for G×E. The justification of the above assumption is as follows. If the gQTLs from two environments are independent from each other and are randomly distributed across the genome, the probability that a gQTL identified in one environment is within 7500 nucleotides from a gQTL identified in the other environment is 1.3%. In fact, an average of 11.0% of gQTLs identified in one environment are within 7500 nucleotides from a gQTL identified in the other environment, suggesting that the vast majority of gQTLs within 7500 nucleotides from each other are not independent but share the same causal mutation.

For each environment pair, we also mapped class II g×eQTLs by considering all SNPs. The method used was the same as mapping gQTLs in an environment, except that growth rate differences between two environments instead of growth rates in one environment were used as phenotypes. We first calculated the difference in end-point colony radius between the two environments for each segregant that has the colony radius measures in both environments, and then followed the same procedure as gQTL mapping to identify class II g×eQTLs. We similarly terminated the search when no more SNP was added to the model. A *Q*-value of 0.005 was used as the cutoff for statistical significance, on the basis of the simulation described below. We counted class II g×eQTLs mapped on chromosomes with no gQTL from either environment. We focused on these chromosomes because it would otherwise be unclear if class I and class II

g×eQTLs reflect the same causal SNPs, owing to strong linkage of SNPs within a chromosome. From the number of class II g×eQTLs on these chromosomes, we extrapolated the number of class II g×eQTLs in the entire genome on the basis of the relative sizes of the chromosomes, under the assumption that class II g×eQTLs are evenly distributed across the genome. Extrapolated class II g×eQTLs were used only to estimate the g×eQTLs missed by class I g×eQTL mapping.

2.3.4 Computer simulation for determining the *Q*-value cutoff

We converted *P*-values to *Q*-values according to the method of Storey and Tibshirani (STOREY and TIBSHIRANI 2003), because it is in theory ~1000 times faster than obtaining *Q*-values from the permutation test used in the original analysis of this dataset (BLOOM *et al.* 2013). We used computer simulation to compare the performance of our method with the one previously used (BLOOM *et al.* 2013) in order to choose a proper *Q*-value cutoff. To save computational time, we simulated three chromosomes instead of all 16 chromosomes in the yeast genome, using parameters appropriate for average-size yeast chromosomes. Each simulated chromosome carried 1500 SNPs, and two recombination events were randomly allocated per chromosome in each segregant on the basis of 90.5 crossovers per yeast meiosis (MANCERA *et al.* 2008). We randomly assigned three SNPs that are >30 SNPs away from one another to be gQTLs. Phenotypic noise is simulated using the standard normal distribution. In the first simulation, each of the three gQTLs has an effect size of 1, and one of the two alleles at a gQTL is randomly picked to be the fitter allele. The narrow-sense heritability $h^2 = 3\times1^2/(3\times1^2+1) = 0.75$. In the second and third simulations, we used the effect size of 0.75 and 0.5, respectively, corresponding to $h^2 = 0.63$ and 0.43, respectively. These h^2 values match approximately the

observed h^2 values in our data. Each simulation generated 1000 segregants. We then mapped gQTLs using different Storey and Tibshirani *Q*-value cutoffs (0.05, 0.02, 0.01, 0.005, 0.002, and 0.001) in our method, and compared our results with those of Bloom et al. (2013) that were based on the permutation *Q*-value of 0.05. The false discovery and false negative rates were estimated for both methods. We found that, for both methods, the false discovery rates were greater than what the *Q*-values suggested, but false negative rates were negligibly small. The false discovery rate of our method under *Q*-value of 0.01 and 0.005 was comparable to that of Bloom et al.'s (2003) method. We thus chose the more stringent *Q*-value cutoff of 0.005 in our mapping.

We also simulated an environment pair with the parameters used above. That is, three gQTLs existed in each environment but they had no effect in the other environment. We then mapped gQTLs with a *Q*-value cutoff of 0.005, followed by class I g×eQTLs mapping with a *P*-value cutoff of 0.05. The obtained results are presented in Table C-1. Because our detection of gQTLs had very low false negative rates (Table C-1), we were not able to study the performance of identifying class II g×eQTLs by our simulation. One type of gQTLs not considered in the above simulation is those that have the same effects in two environments. Such gQTLs could be erroneously identified as g×eQTLs. To examine the probability of this error, we simulated three gQTLs with the same effects in two environments. We found that this type of false positive error hardly increases the overall false discovery rate of g×eQTLs and therefore did not include it in Table C-1.

2.4 Results

2.4.1 Identification of QTLs that interact with environments

Because we aim to identify G×E in all 47×46/2 = 1081 environment pairs, a computationally efficient mapping method is needed. To this end, we developed a customized rapid mapping method with a false discovery rate comparable to that of a previous method (BLOOM *et al.* 2013), and validated its performance by computer simulation (Table C-1; see Materials and Methods). With the new method, we first identified QTLs underlying the among-segregant growth rate variation in each environment using the genotype and phenotype data of the 1005 segregants. The identified QTLs are denoted as gQTLs, where "g" stands for growth rate. We were able to identify gQTLs in 45 of the 47 environments (File C-1). The number of gQTLs ranges from 0 to 22 across the 47 environments, with the mean equal to 10.3. We calculated the similarity between two environments by the across-segregant rank correlation between growth rates in the two environments. The higher the similarity between two environments, the smaller the difference in the number of gQTLs mapped in these environments (Spearman's ρ = -0.26, $P < 10^{-17}$).

We then attempted to identify loci exhibiting G×E (g×eQTLs) for each of the 1081 environment pairs. We used the gQTLs identified from each of the two environments under consideration and tested if a gQTL has significantly different effects in the two environments. This approach is based on the premise that a g×eQTL should have a phenotypic effect (though not necessarily significant) in at least one of the two environments compared. We used this approach rather than directly testing each single nucleotide polymorphism (SNP) for G×E, because the former is expected to have a higher signal to noise ratio such that the identified g×eQTLs are more likely to be genuine. This expectation was confirmed by computer simulation. Specifically, the false discovery rate was lower and the identified g×eQTLs were closer to the causal SNPs when comparing our approach with directly testing all SNPs for G×E

(Tables S1, S2; see Materials and Methods). Nevertheless, if the phenotypic effects of a locus in two environments are both small, the locus may be detected as a gQTL in neither environment. Thus, even if the locus has a significant G×E effect, it may be missed by our approach. To rectify this problem, we also directly mapped G×E for all SNPs but considered only those that are on chromosomes where no gQTL in the relevant environments was found by the first approach (see Materials and Methods). We focused on these chromosomes because it would otherwise be unclear if g×eQTLs identified by the two approaches reflect the same causal SNPs, owing to strong linkage of SNPs within a chromosome, and because the performance in detecting g×eQTLs is better for the first approach than the second approach. The g×eQTLs identified by the two approaches are respectively referred to as class I and class II g×eQTLs. Considering the total length of chromosomes where class II g×eQTLs are considered and the total length of all yeast chromosomes, we extrapolated the expected number of class II g×eQTLs for the entire genome from that of the considered ones. They are respectively referred to as the extrapolated number and the observed number of class II g×eQTLs.

2.4.2 Class I g×eQTLs outnumber class II g×eQTLs

As an example, let us examine the gQTLs respectively identified under two environments: hydrogen peroxide (HydPer) medium and indoleacetic acid (IndAci) medium, as well as the g×eQTLs identified for this pair of environments (Fig2-1A). There are 9 gQTLs identified in HydPer and 13 identified in IndAci. The RM allele is fitter than the BY allele at 13 gQTLs, while the opposite is true at the other 9 gQTLs. We identified 8 class I g×eQTLs and observed 1 class II g×eQTL. Some clear examples of various types of G×E, not necessarily from the above environment pair, are shown in Fig2-1B-F. In these examples, g×eQTLs are found on

chromosomes with at most one mapped gQTL, so the difference in mean growth rate between genotypes of alternative alleles likely represents primarily the g×eQTL effect without influences from linked gQTLs. For instance, Fig2-1B shows a gQTL identified from both 5-fluorouracil (5FluUra) and calcium chloride (CalChl) but with alternative fitter alleles. Not surprisingly, it is a class I antagonistic g×eQTL (i.e., the effects of an allele in the two environments are of opposite directions). Fig2-1C shows a gQTL identified from both 5FluUra and Xylose. Although the RM allele is the fitter allele in both environments, the effect size differs; this QTL is thus a concordant class I g×eQTL (i.e., the effects of an allele in the two environments are of the same direction). Fig2-1D shows a gQTL identified in only one of the two environments (lithium chloride, or LitChl), and it is a class I antagonistic g×eQTL. Fig2-1E shows a gQTL identified in 5FluUra but not in 5-fluorocytosine (5FluCyt), and it does not have a significant G×E effect between the two environments. Fig2-1F shows a locus that is not a gQTL in either 5FluCyt or hydrogen peroxide (HydPer), but is a class II g×eQTL.

The numbers of gQTLs, class I g×eQTLs, and observed class II g×eQTLs found in each 3cM (7500-nucleotide or 4-gene) segment along the yeast genome for all environments and environment pairs considered are presented in Fig2-2. The total number of gQTLs identified from 47 environments in a 3cM segment ranges from 0 to 17 (Fig2-2A). The number of class I g×eQTLs from all environment pairs in a 3cM segment ranges from 0 to 374 (Fig2-2B), while the corresponding number of observed class II g×eQTLs ranges from 0 to 13 (Fig2-2C). The numbers of gQTLs and class I g×eQTLs across 3cM segments are highly correlated (Pearson's $r = 0.901$, $p < 10^{-250}$), while those of gQTLs and class II g×eQTLs are distinct ($r = 0.011$, $p = 0.67$) (Fig2-2). On average, there are 9.2 class I g×eQTLs but only 0.37 observed class II g×eQTLs per environment pair, the former being significantly greater than the latter ($p < 10^{-250}$). The same

trend is observed when extrapolated instead of observed class II g×eQTLs are considered ($p < 10^{-161}$). We tested three genes (*HAP1*, *MKT1*, and *IRA2*) that accounted for much of the deviation from null in a previous gene expression G×E study of the same strain pair between glucose and ethanol environments (SMITH and KRUGLYAK 2008). Interestingly, these genes locate in 3cM segments frequently harboring gQTLs and class I g×eQTLs in our study as well. Specifically, *IRA2*, encoding a GTPase-activating protein that modulates the metaphase to anaphase transition during yeast mitosis (LUO *et al.* 2014), overlaps with the segment that has the highest numbers of gQTLs and class I g×eQTLs among all segments (Fig2-2). All class I g×eQTLs mapped are listed in File C-2.

For each environment pair, we computed the ratio between the number of class I g×eQTLs and the total number of unique gQTLs (i.e., shared gQTLs between the environments are counted only once) identified (Fig2-3A). The ratio averages 0.45 across all environment pairs. Many human studies tested G×E by considering candidate genes that are previously known or predicted to have effects in at least one of the environments compared (DUNCAN and KELLER 2011). Across environment pairs in our data, on average 87% of all g×eQTLs (i.e., class I g×eQTLs plus extrapolated class II g×eQTLs) are class I (Fig2-3B), supporting the validity of this practice. The number of g×eQTLs for a pair of environments is on average 0.58 times the total number of unique gQTLs in these environments (Fig2-3C), indicating the high abundance of G×E.

2.4.3 Antagonistic G×E is uncommon

Previous case studies in *Escherichia coli*, *Drosophila melanogaster*, and *Arabidopsis thaliana* suggested the scarcity of antagonistic G×E involving natural genetic polymorphisms

(FRY *et al.* 1998; EL-SODA *et al.* 2014; DILLON *et al.* 2016), but the data were all small and thus the generality of these observations is unclear. The large yeast data analyzed here appear to show the same pattern. A g×eQTL is considered antagonistic between two environments if the BY allele is fitter than the RM allele in one environment while the RM allele is fitter than the BY allele in the other environment, even if the difference is statistically significant in neither environment. Otherwise, the g×eQTL is considered concordant between the two environments. Thus, purely by chance, we would expect a g×eQTL to be equally likely to be antagonistic and concordant. However, on average only 28% of class I g×eQTLs are antagonistic, significantly lower than the null expectation ($P < 10^{-250}$, binomial test; Fig2-4A). Among the observed class II g×eQTLs, 94% are antagonistic, which is not unexpected, because a concordant g×eQTL should have a significant effect in at least one of the environments and thus is unlikely to be of class II. Because class I g×eQTLs substantially outnumber class II g×eQTLs (Fig2-2), only 37% of all g×eQTLs are antagonistic ($P < 10^{-171}$, binomial test), under the assumption that antagonism is equally frequent among the observed and extrapolated class II g×eQTLs.

2.4.4 Large-effect QTLs are more likely than small-effect QTLs to be antagonistic

A previous study of yeast gene deletions identified many antagonisms between environments (QIAN *et al.* 2012), seemingly contrasting the scarcity of antagonism of natural polymorphisms surveyed in the present study. Because gene deletions should on average have larger phenotypic effects than natural polymorphisms, a potential explanation of the disparity in the frequency of antagonism may be that large-effect mutations are more likely than small-effect mutations to be antagonistic. To directly test this hypothesis, for each gQTL, we counted the number of environments where its effect is opposite to the effect in the environment where the

gQTL was detected. Indeed, the larger the effect of a gQTL, the higher the likelihood that it has an antagonistic effect in another environment ($\rho = 0.14$, $P < 10^{-4}$; Fig2-4B).

2.4.5 Prevalence of antagonism varies among environments

To study whether antagonism is enriched in certain environments, for each pair of environments, we calculated the fraction of class I g×eQTLs that are antagonistic. If this fraction is 0, we say that this pair of environments is non-antagonistic to each other. Similarly, if this fraction ≥ 0.5, these two environments are highly antagonistic to each other. We counted the number of times that each environment is said to be non-antagonistic and the number of times that it is said to be highly antagonistic to another environment. We then respectively computed the mean number of times that an environment is non-antagonistic and the mean number of times that an environment is highly antagonistic. Environments showing two or more times the mean number of non-antagonism are galactose, caffeine, 4-hydroxybenzaldehyde, calcium chloride, mannose, menadione, and YNB (Fig2-4C), whereas those exhibiting two or more times the mean number of high antagonism are cadmium chloride, copper, hydrogen peroxide, and cycloheximide (Fig2-4D). A potential explanation of the among-environment variation in the prevalence of antagonism is that antagonisms may have been resolved by natural selection in commonly encountered environments but not so in rarely encountered environments (QIAN *et al.* 2012). However, to what extend the environments in Fig2-4C are more common than the environments in Fig2-4D is unknown, due to the paucity of the ecological information of yeast. Another possibility, non-mutually exclusive from the above, is that some environments are more dissimilar to other environments and hence exhibit more antagonism. In support of the latter

hypothesis, the fraction of antagonistic class I g×eQTLs between two environments negatively correlates with their environment similarity (ρ = -0.61, $P < 10^{-110}$).

2.4.6 Distributions of gQTLs and g×eQTLs across the genome

To understand the molecular basis of G×E, we first categorized all 28,220 SNPs between BY and RM strains into coding SNPs, intronic SNPs, and intergenic SNPs. We merged gQTLs from all environments and merged class I g×eQTLs from all environment pairs. A gQTL or g×eQTL is counted as many times as it appears in the merged list. Table 2-1 summarizes the results of enrichment tests for each genomic category. Compared with all SNPs, gQTLs are not significantly different in frequency distribution among coding, intronic, and intergenic regions (Table 2-1). Relative to gQTLs, class I g×eQTLs are two-fold more likely to be in introns ($P = 2.2\times10^{-7}$; Table 2-1), suggesting that yeast introns are more important in regulating environment-dependent growth rates than environment-independent growth rates.

We also analyzed the distributions of gQTLs and g×eQTLs among synonymous, nonsynonymous, and nonsense SNPs within coding regions. A synonymous SNP does not alter the amino acid encoded by the codon where the SNP resides, whereas a nonsynonymous SNP alters the amino acid. A nonsense SNP changes a sense codon in one strain to a stop codon in another. Relative to all SNPs, gQTLs are more likely to occur at nonsynonymous SNPs (1.125 fold, $P = 0.03$) and are less likely to occur at synonymous SNPs (0.896 fold, $P = 0.02$). This observation is not unexpected, because nonsynonymous mutations are more likely than synonymous mutations to have phenotypic effects. Relative to gQTLs, g×eQTLs are more likely to occur at synonymous SNPs (1.070 fold, $P = 9.6\times10^{-9}$), but are less likely to occur at nonsynonymous (0.935 fold, $P = 2.5\times10^{-7}$) and nonsense (0.826 fold, $P = 0.0265$) SNPs,

suggesting that nonsynonymous and nonsense mutations tend to have universal rather than environment-specific growth effects, when compared with synonymous mutations. Among all g×eQTLs, we analyzed only class I g×eQTLs here, because the number of class II g×eQTLs is small and because our simulation (Table C-1) showed that mapping is less precise for class II g×eQTLs.

Note that because the gQTLs and g×eQTLs identified may not be causal SNPs but are simply linked with causal SNPs, the above analysis has a lower statistical power than when causal SNPs are used in the analysis. In our simulation, >31% of gQTLs and >29% of class I g×eQTLs are mapped to causal SNPs (Table C-1), suggesting that a sizable proportion of mapped sites are causal, explaining why our test is not entirely powerless. Thus, the significant results obtained are likely to be genuine and the conclusions conservative.

2.4.7 Different GO distributions of gQTLs and g×eQTLs

Gene ontology (GO) annotation is organized into three domains: cellular component, molecular function, and biological process (ASHBURNER *et al.* 2000). Each domain contains many GO terms, which may be a word or string of words related to gene function. A gene is annotated for all three domains and one to many terms in each domain on the basis of its product and function. We examined the enrichment of gQTLs and g×eQTLs for GO domains and terms (Table 2-2). Note that intergenic SNPs were assigned to their closest genes. We compared gQTLs to the background of all SNPs and compared class I g×eQTLs to the background of all gQTLs, using binomial tests followed by Bonferroni corrections with a corrected $P = 0.05$ as the cutoff. Compared with all SNPs, gQTLs are not enriched in any GO domain but are significantly enriched in 24 GO terms (File C-3). gQTLs are not underrepresented in any GO domain or GO

term. These results suggest that gQTLs are overall annotated with more functions than average SNPs. Relative to gQTLs, class I g×eQTLs are enriched in the GO domain cellular component ($P = 0.028$), suggesting that proteins encoded by g×eQTLs have relatively more locations in the cell or are relatively better annotated for cellular component. Class I g×eQTLs are significantly underrepresented in biological process ($P = 4.2\times10^{-6}$) and molecular function ($P = 3\times10^{-6}$), when compared with gQTLs. Strikingly, of the 848 GO terms that contain at least one gQTL, g×eQTLs are enriched in 137 of them and are underrepresented in 139 (File C-3). Of the GO terms enriched in gQTLs, 4 terms are further enriched in g×eQTLs (Table 2-2), and four are underrepresented. Thus, the functional distributions of gQTLs and class I g×eQTLs are quite different, despite that the latter constitutes a large subset of the former. One potential bias in the above GO enrichment analysis of gQTLs is that SNPs are not evenly distributed along genes and chromosomes. To rectify this problem, we also tested GO enrichment of gQTLs against all genes instead of all SNPs, by assigning each gQTL to its closest gene. The enriched GO terms (File C-4), however, remained largely the same.

2.4.8 Antagonistic and concordant g×eQTLs have different genomic and functional enrichments

Comparing antagonistic and concordant class I g×eQTLs, we found no significant difference in their frequency distributions among coding, intronic, and intergenic regions (Table 2-3). However, within coding regions, antagonistic g×eQTLs are enriched at synonymous ($P = 9.7\times10^{-9}$, chi-squared test) and nonsense SNPs ($P = 2.7\times10^{-7}$) but underrepresented at nonsynonymous SNPs ($P = 7.6\times10^{-13}$), when compared with concordant g×eQTLs.

Antagonistic and concordant g×eQTLs show significantly different enrichments for two GO domains, biological process (adjusted $P = 4.2\times10^{-4}$, chi-squared test; File C-5) and cellular component (adjusted $P = 1.4\times10^{-6}$). They are also significantly different in 187 of 907 GO terms that have at least one occurrence in class I g×eQTLs (File C-5). Interestingly, two (ribosomal small subunit biogenesis and 90S preribosome) of the five GO terms significantly enriched in both gQTLs and g×eQTLs are the top two terms that differ significantly between antagonistic and concordant g×eQTLs; they each occur 325 times in concordant g×eQTLs but 0 time in antagonistic g×eQTLs. This result suggests that, although differences in translation underlie g×eQTLs, these differences mostly have concordant G×E effects.

2.4.9 Ignoring G×E causes missing heritability

"Missing heritability" refers to the gap between the phenotypic variance explained by GWAS results and those estimated from classical heritability methods (ZAITLEN and KRAFT 2012) and is a prominent problem in the study of human complex traits that has attracted much attention (MANOLIO *et al.* 2009; EICHLER *et al.* 2010). G×E has been proposed as a potential cause for the missing heritability problem (MANOLIO *et al.* 2009; EICHLER *et al.* 2010). Because heritability is classically estimated from relatives such as by comparing monozygotic (MZ) and dizygotic (DZ) twins, the effect of environmental heterogeneity for a twin is canceled in the comparison between MZ and DZ twins and has no effect on the heritability estimate. However, in human GWAS, the environmental effect and G×E effect are rarely controlled, which could lower the power in identifying the underlying genetic variants and render the estimation of effect size inaccurate. To quantitatively evaluate the contribution of ignoring G×E to the missing heritability problem, we conducted a simulation using the yeast data. That is, for one half of the

segregants, we used their phenotypes measured in one environment, but for the other half of the segregants, we used their phenotypes measured in another environment. We then attempted to identify gQTLs as if all segregants were phenotyped in the same environment. We did this simulation for 100 random pairs of environments. An example is provided in Fig2-5A, where the phenotype data are from YNB at 30°C and YPD at 37°C. Ten and eight gQTLs were identified from 1005 segregants in YNB and YPD, respectively. But only two gQTLs were identified from the mixture of the phenotype data of 502 segregants in YNB and 503 segregants in YPD, although these two gQTLs are a subset of the 18 gQTLs identified from the individual environments. When the phenotype data of the 1005 segregants are all from either YNB or YPD but not both, the identified gQTLs together can explain on average 54% of the total phenotypic variance observed among the segregants. This number reduces to 26% when the mixed phenotype data are used (green dots in Fig2-5A). To distinguish between the environmental effect and G×E effect on gQTL identification, we conducted another analysis, in which the phenotypic value of a segregant in an environment is defined by the difference between its raw phenotypic value and the mean phenotypic value of all segregants in that environment. We then mixed these normalized phenotypic values from two environments to identify gQTLs. We found that such normalization improves gQTL identification, because the number of gQTLs identified rises to six, although this number is still smaller than when homogenous data are used. The total variance of normalized phenotypes explained rises to 42%. The remaining difference between this result (light salmon symbols in Fig2-5A) and the original result (blue and red symbols in Fig2-5A) is attributable to G×E.

On average across the 100 random pairs of environments, the identified gQTLs explain 40% of the total phenotypic variance among segregants under one environment. When mixed

phenotypic data from two environments are used, this number drops to 10% (Fig2-5B). When phenotypic data are normalized by the mean phenotypic value of the environment, the fraction of phenotypic variance explained is 23% (Fig2-5B). Hence, in this dataset, environmental effects and G×E effects have similar amounts of contribution to missing heritability. We also conducted 100 simulations where the phenotype data are generated from 5 and 10 environments, respectively. As the number of environments increases, the amount of missing heritability rises, the contribution of G×E to missing heritability increases, and the contribution of environmental effects decreases (Fig2-5B).

We further calculated the distances between the gQTLs identified using the mixed phenotypes from two environments and the nearest gQTLs identified using phenotypes from individual environments for all 100 random pairs of environments (Fig2-5C**)**. We found that although noise is larger in mixed environments, the identified sites are generally closely linked to the gQTLs identified from individual environments. This is true both with and without controlling the environmental effect. What types of gQTLs are under-detected using mixed phenotype data? On the basis of the same 100 pairs of environments examined, we found that on average 23.6% of gQTLs having the same direction of effect in the two environments and 12.7% of gQTLs having opposite directions of effect were detected using the mixed data, when the environmental effect is uncontrolled ($P = 7.1\times10^{-14}$, *t*-test of equal probability of detection for the two groups of gQTLs). These numbers increase to 52.0% and 33.8%, respectively, upon the control of the environmental effect ($P < 8.5\times10^{-6}$). Thus, while all gQTLs are under-detected using mixed phenotype data, those with opposite effects in the two environments suffer more than those with the same direction of effect.

In human GWAS, larger and larger samples are being used despite that enlarging samples likely increase environmental heterogeneity of the sample. To study this effect, we merged the phenotype data from all 47 environments, resulting in a sample of 42,781 individuals; this number is lower than 47×1005 = 47,325 because not all 1005 individuals had growth data in all 47 environments. Using this very large sample, we were able to mapped 21 gQTLs, more than the number of gQTLs mapped from any one of the 47 environments. Some of the mapped gQTLs overlapped with the gQTLs frequently identified in individual environments (FigC-1), suggesting that using large samples in GWAS might help identify influential loci that have effects in multiple environments. Nevertheless, the fraction of phenotypic variance explained by all mapped sites is only 2.5%, similar to that when a sample of 1005 segregants, each fifth originating from a different environment, is used, and much lower than that when a sample of 1005 segregants from the same environment is used (Fig2-5B). Clearly, the "missing heritability" problem worsens when enlarging samples also increases environmental heterogeneity.

2.5 Discussion

We conducted a systematic analysis of interaction between natural genetic variants and environments in yeast growth, and identified numerous g×eQTLs. The average number of g×eQTLs identified between two environments is 0.58 times the number of unique gQTLs identified in the two environments, indicating a high abundance of G×E. It is debated whether testing all pairs of SNPs or testing only those with effects in at least one of the environments concerned is more suitable for G×E detection (DUNCAN and KELLER 2011; UHER 2014). Our computer simulation showed that using the latter approach has the benefit of lowering the false discovery rate and increasing the chance of finding causal variants. Although our simulation

also indicated that the latter approach has a higher false negative rate than the former approach, our yeast data analysis found that 88% of g×eQTLs could be identified from gQTLs. Similar results were obtained when the larger data of Bloom et al. (2015) were analyzed (FigC-2). Together, these findings support the current practice in human genetics of using genes or QTLs known to have effects in at least one of the environments concerned as candidates in the study of G×E. The gQTL mapping method and G×E detection method developed here are expected to suit other similar large-scale studies of G×E. In our computer simulation, we found that both Storey and Tibshirani *Q*-value (STOREY and TIBSHIRANI 2003) and permutation *Q*-value (DOERGE and CHURCHILL 1996) underestimate the false discovery rate. This underestimation may be a general problem in linkage mapping of complex traits, suggesting the importance of using computer simulation to assess false discovery rates.

We found that most G×E interactions are concordant, suggesting that the fitness landscapes in different environments examined are positively correlated such that a mutation that is beneficial in one tested environment tends to be beneficial in other tested environments. Nevertheless, we detected a few environments with unusually high degrees of antagonistic G×E, such as those with trace minerals or heavy metals. Because we observed a negative correlation between the fraction of antagonistic g×eQTLs and environmental similarity, it is likely that these antagonism-rich environments are relatively dissimilar to the other environments examined. The antagonism-rich environments may also be rarely encountered by yeast in nature such that antagonism has not had chance to be resolved by natural selection. We did not attempt to verify the disparity in antagonism among environments using the data of Bloom et al. (2015), because only 3 of the 11 environments in Fig2-4C and D are included in this dataset. The fact that the extent of antagonism depends on the tested environments illustrates the importance in carefully

choosing environments in testing the potential antagonism of beneficial mutations observed in experimental evolution (OSTROWSKI *et al.* 2005; WENGER *et al.* 2011; BEDHOMME *et al.* 2012; DILLON *et al.* 2016).

We observed that large-effect gQTLs identified in one environment are more likely than small-effect gQTLs to have antagonistic effects in another environment, reminiscent of the common belief and a prediction of Fisher's geometric model (FISHER 1930) that large-effect mutations are more likely than small-effect mutations to be deleterious. Our observation predicts that the prevalence of detected antagonism will decrease with the power of g×eQTL mapping, because, as the power increases, g×eQTLs of smaller and smaller effects are mapped. This prediction is confirmed by using the larger data from Bloom et al. (2015), where the fraction of antagonistic g×eQTLs is found to be even lower (FigC-3). Note, however, that we studied growth rate, a primary component of fitness, in this work. For traits that are irrelevant to fitness, antagonism patterns may be different because they are not subject to natural selection.

We tested the enrichment of different functional sites of the yeast genome as well as different GO categories in g×eQTLs and gQTLs. We found that gQTLs are enriched with nonsynonymous SNPs, similar to the collective finding from human GWAS studies (HINDORFF *et al.* 2009). Relative to gQTLs, g×eQTLs are more likely to occur at intronic SNPs. We confirmed the enrichment of nonsynonymous SNPs in gQTLs and enrichment of intronic SNPs in g×eQTLs (Table C-3) using the data from Bloom et al. (2015). Concordant and antagonistic g×eQTLs also have different distributions among the three categories of coding SNPs, with concordant g×eQTLs enriched at nonsynonymous SNPs and antagonistic g×eQTLs enriched at synonymous and nonsense SNPs. Bloom et al.'s (2015) data showed the same patterns except that the distribution of nonsense SNPs is not significantly different between concordant and

antagonistic g×eQTLs (Table C-4). These results suggest different molecular basis of concordant and antagonistic G×E. We also found g×eQTLs to be enriched in GO terms on ribosome and translation (Table 2-2), which is potentially related to the aforementioned enrichment in introns, because introns are concentrated in ribosomal protein genes in yeast (PARENTEAU *et al.* 2011). The correlation between ribosomal protein gene expression and growth rate is well known (MAGER and PLANTA 1991), and the comparisons between gQTLs and g×eQTLs and between antagonistic and concordant g×eQTLs using the data from Bloom et al. (2015) suggest the possibility that intronic SNPs affect ribosomal protein gene expression, which potentially affects growth rate differently in different environments. Specifically, introns from four genes (*TUB3*, *PFY1*, *RPL34B*, and *RPL40B*) are found to harbor gQTLs. While concordant intronic g×eQTLs are found in all of the four genes, antagonistic intronic g×eQTLs are found only in the two ribosomal protein genes (*RPL34B* and *RPL40B*). Using the data of Bloom et al. (2015), we found that 38 GO terms are enriched in gQTLs while only 1 GO term is underrepresented, confirming that gQTLs are overall annotated with more functions than average SNPs.

Our yeast data-based simulation of mixed environments revealed the importance of considering G×E in QTL mapping and by extension association studies. Neglecting environmental heterogeneity in the data substantially reduces the number of QTLs identified and results in missing heritability. Many human genetic association studies ignore the fact that different individuals have different environments, and our results suggest that failure to account for environmental heterogeneity could be a primary reason underlying the missing heritability phenomenon. Another commonly cited cause of missing heritability is epistasis, or gene by gene interaction (G×G). But recent studies found that failure to consider G×G is not a primary cause

of missing heritability (BLOOM *et al.* 2013; BLOOM *et al.* 2015). In model organism studies, where the environment tends to be well controlled, missing heritability tends to be mild. But, in human GWAS, where environments are hard to control, missing heritability is severe (EICHLER *et al.* 2010). This contrast, coupled with our simulation results, suggests that missing heritability in human GWAS may be primarily due to ignoring environmental factors and/or G×E. We showed in our simulation that using very large samples could help identify more influential loci when compared with small samples of environmental homogeneity, but the "missing heritability" problem is exacerbated if enlarging samples means increasing the environmental heterogeneity of the sample. Although it is impossible to have different human individuals living in exactly the same environment, even partially controlling environments helps identify disease-associated alleles. For example, in GWAS of type II diabetes, controlling for obesity in statistical analysis helps identify new disease-associated variants (ZEGGINI *et al.* 2008). This kind of controlling of environmental/physiological factors will help identify new trait-associated genetic variants and reduce missing heritability. Notwithstanding, because classical estimation of heritability is minimally affected by environmental heterogeneity while modern GWAS is subject to potentially high environmental heterogeneity, the "missing heritability" due to this difference may be considered fictional (HECKERMAN *et al.* 2016). Better estimation of heritability by considering environmental heterogeneity will help gauge the true missing heritability in GWAS (HECKERMAN *et al.* 2016).

2.6 Acknowledgements

We thank the Kruglyak lab for sharing the yeast segregant genotype and phenotype data and Soochin Cho, Wei-Chin Ho, Chuan Li, Jian-Rong Yang, and two anonymous reviewers for

valuable comments. This work was supported by the U.S. National Institutes of Health research grant R01GM103232 to J.Z.

2.7 References

ASHBURNER, M., C. A. BALL, J. A. BLAKE, D. BOTSTEIN, H. BUTLER *et al.*, 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25:** 25-29.

BEDHOMME, S., G. LAFFORGUE and S. F. ELENA, 2012 Multihost experimental evolution of a plant RNA virus reveals local adaptation and host-specific mutations. Mol Biol Evol **29:** 1481-1492.

BLOOM, J. S., I. M. EHRENREICH, W. T. LOO, T. L. LITE and L. KRUGLYAK, 2013 Finding the sources of missing heritability in a yeast cross. Nature **494:** 234-237.

BLOOM, J. S., I. KOTENKO, M. J. SADHU, S. TREUSCH, F. W. ALBERT *et al.*, 2015 Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nat Commun **6:** 8712.

BROWN, J. A., G. SHERLOCK, C. L. MYERS, N. M. BURROWS, C. DENG *et al.*, 2006 Global analysis of gene function in yeast by quantitative phenotypic profiling. Mol Syst Biol **2:** 2006 0001.

BYRD, A. L., and S. B. MANUCK, 2014 MAOA, childhood maltreatment, and antisocial behavior: Meta-analysis of a gene-environment interaction. Biol Psychiatry **75:** 9-17.

CASPI, A., J. MCCLAY, T. E. MOFFITT, J. MILL, J. MARTIN *et al.*, 2002 Role of genotype in the cycle of violence in maltreated children. Science **297:** 851-854.

CASPI, A., T. E. MOFFITT, M. CANNON, J. MCCLAY, R. MURRAY *et al.*, 2005 Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the catechol-O-methyltransferase gene: longitudinal evidence of a gene X environment interaction. Biol Psychiatry **57:** 1117-1127.

CASPI, A., K. SUGDEN, T. E. MOFFITT, A. TAYLOR, I. W. CRAIG *et al.*, 2003 Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science **301:** 386-389.

CHAMAILLARD, M., D. PHILPOTT, S. E. GIRARDIN, H. ZOUALI, S. LESAGE *et al.*, 2003 Gene-environment interaction modulated by allelic heterogeneity in inflammatory diseases. Proc Natl Acad Sci U S A **100:** 3455-3460.

CUBILLOS, F. A., E. BILLI, E. ZÖRGÖ, L. PARTS, P. FARGIER *et al.*, 2011 Assessing the complex architecture of polygenic traits in diverged yeast populations. Mol Ecol **20:** 1401-1413.

DILLON, M. M., N. P. ROUILLARD, B. VAN DAM, R. GALLET and V. S. COOPER, 2016 Diverse phenotypic and genetic responses to short-term selection in evolving Escherichia coli populations. Evolution **70:** 586-599.

DOERGE, R. W., and G. A. CHURCHILL, 1996 Permutation tests for multiple loci affecting a quantitative character. Genetics **142:** 285-294.

DUDLEY, A. M., D. M. JANSE, A. TANAY, R. SHAMIR and G. M. CHURCH, 2005 A global view of pleiotropy and phenotypically derived gene function in yeast. Mol Syst Biol **1:** 2005 0001.

DUNCAN, L. E., and M. C. KELLER, 2011 A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. Am J Psychiatry **168:** 1041-1049.

EHRENREICH, I. M., J. BLOOM, N. TORABI, X. WANG, Y. JIA *et al.*, 2012 Genetic architecture of highly complex chemical resistance traits across four yeast strains. PLoS Genet **8:** e1002570.

EICHLER, E. E., J. FLINT, G. GIBSON, A. KONG, S. M. LEAL *et al.*, 2010 Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet **11:** 446-450.

EL-SODA, M., M. MALOSETTI, B. J. ZWAAN, M. KOORNNEEF and M. G. AARTS, 2014 Genotypexenvironment interaction QTL mapping in plants: lessons from Arabidopsis. Trends Plant Sci **19:** 390-398.

FISHER, R. A., 1930 *The Genetic Theory of Natural Selection*. Clarendon, Oxford.

FLINT, J., and T. F. MACKAY, 2009 Genetic architecture of quantitative traits in mice, flies, and humans. Genome Res **19:** 723-733.

FRY, J. D., S. V. NUZHDIN, E. G. PASYUKOVA and T. F. MACKAY, 1998 QTL mapping of genotype-environment interaction for fitness in Drosophila melanogaster. Genet Res **71:** 133-141.

GAGNEUR, J., O. STEGLE, C. ZHU, P. JAKOB, M. M. TEKKEDIL *et al.*, 2013 Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. PLoS Genet **9:** e1003803.

GERKE, J., K. LORENZ, S. RAMNARINE and B. COHEN, 2010 Gene-environment interactions at nucleotide resolution. PLoS Genet **6:** e1001144.

HECKERMAN, D., D. GURDASANI, C. KADIE, C. POMILLA, T. CARSTENSEN *et al.*, 2016 Linear mixed model for heritability estimation that explicitly addresses environmental variation. Proc Natl Acad Sci U S A **113:** 7377-7382.

HILLENMEYER, M. E., E. FUNG, J. WILDENHAIN, S. E. PIERCE, S. HOON *et al.*, 2008 The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science **320:** 362-365.

HINDORFF, L. A., P. SETHUPATHY, H. A. JUNKINS, E. M. RAMOS, J. P. MEHTA *et al.*, 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A **106:** 9362-9367.

HOOD, L., J. R. HEATH, M. E. PHELPS and B. LIN, 2004 Systems biology and new technologies enable predictive and preventative medicine. Science **306:** 640-643.

HUNTER, D. J., 2005 Gene-environment interactions in human diseases. Nat Rev Genet **6:** 287-298.

KENDLER, K. S., K. SUNDQUIST, H. OHLSSON, K. PALMÉR, H. MAES *et al.*, 2012 Genetic and familial environmental influences on the risk for drug abuse: a national Swedish adoption study. Arch Gen Psychiatry **69:** 690-697.

LI, Y., O. A. ALVAREZ, E. W. GUTTELING, M. TIJSTERMAN, J. FU *et al.*, 2006 Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genet **2:** e222.

LUCK, T., S. RIEDEL-HELLER, M. LUPPA, B. WIESE, M. KÖHLER *et al.*, 2014 Apolipoprotein E epsilon 4 genotype and a physically active lifestyle in late life: analysis of gene–environment interaction for the risk of dementia and Alzheimer's disease dementia. Psychol Med **44:** 1319-1329.

LUO, G., J. KIM and K. SONG, 2014 The C-terminal domains of human neurofibromin and its budding yeast homologs Ira1 and Ira2 regulate the metaphase to anaphase transition. Cell Cycle **13:** 2780-2789.

MAGER, W. H., and R. J. PLANTA, 1991 Coordinate expression of ribosomal protein genes in yeast as a function of cellular growth rate, pp. 181-187 in *Molecular Mechanisms of Cellular Growth*. Springer.

MANCERA, E., R. BOURGON, A. BROZZI, W. HUBER and L. M. STEINMETZ, 2008 High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature **454:** 479-485.

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:** 747-753.

MATSUI, T., and I. M. EHRENREICH, 2016 Gene-Environment Interactions in Stress Response Contribute Additively to a Genotype-Environment Interaction. PLoS Genet **12:** e1006158.

MOFFITT, T. E., A. CASPI and M. RUTTER, 2005 Strategy for investigating interactions between measured genes and measured environments. Arch Gen Psychiatry **62:** 473-481.

OSTROWSKI, E. A., D. E. ROZEN and R. E. LENSKI, 2005 Pleiotropic effects of beneficial mutations in Escherichia coli. Evolution **59:** 2343-2352.

OTTMAN, R., 1996 Gene-environment interaction: definitions and study designs. Prev Med **25:** 764-770.

PADYUKOV, L., C. SILVA, P. STOLT, L. ALFREDSSON and L. KLARESKOG, 2004 A gene–environment interaction between smoking and shared epitope genes in HLA–DR provides a high risk of seropositive rheumatoid arthritis. Arthritis Rheum **50:** 3085-3092.

PARENTEAU, J., M. DURAND, G. MORIN, J. GAGNON, J.-F. LUCIER *et al.*, 2011 Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. Cell **147:** 320-331.

QIAN, W., D. MA, C. XIAO, Z. WANG and J. ZHANG, 2012 The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. Cell Rep **2:** 1399-1410.

RISCH, N., R. HERRELL, T. LEHNER, K.-Y. LIANG, L. EAVES *et al.*, 2009 Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. JAMA **301:** 2462-2471.

SMITH, E. N., and L. KRUGLYAK, 2008 Gene–environment interaction in yeast gene expression. PLoS Biol **6:** e83.

STOREY, J. D., and R. TIBSHIRANI, 2003 Statistical significance for genomewide studies. Proc Natl Acad Sci U S A **100:** 9440-9445.

THORGEIRSSON, T. E., F. GELLER, P. SULEM, T. RAFNAR, A. WISTE *et al.*, 2008 A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature **452:** 638-642.

UHER, R., 2014 Gene-environment interactions in common mental disorders: an update and strategy for a genome-wide search. Soc Psychiatry Psychiatr Epidemiol **49:** 3-14.

UNGERER, M. C., S. S. HALLDORSDOTTIR, M. D. PURUGGANAN and T. F. MACKAY, 2003 Genotype-environment interactions at quantitative trait loci affecting inflorescence development in Arabidopsis thaliana. Genetics **165:** 353-365.

WENGER, J. W., J. PIOTROWSKI, S. NAGARAJAN, K. CHIOTTI, G. SHERLOCK *et al.*, 2011 Hunger artists: yeast adapted to carbon limitation show trade-offs under carbon sufficiency. PLoS Genet **7:** e1002202.

WILKENING, S., G. LIN, E. S. FRITSCH, M. M. TEKKEDIL, S. ANDERS *et al.*, 2014 An evaluation of high-throughput approaches to QTL mapping in Saccharomyces cerevisiae. Genetics **196:** 853-865.

YADAV, A., K. DHOLE and H. SINHA, 2016 Genetic regulation of phenotypic plasticity and canalisation in yeast growth. PLoS One **11:** e0162326.

ZAITLEN, N., and P. KRAFT, 2012 Heritability in the genome-wide association era. Hum Genet **131:** 1655-1664.

ZEGGINI, E., L. J. SCOTT, R. SAXENA, B. F. VOIGHT, J. L. MARCHINI *et al.*, 2008 Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet **40:** 638-645.

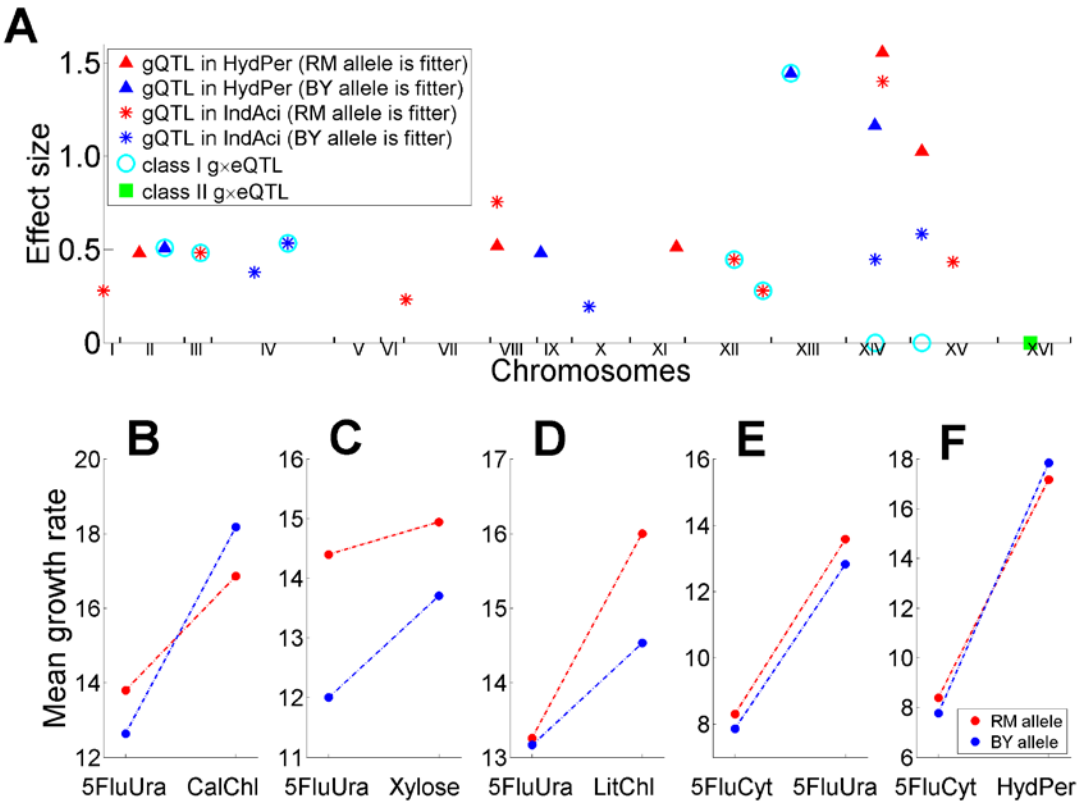


Figure 2-1. Examples of gQTLs and g×eQTLs. (A) Genomic distributions of detected gQTLs in HydPer and IndAci and g×eQTLs between the two environments. The effect size of a gQTL under the environment where it is identified is shown on the Y-axis, while its genomic position is shown on the X-axis. A class I g×eQTL is circled at the triangle if it is a gQTL only in HydPer and circled at the star if it is a gQTL only in IndAci, but is circled on the X-axis if it is a gQTL in both environments. Observed class II g×eQTLs are indicated on the X-axis. (B)-(F) Mean growth rates of segregants carrying the two alternative alleles at various gQTLs or g×eQTLs. Standard errors are too small to see. Panel (B) shows a class I antagonistic g×eQTL that is a gQTL (SNP: 24637) in both 5FluUra and CalChl. Panel (C) shows a class I concordant g×eQTL (SNP: 24651) that is a gQTL in both 5FluUra and Xylose. Panel (D) shows a class I g×eQTL that is a gQTL (SNP: 4821) in LitChl but not 5FluUra. Panel (E) shows a gQTL (SNP: 2277) in

5FluUra that does not show significant G×E. Panel (F) shows a class II antagonistic g×eQTL (SNP: 3512), which is a gQTL in neither 5FluCyt nor HydPer.

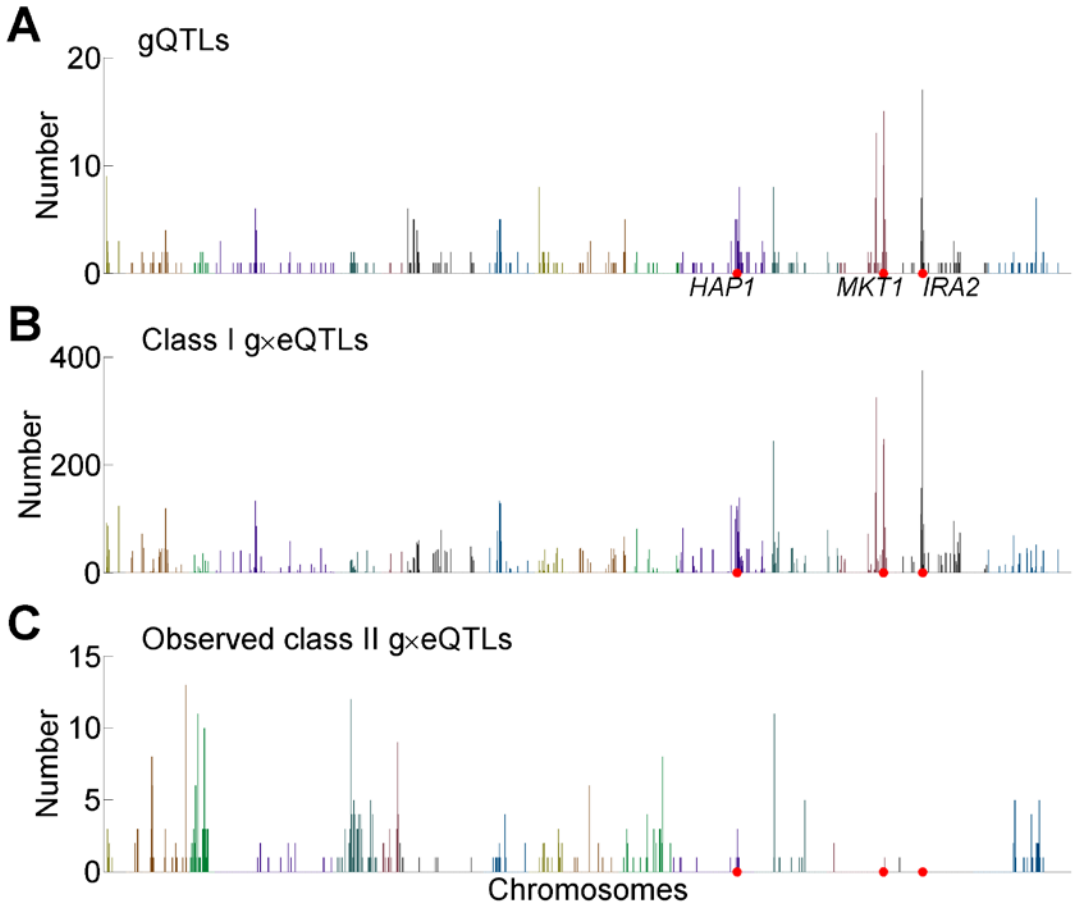


Figure 2-2. Genomic distributions of (A) gQTLs, (B) class I g×eQTLs, and (C) observed class II g×eQTLs. The genome is divided into 7500-nucleotide bins. The total number of gQTLs from all 47 environments, the total number of class I g×eQTLs from all 1081 pairs of environments, and the total number of observed class II g×eQTLs from all 1081 pairs of environments are plotted for each bin. The 16 chromosomes are colored differently. Three genes referred to in the main text are marked according to their genomic locations.

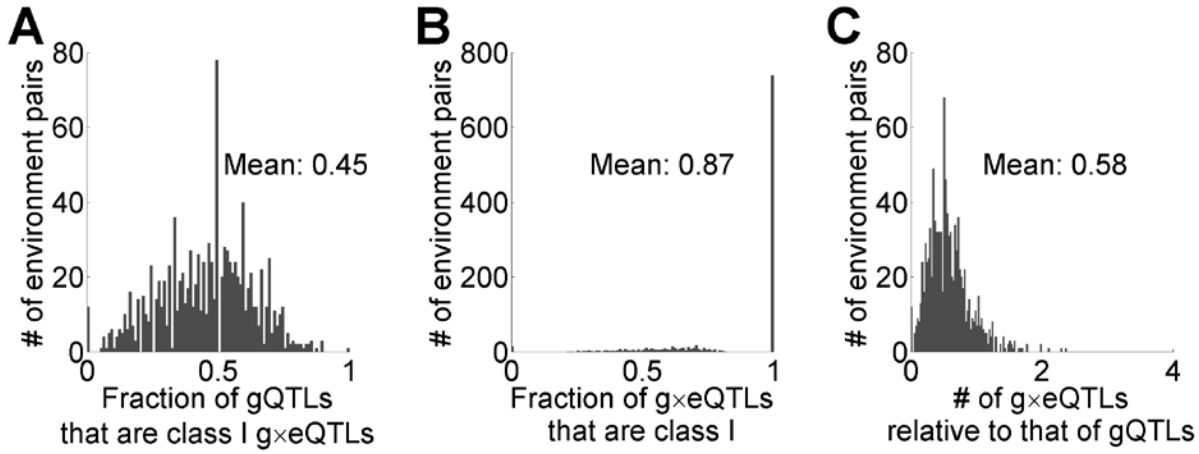


Figure 2-3. Relative numbers of g×eQTLs and gQTLs from all pairs of environments. (A) Frequency distribution of the fraction of unique gQTLs identified from two individual environments that are class I g×eQTLs for the pair of environments. (B) Frequency distribution of the fraction of all g×eQTLs (i.e., class I + extrapolated class II) that are class I. (C) Frequency distribution of the ratio between the number of all g×eQTLs for a pair of environments and the total number of unique gQTLs identified in the two environments.

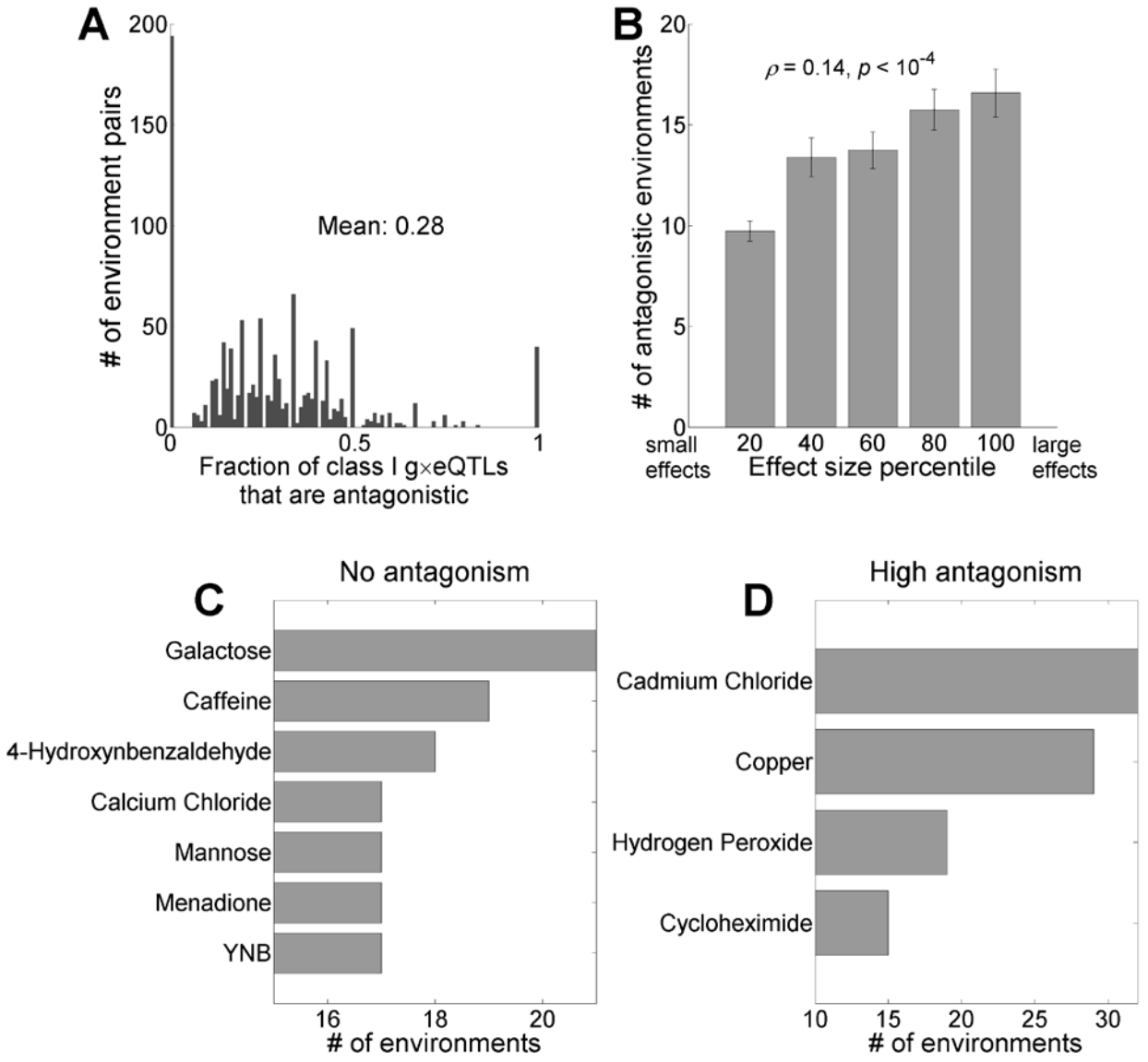


Figure 2-4. Patterns of antagonistic G×E. (A) Frequency distribution of the fraction of class I g×eQTLs that are antagonistic. (B) gQTLs with large effects in the environments where they are identified are more likely than small-effect gQTLs to have antagonistic effects in another environment. Error bars indicate one standard error. The rank correlation ρ and associated *P*-value are based on the unbinned data. (C) Environments that are underrepresented with

antagonistic g×eQTLs with other environments. The X-axis shows the number of environments with which an environment listed on the Y-axis has no antagonistic class I g×eQTL. (D) Environments that are enriched with antagonistic g×eQTLs with other environments. The X-axis shows the number of environments with which an environment listed on the Y-axis has more than 50% of class I g×eQTLs being antagonistic.

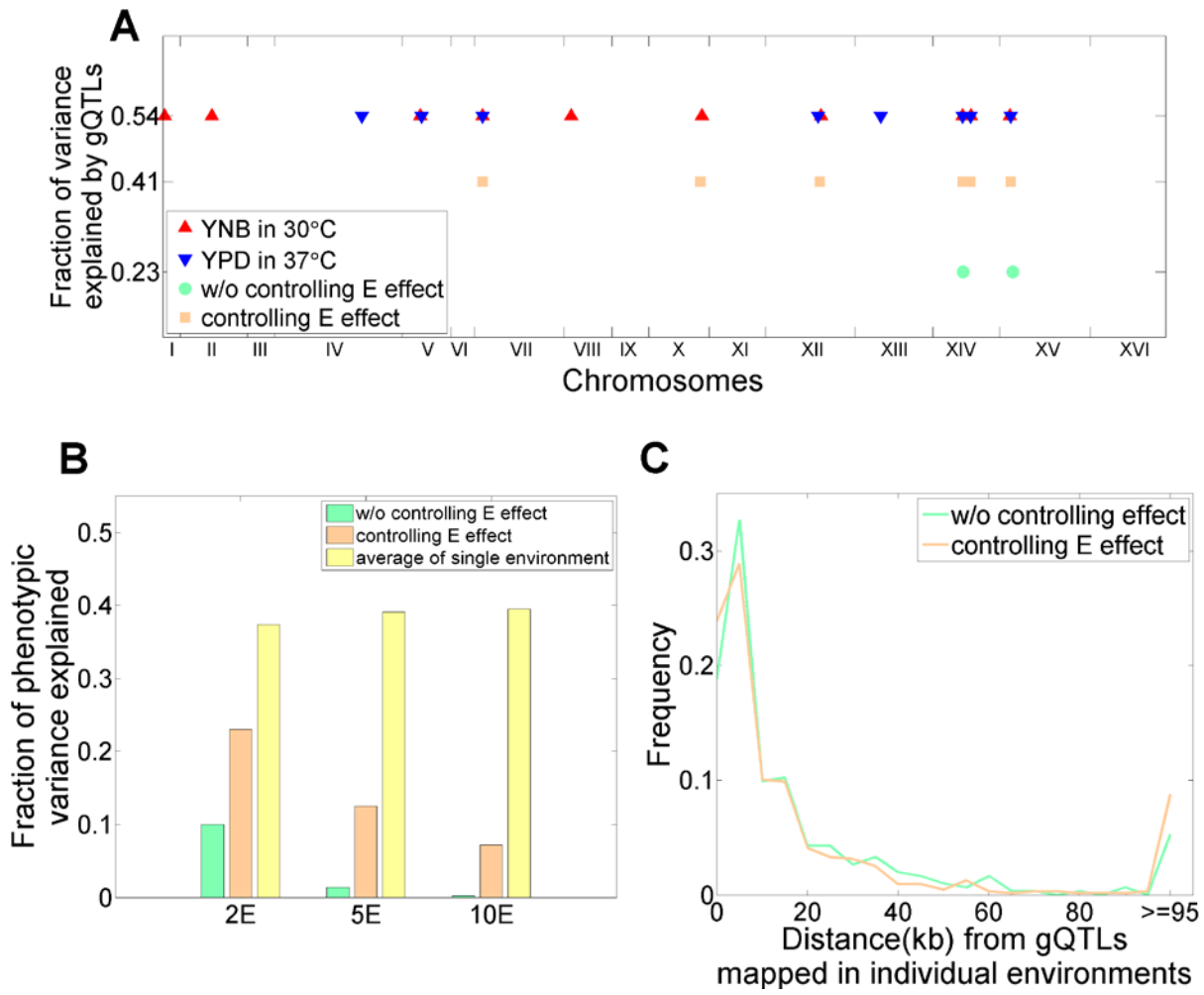


Figure 2-5. Ignoring G×E causes "missing heritability". (A) The genomic distributions of gQTLs identified from phenotypes measured in one environment and those measured in two environments (50% segregants from each environment), respectively. Y-axis shows the fraction of phenotypic variance explained by the identified gQTLs under each mapping scheme. E effect, environmental effect. Without controlling E effect means that neither environmental effect nor G×E is considered in mapping. Controlling E effect means environmental effect but not G×E is considered in mapping. (B) Average faction of phenotypic variance explained by gQTLs (r^2) decreases as the phenotypic data used originate from more environments. The average narrow-

sense heritability is 0.55. 2E, phenotypic data are from a mixture of two environments; 5E, phenotypic data are from a mixture of five environments; 10E, phenotypic data are from a mixture of 10 environments. Results are summarized from 100 random sets of 2, 5, and 10 environments, respectively. (C) Frequency distribution of the distance between gQTLs identified using mixed phenotypes from two environments and those identified using phenotypes from individual environments. The results are summarized from 100 random pairs of environments.

Table 2-1. Distributions of gQTLs and class I g×eQTLs across various genomic regions

Genomic regions	All SNPs[1]	All gQTLs[2]		Class I g×eQTLs[3]	
	Frequency	Frequency	*P*-value[4]	Frequency	*P*-value[5]
Intronic	0.008	0.004	0.2804	0.008	2.2×10^{-7}
Intergenic	0.331	0.344	0.2548	0.324	1.0×10^{-5}
Coding	0.656	0.643	0.2927	0.665	2.3×10^{-6}
Synonymous	0.558	0.500	0.0234	0.535	9.6×10^{-9}
Nonsynonymous	0.425	0.478	0.0265	0.447	2.5×10^{-7}
Nonsense	0.018	0.023	0.1859	0.019	0.0265

[1]Total number of SNPs is 28,220.

[2]Total number of gQTLs is 552.

[3]Total number of class I g×eQTLs is 18,186.

[4]Comparison with all SNPs using a binomial test.

[5]Comparison with all gQTLs using a binomial test.

Table 2-2. Significantly overrepresented gene ontology (GO) domains and terms

GO category	All SNPs[1]	All gQTLs[2]		Class I g×eQTLs[3]	
	Frequency	Frequency	*P*-value[4]	Frequency	*P*-value[4]
GO domains					
cellular component	0.692	0.685	1	0.696	0.028
biological process	0.755	0.790	0.087	0.771	4.2×10^{-6}
molecular function	0.819	0.842	0.23	0.825	3.0×10^{-6}
GO terms[5]					
GDP binding	7.1×10^{-5}	6.2×10^{-3}	1.9×10^{-4}	1.2×10^{-2}	3.2×10^{-9}
sequence-specific DNA binding transcription factor activity	2.5×10^{-2}	6.4×10^{-2}	4.1×10^{-3}	7.5×10^{-2}	1.3×10^{-4}
ribosomal small subunit biogenesis	3.4×10^{-3}	2.7×10^{-2}	7.7×10^{-6}	3.3×10^{-2}	7.3×10^{-3}
90S preribosome	5.3×10^{-3}	2.7×10^{-2}	1.6×10^{-3}	3.3×10^{-2}	7.3×10^{-3}

[1]Total number of SNPs is 28,220.

[2]Total number of gQTLs is 552.

[3]Total number of class I g×eQTLs is 18,186.

[4]Based on a binomial test followed by multiple-testing correction. gQTLs are compared with all SNPs while class I g×eQTLs are compared with all gQTLs.

[5]Shown are GO terms significantly enriched in both gQTLs (relative to all SNPs) and class I g×eQTLs (relative to gQTLs).

Table 2-3. Distributions of antagonistic and concordant class I g×eQTLs across various genomic regions

Genomic regions	Antagonistic		Concordant		*P*-value[1]
	Frequency	Occurrences	Frequency	Occurrences	
Intronic	0.0059	16	0.0084	61	0.2083
Intergenic	0.3370	910	0.3194	2316	0.0939
Coding	0.6556	1770	0.6685	4848	0.2236
Synonymous	0.5927	1049	0.5132	2488	9.7×10^{-9}
Nonsynonymous	0.3740	662	0.4730	2293	7.6×10^{-13}
Nonsense	0.0333	59	0.0138	67	2.7×10^{-7}

[1]Based on a chi-squared test.

Chapter 3

Gene by gene interaction:

patterns and mechanisms of diminishing returns from beneficial mutations

"Knowledge is the only instrument of production that is not subject to diminishing returns."

— John Maurice Clark

3.1 Abstract

Diminishing returns epistasis causes the benefit of the same advantageous mutation smaller in fitter genotypes, and is frequently observed in experimental evolution. However, its occurrence in other contexts, environment-dependence, and mechanistic basis are unclear. Here we address these questions using 1005 sequenced segregants generated from a yeast cross. Under each of 47 examined environments, 63-95% of tested polymorphisms exhibit diminishing returns epistasis. Surprisingly, improving environment quality also reduces the benefits of advantageous mutations even when fitness is controlled for, indicating the inadequacy of the global epistasis hypothesis. We propose that diminishing returns originates from the modular organization of life where the contribution of each functional module to fitness is determined

jointly by the genotype and environment and has an upper limit, and demonstrate that our model predictions match empirical observations. These findings broaden the concept of diminishing returns epistasis, reveal its generality and potential cause, and have important evolutionary implications.

3.2 Introduction

Diminishing returns epistasis refers to a reduction in the benefit of an advantageous mutation when it occurs in a relatively fit genotype compared with that in a relatively unfit genotype(GRIFFING 1950; JERISON AND DESAI 2015). It is believed to explain at least in part why experimental evolution of microbes almost invariantly shows a decreasing speed of adaptation as the fitness of the population rises(WISER *et al.* 2013; COUCE AND TENAILLON 2015). Diminishing returns epistasis has been indirectly inferred from the dynamics of adaptation(MOORE *et al.* 2000; KRYAZHIMSKIY *et al.* 2009; PERFEITO *et al.* 2014; GOOD AND DESAI 2015) and directly demonstrated by engineering the same mutation in multiple strains of different fitnesses(MACLEAN *et al.* 2010; CHOU *et al.* 2011; KHAN *et al.* 2011; KRYAZHIMSKIY *et al.* 2014; WANG *et al.* 2016). While diminishing returns epistasis appears common among fixed mutations in experimental evolution, it is unknown whether it is restricted to experimental evolution, where fixed beneficial mutations are *de novo* and tend to have large effects(ORR 2002; ROKYTA *et al.* 2005), or is also widespread among standing genetic variants. Furthermore, how the pattern of diminishing returns epistasis varies across environments has not been investigated. Most importantly, the underlying cause of diminishing returns epistasis remains elusive. A commonly considered hypothesis termed the global epistasis hypothesis posits that "the effect of each mutation depends on all other mutations, but only through their combined effect on fitness" and that “each individual beneficial mutation provides a smaller advantage in a fitter genetic

background"(KRYAZHIMSKIY *et al.* 2014). Although this hypothesis is currently regarded as the leading description and explanation of diminishing returns epistasis(KRYAZHIMSKIY *et al.* 2014; WANG *et al.* 2016), to what extent it is true and why it may be true remain unanswered. Note that the diminishing returns relationship between the activity of an enzyme and the flux of the relevant metabolic pathway is well explained by the metabolic control theory(KACSER AND BURNS 1981; DYKHUIZEN *et al.* 1987; CHOU *et al.* 2014), but this theory cannot explain diminishing returns epistasis arising from interactions among mutations of different genes.

Here we develop a high-throughput method to investigate diminishing returns epistasis among standing genetic variants. We report widespread diminishing returns epistasis from single nucleotide polymorphisms (SNPs) segregating in budding yeast, discover a novel type of diminishing returns that results from an improvement in environment quality, provide evidence that the origin and patterns of diminishing returns are best explained by the modular structure of life, and discuss evolutionary implications of these findings.

3.3 Results

3.3.1 Quantifying diminishing returns epistasis by comparing mean benefits in multiple genetic backgrounds

Diminishing returns epistasis is conventionally demonstrated by showing that the same mutation causes a smaller growth rate increase in a relatively fit strain than in a relatively unfit strain(MACLEAN *et al.* 2010; CHOU *et al.* 2011; KHAN *et al.* 2011; KRYAZHIMSKIY *et al.* 2014; WANG *et al.* 2016). If the observed diminishing returns epistasis is genuine and general, it should also be testable by comparing the mean benefits of the mutation in two sets of strains that differ in mean growth rate (see Methods). Using this approach allows testing diminishing

returns epistasis for each nucleotide difference between the genomes of two organisms that can be crossed to produce a hybrid and its segregants, as long as the genotypes and growth rates of the segregants can be acquired. For example, for an A/G polymorphism at a site, we can calculate the effect of substituting A with G by comparing the mean growth rate of segregants with genotype A (or AA for diploid segregants) and the mean growth rate of segregants with genotype G (or GG for diploid segregants) at the site, because the A segregants and G segregants are on average equivalent for the rest of their genomes due to random assortment and recombination in meiosis. The above calculation can be separately performed in two sets of strains with different mean growth rates, allowing testing diminishing returns epistasis.

We applied this method to a dataset that includes the genome sequences of 1005 haploid segregants produced from the hybrid between the BY and RM strains of the yeast *Saccharomyces cerevisiae*(BLOOM *et al.* 2013). BY is derived from the widely used laboratory strain S288c, whereas RM is derived from the vineyard strain RM11-1a. The dataset also includes the mean end-point colony radius of each segregant on agar plates in 47 environments, which vary in temperature, pH, carbon source, metal ions, and small molecules(BLOOM *et al.* 2013). We estimated the growth rate of a segregant in each environment using the corresponding colony radius (**Fig D-1**; see Methods).

3.3.2 Widespread diminishing returns epistasis among standing genetic variants

To demonstrate diminishing returns epistasis, we need to show that the mean benefit of a mutation in slow-growth segregants is greater than that in fast-growth segregants. In each environment, we computed for each SNP the mean growth rate (R_{BY}) of the 50 least fit BY-allele-carrying segregants and that (R_{RM}) of the 50 least fit RM-allele-carrying segregants (**Fig 3-**

1a). The effect of the SNP in the 100 slow-growth segregants is $s_L = |R_{BY} - R_{RM}|$. We similarly computed the mean growth rate (R'_{BY}) of 50 fittest BY-allele-carrying segregants and that (R'_{RM}) of 50 fittest RM-allele-carrying segregants (**Fig 3-1a**) and estimated the effect of the same SNP in the 100 fast-growth segregants by $s_H = |R'_{BY} - R'_{RM}|$. We distinguish between two types of diminishing returns. The broad-sense diminishing returns is defined by $s_H < s_L$, while the narrow-sense diminishing returns has the additional requirement that the beneficial allele in the slow-growth segregants is also beneficial in the fast-growth segregants. Results on narrow-sense diminishing returns are qualitatively similar to those on broad-sense diminishing returns and are described in **Fig D-2**. Throughout this work, diminishing returns refers to broad-sense diminishing returns unless noted.

Under the null hypothesis that the benefit of a mutation is independent of the growth rate of the genetic background, a SNP has a 50% chance to exhibit $s_H < s_L$. Strikingly, in each of the 47 environments studied, between g = 63% and 95% of the 28,220 SNPs tested show $s_H < s_L$ (**Fig 3-1b**), with an average of 80%. Although the relationship between s_H and s_L for one SNP is not independent from that for a linked SNP, the estimated g in each environment is unbiased. The non-independence among SNPs, however, makes it difficult to test if g significantly exceeds the chance expectation of 50% in each environment. But, because the growth rates of all segregants were separately measured in different environments, the g values from different environments were estimated independently. The observation that all 47 independently estimated g values exceed 50% has a binomial probability lower than 10^{-14} under the null hypothesis of g = 0.5, strongly suggesting a general presence of diminishing returns epistasis across environments. Relative to the null hypothesis, the excess in the probability of $s_H < s_L$ in our data is $G = g-(1-g) = 2g-1$, which varies from 25% to 90% with a mean of 62% in the 47 environments. We

confirmed that G is positive in each of the 47 environments when 150 or 70 instead of 100 segregants were used to estimate each of s_L and s_H, indicating the robustness of our results. These observations demonstrate that diminishing returns epistasis is widespread among standing genetic variants. We verified that our results are not an artifact of transforming colony radius to growth rate, because repeating the analysis using colony radius yielded similar results. For instance, g varies from 0.49 to 0.89 in the 47 environments and is < 0.5 in only one environment ($P < 10^{-14}$, $N = 47$, binomial test).

Following a recent analysis of the same dataset(WEI AND ZHANG 2017), we mapped quantitative trait loci (QTLs) underlying the growth rate variation among the segregants (at a false discover rate of 0.05) in each of the 47 environments. The number of QTLs identified ranged from 0 to 33 in the 47 environments, with a mean of 15.8 (WEI AND ZHANG 2017). One environment has zero QTL and another has exactly 50% of QTLs exhibiting $s_H < s_L$. Of the remaining environments, 39 showed $s_H < s_L$ in over 50% of QTLs ($P = 3.9\times10^{-8}$, $N = 45$, binomial test) and 27 of them showed $s_H < s_L$ in significantly more than 50% of QTLs (nominal $P < 0.05$). By contrast, only 6 environments showed $s_H < s_L$ in fewer than 50% of QTLs and only one environment showed $s_H < s_L$ in significantly fewer than 50% of QTLs. Thus, the prevalence of diminishing returns epistasis is also evident among SNPs known to have independent growth effects. By bootstrapping the segregants used (see Methods), we confirmed that s_H is significantly smaller than s_L at the nominal P-value of 0.05 for 232 of a total of 741 QTLs (107 significant QTLs after Bonferroni correction of multiple testing per environment).

3.3.3 Fraction of SNPs exhibiting diminishing returns epistasis rises with environment quality

The varying g values among the 47 environments prompted us to investigate a potential role of the environment in influencing the prevalence of diminishing returns, which had not been previously studied(MACLEAN *et al.* 2010; CHOU *et al.* 2011; KHAN *et al.* 2011; KRYAZHIMSKIY *et al.* 2014; WANG *et al.* 2016). We define the quality (Q) of an environment to the population of segregants considered by the mean growth rate of all segregants in the environment. We found a positive association between Q and g (rank correlation $\rho = 0.56$, $P = 6.5\times10^{-5}$; **Fig 3-1b**), indicating that the prevalence of diminishing returns epistasis increases with environment quality. This result is robust to variation in the number of segregants used in estimating s_L and s_H. For instance, $\rho = 0.53$ ($P = 1.8\times10^{-4}$) and 0.53 ($P = 1.3\times10^{-4}$), respectively, when 150 and 70 instead of 100 segregants were used. This correlation also holds when colony radius instead of growth rate was analyzed ($\rho = 0.67$, $P = 2.7\times10^{-7}$). In addition, a closer examination of four YPD environments with different temperatures shows a monotonically increasing relationship between Q and g ($\rho = 1$, $P = 0.083$; **Fig D-3**). Q and g are also correlated when only QTLs are considered ($\rho = 0.46$, $P = 0.0014$) or only QTLs with significant diminishing returns are considered ($\rho = 0.58$, $P < 10^{-4}$).

3.3.4 Prevalence of diminishing returns epistasis rises with environment quality even after the control of growth rate

The positive correlation between Q and g may be caused by a potential among-environment variation in growth rate disparity between the 100 least fit and 100 fittest segregants used for estimating s_L and s_H rather than the environment quality per se, because high-Q environments may have smaller growth rate disparities between the two extreme groups of segregants than those of low-Q environments (**Fig D-1**). To distinguish between these two

scenarios, we calculated the fraction of SNPs exhibiting diminishing returns epistasis after respectively controlling for the median growth rate of the fast- and slow-growth groups of segregants across environments (**Fig 3-1c**). Specifically, we first chose R_H, a relatively high growth rate. For each SNP and in each environment, we picked 25 least fit BY-allele-carrying segregants whose growth rates exceed R_H and 25 fittest BY-allele-carrying segregants whose growth rates are below R_H. We similarly picked 50 RM-allele-carrying segregants with the median growth rate equal to R_H. We then estimated s_H by the difference in mean growth rate between these 50 BY-allele-carrying and 50 RM-allele-carrying segregants. We subsequently chose R_L, a relatively low growth rate, and similarly estimated s_L. This way, the s_H's in different environments were estimated using segregants with the same median growth rate; so were the s_L's. Hence, there is no among-environment difference in the disparity of median growth rate between the two groups of segregants used to estimate s_H and s_L. Because the growth rate range for the 1005 segregants varies among environments (**Fig D-1**), for the specific pair of $R_H = 2.80$ and $R_L = 2.45$ chosen, only 15 environments allowed estimation of s_H and s_L for at least 50% of all SNPs. We estimated the fraction of SNPs exhibiting diminishing returns in each of these environments and referred to it as g'. We found g' to exceed 50% in all 15 environments ($P = 3.1\times10^{-5}$, $N = 15$, binomial test), with a range between 66% and 96% and a mean of 85%. We observed a strong positive correlation between Q and g' ($\rho = 0.81$, $P = 3.9\times10^{-4}$; **Fig 3-1d**), comparable with the correlation between Q and g in the same 15 environments ($\rho = 0.84$, $P = 1.0\times10^{-4}$). Approximately 37% of all SNPs had s_H and s_L estimates in all 15 environments, and all of these SNPs exhibited diminishing returns in each environment. When narrow-sense diminishing returns is considered, g' ranges from 17% to 54% (compared with the chance expectation of 25%) among the 15 environments for these 37% of SNPs and shows a strong

correlation with Q ($\rho = 0.89$, $P < 10^{-250}$). We confirmed the correlation between Q and g' under multiple sets of R_H and R_L (**Table D-1**) that span the range of Q across the 47 environments (2.1 to 3.2) (**Fig 3-1b**). We also verified the correlation between Q and g' across environments for QTLs ($\rho = 0.81$, $P = 1.5\times10^{-4}$). Thus, even when the median growth rates of the less fit and fitter groups of segregants are both fixed across environments, an elevation in environment quality enhances the probability of diminishing returns from beneficial mutations. This is a previously unrecognized characteristic of diminishing returns. The global epistasis hypothesis, asserting that diminishing returns depends solely on the fitness of the genotype, was formulated without considering multiple environments and is obviously inadequate for describing and explaining the observation here.

3.3.5 Benefits of advantageous mutations decrease with environment quality

To examine directly the impact of environment quality on the growth effect of an advantageous mutation, we first measured the effect ($s > 0$) of each SNP in an environment by the absolute value of the difference between the mean growth rate of all BY-allele-carrying segregants and that of all RM-allele-carrying segregants in the environment. If having better environments reduces the benefit of an advantageous mutation, s should decrease as Q rises. Such a negative correlation between Q and s should be common among all SNPs examined if this type of diminishing returns is widespread. Indeed, for 98.1% of SNPs across the genome, we observed a negative rank correlation between Q and s across environments (**Fig 3-2a**).

To verify that the above negative correlation is not simply a byproduct of the canonical diminishing returns epistasis associated with a rise in the growth rate of the background genotype, we again controlled for growth rate in estimating s, similar to what is illustrated in Fig 3-1c.

That is, we first chose a fixed growth rate R for all environments. For each SNP and in each environment, we picked 50 least fit BY-allele-carrying segregants whose growth rates exceed R and 50 fittest BY-allele-carrying segregants whose growth rates are below R. We similarly picked 100 RM-allele-carrying segregants with the median growth rate of R. The effect ($s' > 0$) of the SNP in the environment given R is the absolute value of the difference in mean growth rate between these 100 BY-allele-carrying and 100 RM-allele-carrying segregants. We found the rank correlation between Q and s' to be negative for 90.3% of all SNPs examined using R = 2.5 (**Fig 3-2a**). We repeated this analysis under two other R values (2.2 and 2.8), and found the average fraction of SNPs exhibiting smaller s' in better environments to be 81% for the three R values considered. A total of 600 SNPs were identified as QTLs in one or more environments. For each of these SNPs, we estimated its effect (s and s') in each environment and correlated it with Q. For the 600 correlations between s (or s') and Q across the 47 environments, 94.3% (or 84.9%) are negative. Hence, the diminishing returns from advantageous mutations in better environments, a form of gene-environment interaction (G×E), is distinct from the canonical diminishing returns in fitter genotypes within an environment, a form of gene-gene interaction (G×G).

In the above analyses (**Fig 3-2a**), we did not distinguish which allele is beneficial and which is deleterious. We may make this distinction for each SNP using the environment where the observed absolute effect of the SNP is maximal, which minimizes the chance of misclassification. If the BY allele is beneficial relative to the RM allele in this environment, we estimate s or s' in each environment by subtracting the mean growth rate of RM-allele-carrying segregants from that of BY-allele-carrying segregants, and vice versa. Although now s and s' for an environment can be negative, we found the correlation between Q and s (or s') to remain

negative for 65.7% (or 58.9%) of SNPs (**Fig 3-2b**). Furthermore, we found a negative correlation between *Q* and *s* (or *s*') for 328 (or 357) of the 600 unique QTLs identified, significantly more than expected by chance ($P = 0.025$ or 4×10^{-6}, two-tailed binomial test). These arguably more rigorous analyses verify the environment quality-dependent diminishing returns from beneficial mutations.

3.3.6 The modular life model recapitulates the empirical patterns of diminishing returns

That the same mutation confers different benefits on different genetic backgrounds even when these backgrounds are equally fit contradicts the global epistasis hypothesis and suggests the relevance of the specific genomic compositions of these backgrounds to the fitness effect of the mutation. It is widely accepted that life is organized in a highly modular manner, where each module is a discrete object composed of a group of tightly linked components and performs a relatively independent task(RAFF 1996; HARTWELL *et al.* 1999; IHMELS *et al.* 2002; RAVASZ *et al.* 2002; BARABASI AND OLTVAI 2004; WALL *et al.* 2004; WAGNER *et al.* 2007). Intuitively, diminishing returns epistasis could arise from the modular structure of life. Specifically, our modular life model posits that each module makes a distinct contribution to fitness and that this contribution has an upper limit. Under this model, the same advantageous mutation may contribute to a module and fitness greatly if the functionality of the module is far from its maximum but may contribute only slightly if the module is approaching its maximal functionality. In addition, we assume that the environment contributes differently to the functionalities of various modules and that different environments have different contributions. Because the functionalities of various modules can be different among equally-fit genotypes, under this model, the specific genomic composition of the background genotype matters to the

fitness effect of a mutation. Our model differs from the global epistasis hypothesis where effectively only one module exists. In this one-module model, a diminishing returns curve between the functionality of the module and fitness is assumed rather than explained. Furthermore, the curve should vary from environment to environment for this model to be realistic, but it is unclear how environment modulates the curve in this model. We here explore the modular life model in an attempt to recapitulate the major empirical patterns of diminishing returns.

We started by a computer simulation of the modular life model (**Fig 3-3a** and Methods). We considered three scenarios where the growth rate of a genotype is respectively determined by the geometric mean functionality of all modules, arithmetic mean functionality of all modules, and the lowest functionality of all modules. The third scenario is also known as the barrel effect, because the amount of water storable in a barrel constructed of many wooden staves is dictated by the shorted stave(He *et al.* 2010). The results obtained under the three scenarios are qualitatively similar, and they are respectively presented in the main text (**Fig 3-3**), **Fig D-4**, and **Fig D-5**.

According to the modular life model, we simulated the genotypes and growth rates of 1000 haploid segregants in 50 environments (see Methods). One hundred genes belonging to 10 modules were considered, with each gene harboring one SNP that distinguishes between a fully functional allele and a null allele. We analyzed the simulated data the same way we analyzed the real data. Similar to what was observed in the real data (**Fig 3-1b**, **d**), the simulated data show (i) diminishing returns epistasis for >50% of SNPs in each environment and (ii) a positive correlation between the fraction of SNPs exhibiting diminishing returns epistasis and environment quality, with or without the control for growth rate across environments (**Fig 3-3b**). Furthermore, similar to what was apparent in the real data (**Fig 3-2**), most SNPs in the simulated

data show a negative correlation between growth effect and environment quality, with or without the control for growth rate (**Fig 3-3c**). The similarity between the results from the simulated data and real data indicates that the observed patterns of diminishing returns are explainable by the modular feature of life.

3.3.7 Why effect size decreases with environment quality even after the control for growth rate

Although the canonical diminishing returns epistasis is easily explained by the modular life model, that s' decreases with Q (**Fig 3-2** and **Fig 3-3c**) is puzzling. Furthermore, because we estimated s' from groups of segregants that differ in multiple genes, it is unclear whether the negative correlation between s' and Q holds when s' is estimated by comparing genotypes that differ by a single SNP upon the control of growth rate across environments. To this end, we measured the effect of a beneficial mutation in one genetic background and then averaged this effect across multiple backgrounds in simulated data. Specifically, we simulated 50,000 segregants in 50 environments as in the previous section except that stochastic noise in growth rate is omitted to improve the sensitivity of the analysis. In each environment, we first identified all segregants whose growth rates are in the range of 0.899-0.901. This range is narrower than the maximal growth effect of any beneficial mutation simulated; therefore, the identified segregants are essentially equally fit. We estimated the growth effect of a gene in an environment by averaging the effect of replacing its null allele with functional allele in the above segregants in which the focal gene is occupied by the null allele. We then correlated among environments the growth effect of the gene and Q. For the 100 genes simulated, 63 showed a negative rank correlation ($P = 0.006$, $N = 100$, binomial test). We repeated this analysis using

another growth rate range (0.949-0.951) and found 68 of 100 genes to show negative rank correlations ($P = 2\times10^{-4}$). These results confirm that the negative correlation between s' and Q observed in the simulation is genuine. The cause for this correlation is that, when the growth rate is controlled for, the among-module variance in functionality increases with Q. The reason is that, in this scenario, under a high Q, genotype quality must be relatively low, meaning that it has only a small number of functional alleles distributed among all modules, rendering the among-module variance in functionality relatively high. By contrast, under a low Q, genotype quality must be relatively high, meaning that it has many functional alleles distributed among all modules, rendering the among-module variance in functionality relatively low. Thus, the fraction of modules approaching the upper limit in functionality is greater in good environments than in poor environments, even when the mean functionality per module is the same. Consequently, the growth effect of a beneficial mutation tends to reduce with Q. We confirmed this reasoning using the above simulated data. Specifically, we found that the among-module variance in functionality averaged across all segregants aforementioned correlates positively with Q for both of the growth rate ranges considered (**Fig D-6a**, **b**). The same trend holds when growth rate is defined by the arithmetic mean instead of geometric mean of functionality across modules (**Fig D-6c**, **d**). When growth rate is controlled for in the barrel model, as Q rises, the fraction of modules with saturated functionality increases (**Fig D-7**), lowering the probability that a mutation would improve growth and reducing the average benefit of advantageous mutations.

3.4 Discussion

In this work, we designed a high-throughput method for testing diminishing returns epistasis among standing genetic variants and applied it to 28,220 SNPs as well as 741 QTLs between two yeast strains. We found widespread diminishing returns from beneficial mutations in each of the 47 environments studied, demonstrating that diminishing returns epistasis is abundant among natural genetic variants. There are pros and cons in analyzing QTLs only versus analyzing all SNPs. The QTL-based analysis considers influential SNPs that are independent from one another, but undoubtedly misses many causal SNPs due to the limited statistical power in QTL identification and hence provides an incomplete picture of the entire genome. The analysis of all SNPs provides a complete and unbiased picture of the genome, but because of the linkage among SNPs, some of the statistical tests are difficult. Nevertheless, we found that the two approaches resulted in overall similar findings.

Canonical diminishing returns epistasis is a form of gene-gene interaction, because it is conventionally quantified by comparing the effect of a mutation in genotypes of different fitnesses in the same environment. Our work broadens the concept of diminishing returns to gene-environment interaction, because we found that the effect of a beneficial mutation decreases with environment quality. The results suggest that both types of diminishing returns (gene-gene and gene-environment interactions) are prevalent among standing genetic variants across environments. Our observation supports the common belief that the fitness effects of mutations tend to increase in stressful environments(AGRAWAL AND WHITLOCK 2010) and further demonstrates that this increase also occurs even when the background genotype fitness is controlled.

The prevailing view before this study is that diminishing returns depends on the fitness of the background genotype, as described by the global epistasis hypothesis. Our finding that the

benefit of an advantageous mutation decreases with environment quality even when the fitness of the background genotype remains unchanged indicates that the global epistasis hypothesis is inadequate. This conclusion applies to both the original and broadened concepts of diminishing returns, because a close examination of a previous study(KRYAZHIMSKIY *et al.* 2014) showed that the growth effects of a mutation in several strains of similar growth rates are significantly different even under the same environment (**Table D-2**).

We proposed that diminishing returns can instead be explained by the modular structure of life, where each module contributes to a fitness component and has a maximal possible contribution. Consistently, our computer simulation demonstrates that this modular life model recapitulates the empirical patterns of diminishing returns. Our model is inspired by the modular epistasis model(TENAILLON *et al.* 2012; KRYAZHIMSKIY *et al.* 2014) proposed to explain a phenomenon related to diminishing returns−a reduction in beneficial mutation rate when a population gradually rises in fitness during adaptation(SILANDER *et al.* 2007; TENAILLON *et al.* 2012). This phenomenon may be termed decreasing supplies, because it is about decreasing supplies of beneficial mutation as adaptation progresses. The modular epistasis model asserts that a population has limited ways to adapt and will run out of beneficial mutations if all modules reach their maximal functionalities. It is clear that our modular life model is similar to the modular epistasis model despite that they are proposed to explain different phenomena; one main difference is that our model includes environmental contributions to the functionalities of individual modules, allowing considering both genotype and environment qualities in the study of diminishing returns. It is also obvious that our model is able to explain decreasing supplies, because an advantageous mutation will no longer be visible to selection when its benefit reduces

to a certain level via diminishing returns. This can indeed be seen in our simulation of the modular life model (**Fig D-8**).

It is noteworthy that a previous study disfavored the modular epistasis model(KRYAZHIMSKIY *et al.* 2014). Specifically, Kryazhimskiy et al. evolved *S. cerevisiae* for 240 generations to obtain 64 different founder lines. They then evolved the 64 founders for 500 generations, with 10 replicates per founder. They reasoned that, under the modular epistasis model, the substitutions observed in the 10 replicates from the same founder should have larger overlaps than those observed in the lines from different founders. However, no significant difference was detected. We believe that such negative results do not disprove the modular epistasis model, because it is possible that 240 generations of evolution did not create large enough differences among the 64 founders in the distribution of functionality among modules. It is also possible that only one module could contribute to the specific adaptation studied; therefore all improvements in all founders were in the same module, which would not predict the difference expected by the authors.

In another study(WANG *et al.* 2016), several substitutions observed from an experimental evolution study of *Escherichia coli* were tested on a number of strains picked from the *E. coli* phylogeny. The authors asked whether the higher the ecological similarity between the *E. coli* strains used in the experimental evolution and tested now, the closer the growth effects of the substitutions in the two strains, but found only a marginally significant result. However, because ecological similarity may not correlate well with the similarity in module functionality, this comparison has limited power in testing the modular epistasis hypothesis.

In our simulation of the modular life model, we used the geometric mean functionality, arithmetic mean functionality, or lowest functionality among modules to compute the growth

rate of a genotype. While it is unclear which scenario is more appropriate, the fact that all three simulation schemes qualitatively recapitulated the empirical diminishing returns patterns suggests that the primary cause of these patterns is the gene-gene and gene-environment interactions within modules. Needless to say, our simulation is oversimplified. For instance, antagonistic gene-environment interactions(QIAN *et al.* 2012) have not been considered. Thus, our simulation currently cannot explain how a beneficial allele becomes deleterious upon an environmental change, which is occasionally observed in real data(WEI AND ZHANG 2017). The modular life model is meant to provide the primary mechanism of diminishing returns. Refinement of the model with many more parameters would be necessary for it to explain the specific and detailed features of diminishing returns.

That our modular life model can recapitulate major empirical patterns of diminishing returns does not prove that it is the right model, because the possibility exists that some other models can also explain these patterns. In this context, it is worth mentioning Fisher's geometric model (FGM)(FISHER 1930), because it has been used to explain diminishing returns epistasis during adaptive walks(BLANQUART *et al.* 2014). The FGM depicts a particular, simple phenotype-fitness map without empirical basis. Under the assumption that the phenotypic effect of a mutation is independent of the genetic background, one could show that as the background genotypes become fitter, the benefits of mutations reduce simply because mutations tend to overshoot the optimum, resulting in diminishing returns. However, mutations are highly idiosyncratic under the FGM(TENAILLON 2014), which appears inconsistent with empirical patterns of diminishing returns(KRYAZHIMSKIY *et al.* 2014). In addition, the assumption that the phenotypic effects of mutations are independent of the genetic background is unrealistic. The FGM predicts virtually no change in mean effect size of mutations across environments(MARTIN

AND LENORMAND 2006), which is inconsistent with our observation. It is also worth noting that adaptive trajectories simulated under the NK model show negative epistasis between non-consecutive substitutions and positive epistasis between consecutive substitutions(DRAGHI AND PLOTKIN 2013; GREENE AND CRONA 2014). But the prevalence of diminishing returns epistasis predicted by the NK model is much lower than observed in experimental evolution(WÜNSCHE *et al.* 2017). Whether the NK model can explain our findings from standing genetic variants in single and multiple environments is unknown.

Although our modular life model is designed retrospectively to explain patterns of diminishing returns, it can also explain several reported phenomena of mutational effects in different environments. For instance, Chou et al. tested the growth effects of a novel transporter system that enhances metal uptake in *Methylobacterium extroquens* on various metal-poor (MP) environments(CHOU *et al.* 2009). They observed that the same beneficial mutation had larger effects in better environments. At first glance, this observation appears contradictory to our model. However, the environments considered in our simulation of the modular life model do not have a limiting factor as in their experiment. If we consider metal uptake as a module and if the contributions of all tested MP environments to that module are equally low, our model can explain their observation. Let us assume that the product of functionalities of all modules except the metal uptake module is M_1 in a relatively good environment and M_2 in a relatively poor environment, respectively. Let us further assume that the environmental and genetic contributions to the functionality of the metal uptake module total x for the background genotype in all MP environments. The contribution of the beneficial mutation to the metal uptake module is y. Under the assumption that the growth rate is the geometric mean of all K modules, the growth improvement from the mutation in the relatively good environment is $[M_1(x+y)]^{1/K}$ -

$(M_1x)^{1/K} = M_1^{1/K}[(x+y)^{1/K}-x^{1/K}]$. Similarly, the growth improvement from the mutation in the relatively poor environment is $M_2^{1/K}[(x+y)^{1/K}-x^{1/K}]$. Because M_1 is greater than M_2, the effect size of the mutation increases as the environment gets better. The same trend is predicted by our model when the genotype instead of environment is improved in non-metal uptake modules, as was observed(CHOU *et al.* 2009). The phenomenon that environmental stresses can sometimes decrease the harm of deleterious mutations(KISHONY AND LEIBLER 2003) can be similarly explained by our model. Note that the observations from these experiments cannot be explained if the additive or barrel assumption is made in the modular life model, suggesting that the geometric assumption may be more generally applicable than the additive or barrel assumption.

Our findings about the patterns and mechanistic basis of diminishing returns have several important evolutionary implications. First, the observation that the benefit of an advantageous mutation generally decreases with environment quality Q implies a negative correlation between a population's additive genetic variance in growth rate (V_R) and Q. This is indeed true in the yeast data (ρ = -0.56, $P = 8\times10^{-5}$; see Methods). All else being equal, the growth rate variance among individuals is also expected to decrease as Q rises. Consistently, we observed a negative correlation between the growth rate variance among the 1005 segregants studied here and Q (**Fig 3-4a**). That is, the among-individual variation in growth rate gets larger as the environment becomes harsher, echoing earlier observations made in much smaller datasets(LEWONTIN AND MATSUO 1963; KONDRASHOV AND HOULE 1994; KORONA 1999; SZAFRANIEC *et al.* 2001). Second, Fisher's Fundamental Theorem of natural selection states that the rate with which a population adapts equals the variance of fitness(FISHER 1930). Because the variance of fitness (or growth rate) rises as Q reduces, the same population should adapt faster in harsher environments. Third, related to the above point, evolvability is the ability of a population to

respond to selection(HOULE 1992). Houle(HOULE 1992) showed that evolvability (E) equals additive fitness variance V_F divided by the mean fitness of the population (F). If we regard growth rate as a proxy for fitness, we have $E \approx V_R/Q$. Thus, evolvability rises precipitously as a population moves to harsher environments (**Fig 3-4b**). This prediction is supported by some anecdotes in the literature. For instance, it was reported that the relative fitness gain in the laboratory evolution of an *E. coli* strain is faster in the less preferred temperatures of 32°C and 42°C than in its optimal temperature of 37°C (BENNETT *et al.* 1992). Future studies are required to test this prediction critically and systematically. Fourth, the modular structure of life creates functional redundancy within modules when the functionality of the module approaches its maximum. This redundancy means that when a population is fully adapted to an environment, the population can accumulate genetic variation with little fitness variation, a phenomenon known as phenotypic robustness to mutations(DE VISSER *et al.* 2003; WAGNER 2005). This hidden genetic variance can be useful for adaptation when the environment changes. Thus, via the phenomenon of diminishing returns, the modular structure of life fundamentally impacts both the robustness and evolvability of organisms. It will be of great interest to verify our yeast-based observations in other species.

3.5 Methods

3.5.1 Genotype and phenotype data

We acquired from the Kruglyak lab(BLOOM *et al.* 2013) the genotype data of 1040 segregants from a cross between the BY and RM strains of *S. cerevisiae*, including a total of 28,220 SNPs mapped to the reference genome sequence R64-1-1. We downloaded the genome annotations for R64-1-1 from Ensembl biomart. We also obtained from the Kruglyak lab the

average end-point colony radius of each segregant in each of 47 environments(BLOOM *et al.* 2013). After requiring each segregant to have both genotype and phenotype data in at least one environment, we retained 1005 qualified segregants for subsequent analysis. Note that colonies with ln(radius) > 3.508 had been excluded from the data to minimize the effect of growth saturation on growth rate estimation. We further removed those colonies with ln(radius) < 1.6, because this value approaches the lower limit of colony size measurement. We converted colony radius to average growth rate as described in the next section. Growth rate variance (V) among segregants under each environment was computed from the growth rates of the segregants. We obtained the narrow-sense heritability (h^2) under each environment from Table D-2 of a previous study(BLOOM *et al.* 2013) and computed the additive growth rate variance by $V_R = Vh^2$. Evolvability was calculated using $E \approx Vh^2/Q$ according to Houle(HOULE 1992).

3.5.2 Growth rate estimation from colony size

The original phenotype measured in the data is the mean radius (D) of each colony at the end of T = 48h of growth on solid media. We transformed D to average growth rate in the following way. Let the number of cells in a colony be N, which can be described by

$$N = aD^K , \qquad (1)$$

where K is a constant presumably between 2 (if colonies resemble columns) and 3 (if colonies resemble spheres) and a is a constant representing the number of cells per unit volume. Cell growth can be described by

$$N = N_0 e^{\int_0^T R(t)dt} = N_0 e^{\bar{R}T} , \qquad (2)$$

where N_0 is the number of colonizing cells, N is the number of cells at time T, $R(t)$ is the growth rate at time t, and $\overline{R}$ is the average growth rate from time 0 to T. From Eqs. (1) and (2), we have

$$N_0 e^{\overline{R}T} = aD^K . \quad (3)$$

Eq. (3) can be converted to

$$\overline{R} = (K / T)\ln D + (\ln a - \ln N_0) / T . \quad (4)$$

Because T is constant, N_0 is expected to be constant, and K and a are presumably approximately constant (see below), ln$D$ and $\overline{R}$ have approximately the same linear relationship for all strains. As a result, ln$D$ can be used to represent $\overline{R}$ when comparing $\overline{R}$ values. Throughout this study, we used ln$D$ as a measure of $\overline{R}$.

To verify that K and a are approximately constant, we grew 91 randomly picked segregants on YPD agar plates for 48h. We scanned colonies and measured the pixel number per colony using SGATools(WAGIH *et al.* 2013), allowing quantifying the colony radius D. We then estimated the corresponding cell number N in each colony using flow cytometry (BD Accuri™ C6). If K and a in Eq. (1) are constant across genotypes, ln$N$ should be a linear function of ln$D$. Indeed, our data showed that ln$N$ and ln$D$ have a linear correlation of $r = 0.74$ ($P < 10^{-16}$), supporting approximate constancies in K and a across genotypes.

To verify that the yeast growth did not saturate at 48h, we grew 79 randomly picked segregants on a YPD plate and scanned colonies and estimated D at 13 time points every 2-3h from 15h to 48h of growth. We conducted a linear regression between ln$D$ and time of growth for each colony (**Fig D-9a**), and found that the average adjusted $r^2 = 0.94$, suggesting that $R(t)$ did not change much during the course of 48h growth. Indeed, a quadratic fitting improves the adjusted r^2 only slightly to an average of 0.96, despite that the improvement occurred to most segregants (**Fig D-9a**). Because our formulation (Eq. 4) considers the average growth rate from

0 to 48h, our method is valid as long as the slight saturation is not more pronounced for fast-growth segregants than slow-growth segregants. Indeed, we found no significant correlation among the 79 segregants tested between the growth rate rank at 48h and Δ(adjusted r^2), which is the difference in adjusted r^2 between the quadratic and linear regressions and a measure of saturation (**Fig D-9b**).

3.5.3 Estimating epistasis from growth rate

Let F_{WT}, F_A, F_B, and F_{AB} be the fitness of the wild-type, mutant A, mutant B, and the corresponding double mutant, respectively. It is commonly thought that $(F_{AB}/F_{WT}) = (F_A/F_{WT})(F_B/F_{WT})$ when there is no epistasis. In other words, $\ln(F_{AB}) = \ln(F_A) + \ln(F_B) - \ln(F_{WT})$ under no epistasis. Let R_{WT}, R_A, R_B, and R_{AB} be the growth rates of the wild-type, mutant A, mutant B, and the corresponding double mutant, respectively. The relationship between fitness and growth rate of a genotype is $F = e^{Rt}$, or $\ln F = Rt$, where t is the generation time of the wild-type. Hence, under no epistasis, $R_{AB} = R_A + R_B - R_{WT}$. In other words, epistasis can be estimated by $R_{AB} - (R_A + R_B - R_{WT}) = (R_{AB} - R_A) - (R_B - R_{WT})$, which is the growth effect of mutation B on the background of mutant A minus the corresponding effect on the wild-type background. This is why diminishing returns epistasis is commonly assessed by comparing the growth effect of a mutation on two genetic backgrounds.

3.5.4 Assessing the fitness effect of a mutation in multiple genetic backgrounds

Diminishing returns epistasis is conventionally demonstrated by a higher growth benefit of a mutation in a less fit genotype than in a fitter genotype. Here we show that it can also be demonstrated by a higher growth benefit in a group of less fit genotypes than in a group of fitter

genotypes. Suppose we are interested in assessing the growth effect of mutating allele X_1 to X_2 at a site in two different genetic backgrounds G and H (locus X is not considered part of the genetic background). The growth rate of the genotype with X_1 in background G is $R(G+X_1) = A(G)+A(X_1)+E(G)+E(G, X_1)$, where $A(G)$ is the total additive effect of all alleles in G, $A(X_1)$ is the additive effect of X_1, $E(G)$ is the total epistatic effect among all alleles in G, and $E(G, X_1)$ is the epistatic effect between X_1 and G. Similarly, the growth rate of the genotype with X_2 in background G is $R(G+X_2) = A(G)+A(X_2)+E(G)+E(G, X_2)$. Thus, the growth effect of the mutation in the background of G is $R(G+X_2)-R(G+X_1) = A(X_2)-A(X_1)+E(G, X_2)-E(G, X_1) = A(X_2)-A(X_1)+\Delta E(G, X_2-X_1)$, where $\Delta E(G, X_2-X_1)$ is the difference in epistatic effect between X_2 and X_1 in G and will be referred to as the epistatic effect of the mutation in G. The corresponding growth effect of the mutation in background H is $R(H+X_2)-R(H+X_1) = A(X_2)-A(X_1)+\Delta E(H, X_2-X_1)$. Hence, the difference between the growth effect of the mutation in H and that in G is $\mu = [R(H+X_2)-R(H+X_1)]-[R(G+X_2)-R(G+X_1)] = \Delta E(H, X_2-X_1)-\Delta E(G, X_2-X_1)$, which is the difference in the epistatic effect of the mutation in the two backgrounds. Analysis of diminishing returns is to study μ. Specifically, diminishing returns means that, when $R(H+X_1) > R(G+X_1)$, $\mu = \Delta E(H, X_2-X_1)-\Delta E(G, X_2-X_1) < 0$. In other words, when the genetic background becomes fitter, the epistatic effect of the mutation becomes smaller.

Now let us consider a group of $2k$ relatively unfit random genotypes, of which G_1, G_2, ..., and G_k carry X_1 while G_{k+1}, G_{k+2}, ..., and G_{2k} carry X_2; frequencies of alleles at other loci are not different between the first and last k genotypes. The mean growth effect of muting X_1 to X_2 in the above $2k$ genotypes is

$$\begin{aligned}
&\sum_{i=1}^{k} R(\mathrm{G_{k+i}}+\mathrm{X_2})/k - \sum_{i=1}^{k} R(\mathrm{G_i}+\mathrm{X_1})/k \\
&= \sum_{i=1}^{k} [A(\mathrm{G_{k+i}}) + A(\mathrm{X_2}) + E(\mathrm{G_{k+i}}) + E(\mathrm{G_{k+i}}, \mathrm{X_2})]/k - \sum_{i=1}^{k} [A(\mathrm{G_i}) + A(\mathrm{X_1}) + E(\mathrm{G_i}) + E(\mathrm{G_i}, \mathrm{X_1})]/k \\
&= [A(\mathrm{X_2}) - A(\mathrm{X_1})] + \sum_{i=1}^{k} [E(\mathrm{G_{k+i}}, \mathrm{X_2}) - E(\mathrm{G_i}, \mathrm{X_1})]/k + \sum_{i=1}^{k} [A(\mathrm{G_{k+i}}) - A(\mathrm{G_i}) + E(\mathrm{G_{k+i}}) - E(\mathrm{G_i})]/k.
\end{aligned} \tag{5}$$

There are three terms in the right-hand side of Eq. (5). The first term is the additive effect of the mutation. The second term is the mean epistatic effect of the mutation in the genetic backgrounds concerned. The third term is expected to be 0, because the first and last *k* genotypes are on average the same in additive and epistatic growth effects. Thus, Eq. (5) can be written as

$$\sum_{i=1}^{k} R(\mathrm{G_{k+i}}+\mathrm{X_2})/k - \sum_{i=1}^{k} R(\mathrm{G_i}+\mathrm{X_1})/k = A(\mathrm{X_2}) - A(\mathrm{X_1}) + \Delta\overline{E}(\mathrm{G}, \mathrm{X_2}\text{-}\mathrm{X_1}), \tag{6}$$

where the last term is the mean epistatic effect of the mutation in G backgrounds.

Let us similarly consider a group of 2*k* relatively fit genotypes, of which $\mathrm{H_1}$, $\mathrm{H_2}$, ..., and $\mathrm{H_k}$ carry $\mathrm{X_1}$ while $\mathrm{H_{k+1}}$, $\mathrm{H_{k+2}}$, ..., and $\mathrm{H_{2k}}$ carry $\mathrm{X_2}$. The mean growth effect of mutating $\mathrm{X_1}$ to $\mathrm{X_2}$ in the above *2k* genotypes can be similarly written as

$$\sum_{i=1}^{k} R(\mathrm{H_{k+i}}+\mathrm{X_2})/k - \sum_{i=1}^{k} R(\mathrm{H_i}+\mathrm{X_1})/k = A(\mathrm{X_2}) - A(\mathrm{X_1}) + \Delta\overline{E}(\mathrm{H}, \mathrm{X_2}\text{-}\mathrm{X_1}). \tag{7}$$

Using Eqs. (6) and (7), we can find that the difference between the growth effect of the mutation in the H backgrounds and that in the G backgrounds is

$$\mu' = \Delta\overline{E}(\mathrm{H}, \mathrm{X_2}\text{-}\mathrm{X_1}) - \Delta\overline{E}(\mathrm{G}, \mathrm{X_2}\text{-}\mathrm{X_1}). \tag{8}$$

Thus, it is clear that μ and μ' measure the same thing except that the epistatic effect of the mutation in one genetic background is considered in the former while the mean epistatic effect of the mutation in multiple backgrounds is considered in the latter. Given the stochasticity of epistasis, mean epistasis is presumably more informative than a single epistasis value for studying diminishing returns patterns.

3.5.5 Bootstrap test of the significance of diminishing returns epistasis

We examined whether s_H is significantly smaller than s_L for each QTL by a bootstrap test. We first calculated the observed s_L- s_H. We then generated a bootstrap sample of growth rates from the 50 fitted BY-carrying segregants as well as a bootstrap sample of growth rates from the 50 fitted RM-carrying segregants, allowing the estimation of s_H from the bootstrap samples. We similarly generated bootstrap samples and obtained the estimate of s_L and then s_L- s_H. This process was repeated 10,000 times. *P*-value is estimated by the proportion of bootstrap replications in which $s_L < s_H$.

3.5.6 Analysis of narrow-sense diminishing returns

For a SNP to exhibit narrow-sense diminishing returns, two conditions must be met: (i) $s_H < s_L$ and (ii) the beneficial allele in the 100 slow-growth and that in the 100 fast-growth segregants must be the same. Let g_1 be the fraction of SNPs showing $s_H < s_L$ and having the same beneficial allele in the slow- and fast-growth segregants, and let g_2 be the fraction of SNPs showing $s_L < s_H$ and having the same beneficial allele in the slow- and fast-growth segregants. Under the null hypothesis that the growth effect of an allele is independent of the genetic background, g_1 is expected to equal g_2. If diminishing returns is general, g_1-g_2 should be positive. We estimated g_1 and g_2 under each environment using the method shown in Fig 3-1a. We also estimated them using the method shown in Fig 3-1c. We examined the correlation between g_1 and *Q* using all SNPs or only QTLs.

3.5.7 Simulation of the modular life model

We assume that the growth rate of a genotype in an environment is the combined effects of *C* functional modules. Each module has a functionality value that is the sum of environmental and genetic contributions to the module. The maximum possible functionality of each module is 1 and the minimum is 0. Consequently, further improvement in genotype or environment quality has no contribution to the functionality of a module when it reaches the maximum. Each module has *M* contributing genes, each with one SNP that distinguishes between a fully functional allele and a null allele. There are *N* haploid segregants in a population; the genotype of each segregant is made up of *CM* genes, each carrying the functional allele with a 50% probability.

In our simulation, the specific values of various parameters are not critical to the conclusion, as long as the functionalities of some modules reach the upper limit. Below is the set of parameters used in generating Fig 3-3bc. We used $C = 10$, $M = 10$, and $N = 1000$, and simulated 50 environments. The maximal contributions of the 10 genes to the functionality of a module were set to be 0.11, 0.12, 0.13, …, and 0.2, respectively. Thus, the functional allele of gene 1 contributes 0.11 units of functionality to its module, while the null allele contributes 0 unit. We assumed that the contribution of an environment to a module is a normal random variable with a standard deviation of 0.05. The mean of the normal distribution is 0.2000, 0.2035, 0.2070, …, and 0.3715, respectively, from the 50 environments. We also added a noise term, drawn randomly from the normal distribution of mean = 0 and standard deviation = 0.01, to the growth rate of each simulated genotype in each environment.

3.5.8 Reanalysis of Kryazhimskiy et al.'s data of diminishing returns

We reanalyzed the data from Figure 3 of Kryazhimskiy et al.(KRYAZHIMSKIY *et al.* 2014). The growth rates of all strains were measured using flow cytometry-based competition assays

against the ymCitrine-labelled DivAncCit strain and were represented by percent difference from DivAncCit. *HO*, *GAT2*, *WHI2*, and *SFL1* genes were separately deleted in each of 40 different ancestor strains. The growth rate of each ancestor strain was measured in triplets, and we calculated the mean growth rate and its standard error using the three repeats. For the deletion strains, the growth rates of one to five replicate colonies were measured three times each. For these strains, we first calculated the growth rate of each replicate and then calculated the mean growth rate and its standard error using the replicates. When there was no replication, we calculated the mean growth rate and its standard error using the repeats. We used two-tailed *Z*-test to identify all pairs of strains whose growth rates are not significantly different from each other. For each of these strain pairs, we used a two-tailed *Z*-test to test if the effect sizes of the same mutation are significantly different. The strain pairs with significantly different growth effects for the same mutation after Bonferroni correction are shown in **Table D-2**.

3.6 Acknowledgements

We thank the Kruglyak lab for sharing the yeast segregant genotype and phenotype data and Wei-Chin Ho, Alexey Kondrashov, Chuan Li, Wenfeng Qian, Olivier Tenaillon, Jian-Rong Yang, and four anonymous reviewers for valuable comments on earlier versions. This work was supported in part by the U.S. National Institutes of Health grant R01GM120093 to J.Z.

3.7 References

Agrawal, A. F., and M. C. Whitlock, 2010 Environmental duress and epistasis: how does stress affect the strength of selection on new mutations? Trends Ecol Evol 25**:** 450-458.

Barabasi, A. L., and Z. N. Oltvai, 2004 Network biology: understanding the cell's functional organization. Nat Rev Genet 5**:** 101-113.

Bennett, A. F., R. E. Lenski and J. E. Mittler, 1992 Evolutionary Adaptation to Temperature. I. Fitness Responses of Escherichia coli to Changes in its Thermal Environment. Evolution 46**:** 16-30.

Blanquart, F., G. Achaz, T. Bataillon and O. Tenaillon, 2014 Properties of selected mutations and genotypic landscapes under Fisher's geometric model. Evolution 68**:** 3537-3554.

Bloom, J. S., I. M. Ehrenreich, W. T. Loo, T. L. Lite and L. Kruglyak, 2013 Finding the sources of missing heritability in a yeast cross. Nature 494**:** 234-237.

Chou, H.-H., J. Berthet and C. J. Marx, 2009 Fast growth increases the selective advantage of a mutation arising recurrently during evolution under metal limitation. PLoS Genet 5**:** e1000652.

Chou, H.-H., H.-C. Chiu, N. F. Delaney, D. Segrè and C. J. Marx, 2011 Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science 332**:** 1190-1192.

Chou, H. H., N. F. Delaney, J. A. Draghi and C. J. Marx, 2014 Mapping the fitness landscape of gene expression uncovers the cause of antagonism and sign epistasis between adaptive mutations. PLoS Genet 10**:** e1004149.

Couce, A., and O. A. Tenaillon, 2015 The rule of declining adaptability in microbial evolution experiments. Frontiers in genetics 6**:** 99.

de Visser, J. A., J. Hermisson, G. P. Wagner, L. Ancel Meyers, H. Bagheri-Chaichian *et al.*, 2003 Perspective: Evolution and detection of genetic robustness. Evolution 57**:** 1959-1972.

Draghi, J. A., and J. B. Plotkin, 2013 Selection biases the prevalence and type of epistasis along adaptive trajectories. Evolution 67**:** 3120-3131.

Dykhuizen, D. E., A. M. Dean and D. L. Hartl, 1987 Metabolic flux and fitness. Genetics 115**:** 25-31.

Fisher, R. A., 1930 *The Genetic Theory of Natural Selection*. Clarendon, Oxford.

Good, B. H., and M. M. Desai, 2015 The impact of macroscopic epistasis on long-term evolutionary dynamics. Genetics 199**:** 177-190.

Greene, D., and K. Crona, 2014 The changing geometry of a fitness landscape along an adaptive walk. PLOS Comput Biol 10**:** e1003520.

Griffing, B., 1950 Analysis of quantitative gene action by constant parent regression and related techniques. Genetics 35**:** 303.

Hartwell, L. H., J. J. Hopfield, S. Leibler and A. W. Murray, 1999 From molecular to modular cell biology. Nature 402**:** C47-52.

He, X., W. Qian, Z. Wang, Y. Li and J. Zhang, 2010 Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. Nat Genet 42**:** 272-276.

Houle, D., 1992 Comparing evolvability and variability of quantitative traits. Genetics 130**:** 195-204.

Ihmels, J., G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv *et al.*, 2002 Revealing modular organization in the yeast transcriptional network. Nat Genet 31**:** 370-377.

Jerison, E. R., and M. M. Desai, 2015 Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments. Curr Opin Genet Dev 35**:** 33-39.

Kacser, H., and J. A. Burns, 1981 The molecular basis of dominance. Genetics 97**:** 639-666.

Khan, A. I., D. M. Dinh, D. Schneider, R. E. Lenski and T. F. Cooper, 2011 Negative epistasis between beneficial mutations in an evolving bacterial population. Science 332**:** 1193-1196.

Kishony, R., and S. Leibler, 2003 Environmental stresses can alleviate the average deleterious effect of mutations. Journal of biology 2**:** 14.
Kondrashov, A. S., and D. Houle, 1994 Genotype-environment interactions and the estimation of the genomic mutation rate in Drosophila melanogaster. Proc Biol Sci 258**:** 221-227.
Korona, R., 1999 Genetic load of the yeast Saccharomyces cerevisiae under diverse environmental conditions. Evolution**:** 1966-1971.
Kryazhimskiy, S., D. P. Rice, E. R. Jerison and M. M. Desai, 2014 Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science 344**:** 1519-1522.
Kryazhimskiy, S., G. Tkacik and J. B. Plotkin, 2009 The dynamics of adaptation on correlated fitness landscapes. Proc Natl Acad Sci U S A 106**:** 18638-18643.
Lewontin, R. C., and Y. Matsuo, 1963 Interaction of genotypes determining viability in Drosophila busckii. Proc Natl Acad Sci U S A 49**:** 270-278.
MacLean, R., G. G. Perron and A. Gardner, 2010 Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for rifampicin resistance in Pseudomonas aeruginosa. Genetics 186**:** 1345-1354.
Martin, G., and T. Lenormand, 2006 The fitness effect of mutations across environments: a survey in light of fitness landscape models. Evolution 60**:** 2413-2427.
Moore, F. B., D. E. Rozen and R. E. Lenski, 2000 Pervasive compensatory adaptation in Escherichia coli. Proc Biol Sci 267**:** 515-522.
Orr, H. A., 2002 The population genetics of adaptation: the adaptation of DNA sequences. Evolution 56**:** 1317-1330.
Perfeito, L., A. Sousa, T. Bataillon and I. Gordo, 2014 Rates of fitness decline and rebound suggest pervasive epistasis. Evolution 68**:** 150-162.
Qian, W., D. Ma, C. Xiao, Z. Wang and J. Zhang, 2012 The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. Cell Rep 2**:** 1399-1410.
Raff, R. A., 1996 *The shape of life*. University of Chicago Press, Chicago.
Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabasi, 2002 Hierarchical organization of modularity in metabolic networks. Science 297**:** 1551-1555.
Rokyta, D. R., P. Joyce, S. B. Caudle and H. A. Wichman, 2005 An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nat Genet 37**:** 441-444.
Silander, O. K., O. Tenaillon and L. Chao, 2007 Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. PLoS Biol 5**:** e94.
Szafraniec, K., R. H. Borts and R. Korona, 2001 Environmental stress and mutational load in diploid strains of the yeast Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences 98**:** 1107-1112.
Tenaillon, O., 2014 The utility of Fisher's geometric model in evolutionary genetics. Annual review of ecology, evolution, and systematics 45**:** 179-201.
Tenaillon, O., A. Rodriguez-Verdugo, R. L. Gaut, P. McDonald, A. F. Bennett *et al.*, 2012 The molecular diversity of adaptive convergence. Science 335**:** 457-461.
Wagih, O., M. Usaj, A. Baryshnikova, B. VanderSluis, E. Kuzmin *et al.*, 2013 SGAtools: one-stop analysis and visualization of array-based genetic interaction screens. Nucleic acids research 41**:** W591-W596.
Wagner, A., 2005 *Robustness and Evolvability in Living Systems*. Princeton University Press, Princeton, NJ.

Wagner, G. P., M. Pavlicev and J. M. Cheverud, 2007 The road to modularity. Nat Rev Genet 8**:** 921-931.
Wall, M. E., W. S. Hlavacek and M. A. Savageau, 2004 Design of gene circuits: lessons from bacteria. Nat Rev Genet 5**:** 34-42.
Wang, Y., C. D. Arenas, D. M. Stoebel, K. Flynn, E. Knapp *et al.*, 2016 Benefit of transferred mutations is better predicted by the fitness of recipients than by their ecological or genetic relatedness. Proceedings of the National Academy of Sciences 113**:** 5047-5052.
Wei, X., and J. Zhang, 2017 The Genomic Architecture of Interactions Between Natural Genetic Polymorphisms and Environments in Yeast Growth. Genetics 205**:** 925-937.
Wiser, M. J., N. Ribeck and R. E. Lenski, 2013 Long-term dynamics of adaptation in asexual populations. Science 342**:** 1364-1367.
Wünsche, A., D. M. Dinh, R. S. Satterwhite, C. D. Arenas, D. M. Stoebel *et al.*, 2017 Diminishing-returns epistasis decreases adaptability along an evolutionary trajectory. Nature Ecology & Evolution 1**:** 0061.

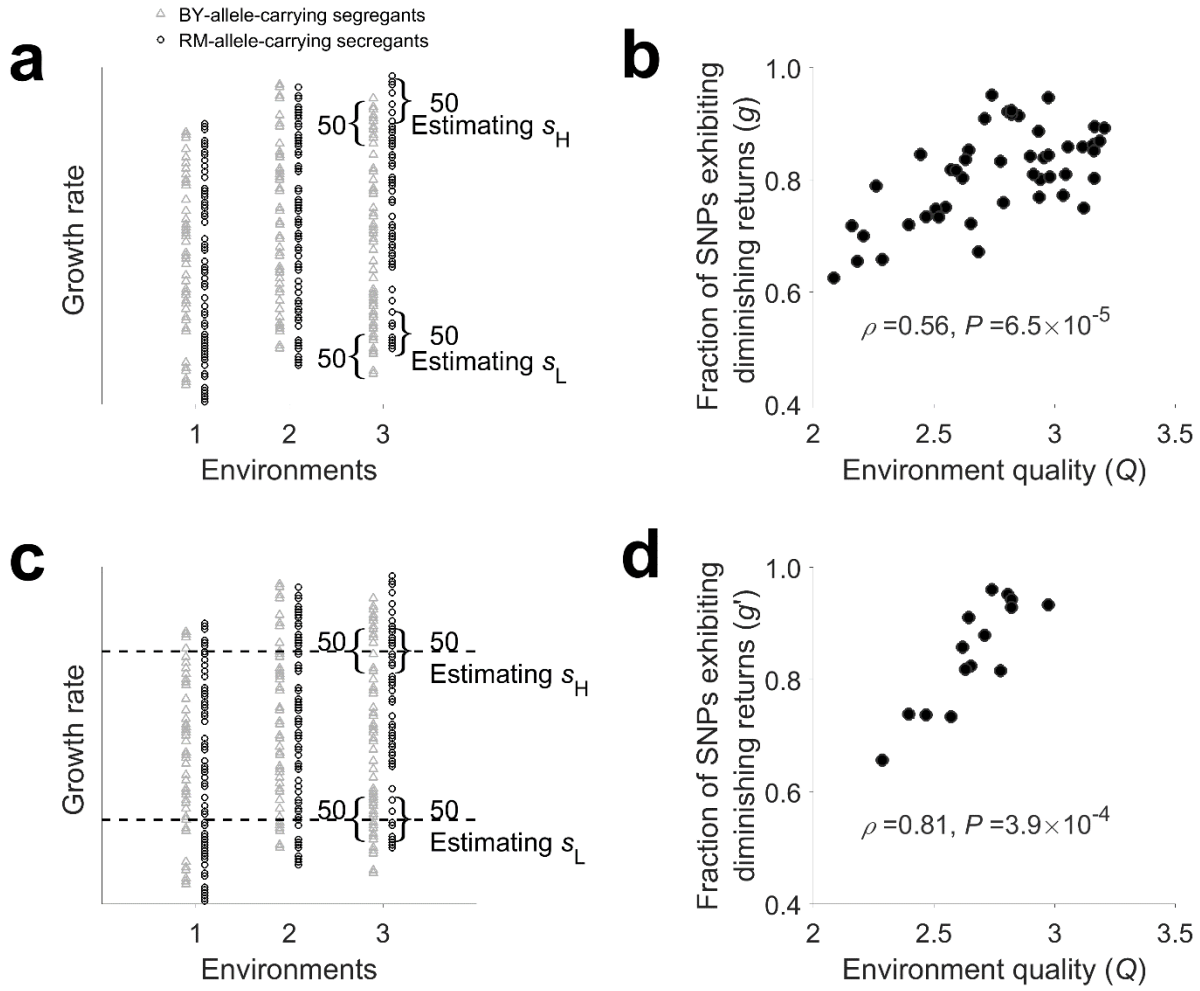


Figure 3-1. Widespread diminishing returns among standing genetic variants in yeast. s_H, growth rate effect of a SNP in fast-growth segregants; s_L, growth rate effect of a SNP in slow-growth segregants. (**a**) Scheme for estimating s_H and s_L. For each SNP under each environment, grey triangles represent BY-allele-carrying segregants, while black circles represent RM-allele-carrying segregants. The 50 fittest BY-allele-carrying and 50 fittest RM-allele-carrying segregants are used to estimate s_H, whereas the 50 least fit BY-allele-carrying and 50 least fit RM-allele-carrying segregants are used to estimate s_L. The data plotted are hypothetical and not all 50 segregants used in each group are shown. (**b**) Fraction (g) of SNPs exhibiting diminishing returns epistasis (i.e., $s_H < s_L$) in an environment increases with the quality of the environment (Q). Spearman's rank correlation and associated P-value are presented. (**c**) Scheme for

estimating s_H and s_L upon the control for median growth rate across environments. For each SNP under each environment, the 50 triangles and 50 circles with the median growth rate indicated by the higher dashed line are used to estimate s_H, whereas the 50 triangles and 50 circles with the median growth rate indicated by the lower dashed line are used to estimate s_L. The data plotted are hypothetical and not all 50 segregants used in each group are shown. (**d**) Fraction (g') of SNPs exhibiting diminishing returns upon the control for median growth rate increases with Q.

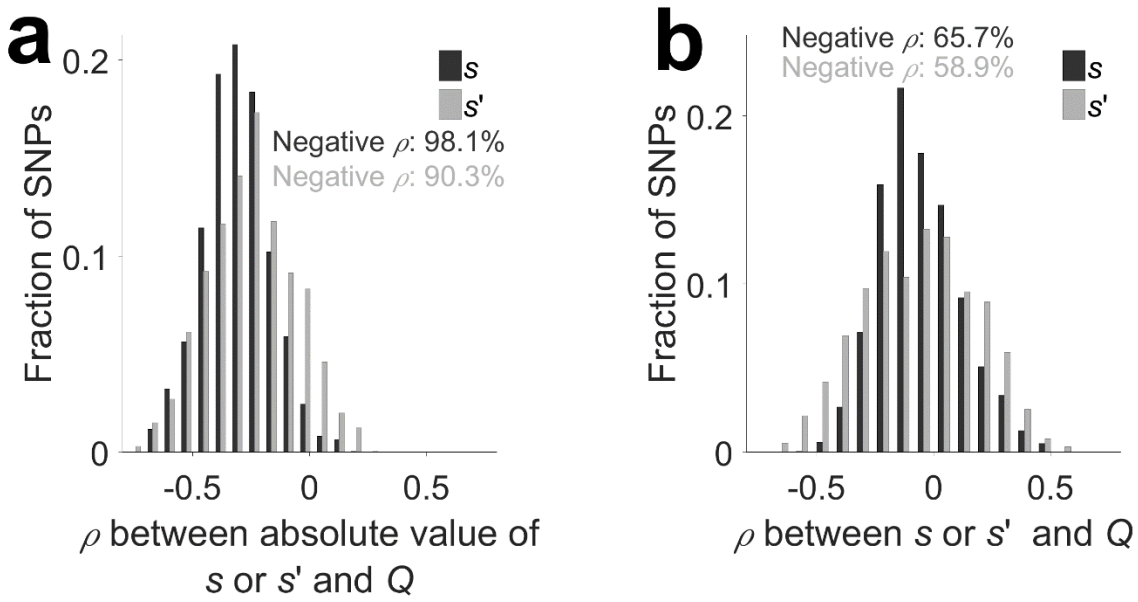


Figure 3-2. Most SNPs show a negative correlation between its effect on growth rate and environment quality (Q). (**a**) Frequency distribution of the rank correlation between Q and the absolute value of the growth rate effect of a SNP measured using either all segregants (s) or a group of segregants with a fixed median growth rate (s'). Here, s and s' are always positive. (**b**) Frequency distribution of the rank correlation between Q and the growth rate effect of a SNP measured using either all segregants (s) or a group of segregants with a fixed median growth rate (s'). Here, s or s' may be negative if the advantageous allele determined from the environment with the largest absolute growth rate effect is less fit than the alternative allele in the environment concerned. In each panel, the fraction of ρ's that are negative is indicated in black and grey for s and s', respectively.

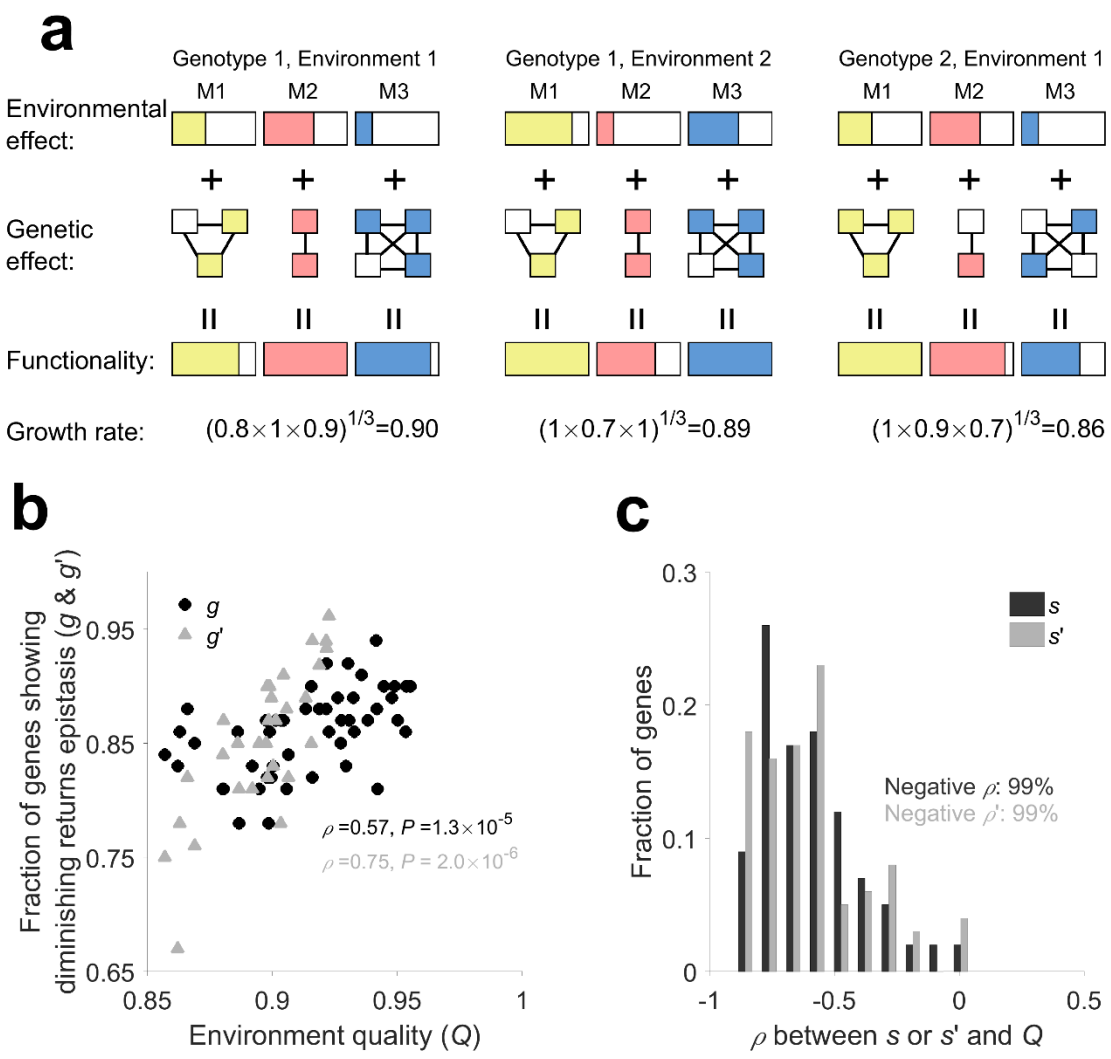


Figure 3-3. Simulation of the modular life model produces diminishing returns patterns resembling empirical observations. (**a**) Simulation scheme under the geometric mean growth rate model. Different modules (M1, M2, and M3) are colored differently. Different environments (Environments 1 and 2) contribute differently to various modules, as illustrated by the three boxes that are filled to different levels. Each module contains a number of genes, each of which could have either a functional allele designated as 1 (filled box) or a null allele designated as 0 (open box). Two genotypes (Genotypes 1 and 2) are shown as examples. The functionality of a module is the sum of environmental and genetic contributions but cannot

exceed 1. The growth rate of each genotype is computed from the functionalities of the individual modules using the formula indicated. See Methods for the parameters used in the simulation. (**b**) Simulation results showing that the fraction of genes exhibiting diminishing returns (g or g') positively correlates with environment quality (Q). Black dots show estimates of g on the basis of the fittest and least fit segregants, whereas grey triangles show estimates of g' from segregants of fixed median growth rates. (**c**) Frequency distribution of the rank correlation (ρ) between Q and the effect of a SNP measured using either all segregants (s; black) or a group of segregants with a fixed median growth rate (s'; grey). The fraction of ρ's that are negative is indicated in black and grey for s and s', respectively. Here, s and s' could be negative if the functional allele is found less fit than the null allele (due to sampling error).

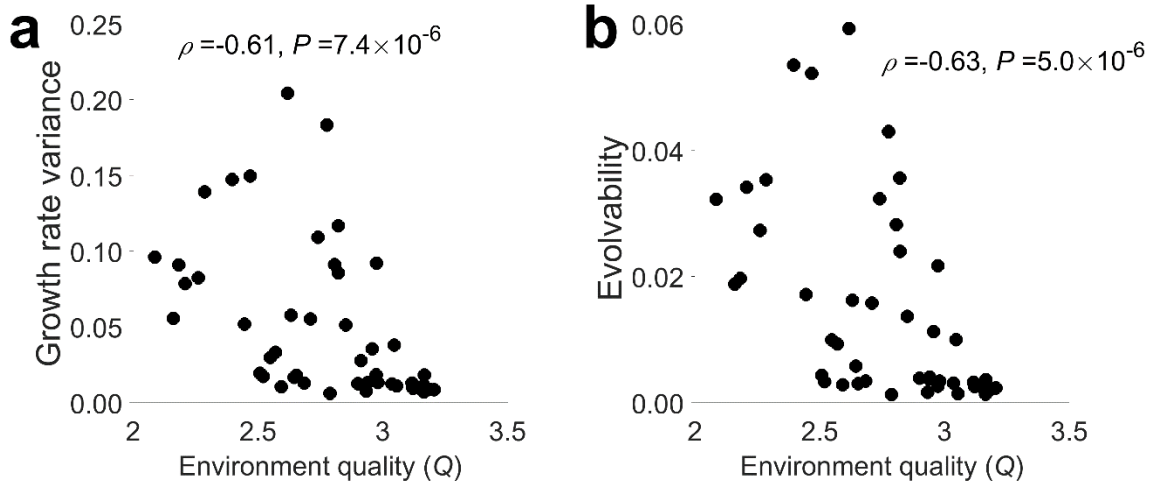


Figure 3-4. Growth rate variance and evolvability of a population increase as the environment quality (*Q*) declines. (**a**) Correlation between *Q* and the growth rate variance among the segregants examined. (**b**) Correlation between *Q* and the evolvability of the population of segregants studied. Spearman's rank correlation and associated *P*-value are presented.

Chapter 4

Allele by allele interaction: a new theory on the cause of genetic dominance

"Questions as to the genetic inter-relations and compositions of varieties can now be definitely answered."

— **William Bateson**

4.1 Abstract

The cause of the widespread dominance of wild-type alleles over deleterious mutant alleles is a subject of long-standing interest and controversy. Fisher's theory that dominance results from selection is now considered untenable. Wright instead argued that dominance is an intrinsic property of metabolic systems, but his theory cannot satisfactorily explain the prevalent dominance in non-enzyme genes. Because dominance means that gaining a wild-type allele at a locus is less beneficial in heterozygous mutants than in homozygous mutants, we hypothesize that dominance is a special case of the phenomenon of diminishing returns epistasis from advantageous mutations. Our previous work established that diminishing returns epistasis results from the modular organization of life where the contribution of each functional module to fitness is determined jointly by the genotype and environment. We use the average fitness of all

genotypes in the environment to measure environmental quality (*Q*), and our model predicts higher dominance in better environments. To test our hypothesis, we used two yeast datasets which provides dominance for growth rates in multiple environments, and both of which showed consistent results with our model prediction. This observation is unexplainable by the existing theories of dominance, but is predicted by the modular life model and is a characteristic of diminishing returns. Furthermore, all previous observations about dominance are consistent with the modular life model. These findings support that dominance is an intrinsic property arising from the modular organization of life.

4.2 Introduction

Dominance is among the first phenomena discovered in genetics (MENDEL 1996), yet its cause remains elusive even after a century of investigation. Fisher first noticed the widespread phenomenon of partial or complete dominance of wild-type alleles to the deleterious alleles (FISHER 1928). This observation has been confirmed in many species (MUKAI *et al.* 1972), including human (WILKIE 1994). Fisher explained genetic dominance by direct selection for modifiers that increase the dominance of the functional allele to defend against repetitive null mutations (FISHER 1928). His theory, if true, explains dominance phenomenon of all genes. However, Wright argued that selection for modifier is too weak to lead to the widespread dominance, and he proposed that the intrinsic property of metabolic systems causes dominance of enzyme genes (WRIGHT 1929).

The debate that whether Fisher (FISHER 1928) or Wright (WRIGHT 1929) correctly explained dominance lasted for more than half a century. Fisher's theory has received many

criticisms and is now considered untenable (CHARLESWORTH 1979; KACSER AND BURNS 1981; ORR 1991), and Wright's theory has gained popularity among biologists. The strongest evidence supporting Wright's idea was contributed by Kacser and Burns; they showed that halving the amount of one enzyme in a multi-enzyme linear pathway barely affects the total system flux (KACSER AND BURNS 1981). According to Kacser and Burns, the dominance of wild type allele occurs intrinsically, requiring no modifier whatsoever. Kacser' and Burns' result, now often referred to as metabolic control theory, standing on years of investigation of enzyme activities in the Kacser lab (KEIGHTLEY 1996), provides significant insights for the origin of genetic dominance as well as the dynamic of enzyme metabolic. Other phenomena, that are inconsistent with Fisher's modifier theory but consistent with Wright's intrinsic theory, were also reported; the two most telling arguments due to Charlesworth and Orr (KEIGHTLEY 1996). According to Charlesworth, Fisher's theory predicts no correlation between *h* and *s*, because the net selection for a modifier with effect of dh equals $2u\frac{\mathrm{d}h}{h}$, independent of *s* (dh is the change of modifier effect, *u* is the mutation rate). However, he observed a negative correlation between effect size *s* and the dominance coefficient *h* of the mutant allele (AA: 1, Aa:1-*hs*, aa: 1-*s*) using the *h* and *s* from different genes (CHARLESWORTH 1979). The negative *h*-*s* correlation by Charlesworth was later further confirmed with larger datasets (PHADNIS AND FRY 2005; MAREK AND KORONA 2016). Orr contributed a stronger evidence against Fisher's theory. He made use of data from artificial diploids of alga *Chlamydomonas reinhardtii*, a typical haploid unicellular organism. He showed that the recessive mutations are as common in the artificial diploid alga as in other natural diploids (ORR 1991). Because selection for dominance modifier cannot act in haploid genome, Orr's finding refutes Fisher's modifier theory. Moreover, Orr's finding demonstrates that dominance occurs intrinsically, consistent with Wright-Kacser-Burns theory. Because the

correct theory of dominance should not require selection on heterozygotes, Orr's finding simultaneously refutes two theories of Haldane (ORR 1991). Haldane suggested two explanations for dominance. One is selection for mutational robustness that wild type alleles which provide safe guard against heterozygous effects of mutations are favorable by natural selection (HALDANE 1930a); another is Haldane's sieve, bias against the establishment of recessive beneficial mutations due to their "invisibility" to selection in heterozygotes (HALDANE 1927; HALDANE 1930b).

Because the obvious caveats in theories that require selection to explain dominance, the Wright-Kacser-Burns theory is accepted as the leading theory on dominance. However, it has limitations. First of all, it predicts that dominance is only prevalent in enzyme genes, even though the widespread dominance in non-enzyme genes was also observed (PHADNIS AND FRY 2005). Lack of theoretical extension to non-enzyme genes raised question about the generality of the Wright-Kacser-Burns theory. Moreover, according to metabolic control theory, all wildtype enzymes in the same metabolic pathway are maintained at intermediate level in order to explain dominance (HARTL *et al.* 1985; WILKIE 1994; MAREK AND KORONA 2016). Marek and Korona showed that in starvation environment, when most enzyme levels are largely unbalanced, dominance remains strong, which suggests that dominance is not necessarily explained by metabolic of enzymes (MAREK AND KORONA 2016). Thus, none of the existing theories satisfactorily explains all patterns of dominance.

Although the progress of dominance theories has been slow moving in recent years, a related phenomenon to dominance, diminishing returns epistasis, has been continuously studied since discovery. A number of experimental evolution studies reported diminishing returns epistasis from advantageous mutations, which refers to the phenomenon that the same

advantageous mutation is less beneficial when occurring in fitter genotypes (CHOU *et al.* 2011; KRYAZHIMSKIY *et al.* 2014; WANG *et al.* 2016; WÜNSCHE *et al.* 2017). Wei and Zhang examined genetic interactions among standing genetic variation in yeast across 47 environments; they showed that a majority of the evaluated SNPs show diminishing returns in all environments and that returns from beneficial mutations also decreases in better environments, supporting modular life model in explaining diminishing returns epistasis (Wei and Zhang, 2018).

Apparently, diminishing returns epistasis and dominance are related. However, because the study of diminishing returns is predominantly in bacteria (CHOU *et al.* 2011; KRYAZHIMSKIY *et al.* 2014; WANG *et al.* 2016; WÜNSCHE *et al.* 2017), these two phenomena has not been discussed together. It is unknown whether share similar underlying mechanisms. Here we propose that genetic dominance is a special case of diminishing returns epistasis, because diminishing returns epistasis implies that gaining a wild-type allele at a locus is less beneficial in heterozygous mutants (fitter) than in homozygous mutants (less fit) irrespective of the function of the gene involved, which is exactly dominance. It is easy to misconceive "diminishing returns epistasis" with the traditional "diminishing returns curve". In previous work of genetic dominance, "diminishing returns" is sometimes used to describe the hyperbolic relationship between enzyme activity and total flux (KLINGENBERG 2004). However, this is different from the "diminishing returns epistasis" discussed here, which refers to the observation that gaining the same beneficial mutation on a fitter genotype background shows smaller benefit than on a less fit genotype background, and the fitter genotype does not necessarily contain more beneficial mutations on the same gene or same pathway.

Our hypothesis predicts that dominance and diminishing returns epistasis share the same underlying mechanism and can be presented by the same model. Under our hypothesis, the

model for diminishing returns epistasis should predict patterns of dominance that resemble the observations. Previous work demonstrated that diminishing returns epistasis originates from the modular organization of life where the contribution of each functional module to fitness is determined jointly by the genotype and environment (Wei and Zhang, 2018). Because dominance theory is about beneficial allele masking the effect of deleterious allele, yeast, as a single cell organism, whose growth rate is often used an unbiased fitness proxy, is good system to study dominance. We went on to test our hypothesis by simulating modular life model in diploid systems and by analyzing two large yeast datasets. We found that the empirical patterns of dominance are similar to previous findings about diminishing returns and can be predicted by the modular life model. In comparison, none of the previous models of dominance can fully explain these empirical results.

4.3 Result

4.3.1 Apply modular life model to diploid system

It is widely accepted that life is organized in a highly modular manner, where each module is a discrete object composed of a group of tightly linked components and performs a relatively independent task (RAFF 1996; HARTWELL *et al.* 1999; IHMELS *et al.* 2002; RAVASZ *et al.* 2002; BARABASI AND OLTVAI 2004; WALL *et al.* 2004; WAGNER *et al.* 2007). Modular life model was previously developed to explain diminishing returns epistasis. It posits that each module makes a distinct contribution to growth rate and the functionality of a module is the sum of all genetic effects and environmental effect, and that the growth rate is the geometric mean of the functionality of all modules (Wei and Zhang, 2018). Under this model, the same advantageous mutation may contribute to a module and growth rate greatly if the functionality of

the module is far from its maximum but may contribute only slightly if the module is approaching its maximal functionality. Furthermore, because the model hasn't any ploidy assumptions, it applies to diploid system as well.

To investigate whether the modular life model is able to recapitulate the existed empirical observations about genetic dominance, we conducted a computer simulation. We assume that the growth rate equals to the combined effects of *K* functional modules. Module *i* has a functionality value M_i which equals to the sum of environmental contribution E_i and genetic contributions G_i to the module and has a maximum functionality level of 1. Each gene has a fully functional allele and a null allele, and each module has *N* genes. The functional allele has an effect β_{ij} each, and with g_{ij} (which could be 0, 1, 2) indicating how many functional alleles there are for module *i* gene *j*. If a diploid genome has one functional allele, it adds β_{ij} to the corresponding module *i*, and if a diploid genome has two functional alleles, it adds 2 β_{ij} to the corresponding module. So growth rate is:

$$R = \prod_{i=1}^{K} M_i^{1/K}. \qquad (1)$$

Eq.1 could also be expended as:

$$R = \prod_{i=1}^{K} \min(\, E_i + \sum_{j=1}^{N} \beta_{ij}\, g_{ij}\,, 1)^{\frac{1}{K}} \qquad (2)$$

We considered that the growth rate of a genotype is determined by the geometric mean functionality of all modules, and a demonstration of this model is in Fig 4-1.

4.3.2 Widespread dominance and *h*-*s* correlation are predicted by modular life model

We first did a simulation based on modular life model to test whether prevalent dominance of wild type allele can be predicted. Since the previous studies showed this trend by

measuring fitness of heterozygous deletion and homozygous deletion on wild type background, we simulated the same process. To this end, environment effect is not considered. Let *x* be the effect of an allele, which could be on any module, and deleting the functional allele will decrease the functionality of the module it belongs to. Let R_{00}, R_{01}, R_{11}, be the growth rate for double deletion, heterozygous deletion, and wild type. We only consider the case where $R_{00} < R_{11}$ because otherwise this deletion is purely neutral.

$$R_{11} = M_j^{1/K} \prod_{i \neq j}^{K} M_i^{1/K} \quad (3)$$

$$R_{01} = \max\left(\left(M_j - x\right), 1 - t\right)^{1/K} \prod_{i \neq j}^{K} M_i^{1/K} \quad (4)$$

$$R_{00} = \max\left(\left(M_j - 2x\right), 1 - t - x\right)^{1/K} \prod_{i \neq j}^{K} M_i^{1/K} \quad (5)$$

Where when $M_j = 1$, due to saturation, the effect of removing one mutation *t* satisfies $0 \leq t \leq x \leq M_j/2 \leq 0.5$. These allow us to calculate the *h* of each deleterious mutation with or without the saturation of M_j.

When there is no saturation for module M_j, we get:

$$h = \frac{R_{11} - R_{01}}{R_{11} - R_{00}} = \frac{M_j^{1/K} - (M_j - x)^{1/K}}{M_j^{1/K} - \left(M_j - 2x\right)^{1/K}} \quad (6)$$

When there is saturation for module M_j, we get:

$$h = \frac{R_{11} - R_{01}}{R_{11} - R_{00}} = \frac{M_j^{1/K} - (M_j - t)^{1/K}}{M_j^{1/K} - \left(M_j - t - x\right)^{1/K}} \quad (7)$$

We proved mathematically that both $h < 0.5$ in both Eq.6 and Eq.7 (see Methods). To demonstrate this result by simulation, we simulated a "wild-type" genotype using a 10-module model. The level of each module is a random number uniformly chosen from 0.6-1, and growth rate of each genotype is calculated with Eq.1. We use the growth rate of each heterozygous deletion and homozygous deletion for each genotype, each module, and each allelic effect. We assume the deletion of each functional allele will decrease the corresponding module level by 0.06, 0.12, 0.18, 0.24, and 0.3. Using Eq. 6, we calculate the *h* for each gene deletion effect for each module on each genotype. We repeated this simulation 100 times. Not surprisingly, $h < 0.5$ is true for all deleterious mutations, meaning modular life model successfully generate the prevalent dominance of wild type allele.

We also proved mathematically that there exist a negative *h*-*s* correlation (see Methods). We used the results from the same 100 simulations to study the predicted correlation between *h* and *s*, where *s* equals to R_{11}-R_{00}. Interestingly, we also observed a strong negative correlation between *h* and *s* ($\rho = -0.997$, $P < 10^{-250}$, Fig 4-2a), meaning modular life model successfully generates the *h*-*s* correlation. Till here, modular life model successfully predicts the two known patterns of genetic dominance.

4.3.3 Modular life model predicts negative *Q*-*h* correlation

Although the origin of dominance has been a long-lasting question, it is not clear whether and how genetic dominance changes across environments due to the absence of systematic comparisons. In our previous study of diminishing returns epistasis, we define environmental quality (*Q*), the average growth rate of many genotypes measured in each environment (Wei and Zhang, 2018). We wonder whether there is a correlation between *Q* and *h* for gene deletions

under modular life model. To test this, we simulated 10 modules and 100 genotypes with modular level randomly chosen from 0.6-1. We simulated six environments each with a uniform contribution to each module with effect 0.05, 0.1, 0.15, 0.2, 0.25, 0.3. We assume all genotypes are homozygous wild-type allele of a gene with 0.1 allelic effect size in each module, and we calculated R_{00}, R_{01}, and R_{11} in each environment with Eqs.2-5. We then calculate the h for each gene in each environment. Q is estimated by the average growth rate $\bar{R}$ of the 100 wildtype genotypes in each environment, which follows the same rank order as the environment contribution we simulated. For each gene and each genotype, we measure h in each environment, and we calculate the rank correlation between Q and h. The resulting 1000 correlations are predominately negative (97%, binomial $P < 10^{-24}$, Fig 4-2b). This is a new pattern that has never been reported before, which, if genuine, suggests that dominance of wild type allele increases as the environment quality improves.

4.3.4 Negative *Q*-*h* correlation for yeast gene deletions

We first test the model predicted *Q*-*h* correlation by reanalyzing the genetic dominance data for a set of yeast nonessential genes generated by Marek and Korona (MAREK AND KORONA 2016). Two different growth conditions were used to measure dominance in their study, YPD and starvation. The maximum growth rates in YPD and maximum lifespans in starvation condition were measured individually for each genotype, and h were calculated accordingly (MAREK AND KORONA 2016). Because cells do not grow under starvation, it is reasonable to consider starvation environment as the lower Q condition. We used all the genes that have h measured in both conditions to study *Q*-*h* correlation. Among the 369 genes measured in both conditions, 218 genes have smaller h in YPD environment and 151 genes have smaller h in

starvation environment (Binomial P=1.93×10^{-4}), proving that h decreases as Q increases. This is consistent with the modular life model prediction (Fig 4-2C).

4.3.5 Negative *Q*-*h* correlation for yeast polymorphisms

Although dominance is mostly studied for gene deletions or large effect deleterious mutations, it can occur between two alleles of standing variation. Because theory and data both predict negative *hs*-correlation, the h from standing variation should be only slightly smaller than 0.5. Nevertheless, due to scarcity of data with multiple environments, we use genetic polymorphisms data to study dominance. To this end, we test the correlation between Q and h by analyzing the growth rate data of 7310 genotyped diploids of yeast in 9 environments. In this dataset, the number of cells of each genotype was measured continuously between 0 and 72h from growth on agar plate made of 9 different YPD based mediums each with one commonly used chemical. We converted the number of cells at 32h, 40h, and 48h into average growth rate (see Methods). We previously showed by experiment that this conversion is at robust to growth saturation up to at least 48h (Wei and Zhang, 2018). Q for each environment is measured by averaging the average growth rate of all genotypes. Because each diploid genome could be AA, Aa, and aa at each SNP level, we measure R(AA), R(Aa), and R(aa) by averaging the average growth rate of all genotypes with AA, Aa, or aa at each segregating locus. Using gene deletions and polymorphisms are quite different because: 1) the majority of genetic polymorphisms are effectively neutral, and 2) polymorphic sites are likely to have effects in fewer environments than gene deletions. In order to calculate h for beneficial SNPs, we first removed SNPs with small effects to improve signal to noise ratio and then calculate h for the remaining SNPs of each environment (see Methods). Because of the aforementioned reasons, different environments have different remaining SNPs, and direct comparison for the h of each SNP across environments is

difficult. Instead, we calculate the fraction of remaining SNPs showing $h < 0.5$ (g) as the dominance level of each environment.

To validate this approach, we first predict *Q*-*g* correlation under modular life model simulation. We simulated 10 modules with 5 genes per module, assuming each functional allele contribute 0.12 to each module. Each simulated haploid has a random genotype containing 50 genes with either 0 or 1 functional copy. We mate the 86 simulated *a* cell and 85 simulated α cells into 7310 diploid genotypes, resemble the data we used. 9 environments of different quality are simulated, assuming the environment effect to each module follow a normal distribution with mean 0.1, 0.14, 0.18… 0.42 contributions to each module and variance 0.01. We then take similar procedure in estimating *h*, *Q* and *s*. We filtered noise by using allele with estimated effect size larger than 0.01 in each environment to calculate *h* and calculated the *g* using the remaining genes. We found that, 99 out of 100 simulations (binomial $P < 10^{-28}$), the correlation between *g* and *Q* is positive (Fig 4-3AB), suggesting that positive correlation is expected by modular life model when population data is used to calculate genetic dominance.

Because modular life simulation suggests positive *Q*-*g* correlation when using polymorphism data, so we went on to test it with empirical data. Indeed, we observe positive *Q*-*g* correlation (Fig 3C, Fig E-S1), meaning better environment tend to have higher dominance level for functional polymorphisms. The empirical *P*-values (see Method) from linear regression are significant for all three time points we used.

4.3.6 *Q*-*h* correlation is unexpected in the Wright-Kacser-Burns model

By analyzing two different datasets and by performing simulations with modular life model, we confirm a negative *Q*-*h* correlation. Because this correlation has not been reported

before, nor has it been used to test the correctness of theories of dominance, we test whether the Wright-Kacser-Burns model could also predict this correlation. First, we follow the Wright-Kacser-Burns model to get its prediction for *Q*-*h* correlation.

According to metabolic control theory, flux (*F*) equals to growth, and the flux of a linear pathway follows:

$$F = C/ (\sum_1^n 1/Z_i), \qquad (8)$$

where *C* represents the environmental parameters, and *Z* represents the genetically determined parameters of an enzyme (KACSER AND BURNS 1981). According to Eq. 8, environment could have two effects: 1) increase/decrease metabolic reaction by increasing/decreasing substrates, 2) change enzymes' activities. Here, we discuss these two scenarios separately.

Let Z_k be the activity of enzyme *k* when it is homozygous wildtype, then $Z_k/2$ is the activity of enzyme *k* for heterozygous in the same environment. Then the dominance coefficient for the mutant allele h_k follows:

$$h_k = 1- [C [1/ (2/Z_k + \sum_{i\neq k} 1/Z_i)]]/ [C [1/ (1/Z_k + \sum_{i\neq k} 1/Z_i)]]$$

$$= 1- (1/Z_k + \sum_{i\neq k} 1/Z_i)/ (2/Z_k + \sum_{i\neq k} 1/Z_i). \qquad (9)$$

To further simplify Eq.9, we can replace the combined effect of all enzymes $i \neq k$ in the pathway with

$$\sum_{i\neq k} 1/Z_i = 1/Z_{\text{All}}\,, \qquad (10)$$

Combine Eq.9 and Eq.10, we get:

$$h_k = Z_{All}/(Z_k+2Z_{All}). \qquad (11)$$

Assume that Z_k increases to Z'_k during the environment shift, and Z_{All} changes to Z'_{All}, and h_k becomes h'_k. Put the new parameters into Eq.11, we get:

$$h'_k = Z'_{All}/(Z'_k+2Z'_{All}). \qquad (12)$$

Because Eq. 11 and Eq. 12 share the same mathematical expression, and that they are both independent of environmental parameter *C*, the formula does not predict any directional change of *h*. Therefore, the metabolic flux model predicts no correlation between *Q* and *h* and cannot explain the observation that the majority of genes and polymorphisms show the *Q-h* correlation.

4.3.7 Diminishing returns epistasis could not be explained by previous models

Although it is not necessary for a model of dominance to explain diminishing returns epistasis, being able to explain both phenomena makes modular life model more general, so we went on to test whether the other models could also predict diminishing returns epistasis. Diminishing returns epistasis has two general trends: 1) the same beneficial mutation has smaller effects on fitter genotype backgrounds, and 2) the same beneficial mutation has smaller effects on fitter environment backgrounds (Wei and Zhang, 2018). Both Fisher's theory and Haldane's theories rely on selection on heterozygotes, thus are inapplicable to the diminishing returns epistasis in haploid, where functional allele has only 0 and 1 state. To this end, we discussed why the Wright-Kacser-Burns theory could not satisfactorily explain diminishing returns epistasis.

Despite that the Wright-Kacser-Burns model predicts the diminishing returns curve between enzyme activity and flux (KACSER AND BURNS 1981; DYKHUIZEN *et al.* 1987), it fails to

explain diminishing returns epistasis even for enzymes, as we discussed below. First, we discuss the effect of gaining a single beneficial mutation on two genotype backgrounds. Let enzyme *k* in the pathway improves its activity from Z_k to Z'_k by one single mutation, and we combine this with Eq. 8 to calculate the fitness improvement *s* from this single mutation:

$$s = C\,[1/\,(1/Z'_k + \textstyle\sum_{i\neq k} 1/Z_i) - 1/\,(1/Z_k + \sum_{i\neq k} 1/Z_i)]. \qquad (13)$$

To simplify Eq. 13, we replace the combined effect of all enzymes $i \neq k$ in the pathway using Eq. 10, and get:

$$s = C\,[1/\,(1/Z'_k + 1/Z_{All}) - 1/\,(1/Z_k + 1/Z_{All})]. \qquad (14)$$

Background fitness improvement under the Wright-Kacser-Burns theory could be represented by increasing Z_{All} to Z'_{All}. Then the effect of this mutation becomes *s'*, where:

$$s' = C\,[1/\,(1/Z'_k + 1/Z'_{All}) - 1/\,(1/Z_k + 1/Z'_{All})]. \qquad (15)$$

We use Mathematica to simplify Eq. 14-Eq. 15 and to calculate the critical value for *s'- s*. We found that when $Z'_{All} > 0$, *s'- s* increases with Z'_{All} monotonically. Because when $Z'_{All}=Z_{All}$, $s' = s$, so $s' > s > 0$ for all $Z'_{All} > Z_{All} > 0$. Therefore, a beneficial mutation of an enzyme on a fitter genotype with better pathway performance has bigger fitness benefit, which is the opposite of diminishing returns epistasis. The Wright-Kacser-Burns l theory predicts predominant synergistic effect of beneficial mutations for enzymes, contradictory to the first pattern of diminishing returns epistasis.

Now let's consider the second pattern of diminishing returns epistasis regarding the mutational effect in high *Q* and low *Q* environments. Environmental quality increases in the Wright-Kacser-Burns model could be seen as the environmental parameter in Eq. 1 increases

from *C* to *C'* such that the fitnesses of all genotypes increase. The effect size in Eq. 14 will change from *s* to *s'* following:

$$s = C' [1/ (1/Z'_k + 1/Z_{All}) - 1/ (1/Z_k + 1/Z_{All})]. \quad (16)$$

Combining Eq. 14 and Eq. 16, we get:

$$s' = s\, C'/ C, \quad (17)$$

So *s'* increases as environment gets better. Therefore, the Wright-Kacser-Burns model predicts the opposite of the two patterns of diminishing returns epistasis, indicating it is not as general as modular life model.

4.4 Discussion

To summarize, we hypothesize that dominance is a special case of diminishing returns epistasis arriving from interactions among genes of the same functional modules. We extend the modular life model of diminishing returns epistasis to diploid system and use it to predict the patterns of dominance. Simulation using modular life model predicts a negative *Q-h* correlation, which is verified by two large yeast datasets. We find that the Wright-Kacser-Burns model could not predict the negative *Q-h* correlation for genetic dominance, nor could it predict diminishing returns epistasis in haploids. In contrast, modular life model not only predicts all current observations of genetic dominance but also predicts diminishing returns epistasis.

The origin of genetic dominance has been a long-standing question in evolutionary genetics, and finding the correct model/theory is important to revealing the mechanistic causes. We focus on discussing the differences among our model, Fisher's model, Haldane's model and the Wright-Kacser-Burns' model (a summary of the comparison in Table 1), although some

recent attempts has been made by Manna and colleagues using a bivariate Gaussian model (MANNA *et al.* 2011; MANNA *et al.* 2012). The Gaussian model cannot predict the well-known negative *h-s* correlation (MANNA *et al.* 2011; MANNA *et al.* 2012), nor does it provide a mechanistic explanation for using the bivariate Gaussian. We showed that none of the previous theories is sufficient to explain all the current observations of dominance. In contrast, modular life model predicts *h-s* correlation, *h-Q* correlation, dominance, diminishing returns epistasis, overdominance, and using the modular structure of life to explain these phenomena.

We predict and observe the negative *Q-h* correlation meaning higher dominance in better environments, based on modular life model prediction and the analysis of yeast deletion and yeast polymorphisms datasets. This new finding indicates that dominance shares the property of diminishing returns, because not only the returns from gaining a wildtype allele is smaller on the heterozygous background than on the homozygous mutant background (i.e. dominant), but also the returns of an extra wildtype allele becomes even smaller (.e. more dominant) as the environment becomes better. This new finding suggests that dominance changes during adaptation and environment fluctuations, and the level of dominance/diminishing returns reflects how adapted the genome is. Even for conserved genes, the dominant level may increase or decrease according to the genotype and environment. Fisher's theory or Haldane's sieve do not predict higher dominance in better environments, unless we assume the population has adapted to all tested environments and they are more adapted to high fitness environments than low fitness environments. However, the environments used in the yeast datasets are quite arbitrary, and high environmental quality can be a feature of the environment rather than adaptation. Moreover, Fisher's theory was refuted by many other previous observations of genetic

dominance; Haldane's sieve cannot explain Orr's result (ORR 1991) , nor can it explain why new deleterious mutations are partially recessive (MUKAI AND YAMAZAKI 1968).

Under modular life model, genetic dominance couples with selection for high fitness, so genetic dominance arises intrinsically during adaptation. Because of this coupling effect, it avoids the problem of using selection for weak effects (WRIGHT 1929) (Fisher's modifier theory and Haldane's robustness explanation) to explain the prevalence of genetic dominance. Moreover, the arrival of diminishing returns/dominance is unavoidable (also intrinsic) under this model, because as long as historical contingency exist, the genotype is unlikely to be maladapted for all modules. The intrinsic origin is a pivotal advantage for the Wright-Kacser-Burns theory, but because their model requires all enzymes at intermediate level, selection has been used to explain why enzyme activities are neither too high nor too low (WILKIE 1994). Because selection does not directly act on enzyme activity, explaining the intermediate enzyme activity by selection is probable but somewhat difficult. By coupling genetic dominance with selection on main mutational effect *s*, our model bypasses the difficulty of explaining dominance by selection and allows dominance to exist for all genes.

A big advantage of modular life model is that it was not designed retrospectively to explain genetic dominance as were both Fisher's and Wright's models. Even so, it more satisfactorily explains all current observations of genetic dominance, compared to the previous retrospective models. Moreover, it provides the connection between the two widespread phenomena in genetics and evolution, dominance and diminishing returns epistasis. Neither Fisher's model nor Wright's model is able to explain diminishing returns epistasis. Although they were not retrospectively built to explain diminishing returns, they do not share the generality of modular life model. Given the high similarity between dominance and diminishing

returns epistasis, a model that sufficiently explains both phenomena is superior. This and other results suggest that modular life model might be generally applicable in explaining the effect sizes of mutations and genotype-phenotype mapping.

Generality and specificity sometimes do tradeoff. Note that, metabolic control theory is formalized based on enzyme pathway activities, and it is good at explaining the enzyme metabolic flux (DYKHUIZEN *et al.* 1987; NIEDERBERGER *et al.* 1992). The evidences in this work only show its consistentency with all patterns of genetic dominance thus should not be the model of genetic dominance. Refuting its prediction power for genetic dominance does not contradict it being a model for metabolic flux. Similarly, just because modular life model provides a simple explanation for all genes and it is compatible with all current patterns do not mean it can provide specific prediction for a specific group of genes.

In this paper, we assume one gene only improve one module, while the reality could be more complicated. We find that modular life model can successfully explain overdominance (see Supplementary Materials) assuming the two alleles slightly differ in their functions. Future work could explore the possibility of using modular life model to explain more complicated mutational effect.

Although modular life model seems to be very general, the predictions it made are testable predictions thus it is refutable and has the potential to be falsified and refined. Some other predictions of it can be tested in the future. For example, the model predicts transitive relation of dominance (assuming that one gene only contributes to one module), such that if gene A has three alleles, A_1 is dominant to A_2, and A_2 is dominant to A_3, then A_1 is dominant to A_3. If future studies found results mostly consistent with modular life model predictions, the model will

be further supported. Moreover, it predicts that, in the absence of genetic incompatibility, the hybrid between two homozygous diploid genotypes should not be lower than the less fit parent's growth rate, but the hybrid growth rate could be better than both parents or anywhere in between.

It is possible that some future models could also explain all the phenomena modular life model explains, but such models are currently unavailable. Thus, it is worthwhile to explore more of this model especially in the light of molecular mechanism. The molecular mechanisms of dominance has been discussed, where dosage change, structural alternation, toxic, and functional change mutations were discussed for dominance at different phenotypic levels (WILKIE 1994). The molecular mechanism of diminishing returns epistasis has not been reviewed, but our work suggests diminishing returns expistasis may share the mechanisms of dominance. Future work may combine the modular life model with molecular mechanisms to justify its usability as a model for mutational effects.

4.5 Material and methods

4.5.1 Genome and phenotype data in yeast gene deletion

We downloaded the supplementary data from Marerk and Korona (MAREK AND KORONA 2016). We chose only the genes with *s* and *h* measured in both regular and starvation environments in our analysis, which restrict it into 369 total gene deletions.

Genotype and average growth rate for diploid yeast hybrids

We acquired from the Hallin et al the genotype data and of 7310 diploids from a cross between 86 *MATa* and 86 *MATα* strains haploid of *S. cerevisiae* (HALLIN *et al.* 2016). The haploids were randomly drawn from a twelfth generation two-parent intercross pool which is mated from two wild strains sampled in North America and West Africa (HALLIN *et al.* 2016).

For each genomic region that several SNPs are completely in linkage with each other with no recombination in any diploid genome, we keep only the middle SNP. This way, we have 13350 remaining SNPs in our analysis.

We also acquired the unsmoothed cell numbers at the time points (between 0 and 72h) of their measurements for each of the diploid hybrid for all the nine environments they used. The cell number is measured based on the cell growth on agar plates (ZACKRISSON *et al.* 2016) and each diploid contains 8 replicate measurements.

For this yeast polymorphism data we used, we have cell numbers measured at different time points based on their growth on solid medium. We follow the following formulas to get average growth rate of each genotype from cell number. Cell growth can be described by

$$N = N_0 e^{\int_0^T R(t)dt} = N_0 e^{\bar{R}T}, \qquad (16)$$

where N_0 is the number of colonizing cells, N is the number of cells at time T, $R(t)$ is the growth rate at time t, and $\bar{R}$ is the average growth rate from time 0 to T. From Eq. 16, we have

$$\bar{R} = \frac{1}{T}\ln\frac{N}{N_0}, \qquad (17)$$

Because N_0 could be seen as a constant when there are 8 replicated measurements for each genotype, we use $\frac{1}{T}\ln N$ as growth rate. We use the cell numbers from 3 intermediate time points: 32h, 40h, and 48h. If a diploid hasn't been measured in one environment, it will be removed from the analysis in that environment. If multiple replicates are available, we average the $\bar{R}$ of all the replicates.

If a SNP has no effect, then $\overline{R_{\mathrm{AA}}}$, $\overline{R_{\mathrm{Aa}}}$, $\overline{R_{\mathrm{aa}}}$ are random numbers, so their h has 67% chance to be outside 0 and 1. So filtering out SNPs with small effects could reduce noise, and the fraction of remaining SNPs with h between 0 and 1 should increase. Because the majority of

SNPs do not have fitness effects and because we want to only calculate *h* for SNPs with *h* between 0 and 1, we filter out SNPs with smaller effects and SNPs whose *h* are outside 0 and 1. For each condition, we use different cutoffs for $s = |\overline{R_{AA}} - \overline{R_{aa}}|$ to filter out the SNPs with small effects (due to noise) and then calculate the fraction of remaining SNPs with heterozygotes having intermediate fitness. We find that, as we increase the cutoff from 0 to 0.065, the fraction of such SNPs increases from about 90% to about 98% suggesting the noise significantly decreases and all conditions have at least 98% remaining, but further increasing the cutoff from 0.065 to 0.1 does not improve the fraction of such SNPs (Fig E-S2). We therefore used 0.065 as the cutoff for all conditions and all time points.

4.5.3 Modular life model predicts dominance mathematically

We first show that under modular life model $h < 0.5$ is true when there is no saturation. We can rewrite Eq. 6 in the following form:

$$h = \frac{M_j^{1/K} - (M_j - x)^{1/K}}{M_j^{1/K} - (M_j - 2x)^{1/K}} = \frac{M_j^{1/K} - (M_j - x)^{1/K}}{[M_j^{1/K} - (M_j - x)^{\frac{1}{K}}] + [(M_j - x)^{\frac{1}{K}} - (M_j - 2x)^{\frac{1}{K}}]} \quad (18)$$

Name a new function $f(M_j)$, which follows:

$$f(M_j) = M_j^{1/K} - (M_j - x)^{1/K} \quad (19)$$

Take Eq. 19 into Eq. 18, we get:

$$h = \frac{f(M_j)}{f(M_j) + f(M_j - x)} \quad (20)$$

Because the derivative of $f(M_j)$ follows

$$f'(M_j) = \frac{1}{K}(M_j^{\frac{1}{K}-1} - (M_j - x)^{\frac{1}{K}-1}) = \frac{1}{K}((\frac{1}{M_j})^{1-\frac{1}{K}} - (\frac{1}{M_j - x})^{1-\frac{1}{K}}) \quad (20)$$

Because $f'(M_j) < 0$ under the condition of modular life model that $0 \le x \le M_j/2 \le 0.5$ and that $K > 1$, $f(M_j) < f(M_j - x)$. Therefore,

$$h = \frac{f(M_j)}{f(M_j) + f(M_j - x)} < \frac{f(M_j)}{f(M_j) + f(M_j)} = 0.5 \qquad (21)$$

We then show that under modular life model $h < 0.5$ is true when there is saturation. We can rewrite Eq. 7 in the following form:

$$h = \frac{M_j^{1/K} - (M_j - t)^{1/K}}{M_j^{1/K} - (M_j - t - x)^{1/K}} = \frac{M_j^{1/K} - (M_j - t)^{1/K}}{[M_j^{1/K} - (M_j - t)^{\frac{1}{K}}] + [(M_j - t)^{\frac{1}{K}} - (M_j - t - x)^{\frac{1}{K}}]} \qquad (22)$$

Name a new function $g(M_j)$, which follows:

$$g(M_j) = M_j^{1/K} - (M_j - t)^{1/K} \qquad (19)$$

Because $g(M_j)$ and $f(M_j)$ only differs in the t term, and because $0 \leq t \leq M_j/2 \leq 0.5$ and $K > 1$ are true, the derivative of $g(M_j)$: $g'(M_j) < 0$, so $g(M_j) < g(M_j - x)$ is also true. We can then rewrite Eq. 22 as:

$$h = \frac{g(M_j)}{g(M_j) + g(M_j - x)} < \frac{g(M_j)}{g(M_j) + g(M_j)} = 0.5 \qquad (20)$$

Thus, under modular life model, beneficial alleles are dominant with or without saturation effect.

4.5.4 Modular life model predicts *h*-*s* correlation mathematically

We then show *h*-s correlation under modular life model first for no saturation case. To deal with this question, we assume M_j is fixed, and x changes. And, let

$$h(x) = \frac{M_j^{1/K} - (M_j - x)^{1/K}}{M_j^{1/K} - (M_j - 2x)^{1/K}} = \frac{1 - (1 - x/M_j)^{1/K}}{1 - (1 - 2x/M_j)^{1/K}} \qquad (21)$$

when there is no saturation. Let a new $t = x/M_j$. And we get $F(t)$:

$$F(t) = \frac{1 - (1 - t)^{1/K}}{1 - (1 - 2t)^{1/K}} \qquad (22)$$

The effect size s of the beneficial allele is an increasing function of x, because it follows:

$$s = R_{11} - R_{00} = M_j^{1/K} - (M_j - 2x)^{1/K} \qquad (23)$$

Within modular life model's parameter range, the derivative of $F(t)$ follows:

$$F'(t) = \frac{\frac{1}{K}(1-t)^{\frac{1}{K}-1}\left(1-(1-2t)^{\frac{1}{K}}\right)-\frac{2}{K}(1-2t)^{\frac{1}{K}-1}\left(1-(1-t)^{\frac{1}{K}}\right)}{(1-(1-2t)^{1/K})^2}$$

$$= \frac{(1-t)^{\frac{1}{K}-1}\left(1-(1-2t)^{\frac{1}{K}}\right)-2\,(1-2t)^{\frac{1}{K}-1}\left(1-(1-t)^{\frac{1}{K}}\right)}{K(1-(1-2t)^{1/K})^2} \qquad (24)$$

The sign of $F'(t)$ depends only on the numerator part, because the denominator is positive. We want to prove that $F'(t) < 0$, so that *h-s* are negatively correlated.

Divide $F'(t)$ by $(1-t)^{\frac{1}{K}-1}(1-2t)^{\frac{1}{K}-1} > 0$, we get:

$$F'(t) = \frac{(1-2t)^{1-\frac{1}{K}} - (1-2t) - 2\left((1-t)^{1-\frac{1}{K}} - (1-t)\right)}{K(1-(1-2t)^{1/K})^2} \qquad (25)$$

So we need $(1-2t)^{1-\frac{1}{K}} - (1-2t) - 2\left((1-t)^{1-\frac{1}{K}} - (1-t)\right) < 0$, let

$$L(t) = (1-2t)^{1-\frac{1}{K}} - (1-2t) - 2\left((1-t)^{1-\frac{1}{K}} - (1-t)\right)$$

$$= (1-2t)^{1-\frac{1}{K}} - 2(1-t)^{1-\frac{1}{K}} + 1 \qquad (26)$$

Calculate the derivative of $L(t)$, we get:

$$L'(t) = -2\left(1-\frac{1}{K}\right)(1-2t)^{-\frac{1}{K}} + 2\left(1-\frac{1}{K}\right)(1-t)^{-\frac{1}{K}}$$

$$= 2\left(1-\frac{1}{K}\right)\left((1-t)^{-\frac{1}{K}} - (1-2t)^{-\frac{1}{K}}\right) \qquad (27)$$

Therefore, $L'(t) < 0$ when $t > 0$. So $F'(t) < 0$ when $t > 0$. So *h-s* are negatively correlated. When there is saturation in a module, the *h-s* has no correlation unless with specific parameter assumptions. It does not have a mathematical solution. But because the majority of genes with effect are in non-saturated modules, we expect to see *h-s* correlation even when some modules are saturated.

4.6 References

Barabasi, A. L., and Z. N. Oltvai, 2004 Network biology: understanding the cell's functional organization. Nat Rev Genet 5**:** 101-113.

Charlesworth, B., 1979 Evidence against Fisher's theory of dominance. Nature 278**:** 848-849.

Chou, H.-H., H.-C. Chiu, N. F. Delaney, D. Segrè and C. J. Marx, 2011 Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science 332**:** 1190-1192.

Dykhuizen, D. E., A. M. Dean and D. L. Hartl, 1987 Metabolic flux and fitness. Genetics 115**:** 25-31.

Fisher, R. A., 1928 The possible modification of the response of the wild type to recurrent mutations. The American Naturalist 62**:** 115-126.

Haldane, J., 1930a A note on Fisher's theory of the origin of dominance, and on a correlation between dominance and linkage. The American Naturalist 64**:** 87-90.

Haldane, J. B. S., 1927 A mathematical theory of natural and artificial selection, part V: selection and mutation, pp. 838-844 in *Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge University Press.

Haldane, J. B. S., 1930b A mathematical theory of natural and artificial selection.(Part VI, Isolation.), pp. 220-230 in *Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge University Press.

Hallin, J., K. Märtens, A. I. Young, M. Zackrisson, F. Salinas *et al.*, 2016 Powerful decomposition of complex traits in a diploid model. Nature communications 7.

Hartl, D. L., D. E. Dykhuizen and A. M. Dean, 1985 Limits of adaptation: the evolution of selective neutrality. Genetics 111**:** 655-674.

Hartwell, L. H., J. J. Hopfield, S. Leibler and A. W. Murray, 1999 From molecular to modular cell biology. Nature 402**:** C47-52.

Ihmels, J., G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv *et al.*, 2002 Revealing modular organization in the yeast transcriptional network. Nat Genet 31**:** 370-377.

Kacser, H., and J. A. Burns, 1981 The molecular basis of dominance. Genetics 97**:** 639-666.

Keightley, P. D., 1996 A metabolic basis for dominance and recessivity. Genetics 143**:** 621-625.

Klingenberg, C. P., 2004 Dominance, nonlinear developmental mapping and developmental stability. The biology of genetic dominance**:** 37-51.

Kryazhimskiy, S., D. P. Rice, E. R. Jerison and M. M. Desai, 2014 Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science 344**:** 1519-1522.

Manna, F., R. Gallet, G. Martin and T. Lenormand, 2012 The high-throughput yeast deletion fitness data and the theories of dominance. Journal of evolutionary biology 25**:** 892-903.

Manna, F., G. Martin and T. Lenormand, 2011 Fitness landscapes: an alternative theory for the dominance of mutation. Genetics 189**:** 923-937.

Marek, A., and R. Korona, 2016 Strong dominance of functional alleles over gene deletions in both intensely growing and deeply starved yeast cells. Journal of evolutionary biology 29**:** 1836-1845.

Mendel, G., 1996 Experiments in plant hybridization (1865). Verhandlungen des naturforschenden Vereins Brünn.) Available online: www. mendelweb. org/Mendel. html (accessed on 1 January 2013).

Mukai, T., S. I. Chigusa, L. Mettler and J. F. Crow, 1972 Mutation rate and dominance of genes affecting viability in Drosophila melanogaster. Genetics 72**:** 335-355.

Mukai, T., and T. Yamazaki, 1968 The genetic structure of natural populations of Drosophila melanogaster. V. Coupling-repulsion effect of spontaneous mutant polygenes controlling viability. Genetics 59**:** 513.

Niederberger, P., R. Prasad, G. Miozzari and H. Kacser, 1992 A strategy for increasing an in vivo flux by genetic manipulations. The tryptophan system of yeast. Biochem J 287 (Pt 2)**:** 473-479.

Orr, H. A., 1991 A test of Fisher's theory of dominance. Proceedings of the National Academy of Sciences 88**:** 11413-11415.

Phadnis, N., and J. D. Fry, 2005 Widespread correlations between dominance and homozygous effects of mutations: implications for theories of dominance. Genetics 171**:** 385-392.

Raff, R. A., 1996 *The shape of life*. University of Chicago Press, Chicago.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabasi, 2002 Hierarchical organization of modularity in metabolic networks. Science 297**:** 1551-1555.

Wagner, G. P., M. Pavlicev and J. M. Cheverud, 2007 The road to modularity. Nat Rev Genet 8**:** 921-931.

Wall, M. E., W. S. Hlavacek and M. A. Savageau, 2004 Design of gene circuits: lessons from bacteria. Nat Rev Genet 5**:** 34-42.

Wang, Y., C. D. Arenas, D. M. Stoebel, K. Flynn, E. Knapp *et al.*, 2016 Benefit of transferred mutations is better predicted by the fitness of recipients than by their ecological or genetic relatedness. Proceedings of the National Academy of Sciences 113**:** 5047-5052.

Wilkie, A. O., 1994 The molecular basis of genetic dominance. J Med Genet 31**:** 89-98.

Wright, S., 1929 Fisher's theory of dominance. The American Naturalist 63**:** 274-279.

Wünsche, A., D. M. Dinh, R. S. Satterwhite, C. D. Arenas, D. M. Stoebel *et al.*, 2017 Diminishing-returns epistasis decreases adaptability along an evolutionary trajectory. Nature Ecology & Evolution 1**:** 0061.

Zackrisson, M., J. Hallin, L.-G. Ottosson, P. Dahl, E. Fernandez-Parada *et al.*, 2016 Scan-o-matic: high-resolution microbial phenomics at a massive scale. G3: Genes| Genomes| Genetics 6**:** 3003-3014.

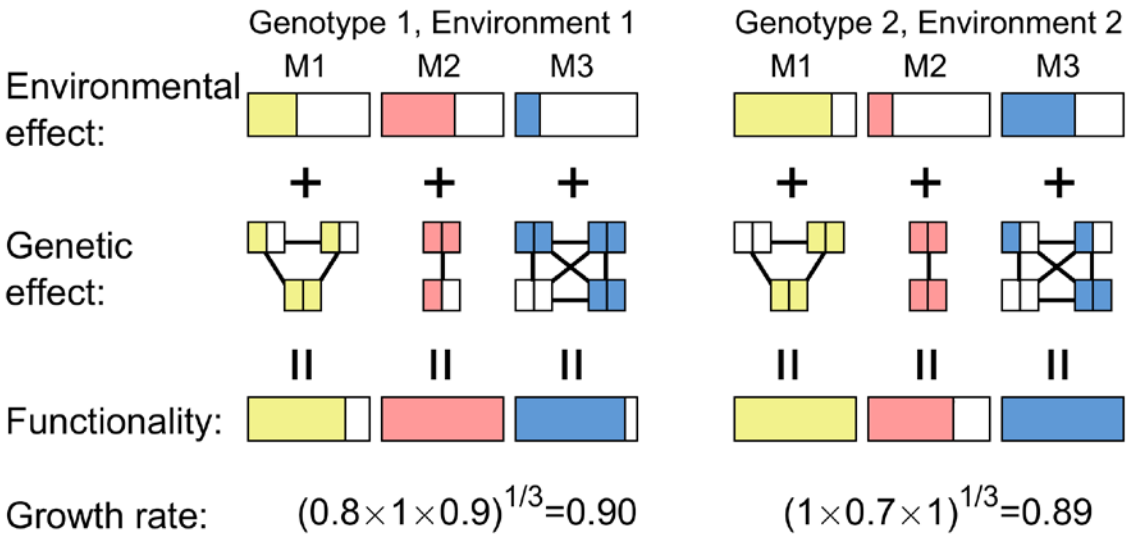


Figure 4-1. Modular life model in diploid systems. Different modules (M1, M2, and M3) are colored differently. Different environments (Environments 1 and 2) contribute differently to various modules, as illustrated by the different sizes of the three color-filled boxes. Each module contains a number of biallelic genes, shown as two connected boxes, each of which could have either a functional allele designated as 1 (filled box) or a null allele designated as 0 (open box). If both boxes are filled, the genotype has two functional alleles of the gene; if only one box is filled, the genotype has one functional allele; if zero box is filled, the genotype has no functional allele of the gene. Two genotypes (Genotypes 1 and 2) are shown as examples. The functionality of a module has a maximum of 1, and is the sum of environmental and genetic contributions. The growth rate of each genotype is computed from the functionalities of the individual modules using the formula indicated, which equals the geometric mean of all modules.

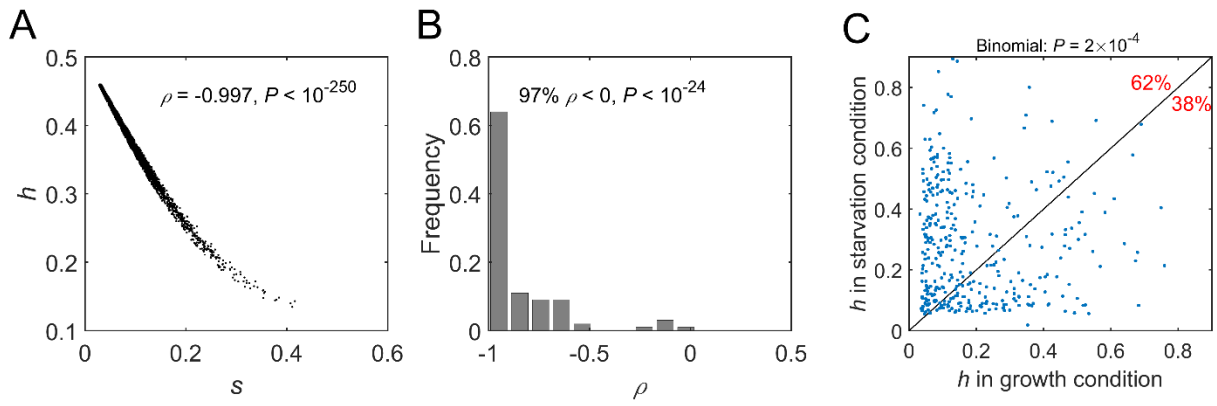


Figure 4-2. The inferences about dominance from modular life model. (A) *h* decreases as *s* increases. X-axis if the effect size *s* of a gene and y-axis is the *h* of deleting the gene on one genotype background. Each dot represents the deletion effect of one gene on one background. ρ, the spearman correlation. (B) The distribution of *Q*-*h* correlations. This is based on 100 simulations, and the x-axis is the ρ, the spearman correlation of the correlation. (C) The *Q*-*h* correlations in yeast dataset I. Each dot represents one gene.

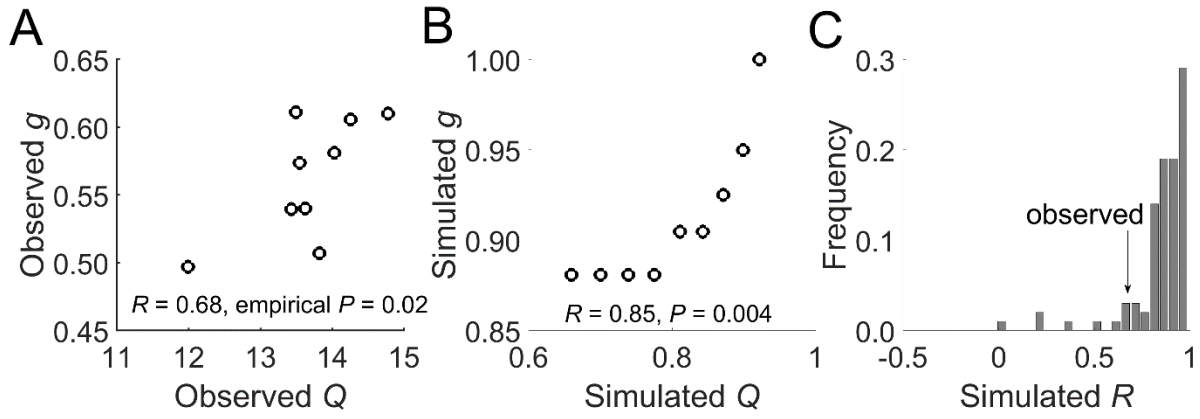


Figure 4-3. Positive correlation between environmental quality and the fraction of genes/SNPs showing $h < 0.5(g)$. (A) The observed result at 40h between Q and g. Each dot represents one environmental condition. Linear correlation coefficient R and empirical P (from 1000 random shuffling of x- and y- axes numbers) are listed. (B) An example of the correlation between Q and g in modular life model simulation. Each dot represents one environmental condition. Linear correlation result is listed. (C) The distribution of all R from 100 simulations. The arrow pointed place is the observed R from data.

Chapter 5

Environment-dependent pleiotropic effects of mutations on growth rate and carrying capacity of population growth

"The cost of a thing is the amount of what I will call life which is required to be exchanged for it, immediately or in the long run."

— **Henry David Thoreau**

5.1 Abstract

Growth rate (*r*) and carrying capacity (*K*) are key life history traits that together characterize the density-dependent population growth, and therefore are crucial parameters of many ecological and evolutionary theories. Although *r* and *K* are generally thought to be negatively correlated, both *r*-*K* tradeoffs and tradeups have been observed. However, neither the conditions under which each of these relationships occur nor the causes of these relationships are fully understood. Here we address these questions using genetic mappings of *r*-QTLs and *K*-QTLs followed by mathematical modeling. We estimated *r* and *K* using the growth curves of more than 7000 yeast recombinant diploid genotypes in nine lab environments and found that the *r*-*K* correlation

among genotypes changes from 0.53 to -0.52 with the rise of the environment quality, measured by the mean *r* of all genotypes in the environment. Many QTLs simultaneously influence *r* and *K*, but the directions of their effects are environment-dependent such that a QTL could show concordant effects on the two traits in a poor environment but antagonistic effects in a rich environment. We propose that these varying trends are generated by the relative impacts of two factors: the tradeoff between the speed and efficiency of ATP production and the energetic cost of cell maintenance relative to reproduction, and demonstrate a good agreement between model predictions and empirical observations. Together, these results reveal and explain the complex environment-dependency of the *r*-*K* relationship, which bears on many ecological and evolutionary phenomena.

5.2 Introduction

In the past, evolutionary biologists view growth rate *r* as fitness proxy while ecologists prefer carrying capacity *K* as fitness proxy (MACARTHUR AND WILSON 2016). Because *r* and *K* are important characters of density-dependent growth, the studies of *r*-*K* relationship trace back to the rich literature in evolutionary ecology. MacArthur and Wilson proposed the *r*-selection and *K*-selection theory based on their work on island biogeography (MACARTHUR AND WILSON 2016). By connecting these two fitness proxies with the environment, they explained the relative importance of *r* and *K* for fitness. They also envisioned tradeoff between *r*-*K* in *r*-selected and *K*-selected species in their book (MACARTHUR AND WILSON 2016). At about the same time, George Williams proposed antagonistic pleiotropy (or tradeoff) and discussed whether reproductive success of an individual (which could be measured by *r* under MacArthur and Wilson framework) necessarily extend to the success of the population (which could be

measured by *K* under MacArthur and Wilson framework) (WILLIAMS 1966b; WILLIAMS 1966a). Triggered by MacArthur's, Wilson's, and Pianka's work (PIANKA 1970) as well as Williams's influence on tradeoff, *r-K* tradeoff along with *r-K* selection theory was once most fashionable topic in ecology, but it is highly criticized later when empirical studies showed mismatched results (STEARNS 1977); the essence of the *r-K* theories later blended into other life-history models (REZNICK *et al.* 2002).

Studying *r-K* selection and *r-K* tradeoff with evolutionary ecology approaches can be difficult, because 1) the intrinsic nature of tradeoff is not clear, 2) initial environment is usually unknown, 3) natural environment is hard to manipulate, and 4) number of replicates and species is insufficient most of the time (STEARNS 1977). Recent studies in *r-K* focuses on experimental tests of *r-K* trade-off, or rate-yield (growth rate *r* and number of cells produced per mol of resource) trade-off with microbes (NOVAK *et al.* 2006; FITZSIMMONS *et al.* 2010; BEARDMORE *et al.* 2011; MEYER *et al.* 2015; REDING-ROMAN *et al.* 2017). Many of the microbial studies used experimental evolution to specific environment (NOVAK *et al.* 2006; REDING-ROMAN *et al.* 2017) Although these microbial studies provides the benefit of manipulated environment and replicates to confirm the observed correlation, these studies are small in scale (both in terms of number of genotypes, and in terms of number of environments), and the *r-K* tradeoff is not consistently found across experiments (NOVAK *et al.* 2006; FITZSIMMONS *et al.* 2010; BEARDMORE *et al.* 2011; MEYER *et al.* 2015; REDING-ROMAN *et al.* 2017). It is unclear under what condition the *r-K* relationship should be negative and under what condition *r-K* relationship should be positive.

Despite the criticisms by Sterns (STEARNS 1977), *r-K* tradeoff is believed because of some biochemical laws. adenosine triphosphate (ATP) production between rate (moles of ATP per unit of time) and yield (moles of ATP per mole of substrate) ($ATP_{rate\text{-}yield}$ to distinguish from the rate-

yield in terms of growth rate) is believed general for heterotrophic organisms (PFEIFFER *et al.* 2001). For example, tradeoff happens during sugar degradation, because unlike respiration, fermentation is not restricted by oxygen and sugar supply, thus the use of fermentation in addition to respiration increases the rate and decreases the ATP yield (POSTMA *et al.* 1989; PFEIFFER *et al.* 2001). Moreover, for some fundamental thermodynamic reasons, this $ATP_{rate\text{-}yield}$ tradeoff holds even without sugar degradation, because some of the free energy can be used to drive the reaction rather than to convert into ATP (WADDELL *et al.* 1999; PFEIFFER *et al.* 2001). Supported by the general tradeoffs in ATP production, *r* and *K* are believed to tradeoff. However, it is still unknown whether simple biochemical laws could explain the mixed *r*-*K* relationship.

The genetic effect of *r*-*K* relationship is rarely discussed (REDING-ROMAN *et al.* 2017), yet of great value to understand the underlying relationship. Charlesworth demonstrated that pure phenotypic correlations among life-history variables are unlikely to provide useful information on trade-offs, because selection and environmental effects may generate positive correlation between traits even when they have negative underlying correlations, and he pointed out that studying genetic correlations can help understand evolutionarily relevant tradeoff and predict evolutionary response to new selection pressures (CHARLESWORTH 1990). It is unknown how *r* and *K* are affected by mutations, and how likely there is genetic by environment interactions in terms of *r*-*K*-pleiotropy. Moreover, a recent study demonstrated that introducing genetic variability in experiment increases reproducibility for ecological study and can help solve 'reproducibility crisis' of scientific findings (MILCU *et al.* 2018). Therefore, studying many different genotypes in controlled environments will help understand the *r*-*K* relationship at mutation level and provide more confident results.

Because *r-K* relationship is hyper-interdisciplinary, understanding it not only help understand two important fitness proxies and life-history evolution, but also improve the understanding of pleiotropy, plasticity, as well as how biochemical laws constrain or facilitate cellular and organismal growth. In order to provide a mechanistic explanation for *r-K* relationship, we need to know the effects at genotype level, mutation level, and environment level. We take advantage of the budding yeast system in which different genotypes could be generated by recombination and the same genotype could be measured in multiple environments. We would like to study the patterns of *r-K* relationship by conducting a large-scale genome-wide and environment-wide analysis and to explain the patterns of *r-K* relationship with biochemical and biological insights.

5.3 Results

5.3.1 *r-K* correlation among genotypes is more negative in better environment

We acquired from the Hallin et al the genotype data and the unsmoothed growth data of 7310 diploids from a cross between 86 recombinant *MATa* and 86 recombinant *MATα* haploid strains of *S. cerevisiae* (HALLIN *et al.* 2016). 9 different YPD based growth medias were used, each with a different commonly used substrate. Each diploid genotype was grown on solid media with 4 replicates and the cell numbers of each replicates are measured with high resolution from 0 and 72h at 20min by colony scan-o-matic (ZACKRISSON *et al.* 2016). We first estimated the *r* and *K* for each replicate of each of the 7310 genotypes by fitting a logistic curve. We then calculated the average *r* and *K* for each genotype using all replicates that pass our quality control (see Material and Methods) and the average coefficient of determination for each genoptype r_g^2. The growth of yeast tightly follow logistic curve, resulting a median r_g^2 among all measured

genotypes in these 9 environment 0.979-1.000. Our estimated *r* is the growth rate per hour, and our estimated *K* is the carrying capacity in terms of cell number.

In each environment, we correlate all *r* and all *K* among genotypes. In three environments, we found significant positive correlations (Spearman correlation, $\rho_{r\text{-}K} \geq 0.32$, $P < 10^{-250}$), and in nine environments, we found significant negative correlations ($\rho_{r\text{-}K} \leq -0.08$, $P < 10^{-11}$). The correlations have a range of -0.52 to 0.53. Because the same genotypes are used across environments, suggesting environment has substantial effect on *r*-*K* correlation. To exclude the possibility that the *r*-*K* correlation is not due to biased estimation, we conducted a simulation where *r* and *K* are not correlated. The simulated data mimic the empirical data all other aspects such as the number of replicates, genotypes, and environments, the number of time points, the range of *r*, and the range of *K*, and the goodness of fitting (see Materials and Methods). We process the simulated data the same way as the empirical one. In none of the 9 simulated environment, *r* and *K* are correlated. Moreover, the estimated parameters are sufficiently accurate when compare to the simulated parameters (see Materials and Methods).

To investigate what causes change of sign and magnitude of these *r*-*K* correlations, we calculated the average growth rate of each environment as E_r and the average carrying capacity of each environment as E_K. For each of the nine environments, we have one $\rho_{r\text{-}K}$, one E_r and one E_K measured. We found that that E_r and $\rho_{r\text{-}K}$ are negatively correlated (Fig 5-1A, $\rho = -0.88$, $P < 10^{-11}$), but E_K and $\rho_{r\text{-}K}$ are not correlated (Fig 5-1B, $\rho = 0.23$, $P = 0.56$). Therefore, as environment gets better such that the majority of genotypes acquire faster growth rate, the *r*-*K* correlation continuously changes from positive to negative. This result suggests that the *r*-*K* correlation is mostly determined by *r*.

5.3.2 *r* and *K* are affected by shared genetic component

To study whether *r* and *K* are affected by shared genetic component, we mapped quantitative trait loci (QTLs) for *r* (rQTLs) and for *K* (KQTLs) in each environment (see Material and Methods). For each trait (*r* and *K*) in each environment, mapped 93-96 QTLs. For the later purpose of studying pleiotropic QTLs with high confidence, we want to avoid having too many QTLs. Therefore, we removed small effect QTLs (see Material and Methods) until the total explained variances by QTLs and the total explained variances by the same number of random SNPs are maximized. We use the most significant 36 QTLs to assay how much of the total variance could be explained by the large effect QTLs, and whether rQTLs could explain *K* more than by chance, and whether KQTLs could explain *r* more than by chance.

We found that 36 rQTLs explains 65%-81% of the total variance of *r*, and KQTLs explains 53%-77% of the total variance of *K*. Moreover, 27%-66% of the total variance of *r* could be explained by the KQTLs of the same environment, and 21%-60% of the total variance of *K* could be explained by the rQTLs of the same environment. These fractions, although smaller than the fractions explained by QTLs for each trait, is much larger than the fraction explained by 36 random sampled sites in all environments (Fig 5-2 AB). This result suggests that, a lot of the total variances of *r* and *K* are controlled by sites with pleiotropic effect.

5.3.3 *r*-*K* correlation among QTLs is more negative in better environment

We next ask whether the change from positive *r*-*K* correlation to negative *r*-*K* correlation as E_r increases also exist at QTL level. To this end, we use linear regression to estimate the effect of rQTL on *r* and on *K* in each environment. If the same rQTL allele increases *r* but decreases *K*, it is a tradeoff-rQTL. Otherwise, it is a tradeup-rQTL. We then have the fraction of rQTLs

showing tradeoff effect for each environment (F_{rQTL}). Out of the 9 environments, we found in 7 environments, the majority of rQTLs are tradeoff-rQTL ($F_{rQTL} > 0.5$), and in 1 environment, and the majority of rQTLs are tradeup-rQTL ($F_{rQTL} < 0.5$). The remaining environment has $F_{rQTL} = 0.5$. The number of environments with $\rho_{r\text{-}K}$,< 0 and the number of environments with $F_{rQTL} > 0.5$ are not exactly the same. This could be due to that $\rho_{r\text{-}K}$ is affected both the signs and the effect sizes of QTLs. Similar to the correlation observed between ρr-K and Er, we found F_{rQTL} and E_r are positively correlated (Fig 5-2C, $\rho = 0.91$, $P = 0.0013$), which suggest that high growth rate environment also has more tradeoff rQTLs. We also measured the effect of KQTL for *K* and for *r* for each environment and calculated the fraction of KQTLs showing tradeoff effect for each environment (F_{KQTL}). Similarly, we found two low E_r environments showing $F_{KQTL}<0.5$, and the rest 7 showing $F_{KQTL}>0.5$. Again, F_{KQTL} and E_r are positively correlated (Fig 5-2C, $\rho = 0.74$, $P = 0.027$). Moreover, neither of the F_{rQTL} and F_{KQTL} is correlated with E_K (Fig 5-2D).These results from QTL mapping provides genetic evidence for the among genotype observations.

5.3.4 Pleotropic QTLs can show *r*-*K* trade-up and trade-off depending on the environment

Because we found that *r* and *K* are controlled by sites with pleiotropic effect, we want to see if there exists pleiotropy by environment interactions. Gene-environment interaction refers to the phenomenon that the same mutation has different phenotypic effects in different environment, and it is often discussed in quantitative genetics, evolutionary genetics, and personalized medicine (WEI AND ZHANG 2017). In theory, when a mutation has pleiotropic effect, such that it changes multiple phenotypes, without pleiotropy by environment interaction, changing environment will not change the effect of it on different phenotypes. However, with pleiotropy by environment interactions, it may change those phenotypes in completely different ways. To our knowledge, pleiotropy by environment interaction for QTL has only been documented with

one example (VASSEUR *et al.* 2012). Because we have already found interactions with environment changes *r-K* correlation, we want to invest the general possibility of pleiotropy by environment interactions using or rQTLs and KQTLs. In particular, we are interested in antagonistic-pleiotropy by environment interactions, such that a site may increase *r* and *K* together in one environment but then flip the sign of effect for at least one trait in another environment.

We used the most significant 36 rQTLs and KQTLs in each environment to find the enriched regions. If a 3kb region in the genome show up 4 or more times as either rQTL or KQTLs in the 9 environments, it is enriched. We found 21 such regions. By chance, we expect to observe only 0.83 region (based on the average of 100 simulations), result in FDR = 4%. Among the 21 regions, 18 regions are sometimes rQTLs and sometimes KQTLs when we use only the most significant 36 rQTL and KQTL. We surveyed these 18 regions based on their effects in all 9 environments. For the QTL region with clear antagonistic pleiotropy by environment interactions, we highlighted the environments showing such effect in Fig 5-3A-K; for those QTL regions without clear antagonistic pleiotropy by environment interactions, effects in all environments are shown (Fig 5-3L-R).

5.3.5 Explaining *r-K* relationship by a cell division energy cost model with two tradeoffs

It is surprising that there is a clear pattern that $\rho_{r\text{-}K}$ changes with E_r but is unaffected by E_K, and similar results are also observed at QTL level. Because the tradeoff between $ATP_{rate\text{-}yield}$ can only explain the *r-K* tradeoff when E_r is large, it requires another biological process to overcome the $ATP_{rate\text{-}yield}$ tradeoff when E_r is small to explain the empirical observations. Therefore, we looked into possibilities that could increase the energy cost when *r* becomes

smaller. In microbes, generation time (time per cell division) G_T of a cell is proportional to 1/r. If a cell needs some energy per time just to maintain its healthy state, then such cost is linear with time. Indeed, as early as 50 years ago, Prit showed in multiple organisms that the extra substrates (glucose or glycerol) needed to produce the same amount of dry weight increases linearly with 1/r (PIRT 1965), suggesting the maintenance energy a cell needs is proportional to time. If we consider both $ATP_{rate\text{-}yield}$ tradeoff and maintenance cost, we may reconcile the mixed results for *r-K* relationship (LIPSON 2015).

Based on the $ATP_{rate\text{-}yield}$ tradeoff and maintenance cost, we derive the total cost of energy per cell division. Let α be an environment specific cost factor that is larger than 0, because the extra energy to maintain healthy state of a cell during one cell division is proportional to 1/r, we have α/r as the energy cost per cell division. Now assume the energy needed to produce new material for cell division is C, constant in all environments, and the energy wasted due to $ATP_{rate\text{-}yield}$ per cell division is $f(r)$. Because the nature of ATP production tradeoff, $f(r)$ is a monotonic increasing function with r, such that the first derivative of it, $f'(r)$ is larger than 0 for all valid r. Therefore, the total cost per cell division for a single cell (C_{Total}) is the sum of all three costs, which is

$$C_{Total} = C + f(r) + \alpha/r \qquad \text{Eq.1}$$

Take derivative of Eq.1, we get

$$\frac{dC_{Total}}{dr} = f'(r) - \frac{\alpha}{r^2} \qquad \text{Eq. 2,}$$

where $f'(r)$ is an unknown positive function which may or may not depend on r.

In a simple case where $f'(r)$ is independent of r, we have $\frac{dC_{Total}}{dr} < 0$ when $r \subseteq (0, \sqrt{\frac{\alpha}{f'(r)}})$ and $\frac{dC_{Total}}{dr} > 0$ when $r \subseteq (\sqrt{\frac{\alpha}{f'(r)}}, +\infty)$. Therefore, C_{Total} first decreases with r and then increases with r. When total resource is fixed, K should be a decreasing function of C_{Total} (although r may also affect K independent of C_{Total}), such that when C_{Total} increase, K decrease, and when C_{Total} decreases, K increases. As r increases from 0 to $\sqrt{\frac{\alpha}{f'(r)}}$, C_{Total} decreases therefore K increases, and as r further increases, C_{Total} increases, and K decreases. The turning point is $\sqrt{\frac{\alpha}{f'(r)}}$. Therefore, when environment gets better such that E_r is larger than the turning point, we expect to see negative *r*-*K* correlation; when environment is poor and E_r is much smaller than the turning point, we expect to see positive *r*-*K* correlation. For environment where growth rates of genotypes enclose the turning point, the sign of correlation depends on the majority, and the Spearman correlation should be weaker. This prediction matches the observations.

In the more complicated case where $f'(r)$ is still a function of r, $\frac{dC_{Total}}{dr}$ is negative when r is smaller than the first positive root of Eq.2, and it is positive when r further increases. Because it is possible to have more than one positive root, the dynamics can be more complicated. However, because there were only one transition from tradeup to tradeoff in our empirical result, it is more likely that there is only one positive root even when $f'(r)$ is still a function of r.

5.3.6 Explaining *f*(*r*) by fermentation and respiration pathway in yeast

The most simple tradeoff cost formula $f(r)$ could be $f(r) = \beta r$, where β is a constant and $f'(r) = \beta$. This turns out to be the energy tradeoff function for yeast when fermentation versus

respiration strategy differs across environment and among genotypes (see Method). This formula should also work in general when any faster but inefficient alternative pathway is used.

Because for $f(r) = \beta r$, $f'(r) = \beta$ is independent of r, we have $\frac{dC_{Total}}{dr} < 0$ when $r \subseteq (0, \sqrt{\frac{\alpha}{\beta}})$, and $\frac{dC_{Total}}{dr} > 0$ when $r \subseteq (\sqrt{\frac{\alpha}{\beta}}, +\infty)$. So there is only one turning point theoretically and empirically. When environment is good such that E_r is larger than $\sqrt{\frac{\alpha}{\beta}}$, we expect to see negative *r*-*K* correlation; and when environment is poor and E_r is smaller than $\sqrt{\frac{\alpha}{\beta}}$, we expect to see positive *r*-*K* correlation.

5.3.7 Testing model predictions with empirical data

The per cell division energy cost model provides us two extra testable predictions that could be verified with our data.

First, because within an environment, *r* is determined by genotypes, the change from *r*-*K* tradeup to *r*-*K* tradeoff should be seen among genotypes within an environment if the *r* of different genotype spread around the turning point. Based on this prediction, in each environment, we divide the genotypes into small bins based on their *r* (each bin has 500 genotypes). We then calculate the average *K* for each bin. The average *r* of the bins showing maximum *K* of each environments is 0.1076, shown by the black vertical line in Fig 5-4. We found that for the environments with many genotypes around 0.1076, there is a clear pattern of *K* increases and then decreases as *r* increases. In almost all environments, *K* is maximized at intermediate *r* (Fig 5-4A-G), suggesting the turning point is close to 0.1076. We found almost the same *r* as turning point for all environments, even though the genotypes in each bin change

from environment to environment. This result strongly supports that *K* depends on *r*, and is consistent with the model prediction.

Second, because our model suggests that *K* depends on *r*, rather than the opposite, we expect to see the rQTLs explain *K* better than KQTLs explain *r*. To this end, we calculated the total variance explained for *r* and *K* in each environment using all the significant rQTLs and KQTLs mapped from the 6 rounds of mapping (93-96 for each trait). We found that KQTLs explain an extra of 4.8% -17.1% of the total variance of *K* than rQTLs; rQTLs explain an extra of 8.9% -27.0% of the total variance of *r* than KQTLs. In 8 out of 9 environments (Binomial $P = 0.0039$), the rQTLs explain *K* better than KQTLs explain *r*, which is consistent with our model prediction.

5.4 Discussion

Charlesworth suggested that studying genetic correlations can help understand evolutionarily relevant tradeoff and predict evolutionary response to new selection pressures (CHARLESWORTH 1990). We provided the largest test for *r*-*K* relationship based on more than 7000 genotypes and 9 environments. We showed that $\rho_{r\text{-}K} > 0$ in low E_r and $\rho_{r\text{-}K} < 0$ in high E_r environment at both genotype level and QTL level. Because the genotypes in our study are all recombinants from two divergent strains which do not exist in nature, and because we also observe similar result at QTL level, suggesting *r*-*K* tradeup and tradeoff are intrinsic. Moreover, because the lab environments examined are random environments, to which the segregants have not adapted, we can treat the observed patterns as intrinsic to predict post-selection *r*-*K* relationship. First of all, if a genetically diverse population start from a new environment where initial E_r is low, because *r*-*K* tradeup, the population increases *r* and *K* together despite selection

may prefer only one of these two traits. As the population approach intermediate *r*, *K* reaches its highest potential. At this time, if selection prefers high *K* genotypes, then *K* will increase via decreasing the constant cost *C* in Eq. 1 and *r* may shifts up and down a bit due to relaxation of selection; if selection prefers high r genotypes, then *r* will continue to increase, and *r*-*K* tradeoff causes *K* to decrease. The adaptation dynamics predicted by intrinsic *r*-*K* relationship and our model, is different from MacArthur and Wilson's prediction (MACARTHUR AND WILSON 2016), which did not consider mutational level *r*-*K* relationship. The discrepancy demonstrated the importance of considering genetic correlations for understanding and predicting life history evolution. In fact, knowing and counting the mutational relationship between traits is always important for studying phenotype evolution, and the fail of which largely explains why Pianka's extention of *r*-*K* selection to predict life histories (PIANKA 1970) does not work well (STEARNS 1977).

We explain our observed *r*-*K* relationship by cell division energy cost model, which combines the effect of cost from maintenance energy for cell survival and cost of using fast but inefficient metabolic pathways. The mathematical part of this model suggests that *K* depends on *r* rather than the opposite, which makes biological sense, because *r* is an individual measurable parameter and *K* is only measurable at population growth level. Our model considers two kinds of tradeoff, one is the extra cost of maintenance when *r* is small, and another is the extra waste of resource when *r* is large. David Lipson proposed that if maintenance cost is considered, then *r*-*K* should tradeup in slow growth environment and tradeoff in fast growth environment (LIPSON 2015). We demonstrated both mathematically and empirically that this is true. Moreover, because our model not only suggests that *r*-*K* relationship changes with environment, it also suggests the same trend among genotypes within an environment. Indeed, we showed that in

environments where growth rates cover the transition point from tradeup to tradeoff, slow growth genotypes show *r-K* tradeup, while fast growth genotypes show *r-K* tradeoff (Fig 5-4). Because the factors in our model does not restrict to our study system, we believe this model generally applies to previous studies of *r-K* relationship in microbes. In fact, our model suggests those mixed results (NOVAK *et al.* 2006; FITZSIMMONS *et al.* 2010; BEARDMORE *et al.* 2011; MEYER *et al.* 2015; REDING-ROMAN *et al.* 2017) are expected rather than surprising.

In a recent paper, Reding-Roman et al showed that *r* and *K* could trade-up or tradeoff depending on the glucose concentration (REDING-ROMAN *et al.* 2017). Based on 6 *E. coli* genotypes which differ in ribosomal gene copy number, Reding-Roman et al showed that *r* first increases as *K* increases and then decreases as *K* further increases (REDING-ROMAN *et al.* 2017). This is different from the trend we observed because their their *K* (or yield) can be maximized or minimized when r is smallest, but in our case, *K* is maximized when *r* is at intermediate level. Moreover, the model they propose is based on Monod function (MONOD 1949), which neglects maintenance cost when there is significant maintenance cost even in bacteria (PIRT 1965). Because they only used six genotypes and the replicates vary a lot, it is quite probable that their observed trend is statistically insignificant. In any case, the model of Reding-Roman et al cannot explain our large-scale observations.

Understanding pleiotropy by environment interactions is important for studying phenotype evolution, especially for fluctuating environments. We showed that pleiotropy by environment interactions is common in the case of *r-K*. Moreover, we observed antagonistic pleiotropy by environment interactions at QTL level. There are alleles that always increase *K* showing opposite effects on *r* in different environments (Fig 5-3BCGHJ), alleles that always increase r showing opposite effects on K in different environments (Fig 5-3EFK), and more

complicated case (Fig 5-3ADI). In our analysis, we used 3kb region to determine whether the mapped QTL for different traits belongs to the same causal place (FDR = 0.04). Because the SNP density used for QTL mapping is 1.01 per kb, and on average, there is one ORF in every 2 kb region in yeast genome. A 3kb region only incorporates an average of 3 SNPs used for mapping and 1.5 ORFs. Because the majority of SNPs have little or no effects on traits, the strong antagonistic pleiotropy by environment interactions observed are most likely true signal than the combined effect of multiple linked SNPs.

Although we present our study in *r-K* framework rather than rate-yield framework (i.e. growth rate – dry weight produced per mol. substrate), these two relationships are synonymous in our case. It is because the *r-K* relationship measured in each environment has fixed environmental resource for all genotypes and *K* rather than yield is directly estimated from the data that we present this way. The tradeup and tradeoff region based on growth rate applies for rate-yield relationship as well; especially, because our model does not convert the per cell division energy cost and total amount of resource into *K*, it is in fact more of a model for rate-yield than *r-K*. Among the nine environment we tested, we did not observe a change in the turning point from tradeup to tradeoff (Fig 5-4), it might be interesting to examine more environments and species to see how general this observation is.

For therapeutic reason, *r-K* relationship is sometimes discussed in cancer progression (AKTIPIS *et al.* 2013; KOROLEV *et al.* 2014). Our observed *r-K* relationship also affect our understanding of antibiotic resistance.

5.5 Materials and Methods

5.5.1 Genotype and growth data for diploid yeast hybrids

We acquired from the Hallin et al the genotype data and the unsmoothed growth data of 7310 diploids from a cross between 86 *MATa* and 86 *MATα* strains haploid of *S. cerevisiae* (HALLIN *et al.* 2016). The haploids were randomly drawn from a twelfth generation two-parent intercross pool which is mated from two wild strains sampled in North America(NA) and West Africa(WA) (HALLIN *et al.* 2016). The NA genome and WA genome differs by 0.53%. The cell number of each of the diploid genotype is measured at 217 time points (between 0 and 72h at 20min interval) with 4 replications by scan-o-matic, a high-resolution automatic microbial growth phenotyping approach (ZACKRISSON *et al.* 2016). Because the cell number estimation is based on colony scan, the estimated *K* reflects the true yield and it is robust to cell size. The genotypes were grown in 9 different growth environments, allantoin, caffeine, galactose, glycine, hydroxyurea, isoleucine, NaCl, phleomycin, and rapamycin.

Before QTL mapping, we first code the genotype of each SNP with 0, 1, or 2, if it is homozygous for WA allele, heterozygous, or homozygous for NA allele, respectively. We then filtered the SNPs that contain redundant information such that only the middle SNP is maintained when several neighboring SNPs have exactly the same allele in all hybrid genotypes. This results in 13350 remaining SNPs.

5.5.2 QTL mapping

We mapped rQTLs and for KQTLs in each environment with the same approach. We first mapped QTLs underlying the growth rate variation among the segregants in each of the 9 environments at a false discover rate (FDR) of 0.05 follow the approach of a recent study (BLOOM *et al.* 2013). In short, this approach takes multiple rounds of mapping, and in each round, at most one most significant SNP of each chromosome will be mapped as QTLs, the residues from fitting all mapped QTLs from all previous rounds will be used for next round of mapping.

FDR is calculated by permutation test. We stopped the program at 6th round which results in 93-96 QTLs for each trait and we calculated the total r^2 explained by all mapped QTLs. We then remove the QTL that has the smallest effect on total r^2, and recalculate the total r^2 with all remaining QTLs. We repeat this process and remove small effect QTLs one by one until we have 48 QTLs (QTL48), 36 QTLs (QTL36), 24 QTLs (QTL24), or 18 QTLs (QTL18) remaining for each trait. By doing so, we result in equal number of rQTLs and KQTLs for each environment. We also calculate the total explained variance (r^2_{SNPs}) by 96 SNPs, 48 SNPs, 36 SNPs, 24 SNPs and 18 SNPs as comparison. When we maintain QTL48, the averaged r^2 for all traits is 0.738 (r^2_{QTL48}). The averaged r^2 reduces to 0.703 (r^2_{QTL36}) when we maintain QTL36. After QTL36, the averaged r^2 dropped very fast, and the difference between r^2_{SNPs} and r^2_{QTL} is maximized at QTL36. Having slightly fewer but large effect QTLs allow us to study pleiotropy by environment interaction with high confidence, because many small effect QTLs are very randomly located across the genome, making it difficult to get a low FDR region.

5.5.3 Estimating *r* and *K*

The logistic equation was derived to describe density-dependent growth (VERHULST 1838), and it was popularized by Raymond Pearl and Lowell Reed when they substituted *r* and *K* into the Verhulst Model (REED AND PEARL 1927). In as early as 1913, the logistic growth of yeast was demonstrated by Carlson (CARLSON 1913). Our estimation of *r* and *K* from growth data is based on logistic equation.

$$\frac{dN}{dt} = rN(1 - \frac{N}{K}) \qquad \text{Eq. 8}$$

The integral of Eq. 8:

$$N = \frac{K}{1+(\frac{K}{N_0}-1)e^{-rT}} \qquad \text{Eq. 9}$$

5.5.4 Goodness of logistic fitting

We first estimate *r* and *K* for each replicate of each genotype in each environment individually by fitting Eq. 9 with with NonLinearModel.fit function in Matlab using cell number *N* and time of measurement *T*. We then removed the low quality replicates by experience. We assume that values that are far from the nearest neighbors are outliers and set cutoffs based on the fold difference between outliers and median. Because *K* has a wider range than *r*, different cutoffs for *r*-*K* are used. In practice, we removed the replicates whose estimated *r* is larger than 2-fold or smaller than ½ of the median *r* from all measurements of all genotypes in the same growth condition and the replicates whose estimated K is larger than 4-fold or smaller than ¼ of the median *K* from all measured genotypes in the same growth condition. The majority of removed replicates are extreme outliers, who have either negative *r*-*K* or estimated *r*-*K* estimation hundreds fold bigger than nonoutliers. While enlarge the fold number from ½ to 2 into 1/3 to 3 for *r* or from ¼ to 4 to 1/5 to 5 for *K*, will affect less than 1% of the total remaining replicates, shrinking it slightly start to exclude much more replicates. After quality control, in each environment, 93.2-100% of the genotypes have at least 3 out of 4 replicates measured. We calculate the average the *r* and *K* using all remaining replicates of each genotype as the *r* and *K* of the genotype. We also have one r^2 showing the goodness of logistic fitting for each replicate. The average r^2 using all remaining replicates of the same genotype, r_g^2, represents the goodness of logistic fitting for that genotype. In each environment, 97.6-100% of the genotypes have r_g^2 larger than 0.97; 75.5-100% of the genotypes have r_g^2 larger than 0.98. The median r_g^2 among all measured genotypes in these 9 environment are 0.979-1.000. These goodness of fitting

measurements do not correlate with E_r and E_K. We calculated the standard deviation using the replicates (SD_{rep}) of each genotype in each environment, and found the median SD_{rep} for *r* is 0.0034 to 0.013, and the SD_{rep} for *K* is 1.2×10^5-2.6×10^5 across 9 environments. The median SD_{rep} of *r* and *K* are also independent of E_r and E_K. We also calculated the SD of genotypes (SD_{geno}) of *r* and *K* among genotypes for each environment for simulation.

To exclude the possibility that our logistic fitting has no bias, we did a simulation and estimated the simulated *r*-*K* correlation as well as *r*-*K* estimation sensitivity. We simulated the growth of 7000 genotypes for 9 environments to best mimic the real data. The *r* and *K* of genotypes follow normal distribution with mean as observed E_r and E_K and SD as SD_{geno} of *r* and *K* in that environment. We then calculate the cell number using the logistic curve from 0 to 72h at 20min interval. Each genotype has 4 replicates sharing same *r* and *K* but independent noise. The random growth noise added at each time point follows a normal distribution with mean 0 and variance equals (median 1- r_g^2 of each environment, four digits) × SST (i.e., the total sum of square of cell numbers for each replicates). By doing so, our median fitted r_g^2 from simulation equals the empirical median r_g^2. After adding random noise, we follow the exact same process as we do to the empirical data to estimate the simulated *r* and *K* for each replicate and each genotype. Because both *r* and *K* follow normal distribution in each environment, the simulated data has the same range of *r* and *K* as the empirical one but *r*-*K* are not correlated. In each simulated environment, 95.1-99.9% of the total simulated genotypes have *r* and *K* estimated. Among the measured genotypes in each environment, 71.6-74.6% of the genotypes deviate less than 1% from the simulated value of *r* and *K*; 93.0-97.6% of the genotypes deviate less than 20% from the simulated value of *r* and *K*, proving the logistic fitting is accurate in estimating the true value. Out of the 9 simulated environments, none has significant *r*-*K* correlation after multiple

testing correction. Given the accuracy of logistic fitting, it is impossible to generate the strong positive and negative *r*-*K* correlations that we observe in the real data. Thus, the observed *r*-*K* relationship must be true.

5.5.5 The cost of using energy inefficient pathway

If we assume *r* increases linearly with the fraction of total resource used by fermentation pathway. Let p_F be the fraction of substrates used by fermentation. p_F represents a weighted value, which could be either the total amount of time a cell uses fermentation during one cell cycle, or the total amount of cells with the genotype that use fermentation due to bet-hedging. Suppose that the same amount of resource (here, glucose) used by respiration pathway produces ATP at rate γ_1 per second, and the same total resource if used by fermentation pathway produces ATP at rate γ_2 per second ($\gamma_2 > \gamma_1$). Then the rate of ATP production equals

$$\gamma = (1 - p_F)\,\gamma_1 + p_F\,\gamma_2 = \gamma_1 + p_F\,(\gamma_2 - \gamma_1) \qquad \text{Eq. 3,}$$

The γ minus maintenance cost α determines *r*, so that

$$r = c_R(\gamma - \alpha) \qquad \text{Eq.4}$$

, and that c_R is a constant that convert per second free energy to growth rate.

Now let ι be the extra energy produced by respiration as compare to fermentation, we can calculate the energy waste because of using some fermentation:

$$f(r) = \iota p_F \qquad \text{Eq.5,}$$

Combine Eqs. 3 and 4, we get

$$p_F = \frac{r/c_R + \alpha - \gamma_1}{\gamma_1 - \gamma_2} \qquad \text{Eq.6}$$

Put Eq. 6 into Eq. 5, we get the total extra energy cost

$$f(r) = \iota\frac{r/C_R+\alpha-\gamma_1}{\gamma_1-\gamma_2} = r\frac{\iota}{(\gamma_1-\gamma_2)C_R} + \frac{\alpha-\gamma_1}{\gamma_1-\gamma_2} \qquad \text{Eq. 7}$$

Because the term $\frac{\alpha-\gamma_1}{\gamma_1-\gamma_2}$ in Eq. 7 is a constant, we can put it into *C* of Eq. 1. The coefficient of *r* in Eq. 7, $\frac{\iota}{(\gamma_1-\gamma_2)C_R}$, is a constant. Let it equals β, then we have $f(r) = \beta r$.

Therefore, this formula $f(r) = \beta r$ makes sense for yeast, because yeast has both fermentation pathway and respiration pathway. The formula should also be true whenever a faster but inefficient metabolic pathway is used.

References

Aktipis, C. A., A. M. Boddy, R. A. Gatenby, J. S. Brown and C. C. Maley, 2013 Life history trade-offs in cancer evolution. Nat Rev Cancer 13**:** 883-892.

Beardmore, R. E., I. Gudelj, D. A. Lipson and L. D. Hurst, 2011 Metabolic trade-offs and the maintenance of the fittest and the flattest. Nature 472**:** 342-346.

Bloom, J. S., I. M. Ehrenreich, W. T. Loo, T. L. Lite and L. Kruglyak, 2013 Finding the sources of missing heritability in a yeast cross. Nature 494**:** 234-237.

Carlson, T., 1913 Über geschwindigkeit und grösse der hefevermehrung in würze. Biochem. Z 57**:** 313-334.

Charlesworth, B., 1990 Optimization models, quantitative genetics, and mutation. Evolution 44**:** 520-538.

Fitzsimmons, J. M., S. E. Schoustra, J. T. Kerr and R. Kassen, 2010 Population consequences of mutational events: effects of antibiotic resistance on the r/K trade-off. Evolutionary ecology 24**:** 227-236.

Hallin, J., K. Märtens, A. I. Young, M. Zackrisson, F. Salinas *et al.*, 2016 Powerful decomposition of complex traits in a diploid model. Nature communications 7.

Korolev, K. S., J. B. Xavier and J. Gore, 2014 Turning ecology and evolution against cancer. Nat Rev Cancer 14**:** 371-380.

Lipson, D. A., 2015 The complex relationship between microbial growth rate and yield and its implications for ecosystem processes. Frontiers in microbiology 6.

MacArthur, R. H., and E. O. Wilson, 2016 *The theory of island biogeography*. Princeton university press.

Meyer, J. R., I. Gudelj and R. Beardmore, 2015 Biophysical mechanisms that maintain biodiversity through trade-offs.

Milcu, A., R. Puga-Freitas, A. M. Ellison, M. Blouin, S. Scheu *et al.*, 2018 Genotypic variability enhances the reproducibility of an ecological study. Nat Ecol Evol.

Monod, J., 1949 The growth of bacterial cultures. Annual Reviews in Microbiology 3**:** 371-394.

Novak, M., T. Pfeiffer, R. E. Lenski, U. Sauer and S. Bonhoeffer, 2006 Experimental tests for an evolutionary trade-off between growth rate and yield in E. coli. Am Nat 168**:** 242-251.
Pfeiffer, T., S. Schuster and S. Bonhoeffer, 2001 Cooperation and competition in the evolution of ATP-producing pathways. Science 292**:** 504-507.
Pianka, E. R., 1970 On r-and K-selection. The American Naturalist 104**:** 592-597.
Pirt, S., 1965 The maintenance energy of bacteria in growing cultures. Proceedings of the Royal Society of London B: Biological Sciences 163**:** 224-231.
Postma, E., C. Verduyn, W. A. Scheffers and J. P. Van Dijken, 1989 Enzymic analysis of the crabtree effect in glucose-limited chemostat cultures of Saccharomyces cerevisiae. Applied and environmental microbiology 55**:** 468-477.
Reding-Roman, C., M. Hewlett, S. Duxbury, F. Gori, I. Gudelj *et al.*, 2017 The unconstrained evolution of fast and efficient antibiotic-resistant bacterial genomes. Nat Ecol Evol 1**:** 50.
Reed, L. J., and R. Pearl, 1927 On the summation of logistic curves. Journal of the Royal Statistical Society 90**:** 729-746.
Reznick, D., M. J. Bryant and F. Bashey, 2002 r-and K-selection revisited: the role of population regulation in life-history evolution. Ecology 83**:** 1509-1520.
Stearns, S. C., 1977 The evolution of life history traits: a critique of the theory and a review of the data. Annual Review of Ecology and Systematics 8**:** 145-171.
Vasseur, F., C. Violle, T. Bontpart, B. J. Enquist, F. Tardieu *et al.*, 2012 Genetic variability in plant allometries under combined water deficit and high temperature. Réponses intégrées des plantes aux contraintes hydriques et thermiques: du gène au phénotype chez Arabidopsis thaliana**:** 123.
Verhulst, P.-F., 1838 Notice sur la loi que la population suit dans son accroissement. correspondance mathématique et physique publiée par a. Quetelet 10**:** 113-121.
Waddell, T. G., P. Repovic, E. Meléndez-Hevia, R. Heinrich and F. Montero, 1999 Optimization of glycolysis: new discussions. Biochemical education 27**:** 12-13.
Wei, X., and J. Zhang, 2017 The Genomic Architecture of Interactions Between Natural Genetic Polymorphisms and Environments in Yeast Growth. Genetics 205**:** 925-937.
Williams, G. C., 1966a Adaptation and natrual selection: a critique of some current evolutionary thought, pp.
Williams, G. C., 1966b Natural selection, the costs of reproduction, and a refinement of Lack's principle. The American Naturalist 100**:** 687-690.
Zackrisson, M., J. Hallin, L.-G. Ottosson, P. Dahl, E. Fernandez-Parada *et al.*, 2016 Scan-o-matic: high-resolution microbial phenomics at a massive scale. G3: Genes, Genomes, Genetics 6**:** 3003-3014.

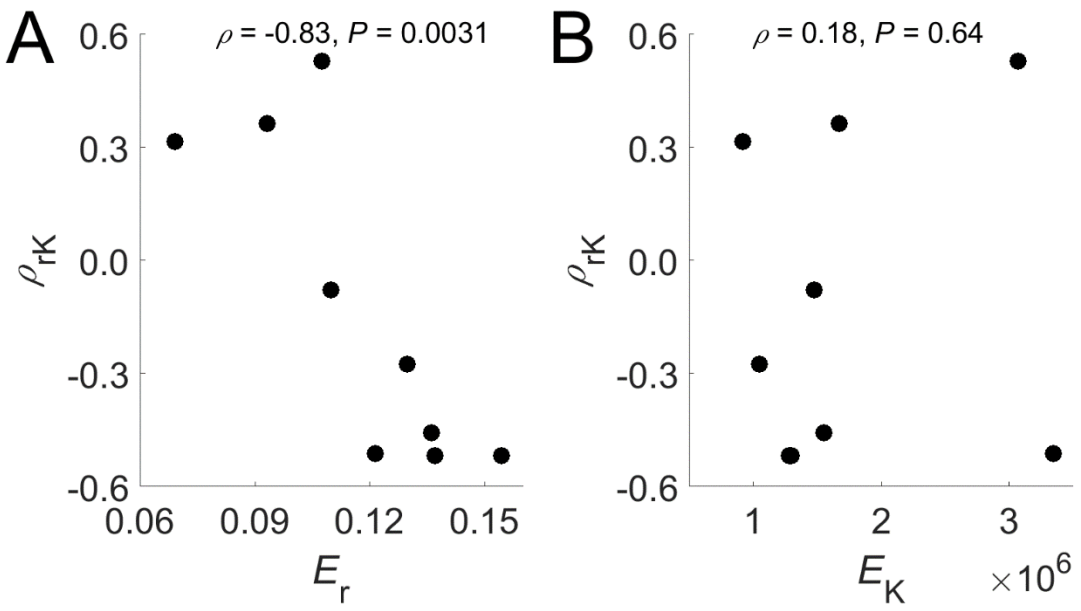


Figure 5-1. *r*-*K* correlation depends on environmental effects on E_r but not E_K. (A). Negative correlation between E_r and $\rho_{r\text{-}K}$. (B). No correlation between E_K and $\rho_{r\text{-}K}$. Each dot shows the E_K and $\rho_{r\text{-}K}$ one environment. $\rho_{r\text{-}K}$ is the Spearman correlation between *r* and *K* of all measured genotypes in an environment. E_r is the average *r* of all genotypes in an environment, and E_K is the average *K* of all genotypes in an environment. Each dot shows the E_r (A) or E_K (B) and $\rho_{r\text{-}K}$ of one environment. Spearman correlation between E_r (A) or E_K (B) and $\rho_{r\text{-}K}$ is shown on the graph.

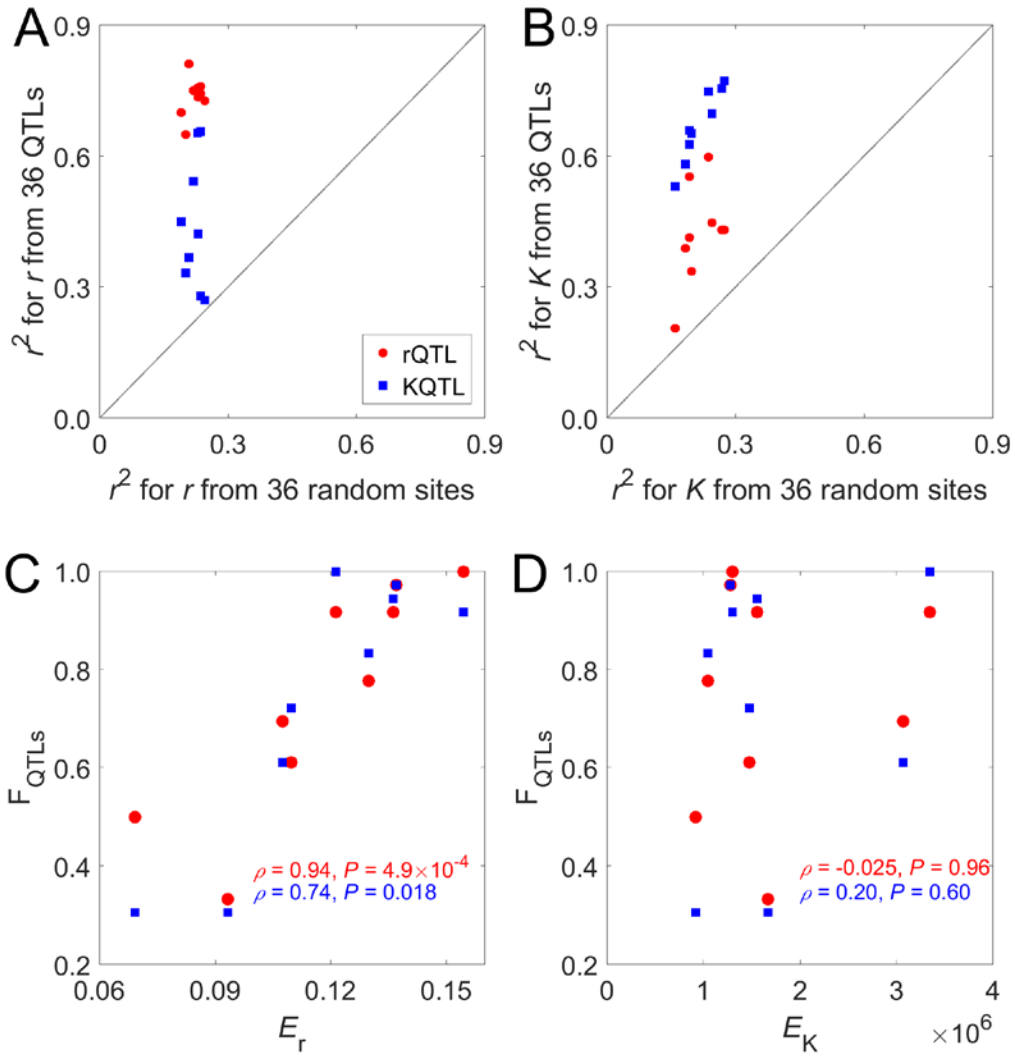


Figure 5-2. *r*-*K* correlation from on QTL mapping results. (A-D). Each dot is the result from one environment. Red circles show the results from rQTLs, and blue squares show the results from KQTLs. (A-B) X-axis is the expected total variance explained (r^2) based on 36 random sites. Y-axis is the r^2 based on 36 QTLs. (A) KQTL explain *r* better than random sites in all environments. (B) rQTL explain *K* better than random sites in all environments. (C-D). F_{QTLs} measures the fraction of QTLs showing opposite effects on *r* and *K*. Spearman correlation is listed. (C) F_{QTLs} and E_r are positively correlated. (D) F_{QTLs} and E_r are not correlated.

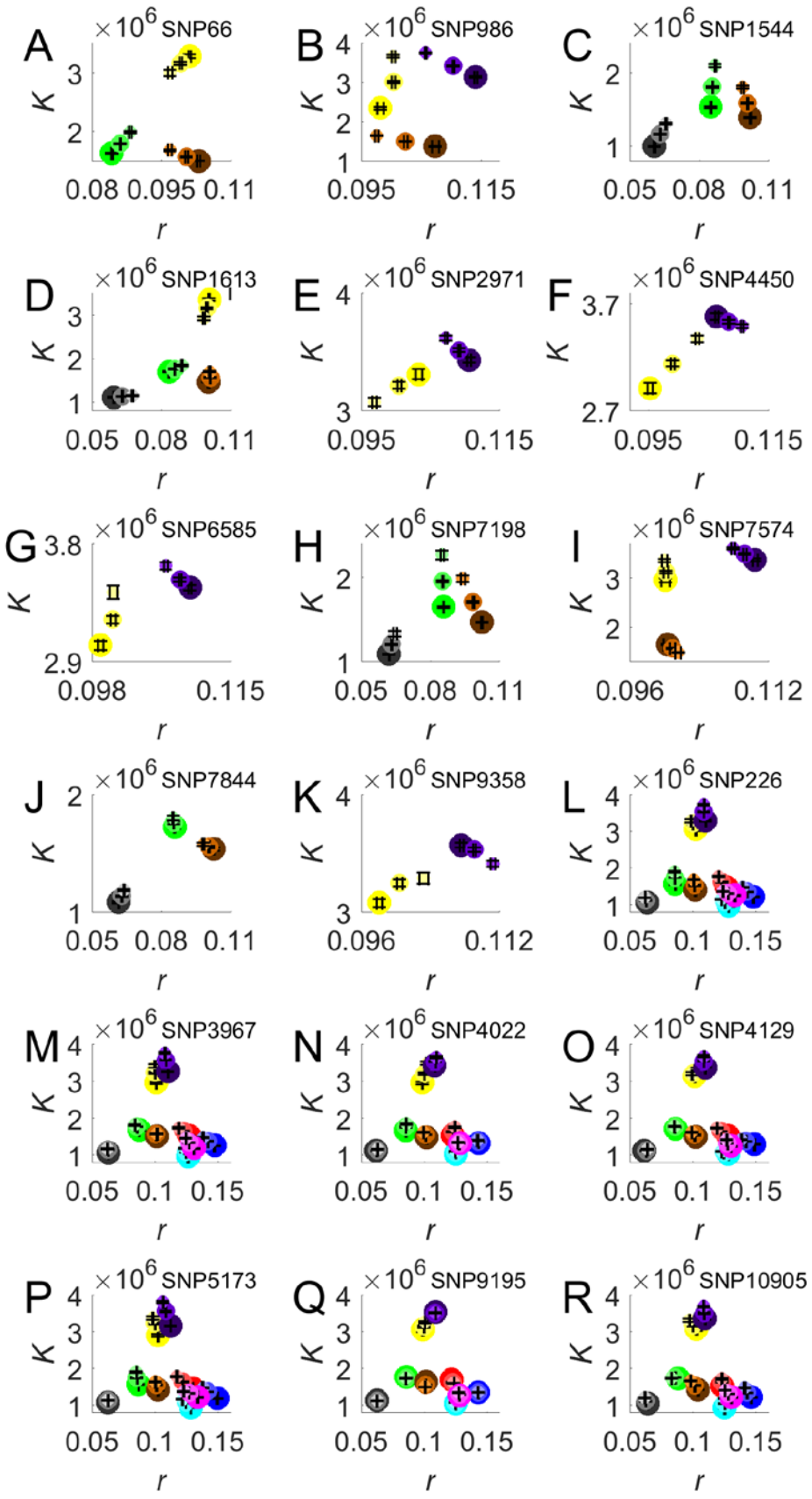

A
×10^6 SNP66
K
r
B
×10^6 SNP986
C
×10^6 SNP1544
D
×10^6 SNP1613
E
×10^6 SNP2971
F
×10^6 SNP4450
G
×10^6 SNP6585
H
×10^6 SNP7198
I
×10^6 SNP7574
J
×10^6 SNP7844
K
×10^6 SNP9358
L
×10^6 SNP226
M
×10^6 SNP3967
N
×10^6 SNP4022
O
×10^6 SNP4129
P
×10^6 SNP5173
Q
×10^6 SNP9195
R
×10^6 SNP10905

Figure 5-3. Pleiotropy by environment interaction QTLs. Each color represent one environment. X-axis is the average *r*, and y-axis is the average *K*. Each circle shows the *r*-*K* of genotypes with particular allele, with error bar showing the standard error. Small circle with lighter color represents the homozygotes of NA allele; intermediate circle shows the heterozygotes, and large circle with darker color shows the homozygotes of WA allele. The SNP number is labelled at right upper corner. (A-K) Examples showing antagonistic pleiotropy by environment interaction. The interaction part is highlighted. (L-R) Examples showing pleiotropy by environment interactions.

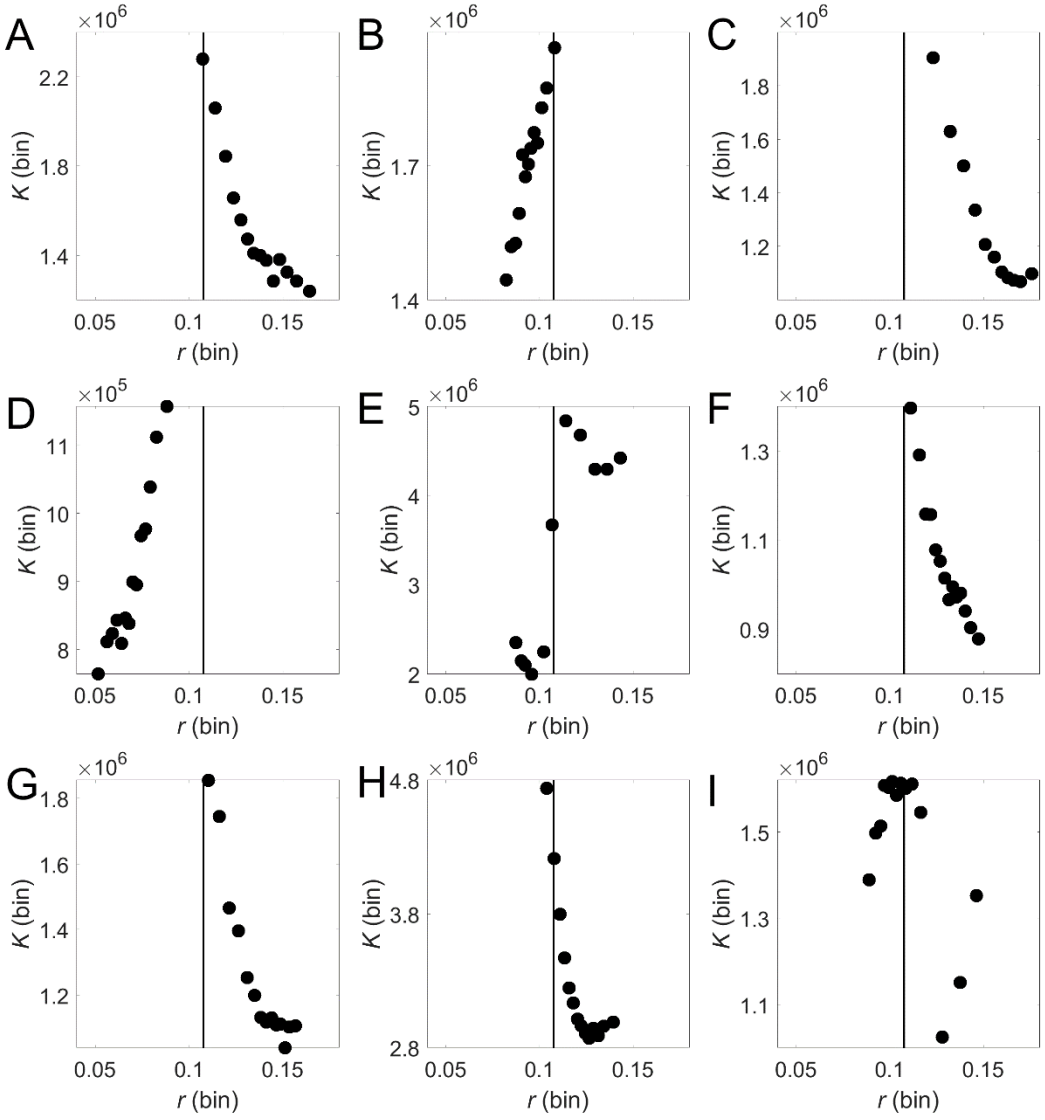


Figure 5-4. Same transition point between *r*-*K* tradeup and *r*-*K* tradeoff. Each environment is shown in one panel. Each bin contains 500 genotypes, grouped from small *r* to large *r*. Each dot shows the average *r* and *K* of each bin. The same black line *r* = 0.1074 is plotted on all panels. (A) hydroxyurea. (B) NaCl. (C) allantoin. (D) caffeine. (E) galactose. (F) glycine. (G) isoleucine. (H) phleomycin. (I) rapamycin.

Chapter 6

All interactions: The optimal mating distance resulting from heterosis and genetic incompatibility

"It is the things for which there is no evidence that are believed with passion."

— Bertrand Russell

6.1 Abstract

The genetic distance between the two parents of an individual, or mating distance, influences the individual's fitness via two competing mechanisms. On the one hand, increasing the mating distance is beneficial because of the phenomenon of heterosis. On the other hand, too large of a mating distance is harmful owing to genetic incompatibility. It is thus believed that the fitness of a genotype is a hump-shaped function of the mating distance, culminating at an intermediate distance referred to as the optimal mating distance (OMD). However, decades of research has generally failed to validate this belief or identify the OMD. Here we address this question using large datasets from the plant *Arabidopsis thaliana*, fungus *Saccharomyces cerevisiae*, and animal *Mus musculus*, including phenotypic measures of multiple fitness-related

traits from tens to hundreds of crosses and whole-genome sequence-based mating distance estimates. In each species, we find the hybrid phenotypic value a humped quadratic polynomial function of the mating distance for the vast majority of traits examined, with different traits exhibiting similar OMDs. OMDs are generally slightly greater than nucleotide diversities but smaller than the maximal observed genetic distances within species. Hence, the benefit of heterosis is at least partially offset by the harm of genetic incompatibility even within species. These results have implications for speciation, conservation, agriculture, and human health.

6.2 Introduction

Numerous studies attempted to verify a hump-shaped relationship between an individual's fitness (or its proxy) and mating distance (*D*)(MOLL *et al.* 1965; LYNCH 1991; MORAN *et al.* 1995; XIAO *et al.* 1996; EDMANDS 1999; AMOS *et al.* 2001; WILLI AND VAN BUSKIRK 2005; GONZALEZ *et al.* 2007; MCCLELLAND AND NAISH 2007; STOKES *et al.* 2007; ROBINSON *et al.* 2009; JAGOSZ 2011; HUNG *et al.* 2012; PEKKALA *et al.* 2012; PLECH *et al.* 2014; STELKENS *et al.* 2014; YANG *et al.* 2017), but all failed except two. In the first exception(LYNCH 1991), however, *D* was approximated by geographic distance(MOLL *et al.* 1965), and genetic incompatibility was detected only under the smallest *D* (LYNCH 1991), rendering the conclusion uncertain. In the second exception, *D* was estimated using the electrophoretic data of only eight allozyme loci; the low resolution prevented an unequivocal assessment of the OMD relative to the level of intraspecific genetic diversity(WILLI AND VAN BUSKIRK 2005). We hypothesize that the lack of support for OMD were contributed by the lack of reliable *D* estimates. Furthermore, given *D*, the fitness of a hybrid presumably varies greatly depending on its genotype. Hence, a large number of crosses are required to estimate accurately the expected hybrid fitness at each *D*.

Given these considerations, we collected from the literature large sets of relevant genotype and phenotype data in an attempt to verify the humped relationship between *D* and hybrid fitness and to estimate the OMD.

6.3 Results

Fitness is a compound trait consisting of multiple components. Most studies measure one to several key components of fitness such as the maximum growth rate of microbes, shoot weight of plants, and body weight of animals. The phenotypic value of a fitness-related trait is commonly referred to as "performance". To allow among-cross comparisons, for a given trait, we examined the fractional increase in hybrid performance relative to the average performance of its homozygous parents by $F = (H - \frac{P_1+P_2}{2})/(\frac{P_1+P_2}{2})$, where H is the performance of the hybrid, and P_1 and P_2 are the performances of the two parents, respectively. When $D = 0$, the hybrid and the two parents are isogenic and hence $F = 0$. Under pure genetic additivity, H is expected to equal the average of P_1 and P_2, resulting in $F = 0$ regardless of D. Heterosis arises from genetic interactions between the paternal and maternal alleles of the same loci (via dominance and overdominance) and/or different loci (via positive intergenic epistasis)(LIPPMAN AND ZAMIR 2007). Genetic incompatibility similarly originates from allelic interactions at the same loci (via underdominance) and/or different loci (via negative intergenic epistasis). At any locus, if the paternal and maternal alleles differ, either both of them are derived from their common ancestral allele or only one of them is derived whereas the other is ancestral. In the hybrid, the number of interactions between an ancestral allele from one parent and a derived allele from the other parent is expected to rise linearly with D, whereas the number of interactions between two derived alleles is expected to rise in proportion to D^2. It can be shown that, dominance most likely occurs between one ancestral and one derived alleles, whereas the other interactions

mentioned most likely occur between two derived alleles (see Methods). Therefore, the expected number of dominance interactions is proportional to D, while the expected numbers of overdominance, underdominance, positive intergenic epistasis, and negative intergenic epistasis are proportional to D^2. High-order interactions are ignored here because the contribution of high-order interactions to quantitative traits is much smaller than those of additive effects and two-way interactions(BLOOM *et al.* 2015) and because considering high-order interactions substantially increases the complexity of the model and difficulty in model selection. Because the effect size of an interaction is expected to be independent of D, the joint effect of heterosis and genetic incompatibility is expected to result in $F = aD + bD^2$, where the first term reflects heterosis due to dominance while the second term reflects the combined effect of heterosis arising from overdomiance and positive intergenic epistasis and genetic incompatibility arising from underdominance and negative intergenic epistasis. If $|aD| >> |bD^2|$, $F \approx aD$, which monotonically changes with D. If $|aD| << |bD^2|$, $F \approx bD^2$, which also monotonically changes with positive D. Under the condition that a is positive, b is negative, and $|aD|$ is comparable with $|bD^2|$, F is a hump-shaped function of D and OMD = $-0.5a/b$.

Based on the above formulation, we considered three competing models: (I) $F = aD$, (II) $F = bD^2$, and (III) $F = aD + bD^2$, where a and b are model parameters to be estimated. Model I has only the linear term, meaning that F is entirely caused by dominance-based heterosis; Model II has only the quadratic term, implying the absence of dominance-based heterosis; and Model III contains both terms. We used R^2 to determine which model best explains a dataset. Because Models I and II are both special cases of Model III, we used likelihood ratio tests (LRTs) to examine if the first two models can be statistically rejected in favor of Model III.

We first analyzed 200 crosses of the model plant *Arabidopsis thaliana*(YANG *et al.* 2017). *D* is measured by the number of single nucleotide polymorphisms (SNPs) between parental genomes divided by the total number of nucleotides in the *A. thaliana* genome (see Methods). Four fitness-related traits were measured for all parents and hybrids: shoot fresh weight, rosette diameter, leaf area, and leaf number at 14 days after sowing(YANG *et al.* 2017). Because the *D* values are not evenly distributed and because *F* varies greatly among crosses of similar *D*, we binned hybrids using a window size of $D = 0.8\times10^{-3}$ and computed the average *F* and average *D* of all hybrids in each window. We then used least squares to fit the binned data to the three models respectively. For each of the four traits, R^2 is negative for Models I and II (**Table 6-1**), indicating that these models, assuming monotonic changes of *F* with *D*, perform even worse than the obviously incorrect null model that *F* is independent of *D*. By contrast, R^2 of Model III is positive for all four traits (**Table 6-1**). Furthermore, for each trait, LRTs showed that Model III fits the data significantly better than the other two models (**Table 6-1**), and the fitted curve under Model III is hump-shaped (**Fig 6-1**). These results are robust to different window sizes (**Table F-1**). Interestingly, the OMDs for the four traits estimated under Model III are within a narrow range of $5.2\text{-}6.2\times10^{-3}$ (**Table 6-1**, **Fig 6-1**), which are close to *A. thaliana*'s genome-wide nucleotide diversity ($\pi = 5.4\times10^{-3}$; see Methods) and are smaller than its maximal intraspecific genetic distance ($D_{max} = 8.5\times10^{-3}$; see Methods).

To examine the generality of the hump-shaped relationship, we analyzed 231 crosses of the yeast *Saccharomyces cerevisiae* that included estimates of the maximum growth rates of all parents and hybrids in 11 different liquid media(PLECH *et al.* 2014). We again estimated *D* by the number of SNPs per site between parental genomes (see Methods). Based on the *D* values of all hybrids, we binned the hybrids using a window size of $D = 10^{-3}$. We first studied the mean *F*

from the 11 environments. Model III has an impressive R^2 of 0.85, whereas the corresponding values are negative for the other two models (**Table 6-2**). LRTs confirmed the significant superiority of Model III over the other two models (**Table 6-2**). Under Model III, a clear hump-shaped relationship is observed between the mean F and D, with the OMD = 4.5×10^{-3} (**Fig 6-2a**). These findings are robust to different window sizes (**Table F-2**). After the exclusion of reproductively isolated Chinese strains(WANG *et al.* 2012), $\pi = 4.3\times10^{-3}$ and $D_{max} = 9.6\times10^{-3}$ in *S. cerevisiae* (see Methods). Therefore, $\pi < \text{OMD} < D_{max}$.

When the data from different environments were separately analyzed, LRTs showed that Model III significantly outperforms the other two models in 10 of the 11 environments (except for the NaCl environment; **Fig 6-2b**). R^2 of Model III is higher than those of the other models in all 11 environments, and R^2 of Model III is positive in 10 of the 11 environments (except for the Y35 medium; **Fig 6-2c**). Intriguingly, however, in the benomyl (Ben) medium, the curve under Model III is not hump-shaped but U-shaped **(Fig F-1)**. Benomyl is a synthetic fungicide that targets microtubules(PLECH *et al.* 2014). It is possible that benomyl penalizes fast-growth strains more than slow-growth strains, resulting in a U-shaped curve. In the 10 environments (except for NaCl) where LRTs finds Model III significantly fitter than the other two models, OMD is in the range of $3.2\text{-}5.3\times10^{-3}$ (**Fig 6-2d**). All of these OMDs are lower than D_{max}, although some are also lower than π.

To verify the above results, we analyzed another yeast dataset(ZORGO *et al.* 2012), which included the measures of three growth traits (growth rate, negative lag time, and growth efficiency) in 56 environments from 28 crosses. Because the number of crosses is relatively small, we averaged F from all environments to minimize the estimation error of F. For each of the three traits, Model III fits the data significantly better than the other two models (**Table F-3**)

and the humped curve is apparent under Model III (**Fig F-2**). The OMDs for the three traits are 6.3, 4.4, and 5.4×10^{-3}, respectively (**Table F-3**), again between π and D_{max}.

We further expanded our analysis to animals by analyzing 28 crosses of the mouse *Mus musculus*(PHILIP *et al.* 2011). Two fitness-related traits, body weight and reproductive rate, were examined (see Methods). For each trait, Model III fits the data significantly better than the other two models (**Table F-4**) and a humped curve is observed under Model III (**Fig F-3**). The OMDs for the two traits are 5.1×10^{-3} and 6.6×10^{-3}, respectively (**Table F-4**), again between π (3.3×10^{-3}) and D_{max} (9.3×10^{-3}) of the species (see Methods).

6.4 Discussion

In summary, we detected the long anticipated hump-shaped relationship between *D* and *F* in each of the three model organisms examined, which represent three of the four kingdoms of eukaryotes. Our finding is also robust to the specific trait, environment, and method of analysis. Our success has a number of contributing factors, the lack of which likely explains previous failures. First, the range of *D* in the data should encompass the OMD; otherwise the humped relationship is easily missed. Second, an accurate measure of *D*, ideally based on genome sequences, is necessary for detecting the hump. Third, the variance of *F* among crosses at a given *D* can be large, requiring the use of many crosses to obtain reliable estimates. Fourth, crossing homozygotes simplifies the expectation and reduces the variance of *F*. Last but not least, having a mathematical model describing the theoretically expected relationship between *D* and *F* helps verify their relation. For instance, without such a model, the original authors of the *A. thaliana* study incorrectly concluded that *F* is independent of *D* on the basis that they are not significantly linearly correlated(YANG *et al.* 2017).

That Model III surpasses the other two models in explaining almost all datasets analyzed has several biological implications. First, it is currently unclear whether heterosis is caused by dominance, overdominance, or positive intergenic epistasis(LIPPMAN AND ZAMIR 2007). While our results do not confirm or refute the roles of overdominance and positive intergenic epistasis, they firmly establish the general contribution of dominance, because *a*, the coefficient of the linear term in Model III is found positive in all three species examined. Second, *b*, the coefficient of the quadratic term, reflects the sum of the incompatibility effect and the heterotic effect other than dominance. Because *b* is found negative while the heterotic effect is by definition nonnegative, the incompatibility effect must be negative. This result, again found in all three species studied, echoes the recent finding in fruit flies(MATUTE *et al.* 2010) and tomatoes(MOYLE AND NAKAZATO 2010) that the number of incompatibilities between two genotypes increases in proportion to D^2, and further demonstrates that fitness-related phenotypic effects of incompatibility also increase in proportion to D^2. Third, while the fly and tomato studies used only interspecific crosses(MATUTE *et al.* 2010; MOYLE AND NAKAZATO 2010), our crosses are all intraspecific. Hence, even within species, genetic incompatibility not only exists(CORBETT-DETIG *et al.* 2013) but also snowballs. Fourth, the net effect of heterosis and incompatibility on hybrid performance rises as *D* increases from 0 to the OMD, but retreats when *D* further increases, and eventually becomes negative when *D* exceeds twice the OMD. Because nonrandom mating and population structure is widespread in nature, the accumulation of genetic incompatibility within species could generate a selective pressure against interbreeding between distantly related conspecifics and initiate speciation. The importance of this process in nature may be tested by examining how often the OMD is below D_{max}. When OMD < D_{max}, as found in all three species examined, studying the incompatibilities between distantly related conspecifics

may shed light on the genetic basis of incipient speciation. It should be noted that the OMD can be recognized even if it exceeds D_{max} because relevant studies often include interspecific crosses(WILLI AND VAN BUSKIRK 2005).

Our findings also have implications for animal and plant breeding. To boost the hybrid performance, one should not only take the advantage of heterosis but also minimize the negative impact of incompatibility. Hence, the best mating distance should be close to the estimated OMD rather than D_{max}, as one might think without considering the impact of intraspecific genetic incompatibility. Further, because we found that the OMDs of multiple fitness-related traits in a given species tend to be similar, using mating distances close to the OMD will likely optimize a suite of fitness-related traits. In conservation biology, it is well appreciated that too small of a *D* is harmful due to inbreeding depression(HEDRICK AND KALINOWSKI 2000), but many studies show that too large of a *D* can cause outbreeding depression and is undesirable either(EDMANDS 2007). Our results suggest that applying the OMD in managing conservation may be most effective. In all three species studied, the OMDs of most traits are greater than π but smaller than D_{max}. This pattern, if further confirmed in additional lineages, suggests the general strategy of using mating distances slightly higher than π to minimize both inbreeding and outbreeding depressions when the OMD is unknown.

It is notable that heterosis has also been reported in humans. For example, an analysis of 35,000 humans from 35 different population samples showed a highly significant association between genome-wide heterozygosity and stature(MCQUILLAN *et al.* 2012). Further, higher levels of genetic heterozygosity are associated with lower blood pressure and total/LDL cholesterol(CAMPBELL *et al.* 2007). Therefore, a positive OMD likely exists in humans. Future estimation of this parameter may help understand relationships between human mating distance

and performances including health. In addition, modern humans interbred with archaic humans multiple times(NIELSEN *et al.* 2017); whether these events immediately increased or decreased the hybrid fitness is an interesting question that can be addressed when the human OMD is estimated.

6.5 Methods

6.5.1 Genetic distance and phenotypic data

We acquired the *Arabidopsis thaliana* phenotypic and genetic distance data from Yang et al.(YANG *et al.* 2017). There are 200 intraspecific hybrids generated by crossing 200 *A. thaliana* accessions with one common maternal accession. The hybrids and their parents were measured for four traits at 14 d after sowing: shoot fresh weight, rosette diameter, leaf area, and leaf number. The genomes of 191 parental accessions had been sequenced(YANG *et al.* 2017). In the original study(YANG *et al.* 2017), the genetic distance between parents was calculated by PLINK based on 722,000 SNPs. *A. thaliana* has a reference genome with a size of ~116.8 Mb. Using genome sequences, we calculated that the genome-wide per nucleotide distance between Col-0 and the commonly used Ler-1 equals 5.4×10^{-3}. Using this information allowed us to convert per SNP distance in the original study to per nucleotide distance for all pairs of accessions. We included all 191 hybrids with available per nucleotide genetic distances in our analysis. Genome-wide nucleotide diversity was estimated using the results of Nordborg et al.(NORDBORG *et al.* 2005). D_{max} was calculated from the maximum distance of 10,000 random pairs of strains from the 1135 genome-sequenced strains provided by the 1001 Arabidopsis Genome Project. Sampling 20,000 random pairs of strains does not increase D_{max}. All Arabidopsis whole-genome VCF files were downloaded from: http://1001genomes.org/data-center.html.

The *Saccharomyces cerevisiae* data were acquired from two sources. Our analysis focused on the data of Plech et al.(PLECH *et al.* 2014), which contained all 231 pairwise mating from 22 haploid parental strains. Plech et al.'s data have a comparable size with the *Arabidopsis* data and the range of genetic distance covered is larger than that in the other yeast dataset(ZORGO *et al.* 2012). Plech et al.'s data included maximum growth rates for the homozygous diploid parents and hybrids in 11 liquid media. They are YPD (nutrient rich with 2% glucose) at 30°C, Gal (nutrient rich with 2% galactose) at 30°C, YPG (nutrient rich with 3% glycerol) at 30°C, SD (synthetic medium with 2% glucose supplemented with uracil) at 30°C, Y20 (YPD at 20°C), Y35 (YPD at 35°C), and five YPD-based media at 30°C with additional chemicals indicated: Ben (benomyl), DM (6% DMSO), Na (2% NaCl), Sal (2% salicylate), and Zn (0.5 mg/ml $ZnSO_4$). Mating distances were from Liti et al.(LITI *et al.* 2009), calculated from 235,127 SNPs. We did not use the distances from a more recent study that sequenced yeast genomes to a higher coverage, due to its underestimation of distances because gaps and missing data were not excluded from the genome size in the distance estimation(MACLEAN *et al.* 2017). But because Liti et al. did not calculate the genome-wide π and included fewer strains than the more recent study(MACLEAN *et al.* 2017), we extrapolate π and D_{max} from the more recent study. Specifically, we regressed the distances between the two studies using all shared strains between them. Based on the linear regression (Pearson's $r = 0.99$, $P = 5.9\times10^{-200}$), we converted π and D_{max} from the more recent study by dividing them by 0.69.

We also analyzed Zorgo et al.'s yeast data, which included 28 pairwise crosses among 8 strains and measures of parent and hybrid phenotypes in growth rate, lag time, and yield in 56 environments(ZORGO *et al.* 2012). Note that because a greater lag time indicates a lower fitness, we used negative lag time as a fitness-related trait. We analyzed the mean F from all

environments to increase the accuracy of *F* estimates because of the relatively small number of crosses performed.

The phenotypic data of *Mus musculus* were acquired from Philip et al.(PHILIP *et al.* 2011). We used body weight and reproductive rate (first litter size divided by the time from first mating to first litter) as fitness-related traits(FLURKEY AND CURRER 2009). Because of the scarcity of data, we did not separate male and female hybrid animals in our analysis. We downloaded the whole-genome SNP data generated by Yalcin et al.(YALCIN *et al.* 2011) for the eight parental strains (ftp://ftp-mouse.sanger.ac.uk/current_snps/strain_specific_vcfs/) and estimated *D* by the number of SNPs per site between parental genomes. We used a window size of $D = 10^{-3}$ to bin the crosses. Because the *D* values of the 28 crosses cluster into four small groups, using a smaller window size such as $D = 0.5\times10^{-3}$ does not give more useful bins. Mouse has a π of 3.3×10^{-3} (FRAZER *et al.* 2007), and we estimated that $D_{max} = 9.3\times10^{-3}$ using the genome sequences of two most diverged subspecies, CAST/EiJ and PWK/PhJ, of *M. musculus*(GOIOS *et al.* 2007).

6.5.2 Causes of heterosis and genetic incompatibility

Heterosis arises from genetic interactions between the paternal and maternal alleles of the same loci (via dominance and overdominance) and/or different loci (via positive intergenic epistasis)(LIPPMAN AND ZAMIR 2007). Genetic incompatibility similarly originates from allelic interactions at the same loci (via underdominance) and/or different loci (via negative intergenic epistasis). At any locus, if the paternal and maternal alleles differ, either both of them are derived from their common ancestral allele or only one of them is derived whereas the other is ancestral.

Because fitter alleles tend to be partially or completely dominant over less fit alleles(FISHER 1928), when homozygous individuals from different populations hybridize, dominance can cause the hybrid to outperform the average of the two parents and result in heterosis. Because the occurrence of heterosis by dominance requires a change from the ancestral state in only one parent, it should rise in proportion to mating distance *D*. Overdominance, underdominance, positive intergenic epistasis, and negative intergenic epistasis can obviously occur in the hybrid between two derived alleles that are respectively homozygous in the two parents. Should overdominance between an ancestral and a derived allele occur, the derived allele will likely stay in the heterozygous state in one population; hence, heterosis is unlikely to occur upon hybridization. Similarly, should positive intergenic epistasis exist between an ancestral and a derived allele, this positive effect is already seen in one parent and thus is not heterotic. Should underdominance or negative intergenic epistasis occur between an ancestral and a derived allele, the derived allele will likely be selectively removed from the population and therefore is unlikely to contribute to genetic incompatibility between the two parents. Therefore, the effects from overdominance, underdominance, positive intergenic epistasis, and negative intergenic epistasis should most likely increase in proportion to D^2.

6.5.3 Parameter estimation

All calculations were performed using MATLAB. We used the function "lsqcurvefit" to perform least-squares estimations of the parameters of our three models. We used the estimated parameters to compute R^2 and conduct LRTs. The confidence interval of OMD is estimated by a bootstrap method. Specifically, we randomly sampled from all crosses with replacement the same number of crosses as in the original data and then estimated the OMD from the sampled

crosses. We repeated this process 1000 times to acquire the 95% confidence interval of the OMD. In our model fitting, only *D* was used as an independent variable. Although better parent heterosis (BPH)(ZORGO *et al.* 2012), which describes the phenotypic difference between the hybrid and the better parent, is also commonly used to study heterosis, there is no clear theoretical relationship between *D* and BPH. Hence, we focused on *F*, which is also known as the heterosis coefficient(ZORGO *et al.* 2012).

6.6 Acknowledgements

We thank Drs. Wei-Chin Ho, Alexey Kondrashov, Haoxuan Liu, Wenfeng Qian, and Jian-Rong Yang for valuable comments. This work was supported by the U.S. National Institutes of Health research grant R01GM103232 to J.Z.

6.7 References

Amos, W., J. W. Wilmer, K. Fullard, T. M. Burg, J. P. Croxall *et al.*, 2001 The influence of parental relatedness on reproductive success. Proceedings of the Royal Society B-Biological Sciences 268**:** 2021-2027.

Bloom, J. S., I. Kotenko, M. J. Sadhu, S. Treusch, F. W. Albert *et al.*, 2015 Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nat Commun 6**:** 8712.

Campbell, H., A. D. Carothers, I. Rudan, C. Hayward, Z. Biloglav *et al.*, 2007 Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Hum Mol Genet 16**:** 233-241.

Corbett-Detig, R. B., J. Zhou, A. G. Clark, D. L. Hartl and J. F. Ayroles, 2013 Genetic incompatibilities are widespread within species. Nature 504**:** 135-+.

Edmands, S., 1999 Heterosis and outbreeding depression in interpopulation crosses spanning a wide range of divergence. Evolution 53**:** 1757-1768.

Edmands, S., 2007 Between a rock and a hard place: evaluating the relative risks of inbreeding and outbreeding for conservation and management. Mol Ecol 16**:** 463-475.

Fisher, R. A., 1928 The possible modification of the response of the wild type to recurrent mutations. The American Naturalist 62**:** 115-126.

Flurkey, K., and J. M. Currer, 2009 *The Jackson Laboratory handbook on genetically standardized mice*. Jackson Laboratory.

Frazer, K. A., E. Eskin, H. M. Kang, M. A. Bogue, D. A. Hinds *et al.*, 2007 A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature 448**:** 1050-1053.
Goios, A., L. Pereira, M. Bogue, V. Macaulay and A. Amorim, 2007 mtDNA phylogeny and evolution of laboratory mouse strains. Genome Res 17**:** 293-298.
Gonzalez, W. L., L. H. Suarez and R. Medel, 2007 Outcrossing increases infection success in the holoparasitic mistletoe Tristerix aphyllus (Loranthaceae). Evolutionary Ecology 21**:** 173-183.
Hedrick, P. W., and S. T. Kalinowski, 2000 Inbreeding depression in conservation biology. Annu. Rev. Ecol. Syst. 31**:** 139-162.
Hung, H. Y., C. Browne, K. Guill, N. Coles, M. Eller *et al.*, 2012 The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. Heredity 108**:** 490-499.
Jagosz, B., 2011 The relationship between heterosis and genetic distances based on RAPD and AFLP markers in carrot. Plant Breeding 130**:** 574-579.
Lippman, Z. B., and D. Zamir, 2007 Heterosis: revisiting the magic. Trends Genet 23**:** 60-66.
Liti, G., D. M. Carter, A. M. Moses, J. Warringer, L. Parts *et al.*, 2009 Population genomics of domestic and wild yeasts. Nature 458**:** 337-341.
Lynch, M., 1991 The Genetic Interpretation of Inbreeding Depression and Outbreeding Depression. Evolution 45**:** 622-629.
Maclean, C. J., B. P. H. Metzger, J. R. Yang, W. C. Ho, B. Moyers *et al.*, 2017 Deciphering the Genic Basis of Yeast Fitness Variation by Simultaneous Forward and Reverse Genetics. Mol Biol Evol 34**:** 2486-2502.
Matute, D. R., I. A. Butler, D. A. Turissini and J. A. Coyne, 2010 A Test of the Snowball Theory for the Rate of Evolution of Hybrid Incompatibilities. Science 329**:** 1518-1521.
McClelland, E. K., and K. A. Naish, 2007 What is the fitness outcome of crossing unrelated fish populations? A meta-analysis and an evaluation of future research directions. Conservation Genetics 8**:** 397-416.
McQuillan, R., N. Eklund, N. Pirastu, M. Kuningas, B. P. McEvoy *et al.*, 2012 Evidence of inbreeding depression on human height. PLoS Genet 8**:** e1002655.
Moll, R. H., J. H. Lonnquist, J. V. Fortuno and E. C. Johnson, 1965 The Relationship of Heterosis and Genetic Divergence in Maize. Genetics 52**:** 139-144.
Moran, P., A. M. Pendas, E. Garciavazquez, J. I. Izquierdo and J. Loboncervia, 1995 Estimates of Gene Flow among Neighboring Populations of Brown Trout. Journal of Fish Biology 46**:** 593-602.
Moyle, L. C., and T. Nakazato, 2010 Hybrid Incompatibility "Snowballs" Between Solanum Species. Science 329**:** 1521-1523.
Nielsen, R., J. M. Akey, M. Jakobsson, J. K. Pritchard, S. Tishkoff *et al.*, 2017 Tracing the peopling of the world through genomics. Nature 541**:** 302-310.
Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol 3**:** e196.
Pekkala, N., K. E. Knott, J. S. Kotiaho, K. Nissinen and M. Puurtinen, 2012 The benefits of interpopulation hybridization diminish with increasing divergence of small populations. Journal of Evolutionary Biology 25.
Philip, V. M., G. Sokoloff, C. L. Ackert-Bicknell, M. Striz, L. Branstetter *et al.*, 2011 Genetic analysis in the Collaborative Cross breeding population. Genome Res 21**:** 1223-1238.

Plech, M., J. A. G. de Visser and R. Korona, 2014 Heterosis is prevalent among domesticated but not wild strains of Saccharomyces cerevisiae. G3: Genes, Genomes, Genetics 4**:** 315-323.
Robinson, S. P., W. J. Kennington and L. W. Simmons, 2009 No evidence for optimal fitness at intermediate levels of inbreeding in Drosophila melanogaster. Biological Journal of the Linnean Society 98**:** 501-510.
Stelkens, R. B., M. Pompini and C. Wedekind, 2014 Testing the effects of genetic crossing distance on embryo survival within a metapopulation of brown trout (Salmo trutta). Conservation Genetics 15**:** 375-386.
Stokes, D., C. Morgan, C. O'Neill and I. Bancroft, 2007 Evaluating the utility of Arabidopsis thaliana as a model for understanding heterosis in hybrid crops. Euphytica 156**:** 157-171.
Wang, Q. M., W. Q. Liu, G. Liti, S. A. Wang and F. Y. Bai, 2012 Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. Mol Ecol 21**:** 5404-5417.
Willi, Y., and J. Van Buskirk, 2005 Genomic compatibility occurs over a wide range of parental genetic similarity in an outcrossing plant. Proc Biol Sci 272**:** 1333-1338.
Xiao, J., J. Li, L. Yuan, S. R. McCouch and S. D. Tanksley, 1996 Genetic diversity and its relationship to hybrid performance and heterosis in rice as revealed by PCR-based markers. Theoretical and Applied Genetics 92**:** 637-643.
Yalcin, B., K. Wong, A. Agam, M. Goodson, T. M. Keane *et al.*, 2011 Sequence-based characterization of structural variation in the mouse genome. Nature 477**:** 326-329.
Yang, M., X. C. Wang, D. Q. Ren, H. Huang, M. Q. Xu *et al.*, 2017 Genomic architecture of biomass heterosis in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 114**:** 8101-8106.
Zorgo, E., A. Gjuvsland, F. A. Cubillos, E. J. Louis, G. Liti *et al.*, 2012 Life History Shapes Trait Heredity by Accumulation of Loss-of-Function Alleles in Yeast. Molecular Biology and Evolution 29**:** 1781-1789.

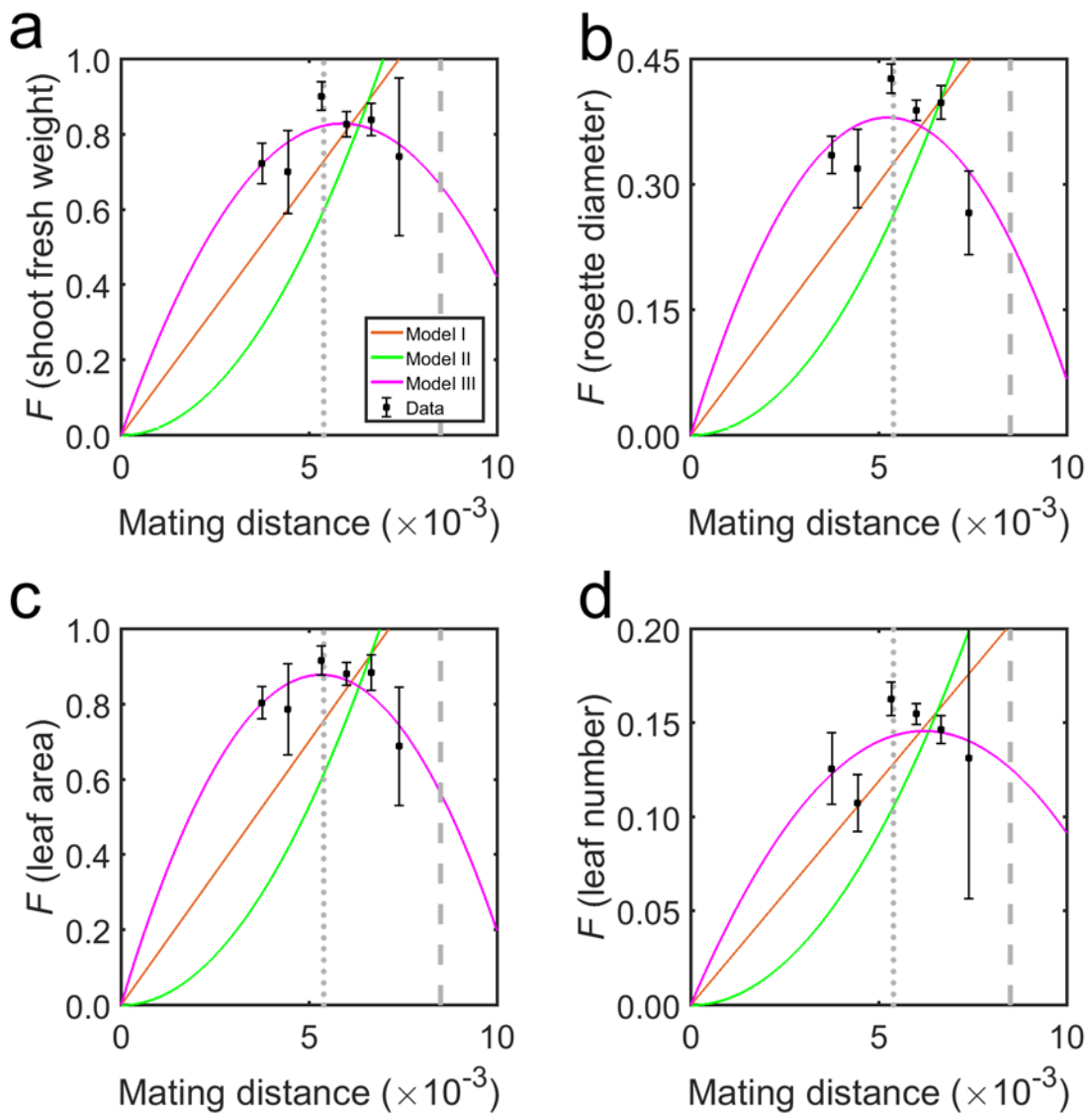


Figure 6-1. Hump-shaped relationship between mating distance (*D*) and hybrid performance (*F*) measured by (**a**) shoot fresh weight, (**b**) rosette diameter, (**c**) leaf area, and (**d**) leaf number in the plant *Arabidopsis thaliana* at 14 d after sowing. The mean and standard error of *F* are respectively shown by black squares and associated error bars. The fitted *D*-*F* curves under different models are shown in different colors. Statistics of model fitting are provided in Table 6-1. Nucleotide diversity (π) and maximal intraspecific genetic distance (D_{max}) are respectively indicated by vertical dotted and dashed lines.

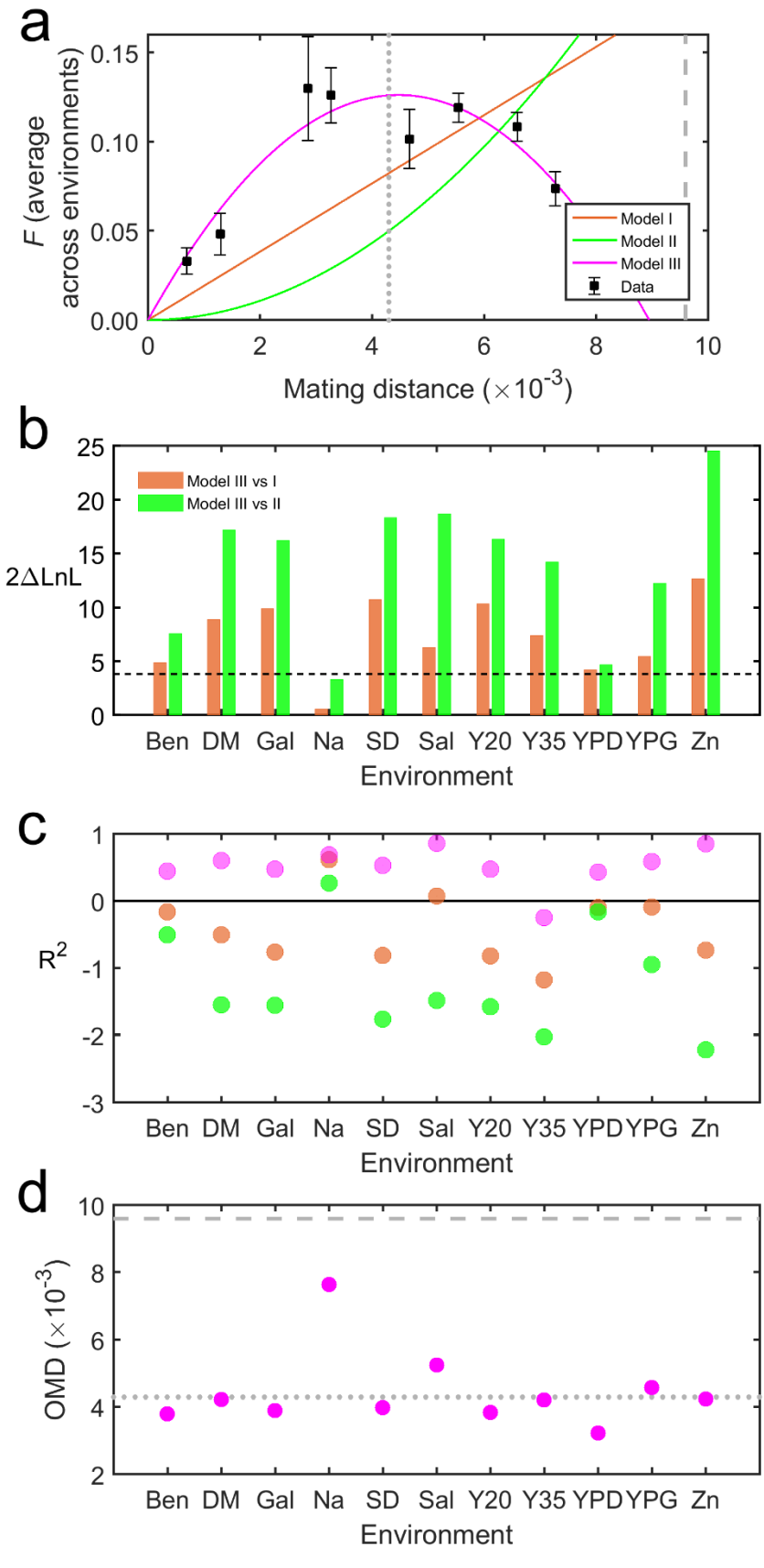

a
F (average across environments)
0.15
0.10
0.05
0.00
0
2
4
6
8
10
Mating distance (×10-3)
Model I
Model II
Model III
Data
b
2ΔLnL
25
20
15
10
5
0
Model III vs I
Model III vs II
Ben DM Gal Na SD Sal Y20 Y35 YPD YPG Zn
Environment
c
R2
1
0
-1
-2
-3
Ben DM Gal Na SD Sal Y20 Y35 YPD YPG Zn
Environment
d
OMD (×10-3)
10
8
6
4
2
Ben DM Gal Na SD Sal Y20 Y35 YPD YPG Zn
Environment

Figure 6-2. Hump-shaped relationship between mating distance (*D*) and hybrid performance (*F*) in the fungus *Saccharomyces cerevisiae* across 11 environments. (**a**) The *D*-*F* relationship when *F* is measured by the average maximum growth rate in 11 environments. The mean and standard error of *F* are respectively shown by black squares and associated error bars. The fitted *D*-*F* curves under different models are shown in different colors. π and D_{max} are respectively indicated by vertical dotted and dashed lines. (**b**) Twice the difference in ln(likelihood) between Model III and Model I (orange) or II (green) under each environment. The larger the difference, the fitter Model III is relative to the model being compared. The horizontal black dashed line shows statistical significance at 5% level. X-axis lists environments, whose details are provided in Methods. (**c**) Model fitting for the *D*-*F* relationship in each of the 11 environments. Color coding is the same as in panel **a**. The higher the R^2, the fitter the model is to the data. The horizontal black line indicates $R^2 = 0$. (**d**) The estimated optimal mating distance (OMD) in each environment. π and D_{max} are respectively indicated by horizontal dotted and dashed lines.

Table 6-1. Fitting of the three models to *A. thaliana* data

Traits	Models	R^2	$2\Delta \ln L$[1]	P-value[2]	OMD [95% CI] ($\times 10^{-3}$)
Shoot weight					
	I	-4.15	28.2	1.1×10^{-7}	
	II	-16.20	100.4	1.2×10^{-23}	
	III	0.54			5.9 [4.8-9.7]
Rosset diameter					
	I	-2.32	16.5	4.7×10^{-5}	
	II	-7.38	46.9	7.4×10^{-12}	
	III	0.44			5.2 [4.7-7.5]
Leaf area					
	I	-6.26	41.3	1.3×10^{-10}	
	II	-19.50	120.7	4.4×10^{-28}	
	III	0.63			5.3 [4.7-7.1]
Leaf number					
	I	-1.34	10.4	1.3×10^{-3}	
	II	-6.50	41.4	1.3×10^{-10}	
	III	0.39			6.2 [-19.9-44.5]

[1]Twice the difference in ln(likelihood) between Model III and the model being compared.

[2]*P*-values of likelihood ratio tests are determined using chi-squared tests with 1 degree of freedom.

Table 6-2. Fitting of the three models to *S. cerevisiae* data averaged across 11 environments

Models	R^2	$2\Delta \ln L$[1]	P-value[2]	OMD [CI 95%] ($\times 10^{-3}$)
Model I	-0.65	12.0	5.3×10^{-4}	
Model II	-2.40	26.0	$3.4. \times 10^{-7}$	
Model III	0.85			4.5 [4.2-4.9]

[1]Twice the difference in ln(likelihood) between Model III and the model being compared.

[2]*P*-values of likelihood ratio tests are determined using chi-squared tests with 1 degree of freedom.

Chapter 7

Discussions and Future Directions

"What we know is a drop, what we don't know is an ocean."

— **Isaac Newton**

7.1 Introduction

The field of evolutionary genetics has progressed fast thanks to many new techniques developed in the last ten years. Although none of the state-of-art techniques is used in my thesis, many empirical and computational studies using those techniques greatly improved our understanding of evolution, genetics, and molecular biology. Though I have not personally involved in those works, I am excited about those achievements. Thanks to the newly developed techniques and the intellectual progress made, I believe right now is a perfect time to study biology.

Each of the main chapter and appendix chapter has their own discussion section about the results and interpretations, which will not be repeated here. In this overall discussion chapter, I discuss some thoughts and opinions I conceived while working on this thesis in an open-ended way, including opinions about genetic interactions, gene-by-environment interactions, more synthetic discussions about the models proposed in the main chapters, opinions about some

unaddressed questions in evolutionary genetics, as well as some new questions and methods in my mind. Some of the sections discussed here have preliminary results; some are discussed in a hypothetical manner. In the end, I will discuss questions that I may work on in the future.

Although the things discussed in this chapter are interesting to me, the "interesting" here is a subjective feeling and does not represent the truly interesting directions in the field of evolutionary genetics. The approaches to deal with some questions proposed here are also based on limited knowledge and incomplete thinking and are likely wrong or distant from the current field's progressing direction. For anyone who accidentally reads this part, please keep an open and critical mind about everything written, and I would best hope a quarter of things discussed are worth to look at.

7.2 Connecting genetic interaction with G×E

In chapter3, genetic interactions and G×E are combined in modular life model through the existence of modules. This bold design is based on our observation that the effect sizes of beneficial mutations decrease as *Q* increases and as genotype quality increases; it indicates that the interactions with environment might be similar to the genetic interactions. Because I have studied both genetic interactions and G×E throughout my chapters, it seems necessary to discuss the connections between them.

I propose that G×G and G×E are similar to each other. This is because it is perhaps difficult for an environment to interact with the product of a gene directly to create G×E. For example, only genes that produce membrane proteins or membrane molecules in single cell organism literally physically interact with the environment. Some chemicals from the environment may enter a cell via diffusion or endocytosis and then have physical interactions

with the products of genes, but this cannot be the case for every factor of an environment, because one could imagine that two environments differ only in a factor that does not enter the cell but still have G×E. If most of the products of genes cannot directly interact with the environment, then G×E observed either must happen through the products of other genes or only exists for a set of special genes that physically interact with the environment. For the former scenario, the existence of G×E interaction for a gene indicates the products of some other genes likely influence this gene's behavior, suggesting that G×E genes are likely to have genetic interactions than non-G×E genes, and there may exist more G×G×E interactions for those G×E genes. If the former is not true, then the genes that show G×E must interact with some parts of the environment. This can be verified when mapping G×E to genes becomes inexpensive.

Another similarity between G×E and genetic interactions is the similarity in the effects of interactions. In chapter 2, I showed that the majority of G×E are concordant G×E (WEI AND ZHANG 2017). This kind of "concordance" is also true for genetic interactions, because underdominance is rare, compared to other allelic interactions (COYNE *et al.* 1991), and sign epistasis is not as common as negative or positive epistasis (KRYAZHIMSKIY *et al.* 2014). This direction of effect similarity again may indicate that G×E and genetic interactions share some underlying mechanisms. Although it is possible for genetic interactions to be one form of G×E, I personal prediction is that G×E often have underlying genetic interactions and happen through genetic interactions. This said, it is possible that antagonistic G×E genes have some epistatic properties different from concordant G×E genes. For example, chapter 2 reported that antagonistic G×E and concordant G×E have different genomic enrichment regions (WEI AND ZHANG 2017). It might provide some mechanistic insights if we could compare and connect types of G×E with different types of G×G using some large datasets.

7.3 Modular life model

7.3.1 The geometric mean in modular life model and in biology

In chapter 3, I propose modular life model to explain diminishing returns epistasis, where I assume that the fitness of an individual can be the geometric mean, arithmetic mean, or the minimal value of all the functionalities of modules. Although I did not discuss which model is closer to the true model, that only the geometric mean one can explain the result from Chou et al study (CHOU *et al.* 2009) is mentioned in the discussion of Chapter 3. In Chapter 4, I also use modular life model to explain the general trend of genetic dominance. It is also true that only when using geometric mean, we could successfully predict all patterns of genetic dominance, the other two approaches cannot predict *h*-*s* correlation for genetic dominance. These results differentiate the three proposed models in Chapter 3 and indicate that the geometric mean of modular life model might be more relevant to real-world biological system and can predict genetic interactions and gene-by-environment interactions better than the other two models.

The geometric mean is also biologically relevant given it uses a multiplicative approach. Geometric mean is more appropriate than the arithmetic mean for describing proportional growth, both exponential growth (constant proportional growth) and varying growth, making it a relevant application to fitness calculation in biology. For example, if the effect of a mutation changes from generation to generation due to the change of biotic or abiotic environment, geometric mean can account for the differences. Another example is that, if we decompose the fitness of an individual into the survival rate at the zygote level, the survival rate at the juvenile level, the fitness at the adult level, and the fitness at germ cell level, the overall fitness of a genotype is a

combination of the four stages. Geometric mean could represent the fitness of a genotype at an average stage.

The logarithm of geometric mean is related to the arithmetic mean, so it is also possible that some unique properties of geometric mean is related to the logarithm properties in biology. While geometric mean in practice is superior in explaining those interactive effects, the true reason is yet to be discovered.

7.3.2 Modular life model for predicting functional modules

Although there are already some works predicting modules in networks based on physical, biochemical, or genetic interactions, the modules in terms of interaction is different from the modules in modular life model. The modules in modular life model are grouped by functional similarity and functional redundancy and it relates to the genotype-phenotype map. According to modular life model, genes that belong to the same module have similar functions, and environment contributes to a module in a similar way. The geometric mean of the functionalities of all modules equals to fitness. We can use this information to identify functional modules.

For example, according to the model, when the environmental contribution to a module varies, the fitness effects of genes within that particular module should increase or decrease together. If environment contribution to modules varies randomly, the effects of genes that belong to different modules should not covariate with each other. Therefore, we can use the effect sizes of some genes and mutations across multiple environments to get the correlation coefficient between the effect sizes of two genes across environments as a score for how likely two genes are from the same module. Using Bloom et al QTL mapping data (BLOOM *et al.* 2013; BLOOM *et al.* 2015), I calculate the effect size of each SNP in each of the 47 environments. I then

calculate the Spearman correlation coefficient and significance level between each pair of SNPs on two different chromosomes. Although there are more than 4×10^8 pairs of SNPs tested, tens of pairs of unlinked regions are significant after multiple testing corrections. Because I only used SNPs pairs locating on different chromosomes to calculate the correlation, linkage cannot be a confounding factor here, and the effect sizes of some mutation pairs truly covariate with each other. The fact that I observe effect size correlations is consistent with the modular life model prediction. However, because it is difficult to use QTL mapping approach to locate to the genes, I cannot determine which gene pairs belong to the same module and how many modules there are. Moreover, that the effect sizes of two genes covariate with each other across environments is also not a direct evidence of them belonging to the same functional module.

Another possible approach to identify the "functional modules" under modular life model is to use the data for identifying essential genes and gene deletion effects. According to modular life model, different environment contributes to different functional modules differently, and the lack of contribution from an environment to a module may create essential genes and bigger deleterious effect for null mutations, higher contribution from an environment to a module may make the null mutations of genes in that module close to neutral. Because in some model organisms, the effect of each null mutation is estimated in many environments, the aforementioned approach as could be used to process this information to infer which sets of genes likely belong to the same module and the minimum number of modules needed to explain the data.

Under the geometric mean modular life model, there is diminishing returns epistasis between two beneficial mutations within a module, but there is also positive (synergistic or widening) epistasis between two beneficial mutations from two different modules. Therefore,

although it is proposed to explain diminishing returns epistasis, it has the property to predict positive epistasis as well. This is again due to the multiplicative nature of geometric mean. Because negative epistasis is likely more general than positive epistasis (COSTANZO *et al.* 2016), the ratio of positive versus negative epistasis may predict the maximum number of modules. Apparently, not every gene is a module; otherwise, there exists only positive epistasis. In addition, there needs more than one module, because otherwise no epistasis is predicted under modular life model.

Combining information from all three approaches may help understand how many modules there are, and how many genes in each module are. A naive guess for the number of modules is on the order of ten. Because the modular life model has some interpretations about the distribution and the effect of genetic epistasis, the data for estimating the genetic interaction map may also be useful for module prediction.

7.3.3. Modular life model for other questions in genotype phenotype mapping

Because modular life model showed good prediction ability in chapters 3 and 4, it is worth to discuss its connection to other genotype-phenotype mapping questions. The current model has very restricted parameters; it might be good to also generalize the current model so that it can work on other types of genotype-phenotype questions.

The modular model can be used to predict phenotype. In diploid, modular life model could be used to predict the fitness of hybrid given the fitness of two homozygotes parents. Because the modular level of each module in a heterozygous hybrid is at least equal to the average modular level of its two homozygous parents (i.e. when there is no saturation for a module in both parents, the hybrid's modular level is the average of the parents; when there is

saturation in one parent, the hybrid's modular level is higher than the average of the parents; when there is saturation for a module in both parents, the hybrid's modular level again equals the average of the parents), predicting the hybrid fitness is at least better than the average of the two parents. This is true in yeast hybridization experiments (PLECH *et al.* 2014). Mathematically, it predicts that the fitness of hybrid is at least the average fitness of the two parents, and it can be better than both of the parents, i.e. hybrid vigor. The current modular life model does not include the effects of incompatibility, therefore could only predict hybrid fitness without the incompatibility effect, serving as an upper bound.

Under the currently proposed model in chapter 3 and 4, each single gene could only contribute to one module; therefore, there is no genetic pleiotropy. It also only considers beneficial mutations to contribute to the level of a module, and null mutation to contribute nothing; in reality, deleterious mutation may decrease the level of a module rather than adding nothing. A more generalized model should allow some genes to contribute one or more modules, and some deleterious mutation to decrease the functionality of a module, and add a lower bound of module level to zero. Lethality can be predicted in the original modular life model when the level of a module reaches zero. It also predicts rescuing mutations, because a mutation that brings back the modular level can rescue the effect of the first deleterious mutation. However, under the current (i.e., non-pleiotropic) model, the beneficial mutation can only be beneficial or neutral across environments. If a mutation could have positive effect in zero or more modules and negative effects in zero or more modules, the effect of the mutation could be antagonistic or concordant depending on how environment contribute to different modules. These two extension of parameters make the model arguably more realistic and more general. Moreover, it will allow two lethal mutations could compensate each other and create sign epistasis (CHEN *et al.* 2016).

Having a more generalized model, although making it slightly more complicated, also provides the possibility for modular life model to explain more phenomena in genotype phenotype mapping.

The modular life model is a special case of an artificial neuron network in deep learning and the structure can be learned and evaluated with QTL mapping data and could be used to predict the amount of explained broad sense heritability. The input layer is the genotypes or mutations that (are known to) have effects on fitness, the second layer is the modules, and the output layer is the fitness of that genotype. Here I propose a way to use modular life model and deep learning to improve QTL mapping result in explaining broad sense heritability. In a large QTL mapping data like Bloom et al (BLOOM *et al.* 2013; BLOOM *et al.* 2015), thousands of genotypes and tens of environments are available. The QTLs mapped in each of the environments could be combined as the input layer. A range of number of module should be explored, such as from two to maximal number of unlinked QTLs. Some loss function to weigh the number of parameters can be added to improve the robustness of the model. The effect of environment is reflected at the hidden layer, so that each environment's effect is being estimated to achieve a maximum likelihood result. The effect of QTLs and the organization from QTLs to modules will also be randomly explored to generate a robust prediction of output fitness. Because some of the genotypes could be held back, the accuracy of different models could be compared. After getting the best model, all data should be used to fit the model, and the total variance explained could be calculated by comparing the predicted fitness to the empirically measured fitness, and the result reflects the total amount of broad sense heritability explained by these QTLs.

7.4 Diminishing returns epistasis of phenotypes

Chapter 4 showed that diminishing returns is also widespread among natural polymorphisms. The comparison between empirical patterns of diminishing returns and modeling results suggests that diminishing returns originates from the modular organization of life where the contribution of each functional module to fitness is determined jointly by the genotype and environment and has an upper limit.

Because diminishing returns epistasis may not be restricted to fitness, we could work on diminishing returns of phenotype the same way as we study diminishing returns of fitness to see if diminishing returns of phenotype also exist. Based on my preliminary study in yeast, when a phenotype is under directional selection, a mutation that changes this phenotype toward its favored direction tends to have a smaller effect when occurring in genotypes already having favored phenotypic values. This result is from a pilot study in yeast. I used 220 yeast cellular phenotypes for segregants from a cross between BY and RM yeast strains (CHUFFART *et al.* 2016) and their growth rates measured in YPD (BLOOM *et al.* 2013) to study phenotype diminishing returns. 1) For each phenotype QTL, I test whether it has fitness effect and whether the BY allele or RM allele increases the growth rate. 2) I measure whether the BY allele increases the phenotype or decreases the phenotype. 3) I measure the phenotypic effect of QTL using large phenotypic value genotypes and small phenotypic value genotypes and test if the former group has smaller or bigger effect than the latter. If an allele increases the phenotypic value, and if it shows a smaller phenotypic effect in large phenotypic value group, it means there is diminishing returns epistasis for that allele. If the majority of QTLs of a phenotype show diminishing returns, then there is the diminishing returns of the phenotype. I found that, when a phenotype has growth rate effect, then it is more likely to show diminishing returns, and when it does not affect growth rate, it does not show diminishing returns of the phenotype ($P < 0.01$). This is a proof of

principle test, suggesting that the diminishing returns pattern of a phenotype tells whether a phenotype has fitness effect/under selection. Because fitness/selection is hard to measure in human, this test could provide information about which human phenotype is under selection.

Studying the diminishing returns epistasis for human phenotypes might be interesting. So here I propose an approach to study it. UK Biobank is a national and international health resource with unparalleled research opportunities, open to all bona fide health researchers. It has been following the health and well-being of 500,000 volunteer participants. The phenotype and genotype data in UK biobank can allow us to test the phenotype diminishing returns prediction thoroughly, and the results will help us infer human phenotypes that are under directional selection.

I propose to first choose a range of quantitative disease/physiology phenotypes based on their heritability (i.e. aspects of cognition, height, lifespan, number of kids, number of siblings, education, cancer, and etc.), and either map the GWAS loci ourselves or search for the GWAS loci from published papers/database (e.g. http://www.ebi.ac.uk/gwas/). For each trait, I can divide the individuals into a high-value group and a low-value group. Then I could calculate the phenotypic effect of each GWAS locus from the high group and from the low group. I then predict the direction of selection based on whether the majority of GWAS loci have larger/smaller effects in the high phenotype group. Using diseases records can also be interesting, for example, the year/age of diagnosis for recurrent/chronic diseases such as diabetes, stroke, and kidney disease, because some diseases which have late onsite is suggested to be invisible to natural selection.

7.5 How to use QTL mapping data for alternative questions

In my chapter 2 to 5, I have shown that QTL mapping data can be used for non-QTL mapping purpose, and it is a valuable source to study mutational effects. These alternative usages of public data are not uncommon for computational evolutionary studies. There are two opposite opinions about such usage, sometimes it is appreciated as an efficient way, sometimes it is accused of harming the research in the long run. In particular, publishing data is hated by some experimentalists (TAICHMAN *et al.* 2016) because the data could be used by other people to refute the original conclusion of people who generated the data. However, I argue even such usages of published data are in the long run healthy because it avoids wasting people's efforts following wrong results. Moreover, published data could be used to address completely different questions, as what I did in my main chapters. Using public data innovatively is also eco-friendly because it avoids the waste of human labor, time, and money to generate similar data again.

Here, I'd like to propose some alternative questions that can be addressed with QTL mapping data. QTL mapping data has several properties, a large number of recombinant genotypes, a large number of phenotypes from different genotypes, and sometimes multiple available environments. Each aspect of these properties could be used to address some questions, and the combination of two or three of these properties could be used to address different questions. Here I provide some examples of how I'd like to use the QTL mapping data.

For example, adaptive walk and fitness landscape could be simulated based on the information from QTL mapping data, where a randomly sampled genotype could be seen as the starting point, and the current segregating SNPs could be seen as the available pool to sample random mutations. The fitness of neighboring genotypes of the starting genotype can be predicted under the assumption of no epistasis, any genotype that is within the predicted mutational steps from the starting genotype could be empirically sampled and the difference

between the predicted and the estimated value is an indicator of how much of epistasis exist between these two genotypes. This approach could be used to address some questions about epistasis, such as whether the further the two genotypes are away from each other, the more different the epistasis is between them. Some genotype might be more epistatic than other genotypes, and some alleles may be more epistatic than the alternative alleles. The properties of epistatic potential for major and minor alleles from population data could be predicted to address the questions about robustness and fixation.

Another example, many of the QTL mapping data generated and sequenced a lot of recombinant genotypes. Questions about recombination could be addressed with QTL mapping data, more likely with even better resolution than with the data generated for the purpose of studying recombination. I am now conducting an analysis about recombination with QTL mapping data.

Another example, some other questions involving next-generation sequencing could be addressed with QTL mapping data. For example, in yeast, the colonies sequenced are often still actively going through cell division and are also quite synchronized because many yeast colonies used in QTL mapping starts from single yeast segregant. Because the ongoing DNA replication and cell division, if a DNA region has early replication and strong DNA replication firing, that region will have twice of the reads than the regions with late DNA replication. This provides good opportunity to extract the DNA replication firing location for different genotypes, which itself creates many new phenotypes for QTL mapping. By comparing the genomic location of the replication origin and the strength of replication firing among genotypes, one could answer how much of these are explained by the DNA level difference, and how much it is cis-regulated versus trans-regulated.

7.6 Evolutionary memories via genetic mutations and "epigenetic mutations"

7.6.1. Molecular clock of "epigenetic evolution" and "epigenetic memory" in adaptation

I've been interested in and thinking about epigenetic evolution for many years, in particular about the evolution of DNA methylation, which might become one direction of my future work. It is perhaps worthwhile to write down what I thought of, for the purpose of discussing future direction. The main conclusion I reached is perhaps many of the tools used to study genetic evolution could be used to study epigenetic evolution. I will illustrate what I mean.

Genetic information has high fidelity and does not plastically change upon the environment shift; any mutation will leave a mark on the information which passes down accurately unless another mutation hits on the same position. It is also known that epigenetic change could pass down generation to generation (HEARD AND MARTIENSSEN 2014). Epigenetic markers depend on both the genetic part and the environment, and just like every biological process has error, it can also change due to "epigenetic mutation", which I define as the random error occurred during the process of copying epigenetic markers to the newly synthesized strand that could pass down to the next generation in a stable environment. Study the evolution of epigenetics is difficult due to its instability and the dependence on both genetic and environment. However, these properties also offer the opportunity for epigenetics to immediately respond to an environment change (BÖRSCH-HAUBOLD *et al.* 2014) and offering potential fitness advantage. Because DNA methylation depends on DNA sequences, it is also a heritable trait, and because methylation change can provide fitness advantage sometimes, it is reasonable to believe that some of the epigenetic changes are adaptive, and epigenetic evolution can be studied and should be studied.

The term “epimutation” was first introduced in 1987 by Robin Holliday to refer to the heritable changes in gene activity due to DNA modification to distinguish from classical mutations (HOLLIDAY 1987). However, “epimutation” is often misused as “epigenetic variation”, which is not what Holliday originally proposed. Some papers tried to define “epigenetic mutation” again, for example, one interesting theoretical paper defines “epigenetic mutation” as a change with higher rate but smaller stability as compared to a genetic mutation (KRONHOLM AND COLLINS 2016). These definitions, though reasonable, has little to do with the molecular nature of “epigenetic mutation”. Here, for the purpose of my discussion, I need to redefine “epigenetic mutation” in a more conservative way. If in a constant environment, some of the DNA methylation (or other DNA modification) changes occur by error or damage that are heritable to newly synthesized DNA, and if there is no DNA level mutation that directly causes the epigenetic level change, then those DNA methylation (or other epigenetic) changes are “epigenetic mutation”. So here, I exclude the plastic “epimutation” due to environmental changes, and the “genetic epimutation” due to classical mutation, and only include the stochastic “epigenetic mutation”. A clearer definition will help study “epigenetic mutation” experimentally.

Given that DNA methylation could be easily measured by MethylC-Capture sequencing (DO *et al.* 2017), and its dependency on DNA sequences, it provides a good opportunity to study the intrinsic “epigenetic mutation” rate. When an environment is constant, the plastic change caused by the environment is minimized. An epigenetic mutation accumulation study could be done in a similar way as a regular mutation accumulation experiment. Because the mutation rate of DNA sequence is low, the majority of DNA methylation changes are caused by epigenetic mutations. Some DNA level mutations happen during epigenetic mutation accumulation process, the epigenetic change linked to a DNA mutation could also be quantified as the epigenetic

change driven by DNA change. The epigenetic mutation rate for DNA methylation can be estimated. Knowing the DNA methylation change linked to DNA mutation allows people to study how much of the DNA methylation difference among individuals of the same species could be explained by the DNA level difference. Besides, comparing to the epigenetic difference between species could help understand whether the majority of epigenetic mutations are being purified by selection.

DNA methylation and other epigenetic modifications may facilitate the adaptation to a new environment. The theoretical work using Fisher's geometric model has been attempted (KRONHOLM AND COLLINS 2016), but empirical work has not yet followed up. Here I discuss the mechanistic reasons why epigenetic mutations may facilitate adaptation. In addition, I discuss the relationship between epigenetic plasticity and adaption, and how to study these epigenetic effects experimentally. Because epigenetics is more plastic upon environment change, and the epigenetic mutations are less stable than genetic mutations, the individuals with beneficial epigenetic mutations or beneficial epigenetic plasticity may be able to fix some genetic mutations that stabilize the adaptive epigenetic effects. Beneficial epigenetic mutation/plasticity may also provide the genotype time and opportunity to accumulate genetic mutations that confer independent benefit. Because natural environment is not stable, it is reasonable to think that the individuals that can best adapt to the environmental changes are also those with the right amount of epigenetic flexibility.

I also want to propose a test about how the stability of epigenetic mutations is optimized. Epigenetic adaptation and epigenetic memory have been discussed and studied experimentally (CASADESÚS AND D'ARI 2002; WOLF *et al.* 2008; NORMAN *et al.* 2013), but all from systems biology perspective. Here I discuss it from the "epigenetic mutation" and evolutionary

perspective, regarding how the stability of "epigenetic mutation" can be selected for or against and how to study the genetic difference of epigenetic memory. Theoretically, the change of epigenetic modifications could be stochastic or could be due to "epigenetic mutation", there may exist some genetic and/or some epigenetic modifiers to modify the stability of "epigenetic mutation" to make it best incorporate with the frequency of environmental fluctuation. Here I propose an experiment to test this idea. Assume we have some different genotypes of the same species and grow them in a fluctuating environment that fluctuates every 6h, 12h, or 24h, for 10 days with many replicates. We then resequencing all the replicates from 10 days (for example) of growth in the fluctuating environment to find the genotypes that are identical to their ancestral genotypes, so they do not evolve genetically. After getting those genotypes, we measure the absolute fitness for both the "epigenetically evolved" genotypes and the ancestral genotypes for one or two fluctuating cycles. If some evolved genotypes are fitter than the ancestral genotypes, this experiment proves that the "epigenetic evolution" allows the genotype to adapt to new environments. Moreover, given "epigenetic adaptation" exists, and if different "epigenetically evolved" genotypes have a different amount of fitness improvements, we continue to evolve after 10 days of fluctuation and test if those genotypes with better "epigenetic adaptability" are better evolved later due to adaptation at DNA level. We could also conduct DNA methylation sequencing to identify the changes in DNA methylation level and estimate the effect sizes of epigenetic mutations. Moreover, different genotypes may adapt to different fluctuation frequency at different rates, some may have their preferred frequency of change; this could be tested by having multiple pairs of fluctuation environments. For genotypes that experience a frequency fluctuating environment for a long time, they may have less stable epigenetic mutations, and for genotypes that experience low frequency of fluctuating environments, they may have more stable

epigenetic mutations. The genotypes that adapt to a constant environment for a long time will have the most stable epigenetic mutations. This tuning of epigenetic mutations' stability needs to be distinguished from epigenetic plasticity due to switching environments, as the former stay the same once changed, and the latter can mutate back and forth even in a constant environment. I predict that how good epigenetic memory depends on the close history of the genotype, and adaptation could be facilitated by epigenetic memory or prohibited by it depending on the situation. The frequency of "environment fluctuation" may be an important factor for the stability of epigenetic mutations.

7.6.2 "Genetic memory" in adaptation

In the previous section, I discussed why epigenetic memory can and should exist. Here I propose that evolutionary memory can also exist at genetic level for a longer time scale. To my current knowledge, this model is perhaps new. Here follows the model explaining why genetic memory for fitness exists and why every species can remember not only the previously adapted environment but also many previously adapted environments.

If the ancestor (genotype G_0) of a species first experience and adapt to environment A (genotype G_{0A}), then switch to and adapt to environment B (genotype G_{0AB}), we may expect that this genotype G_{0AB} could perform better than the ancestor G_0 in environment A. This is under the assumption and my observation that antagonistic G×E is less common than concordant G×E and antagonistic G×E is less common than environment specific genetic effect (WEI AND ZHANG 2017). Let's derive this using logic. During adaptation to environment A, some beneficial mutations and some neutral mutations are fixed, which affects *x* percent (*x* is small) of the genome. After that, the genotype G_{0A} adapt to environment B, and *y* percent (*y* is also small) of

the genome acquire beneficial mutations and neutral mutations to environment B and become G_{0AB}. Because, the genome is large, and the fraction of beneficial mutations fixed is << 1%, there is close to a negligible proportion of the genome that is hit twice by mutations. Under the assumption and the observation that antagonistic G×E is rare, most of the adaptive mutations for environment A still maintained in the genome G_{0AB}, so when the species move back to A from B, G_{0AB} is fitter than the ancestor G_0. Perhaps it is as good as or even better than G_{0A} depending on what proportion of G×E there are concordant between environments A and B. This works only for incomplete adaptation to environment A. If G_0 is already at the global peak of environment A, adaptation to B will not give G_{0AB} higher fitness than G_0.

The memory I talked about here is "recent adaptive memory", primarily related to fitness, but also applies to all the phenotypes positively associated with fitness. I predict that the memory for "fitness" is the strongest, stronger than the memory for fitness associated phenotypes. This is because many organismal level phenotypes are costly the stability of a phenotype could be easily changed by some small change in developmental pathway. Moreover, the loss of a phenotype is irreversible. The fraction of neutral fixations may complicate the situation a little bit because the neutral mutation in environment B might be deleterious in environment A, so we need to also assume most neutral mutations stay neutral across environments. The effects of a neutral mutation in nature and in artificial selection can be different; therefore, the duration of evolutionary memory might be different.

Having this in mind, let us continue with the same logic to derive the evolutionary memory for sequential adaptation to multiple environments. When a species sequentially adapt to environment A, B, C, D, E, F, G, we expect the genotype $G_{0ABCDEFG}$ will perform better than the genotype G_0 in all the environments, which means due to genetic reasons and patterns of

G×E, the genotype "remembers" all previously adapted environments to some extent. Therefore, simple genetic law not only allow evolution to happen but also creates "evolutionary memory", which significantly improves the wellbeing of all natural species. Because of this kind of memory, the plastic response to a reasonably distantly experienced environment can be beneficial.

However, the duration of genetic memory and the number of remembered environments must have an upper bound. It is will "Alzheimer" when the environment experienced is too distant. The memory decay rate, I predict, positively correlates with the number of mutations fixed in environment X, and negatively correlated with the number of fixed mutations happen after X, and needs to correct by the fraction of adaptive mutation versus neutrally fixed mutations or by the functional target size of the genome. The memory decay should also be faster than linear with time. Because the observation about G×E is only made when genotypes are similar to each other, and because G×E depends on the genotype background/could be simultaneously affected by genetic interactions, the "genetic memory" should already fully decay before all the adaptive mutations are turned over. It is reasonable to assume, this "genetic memory" will persist at a different magnitude time scale and persist much longer the previously described "epigenetic memory".

The proposal of genetic memory seems to be contradictory to common knowledge that it is difficult to improve two phenotypes at the same time by artificial selection. This may have something to do with tradeoffs, the design of the experiment, as well as inbreeding depression. However, according to the memory model, if the strategy is to improve one phenotype and fix the beneficial mutation, then improve the other phenotype, it may not be as impossible as

improving two traits at the same time. The success may require reasonable effective population size in the experiment.

7.7 Future missions

In sections of this chapter, I discussed some opinions, preliminary results, questions, and possible solutions. I will probably tackle some of those questions and verify some of the approaches in the next three to five years.

When applying for graduate school, I mentioned some small-scale questions that I wanted to work on during my Ph.D. Some of those I have indeed worked on or touched briefly as planned, which are the evolution of evolvability, the relationship between new functions and regulation networks, mechanisms of recombination, pleiotropy, and speciation. Some questions (i.e. mechanisms of mutation rate and evolution of genome size) that were in my mind before are no longer as interesting to me because I think the current understandings/theories of them are quite plausible. During my Ph.D. study, I found I am also interested in working on some classical genetic questions and collecting results from recently available data to improve the understanding of old questions. I still think this is worth to revisit some classical questions, so I will continue to practice this in the future.

However, if it is about the future in ten to twenty years, I want to talk about some bigger questions. There are one bigger technical question and two bigger evolutionary questions that I want to answer. The technical question is how to understand and study high-dimensional phenotypes (or phenome). The two bigger evolutionary questions that puzzled me and intrigued me most have been around for a while. The first one is how to connect macroevolution with microevolution, and the second one is how to explain evolution by combining the evolution at

the genetic level with the evolution at the epigenetic level. These are my current ultimate research goal. After five years of graduate training, I think I have a clearer idea about how to approach these two questions, but clearer only in the relative sense.

The first question, about high dimensional phenotypes, belongs to a fast-growing and intensively studied subject. There are many attempts in developing more complicated regression models; deep learning has also been applied to questions in this subject (ANGERMUELLER *et al.* 2016). I intend to join the force soon. Although I am not clear about the first step, I do believe that after explainable artificial intelligence (XAI, whose action could be understood by humans) is developed, the high dimensional phenotype data will be better understood. I am interested in using deep learning in my future study, bringing advanced method to basic biological questions. Although I may not personally develop any XAI method, I will watch out for the opportunity to apply it in my future study.

To answer the first evolutionary question, I think one key part is to study genotype-phenotype mapping. However, it is probably insufficient by simply studying GPM using segregating polymorphisms; the likelihood of rare events (rare mutations and large-scale mutations) may be at least as important as small-scale mutations for macroevolution. Therefore, I will work on the cause and consequence of large-scale mutations. Speciation may be important to study, and different modes of speciation may be involved differently for different clades. The environment may have a big impact on adaptation and speciation. Therefore, we also need a deeper understanding of biotic and abiotic environments. Biologists have studied genome for quite a long time, so are transcriptome and metabolome, and less so before but more now, proteome and phenome. Perhaps, soon enough we will need a new term "environmentome", because the "omics" is also a character of the environment, and I believe studying

environmentome helps connect microevolution with macroevolution. Currently, fitness landscape is used for within a species and is measured empirically only at the gene level, but it can be applied across species by connecting fitness landscapes in all environments and creating a "hyperspace" with the many dimensions from genotypes and the many dimensions from the environment. Different species differ in their genotypic distance and environmental distance, so they locate in local landscapes of different ruggedness. The speciation rate should depend on the shape of this hyperspace, the rate of environmental changes, mutation rate, and population size.

To answer the second evolutionary question, I think first we need to understand epigenetic evolution more. In the discussion section about epigenetic evolution, I talked about some of my opinions of how to define "epigenetic mutation" independent of "epigenetic plasticity" due to the environment change and epigenetic change due to genetic mutations. I also discussed how to measure epigenetic mutation rate, why it should facilitate genetic evolution, and how "epigenetic evolution" facilitates adaptation to a new environment at a smaller timescale and "genetic evolution" facilitates adaptation to new environments at a larger timescale. In addition, I discussed the decay of "epigenetic memory" versus "genetic memory". All of these discussions are hypothetical, which I plan to work on computationally and empirically in the future. Eventually, I hope to achieve a model that combines "epigenetic mutation" and "genetic mutation" in predicting the rate of adaptation. I also believe that "genetic-by epigenetic interaction", "epigenetic-by-epigenetic interaction" (or, "epi-epistasis"), and "epigenetic-by-environment interaction" all exist in nature, and are perhaps very prevalent. Here, these effects of interactions are not on the existence of epigenetic marker, but on the phenotypic value. For example, "genetic-by epigenetic interaction" requires at least four phenotypic measurements in a haploid genome, the phenotype with epigenetic modification at locus X and genetic allele A at

locus Y, the phenotype with epigenetic modification at locus X and genetic allele a at locus Y, the phenotype without epigenetic modification at locus X and genetic allele A at locus Y, and the phenotype without epigenetic modification at locus X and genetic allele a at locus Y. Moreover, the interaction effect is defined the same way as genetic interactions. I make this clarification here because a recent review paper used the phrase "genetic epigenetic interaction" to describe mapping of methylation quantitative trait loci (mQTLs) and haplotype-dependent allele-specific DNA methylation (DO *et al.* 2017). Although I believe it is important to study how epigenetics depends on genetics, the meaning of the phrase is different from my aforementioned "genetic-by epigenetic interactions". It is perhaps possible to study these interactions with the approaches developed for studying genetic interactions and G×E. For example, an epigenetic mapping for phenotype could be conducted, and association analysis could be used. I predict that epigenetic evolution is an important part of evolution, and the field may progress swiftly in the next ten years. I would like to work on this area in the future since most of the tools for studying epigenetics are available now.

7.8 References

Angermueller, C., T. Pärnamaa, L. Parts and O. Stegle, 2016 Deep learning for computational biology. Molecular systems biology 12**:** 878.

Bloom, J. S., I. M. Ehrenreich, W. T. Loo, T.-L. V. Lite and L. Kruglyak, 2013 Finding the sources of missing heritability in a yeast cross. Nature 494**:** 234.

Bloom, J. S., I. Kotenko, M. J. Sadhu, S. Treusch, F. W. Albert *et al.*, 2015 Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nature communications 6**:** 8712.

Börsch-Haubold, A. G., I. Montero, K. Konrad and B. Haubold, 2014 Genome-wide quantitative analysis of histone H3 lysine 4 trimethylation in wild house mouse liver: environmental change causes epigenetic plasticity. PloS one 9**:** e97568.

Casadesús, J., and R. D'Ari, 2002 Memory in bacteria and phage. Bioessays 24**:** 512-518.

Chen, P., D. Wang, H. Chen, Z. Zhou and X. He, 2016 The nonessentiality of essential genes in yeast provides therapeutic insights into a human disease. Genome research 26**:** 1355-1362.

Chou, H.-H., J. Berthet and C. J. Marx, 2009 Fast growth increases the selective advantage of a mutation arising recurrently during evolution under metal limitation. PLoS Genet 5**:** e1000652.

Chuffart, F., M. Richard, D. Jost, C. Burny, H. Duplus-Bottin *et al.*, 2016 Exploiting single-cell quantitative data to map genetic variants having probabilistic effects. PLoS genetics 12**:** e1006213.

Costanzo, M., B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons *et al.*, 2016 A global genetic interaction network maps a wiring diagram of cellular function. Science 353**:** aaf1420.

Coyne, J. A., S. Aulard and A. Berry, 1991 Lack of underdominance in a naturally occurring pericentric inversion in Drosophila melanogaster and its implications for chromosome evolution. Genetics 129**:** 791-802.

Do, C., A. Shearer, M. Suzuki, M. B. Terry, J. Gelernter *et al.*, 2017 Genetic–epigenetic interactions in cis: a major focus in the post-GWAS era. Genome biology 18**:** 120.

Heard, E., and R. A. Martienssen, 2014 Transgenerational epigenetic inheritance: myths and mechanisms. Cell 157**:** 95-109.

Holliday, R., 1987 The inheritance of epigenetic defects. Science 238**:** 163-170.

Kronholm, I., and S. Collins, 2016 Epigenetic mutations can both help and hinder adaptive evolution. Molecular ecology 25**:** 1856-1868.

Kryazhimskiy, S., D. P. Rice, E. R. Jerison and M. M. Desai, 2014 Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science 344**:** 1519-1522.

Norman, T. M., N. D. Lord, J. Paulsson and R. Losick, 2013 Memory and modularity in cell-fate decision making. Nature 503**:** 481.

Plech, M., J. A. G. de Visser and R. Korona, 2014 Heterosis is prevalent among domesticated but not wild strains of Saccharomyces cerevisiae. G3: Genes, Genomes, Genetics 4**:** 315-323.

Taichman, D. B., J. Backus, C. Baethge, H. Bauchner, P. W. De Leeuw *et al.*, 2016 Sharing clinical trial data—a proposal from the International Committee of Medical Journal Editors, pp. Mass Medical Soc.

Wei, X., and J. Zhang, 2017 The genomic architecture of interactions between natural genetic polymorphisms and environments in yeast growth. Genetics 205**:** 925-937.

Wolf, D. M., L. Fontaine-Bodin, I. Bischofs, G. Price, J. Keasling *et al.*, 2008 Memory in microbes: quantifying history-dependent behavior in a bacterium. PLOS one 3**:** e1700.

Appendix A:

A simple method for estimating the strength of natural selection on overlapping genes

"Two pairs of genes are coded by the same region of DNA using different reading frames."

— Frederick Sanger

A.1 Abstract

Overlapping genes, where one DNA sequence codes for two proteins with different reading frames, are not uncommon in viruses and cellular organisms. Estimating the direction and strength of natural selection acting on overlapping genes is important for understanding their functionality, origin, evolution, maintenance, and potential interaction. However, the standard methods for estimating synonymous (d_S) and nonsynonymous (d_N) nucleotide substitution rates are inapplicable here because a nucleotide change can be simultaneously synonymous and nonsynonymous when both reading frames involved are considered. We have developed a simple method that can estimate d_N/d_S and test for the action of natural selection in each relevant reading frame of the overlapping genes. Our method is an extension of the modified Nei-

Gojobori method previously developed for non-overlapping genes. We confirmed the reliability of our method using extensive computer simulation. Applying this method, we studied the longest human sense-antisense overlapping gene pair, *LRRC8E* and *ENSG00000214248*. While *LRRC8E* (leucine rich repeat containing 8 family, member E) is known to regulate cell size, the function of *ENSG00000214248* is unknown. Our analysis revealed purifying selection on *ENSG00000214248* and suggested that it originated in the common ancestor of bony vertebrates.

A.2 Introduction

Overlapping genes generally refer to pairs of genes that overlap in their transcribed sequences. In this study, however, overlapping genes refer to pairs of genes that overlap in their protein coding regions but use different reading frames. The first overlapping genes were discovered nearly 40 years ago in bacteriophage ϕX174 (Barrell et al. 1976). Overlapping genes have since been found in numerous viruses and cellular organisms including multicellulars such as humans, and their functional importance has been demonstrated in some case studies (Giorgi et al. 1983; Normark et al. 1983; Chen et al. 1993; Veeramachaneni et al. 2004; Pavesi 2006; Chung et al. 2008; Dornenburg et al. 2010). In theory, two genes may overlap in one of five possible phases (**Fig A-1**), two being sense-sense (ss) and three being sense-antisense (sas). The sas11 phase, in which the second codon position in one gene faces the third codon position in the other gene (**Fig A-1**), was reported to be the most common type (in prokaryotes), likely because this phase minimizes the mutual constraints of the protein sequences of the overlapping genes (Rogozin et al. 2002).

To study the functionality, origin, maintenance, and evolution of overlapping genes, it is often necessary to infer the direction and strength of natural selection acting on them. The

standard approach for studying natural selection acting on protein-coding genes is by estimating the ratio between the rate of nonsynonymous nucleotide substitution (d_N) and that of synonymous nucleotide substitution (d_S). However, because a mutation may be simultaneously synonymous and nonsynonymous in overlapping genes, the commonly used methods for estimating d_S, d_N, and d_N/d_S are inapplicable. Several attempts have been made to estimate selection strengths in overlapping genes. Some authors treated a pair of overlapping genes as two non-overlapping genes and calculated d_N/d_S for each gene independently using the standard methods (Yu et al. 2005; Pavesi 2006; Simon-Loriere et al. 2013). As pointed out long ago (Miyata and Yasunaga 1978), this approach is problematic, because a synonymous mutation to one of the overlapping genes may be nonsynonymous to the other gene and thus may be non-neutral. Realizing that the neutral expectation of d_N/d_S for each overlapping gene may not be 1, Nekrutenko et al. simply calculated d_N and d_S rather than their ratio, but they still applied a standard method directly to each overlapping gene (Nekrutenko et al. 2005). As such, the biological meanings of the estimated d_S and d_N are unclear. Rogozin et al. also noted the impact of one mutation on two genes and hence considered only sites that are fourfold degenerate for one of the overlapping genes. Specifically, they were able to estimate d_N for each gene in gene pairs with the sas11 phase (Rogozin et al. 2002). But this method does not apply to all overlapping genes, and estimating d_S remains difficult (e.g., Rogozin et al. estimated d_S from non-overlapping regions). Extending Goldman and Yang's method for non-overlapping coding sequences (Goldman and Yang 1994), Sabath et al. developed a maximum likelihood (ML) method for simultaneous estimation of the selection intensity in each of two overlapping genes (Sabath et al. 2008). However, as currently implemented, the method cannot test whether d_N/d_S

significantly differs from 1 for either gene (Sabath et al. 2008; Sabath et al. 2012), rendering the utility of the method limited.

Here we describe a simple method that estimates the selection strength of each of the two overlapping genes by separating the effects of each mutation on the two genes. Our method also estimates the associated variance, allowing a test of neutrality for each gene. We evaluate the performance of our method using computer simulation, and illustrate its utility by analyzing the human sense-antisense gene pair with the longest overlapping region.

A.3 Materials and methods

Computer simulation

Our new method for estimating the selection strengths in overlapping genes is described in Results. Here we describe the simulation used to evaluate the performance of our method. To generate a pair of overlapping genes, we set the following parameters: the overlapping phase, the length of the overlapping region *l*, the ratio (*R*) between the number of transitions and number of transversions, the distance (*d*) between two sequences defined by the expected number of substitutions per neutral site, selection strength on ORF1 (ω_1), and selection strength on ORF2 (ω_2). We generated an ancestral sequence that contained overlapping ORFs by first randomly choosing sense codons for the first ORF and then removing all stop codons until no stop codon is found in each ORF. We then introduced mutations following Kimura's two-parameter model (Kimura 1980) with a preset *R*. The fixation probability of a mutation is determined jointly by ω_1 and ω_2. Specifically, if the mutation is synonymous in both ORFs, its fixation probability is

set to be a ($0< a <1$); if the mutation is synonymous to ORF1 but nonsynonymous to ORF2, its fixation probability is $a\omega_2$; if the mutation is synonymous to ORF2 but nonsynonymous to ORF1, its fixation probability is $a\omega_1$; if the mutation is nonsynonymous to both ORFs, its fixation probability is $a\omega_1\omega_2$. The parameter a must be small enough so that $a\omega_1$, $a\omega_2$, and $a\omega_1\omega_2$ are all smaller than 1. Under this scheme, both positive and negative selection can be simulated. When negative selection is simulated for both ORFs, a can take any value between 0 and 1, but we assigned 0.9 to a to decrease the computational time. When positive selection is simulated for ORF1 but negative selection is simulated for ORF2, $0.9/\omega_1$ was assigned to a. If both ORFs are under positive selection, $0.9/(\omega_1\omega_2)$ was assigned to a. Each ancestral sequence was evolved independently to produce two derived sequences, by either accepting or rejecting the randomly generated mutations. Simulation ended when the number of mutations introduced equals the preset number (dl/a). ω_1 and ω_2 were then estimated by comparing the two simulated derived sequences. The scripts used for simulating overlapping genes and for estimating ω were written with Perl and are available at http://www.umich.edu/~zhanglab/download.htm.

Case study

Annotation for human protein coding genes and sequences used in the selection analysis were downloaded from Ensembl GRCh37 (http://useast.ensembl.org/). Overlapping genes were identified by comparing exon start and end positions of each gene on the same chromosome. For example, if exon 2 of gene A starts at position 13,780 and ends at 13,942 on Chromosome 1, and exon 5 of gene B starts at 13,950 and ends at 13,820 on the same chromosome, we can infer that these two genes form a pair of sense-antisense overlapping genes and that the overlapping region

between the two exons has (13942-13820+1) = 123 bp. The overlapping genes analyzed were identified from Ensembl annotations using a Python script. Sequences were aligned using an online version of clustralw2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/). Transition/transversion ratio was calculated using MEGA5 (Tamura et al. 2011). The protein expression levels were from ProteomicsDB at https://www.proteomicsdb.org/ (Wilhelm et al. 2014). The GenBank accession numbers of LRRC8 genes analyzed are provided in Tables A-S1 and S2. We used MEGA5 to reconstruct the neighboring-joining tree of LRRC8 genes using protein *p*-distances.

A.4 Results

A new method for estimating the selection strength in overlapping genes

Because most species use double-stranded DNA, one segment of DNA can harbor at most six different open reading frames (ORFs). However, very rarely do all six ORFs coexist. Even in cases where all six ORFs coexist, it is unclear whether all ORFs code for actual proteins (Menon et al. 1990). The simplest and most common overlapping coding regions harbor two different ORFs, which can be either on the same strand (sense-sense overlap) or on opposite strands (sense-antisense overlap) (**Fig A-1**). The two types of sense-sense overlap are in fact equivalent, because they both have the third codon positions of one ORF facing the first codon positions of the other ORF (**Fig A-1**). Here we use the sense-sense overlap as an example to describe our method, but the same applies to all overlaps between two ORFs.

Our method is an extension of the modified Nei-Gojobori (mNG) method for estimating d_S and d_N in non-overlapping genes (Nei and Gojobori 1986; Zhang et al. 1998), but considers

the complication that one mutation simultaneously affects two ORFs, often with different effects. Let us consider a pair of homologous DNA sequences (e.g., respectively from human and mouse) that harbor overlapping ORF1 and ORF2. Our method for quantifying the selection strength in ORF1 and that in ORF2 involves the following four steps.

In the first step, we classify human nucleotide sites in the overlapping region into four categories depending on the impacts of potential mutations on the two ORFs. The four categories are referred to as NN, NS, SN, and SS sites, respectively, where N stands for nonsynonymous and S stands for synonymous. That is, if all potential mutations at a site cause nonsynonymous changes in both ORFs, it is an NN site, and so on. A site may belong to multiple categories and be called, for example, 1/3 NN site and 2/3 NS site, if one third of potential mutations at the site cause nonsynonymous changes in both ORFs and two thirds of potential mutations at the site cause nonsynonymous changes in ORF1 but synonymous changes in ORF2. When considering potential mutations, it is important to separate transitions from transversions because they typically have different mutation rates and have different probabilities of causing nonsynonymous changes (Zhang 2000). Let R be the ratio between the number of transitional mutations and that of transversional mutations and be estimated from external information (e.g., from non-overlapping regions or other genes). Hence, we consider a fraction of $R/(1+R)$ mutations to be transitions and the rest transversions (Zhang et al. 1998) in determining to which of the above four categories a site belongs. For instance, if the transitional mutation at a site causes a synonymous change in both ORFs and the two transversional mutations both cause a synonymous mutation in ORF1 and a nonsynonymous mutation in ORF2, this site is counted as $R/(R+1)$ SS site and $1/(R+1)$ SN site. We then calculate the total number of sites in the human overlapping region belonging to each of the four categories. The

corresponding values are also calculated for the mouse sequence, and the averaged value from the two sequences for each category (L_{NN}, L_{NS}, L_{SN}, and L_{SS}) will be used subsequently.

In the second step, we classify all nucleotide differences between the two sequences into four categories: NN, NS, SN, and SS. That is, if a difference is nonsynonymous in both ORF1 and ORF2, it belongs to the NN group, and so on. When a nucleotide difference is in isolation, meaning that in neither ORF is there another difference in the same codon as the focal difference, the classification is straightforward. But when a codon (in either ORF) harbors two or more differences, the situation becomes complicated, because to determine the categories of the multiple differences, one has to consider all possible evolutionary pathways that can give rise to the observed nucleotide differences. In the case of non-overlapping ORFs, there are two equally shortest evolutionary pathways between a pair of codon sequences with two differences (e.g., to evolve from AAA to AGG, one can go through AAG or AGA) and six equally shortest pathways when it harbors three differences (Nei and Gojobori 1986). For overlapping ORFs, however, one may need to consider a lot more pathways, because a codon in ORF1 overlaps with a codon in ORF2, which overlaps with another codon in ORF1, and so on. Thus, we need to find a segment of DNA in which each codon (defined by both ORFs) has multiple nucleotide differences with the exception of the codon at each end of the segment (**Fig A-2**). When this segment has a total of m nucleotide differences between the pair of homologous sequences, a total of $m!$ pathways should be considered, each of which contains a unique order of m nucleotide changes. For each pathway, we count the number of nucleotide changes belonging to each of the four categories (NN, NS, SN, and SS). We average these numbers across all open pathways, which are pathways with no intermediate sequences that contain stop codons. An example is provided in

Fig A-S1. After classifying all nucleotide differences between the pair of homologous sequences into the four categories, we count their numbers (M_{NN}, M_{NS}, M_{SN}, and M_{SS}, respectively).

In the third step, we calculate the proportion of sites with nucleotide differences by $p_{NN} = M_{NN}/L_{NN}$, $p_{NS} = M_{NS}/L_{NS}$, $p_{SN} = M_{SN}/L_{SN}$, and $p_{SS} = M_{SS}/L_{SS}$ for NN, NS, SN, and SS sites, respectively. The Jukes-Cantor formula (Jukes and Cantor 1969) may be used to correct for multiple hits. For instance, the number of nucleotide substitutions per site at NN sites can be estimated by $d_{NN} = -\frac{3}{4}\ln(1 - \frac{4p_{NN}}{3})$; d_{NS}, d_{SN}, and d_{SS} can be similarly estimated. Here we used the Jukes-Cantor correction instead of more complex corrections such as Kimura's two-parameter model (Kimura 1980) or Tamura-Nei model (Tamura and Nei 1993), because overlapping regions are usually so short that the variance of a distance estimate would be large under complex corrections (Nei and Kumar 2000).

In the fourth step, we propose that the strength of natural selection acting on ORF1 be estimated by $\omega_1 = d_{NN}/d_{SN}$ and that acting on ORF2 be estimated by $\omega_2 = d_{NN}/d_{NS}$. This formulation is based on two assumptions. First, synonymous mutations are neutral. Although not all synonymous mutations are neutral due to their potential impacts on DNA-protein interaction, pre-mRNA splicing, mRNA folding, translational efficiency, translational accuracy, and other aspects of cell biology (Chamary and Hurst 2005; Pagani et al. 2005; Warnecke and Hurst 2007; Qian et al. 2012; Park et al. 2013; Yang et al. 2014), most synonymous mutations may be considered largely neutral when compared with nonsynonymous mutations, especially in species with small effective population sizes (Li 1987; Ohta 1992). Second, the two overlapping genes do not have genetic interaction, such that the probability that a mutation gets fixed is the product of the probability with which it gets fixed in the absence of ORF1 and the probability

with which it gets fixed in the absence of ORF2. This assumption implies that (i) NN-type mutations and SN-type mutations have comparable average effects on ORF2 and (ii) NN-type mutations and NS-type mutations have comparable average effects on ORF1. Hence, ω_1 can be estimated by d_{NN}/d_{SN} and ω_2 can be estimated by d_{NN}/d_{NS}. In theory, we could also estimate ω_1 by d_{NS}/d_{SS} and estimate ω_2 by d_{SN}/d_{SS}. But, such estimates are usually subject to large sampling errors, because with the exception of the sas12 overlap that has a sizeable fraction of SS sites (**Fig A-1**), overlapping regions typically have few SS sites. Thus, unless otherwise noted, we do not use d_{SS} in this study. It is sometimes of interest to compare the selective pressures acting on the two overlapping genes. For this purpose, we can compute ω_1/ω_2, which equals d_{NS}/d_{SN}.

To calculate the variances of d_{NN}, d_{NS}, d_{SN}, and d_{SS}, the commonly used bootstrap method (Nei and Kumar 2000) is inapplicable because of the difficulty in bootstrapping codons from one ORF while maintaining the other ORF. We therefore extend an approximate analytical method previously developed for estimating the variances of d_S and d_N in the Nei-Gojobori method (Nei 1987), which is known to be quite accurate (Ota and Nei 1994). Following this method, we calculate the variance of d_{NN} by $\mathrm{Var}(d_{NN}) = \mathrm{Var}(p_{NN})/(1-4p_{NN}/3)^2$, where the variance of p_{NN} is given by $\mathrm{Var}(p_{NN}) = p_{NN}(1-p_{NN})/L_{NN}$. Variances of d_{NS}, d_{SN}, and d_{SS} can be similarly estimated. Standard deviations (SDs) of d_{NN}, d_{NS}, d_{NS}, and d_{SS} are then estimated by taking the square root of their variances, respectively. The hypothesis of neutral evolution of ORF1 can be tested by a Z-test of the equality between d_{NN} and d_{SN}. That is, we can conduct a Z-test using $Z = (d_{NN} - d_{SN})/(\mathrm{Var}(d_{NN})+\mathrm{Var}(d_{SN}))^{1/2}$. Similarly, the neutral evolution hypothesis for ORF2 can be tested by a Z-test of the equality between d_{NN} and d_{NS}. We can also test if the strengths of natural selection acting on the two ORFs are equal by a Z-test of the equality between d_{SN} and d_{NS}.

Performance of the new method in estimating the selection strengths in overlapping genes

To examine the performance of the new method, we conducted extensive computer simulation of overlapping genes of each phase. The overlapping region had 3000 nucleotides, and the simulation was repeated 100 times under each parameter set. We used exceptionally long overlapping regions to minimize the sampling error such that potential biases of our estimators became more readily detectable. We start by describing the results obtained under the sense-sense overlap. We first examined the situation that both overlapping genes are under purifying selection. We fixed $\omega_1 = 0.2$ and $\omega_2 = 0.5$ and studied how the distance between a pair of homologous sequences affects the accuracy of estimation (**Fig A-3A**), where the distance is defined by the expected number of substitutions per neutral site between the two homologous sequences (*i.e.*, the expected value of d_{SS}). We found that the mean ω_1 estimate and the mean ω_2 estimate are both slightly greater than their true values, and this excess in the estimated ω value appears unrelated to the distance. This bias may be due to the fact that we simulated sequence evolution using Kimura's two-parameter model, but estimated d_{NN}, d_{NS}, and d_{SN} using the Jukes-Cantor correction, which is known to undercorrect multiple hits in this scenario. When ω_1 and ω_2 are lower than 1, d_{SN} and d_{NS} are greater than d_{NN}, making the undercorrection more severe for the former than the latter and the resultant ω_1 and ω_2 upward biased. Nevertheless, the biases appear to be generally lower than 10%. By contrast, if we estimate ω_1 and ω_2 by the mNG method without considering the mutual influences between overlapping genes, the estimates are much higher than their respective true values (**Fig A-3A**). This is because some synonymous mutations to one ORF are nonsynonymous to the other ORF and hence have been removed by

purifying selection, causing overestimation of ω_1 and ω_2. Because the true $\omega_1 < \omega_2 < 1$, ω_2 is overestimated to a larger extent than ω_1 (**Fig A-3A**).

Next, we examined the situation that one overlapping gene is under positive selection ($\omega_1 = 3$) while the other is under purifying selection ($\omega_2 = 0.2$). We again found the mean estimates of ω_1 and ω_2 by our method to be close to their respective true values, for all levels of distance considered (**Fig A-3B**). When the mNG method is used, ω_2 is slightly underestimated (**Fig A-3B**), likely because some synonymous mutations to ORF2 are beneficial to ORF1 and are fixed by positive selection. By contrast, ω_1 is grossly overestimated by mNG (**Fig A-3B**), for the reason mentioned in the previous paragraph.

We next examined the impact of the transition/transversion ratio *R* on estimates of ω_1 and ω_2 when their true values are 0.2 and 1, respectively (**Fig A-3C**). We found both ω_1 and ω_2 slightly overestimated. This becomes moderately severe for ω_2 when $R \geq 10$, probably due to the aforementioned undercorrection of multiple hits by the Jukes-Cantor formula that is more serious when *R* gets higher. The mNG method performs similarly well as the new method in estimating ω_1 (**Fig A-3C**), likely because of the lack of any selection on ORF2. But ω_2 is grossly overestimated by mNG (**Fig A-3C**). Because ORF2 itself is not under any selection, the above phenomenon must be due to the fact that synonymous mutations to ORF2 are more likely than nonsynonymous mutations to ORF2 to be deleterious to ORF1.

We next varied ω_1 from 0.2 to 3.0 while keeping ω_2 at 0.2. We found estimates of ω_1 and ω_2 by our method to be generally reliable (**Fig A-3D**). By contrast, ω_1 is consistently and

grossly overestimated by mNG, whereas ω_2 is overestimated when $\omega_1 < 1$ and underestimated when $\omega_1 > 1$, as expected (**Fig A-3D**).

In addition to the sense-sense overlap, we also examined the three sense-antisense overlapping phases with different parameter sets. We found that our method generated reliable results under all phases (**Fig A-4**). By contrast, the mNG method can make grossly wrong estimates, and the direction and extent of the error depends on ω_1, ω_2, and the specific overlapping phase (**Fig A-4**). For phase sas12, third codon positions in ORF1 overlap with third codon positions in ORF2. Consequently, the fraction of SS sites is higher than that in other phases, allowing the possibility of estimating natural selection using SS sites. We thus also estimated ω_1 by d_{NS}/d_{SS} and estimated ω_2 by d_{SN}/d_{SS} for phase sas12 (see sas12* in **Fig A-4**). The results showed that these estimates are either similar to or slightly better than those using NN sites (see sas12 in **Fig A-4**).

Because the analytical formulas for standard deviations are approximate, we used computer simulation to investigate their accuracies. For the sense-sense phase, we examined the reliabilities of the analytically computed SD(d_{NN}), SD(d_{NS}), and SD(d_{SN}), but could not examine SD(d_{SS}) because of the paucity of SS sites. We conducted 100 simulation replications under each set of parameters. We then compared the SD among the 100 d_{NN} values obtained and the mean of SD(d_{NN}) analytically calculated using the data from each simulation. The same was done for d_{NS} and d_{SN}. We found the analytically calculated SD values to be overall similar to the simulation observations, with statistically insignificant differences (**Fig A-5**).

Evolutionary analysis of the human gene pair with the longest sense-antisense overlapping region

To illustrate the utility of our method, we searched for an appropriate pair of overlapping genes from Ensembl for detailed analysis. We found that Ensembl annotates most sense-sense overlapping genes with different reading frames as alternative splicing (Curwen et al. 2004), greatly underestimating the prevalence of sense-sense overlapping genes. We thus focused on sense-antisense overlapping and identified the longest sense-antisense overlapping coding region in the human genome, containing 732 bases. The involved genes are *LRRC8E* (leucine rich repeat containing 8 family, member E) and an uncharacterized gene with an Ensembl Gene ID of *ENSG00000214248*. The structure of this gene pair (**Fig A-6A**) shows that the entire 243 amino acid coding region of *ENSG00000214248* lies within the second exon of *LRRC8E*, with the sas12 overlapping phase. It was recently discovered that LRRC8E functions as an essential component of the cell volume-regulated anion channel VRAC (Voss et al. 2014), but whether *ENSG00000214248* encodes a functional protein and what its function is are unknown.

We found from the recently published human proteomic data (Wilhelm et al. 2014) that *ENSG00000214248* is not only transcribed but also translated in coronary sinus and blood platelet (**Fig A-6B**). The protein expression sites of *ENSG00000214248* and those of *LRRC8E* overlap in blood platelet but are otherwise distinct (**Fig A-6B**). The expression levels of the two proteins are generally comparable (**Fig A-6B**). We acquired the sequences of the orthologous genes of human *ENSG00000214248* and *LRRC8E* from the macaque genome sequence. Using our method, we estimated the ω values for the two genes in the overlapping region as well as the ω in the non-overlapping region of *LRRC8E*. *R* was estimated to be 3.61 from the non-overlapping region of *LRRC8E* using Kimura's two-parameter model (Kimura 1980). We found

that the overlapping region and non-overlapping region of *LRRC8E* have been under similar levels of purifying selection, with ω = 0.08 and 0.09, respectively. The ω for *ENSG00000214248* is 0.20, significantly lower than the neutral expectation of 1 ($P < 0.002$, two-tail Z-test), suggesting that this uncharacterized gene has been under purifying selection at least since the divergence between human and macaque. For the overlapping region, we used SS sites in the above estimation of ω values for *ENSG00000214248*, because there was no substitution at NS sites.

Because *ENSG00000214248* is entirely within *LRRC8E*, we traced the origin of *ENSG00000214248* by examining its presence in *LRRC8E* of various species. We were able to identify *LRRC8E* in all bony vertebrate genome sequences available at Ensembl and NCBI, but not in shark, lamprey, or any invertebrate genome. Interestingly, we also identified the ORF of *ENSG00000214248* within *LRRC8E* in most bony vertebrates, including zebrafish (**Fig A-6C**). Apparently, *ENSG00000214248* already existed in the common ancestor of bony vertebrates, but was pseudogenized several times in subsequent evolution (**Fig A-6C**). Because *LRRC8E* is a member of the *LRRC8* family that contains five genes in human, we reconstructed the phylogeny of this gene family (**Fig A-S2**) to investigate if *ENSG00000214248* originated before *LRRC8E*. We discovered that the closest relative to *LRRC8E* is *LRRC8C*, which can be found in bony vertebrates and shark. However, the presumable *ENSG00000214248* reading frame in *LRRC8C* contains several premature stop codons in each species examined (human, macaque, mouse, rat, zebrafish, and shark), suggesting that the common ancestor of *LRRC8C* and *LRRC8E* did not contain *ENSG00000214248*. Thus, the antisense reading frame probably originated in *LRRC8E* shortly after the birth of *LRRC8E* from the duplication of *LRRC8C*.

A.5 Discussion

Overlapping genes have been identified in many species and are particularly common in bacteria and viruses (Normark et al. 1983; Veeramachaneni et al. 2004), but their evolutionary studies have been hampered by the inapplicability of the standard methods for inferring natural selection acting on overlapping genes. We developed a simple method to estimate the selection strength on each of the overlapping ORFs and demonstrated the reliability of our method by computer simulation. Our method allows testing whether an overlapping gene is under natural selection and hence can be used to identify functional genes from hypothetical overlapping reading frames, as was demonstrated in the example of *ENSG00000214248*.

To more readily detect potential biases of our method, we simulated long overlapping regions (3000 sites). In reality, however, overlapping regions are much shorter. We also performed simulations using overlapping regions of 750 sites and 300 sites, respectively (**Fig A-S3**), based on the parameters used in Fig A-3A and Fig A-3B. When the overlapping region is short and the distance is low, many sequences had no substitution in NS sites or SN sites, making our method inapplicable. For cases where our method did work, the mean ω estimates were reasonably good, although the standard errors were large, as expected (**Fig A-S3**). Thus, accurately estimating ω values of short overlapping regions remains challenging unless the divergence between the two taxa compared is high. Based on current annotations of eukaryotic genomes, there are not many overlapping genes that have long evolutionary histories. However, as in the example studied, although the orthologs of human *ENSG00000214248* are present in many vertebrates, they have not been annotated outside primates. It is likely that much more overlapping genes and long-lasting overlapping genes than currently annotated exist.

Overlapping genes are prevalent in viral genomes. Many viruses have high mutation rates, allowing the use of our methods even for relatively short overlapping regions.

Sabath and colleagues noted that the ML method they developed does not perform well under low distances (mean sequence divergence across sites < 8%) (Sabath et al. 2008). To examine if our method suffers from the same problem, we compared the two methods using the parameters in Fig A-3A and Fig A-3B. The results showed that the two methods are similar in their sensitivity to distance (**Fig A-S4**). However, under both negative (**Fig A-S4a**) and positive (**Fig A-S4b**) selection, our method outperforms the ML method in terms of the accuracy of the ω estimates.

While we introduced our method in the context of estimating the selective strength using interspecific comparisons, our method may also be applied to intraspecific data or comparisons between intraspecific and interspecific data. For instance, let us use D_{NN}, D_{NS}, D_{SN}, and D_{SS} to denote the numbers of the four types of substitutions in a pair of overlapping genes, respectively, and use P_{NN}, P_{NS}, P_{SN}, and P_{SS} to denote the corresponding numbers of the four types of polymorphisms, respectively. We can conduct a selection test similar to the McDonald-Kreitman test (McDonald and Kreitman 1991) for ORF1 by comparing D_{NN}, D_{SN}, P_{NN}, and P_{SN}, because D_{NN}/P_{NN} equals D_{SN}/P_{SN} under the null hypothesis of neutrality. Similarly, we can test selection in ORF2 by comparing D_{NN}, D_{NS}, P_{NN}, and P_{NS}. In addition to studying overlapping genes, our method can also be applied to the study of the functionality of certain alternative splicing. Alternative splicing is generally demonstrated by the existence of various transcripts from a gene, but the existence of a transcript is not a proof that the transcript is functional. For splice variants using alternative reading frames, our method may be used to test if the alternative

reading frame has been under natural selection, which would support the functionality of the splice variant.

In summary, we believe that our development of a simple method for estimating the selective strengths on overlapping genes will facilitate researches toward understanding the origin, evolution, and functionality of overlapping genes.

A.6 Acknowledgements

We thank Wei-Chin Ho and Jian-Rong Yang for valuable comments. This work was supported in part by U.S. National Institutes of Health research grant R01GM103232 to J.Z.

A.7 References

Barrell BG, Air GM, Hutchison CA, 3rd 1976. Overlapping genes in bacteriophage phiX174. Nature 264: 34-41.

Chamary JV, Hurst LD 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol 6: R75.

Chen HS, Kaneko S, Girones R, Anderson RW, Hornbuckle WE, Tennant BC, Cote PJ, Gerin JL, Purcell RH, Miller RH 1993. The woodchuck hepatitis virus X gene is important for establishment of virus infection in woodchucks. J Virol 67: 1218-1226.

Chung BY, Miller WA, Atkins JF, Firth AE 2008. An overlapping essential gene in the Potyviridae. Proc Natl Acad Sci U S A 105: 5897-5902.

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M 2004. The Ensembl automatic gene annotation system. Genome Res 14: 942-950.

Dornenburg JE, Devita AM, Palumbo MJ, Wade JT 2010. Widespread antisense transcription in Escherichia coli. MBio 1: e00024-00010.

Giorgi C, Blumberg BM, Kolakofsky D 1983. Sendai virus contains overlapping genes expressed from a single mRNA. Cell 35: 829-836.

Goldman N, Yang Z 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725-736.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian Protein Metabolism. New York: Academic Press. p. 21-132.

Kimura M 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111-120.
Li W-H 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. Journal of molecular evolution 24: 337-345.
McDonald JH, Kreitman M 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-654.
Menon NK, Robbins J, Peck HD, Jr., Chatelus CY, Choi ES, Przybyla AE 1990. Cloning and sequencing of a putative Escherichia coli [NiFe] hydrogenase-1 operon containing six open reading frames. J Bacteriol 172: 1969-1977.
Miyata T, Yasunaga T 1978. Evolution of overlapping genes. Nature 272: 532-535.
Nei M. 1987. Molecular Evolutionary Genetics. New York: Columbia University Press.
Nei M, Gojobori T 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418-426.
Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. New York: Oxford University Press.
Nekrutenko A, Wadhawan S, Goetting-Minesky P, Makova KD 2005. Oscillating evolution of a mammalian locus with overlapping reading frames: an XLalphas/ALEX relay. PLoS Genet 1: e18.
Normark S, Bergstrom S, Edlund T, Grundstrom T, Jaurin B, Lindberg FP, Olsson O 1983. Overlapping genes. Annu Rev Genet 17: 499-525.
Ohta T 1992. The nearly neutral theory of molecular evolution. Annual Review of Ecology and Systematics: 263-286.
Ota T, Nei M 1994. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. Mol Biol Evol 11: 613-619.
Pagani F, Raponi M, Baralle FE 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci U S A 102: 6368-6372.
Park C, Chen X, Yang JR, Zhang J 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 110: E678-686.
Pavesi A 2006. Origin and evolution of overlapping genes in the family Microviridae. J Gen Virol 87: 1013-1017.
Qian W, Yang JR, Pearson NM, Maclean C, Zhang J 2012. Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet 8: e1002603.
Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV 2002. Purifying and directional selection in overlapping prokaryotic genes. Trends Genet 18: 228-232.
Sabath N, Landan G, Graur D 2008. A method for the simultaneous estimation of selection intensities in overlapping genes. PLoS One 3: e3996.
Sabath N, Wagner A, Karlin D 2012. Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol 29: 3767-3780.
Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M 2011. Global quantification of mammalian gene expression control. Nature 473: 337-342.
Simon-Loriere E, Holmes EC, Pagan I 2013. The effect of gene overlapping on the rate of RNA virus evolution. Mol Biol Evol 30: 1916-1928.
Tamura K, Nei M 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512-526.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739.
Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I 2004. Mammalian overlapping genes: the comparative perspective. Genome Res 14: 280-286.
Voss FK, Ullrich F, Münch J, Lazarow K, Lutter D, Mah N, Andrade-Navarro MA, von Kries JP, Stauber T, Jentsch TJ 2014. Identification of LRRC8 heteromers as an essential component of the volume-regulated anion channel VRAC. Science 344: 634-638.
Warnecke T, Hurst LD 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. Mol Biol Evol 24: 2755-2762.
Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. 2014. Mass-spectrometry-based draft of the human proteome. Nature 509: 582-587.
Yang JR, Chen X, Zhang J 2014. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. PLoS Biol 12: e1001910.
Yu P, Ma D, Xu M 2005. Nested genes in the human genome. Genomics 86: 414-422.
Zhang J 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol 50: 56-68.
Zhang J, Rosenberg HF, Nei M 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A 95: 3708-3713.

SENSE-SENSE

ss12:

ORF1: 1-2-3-1-2-3-1

ORF2: 2-3-1-2-3-1-2

ss13:

ORF1: 1-2-3-1-2-3-1

ORF2: 3-1-2-3-1-2-3

SENSE-ANTISENSE

sas11:

ORF1: 1-2-3-1-2-3-1

ORF2: 1-3-2-1-3-2-1

sas12:

ORF1: 1-2-3-1-2-3-1

ORF2: 2-1-3-2-1-3-2

sas13:

ORF1: 1-2-3-1-2-3-1

ORF2: 3-2-1-3-2-1-3

Figure A-1. Five phases of overlapping genes. Sense-sense overlap is abbreviated as "ss", whereas sense-antisense overlap is abbreviated as "sas". The two sense-sense overlaps are equivalent if one switches the names of the two ORFs.

Figure A-2. Determining the shortest overlapping region for mutational pathway consideration. Shown is an example of the sense-sense overlap. Codons in ORF1 are marked with lines above the sequences, whereas codons in ORF2 are marked with lines below the sequences. Differences between the two species are in black, whereas identical nucleotides are in grey. The boxed region is the shortest region for mutational pathway consideration.

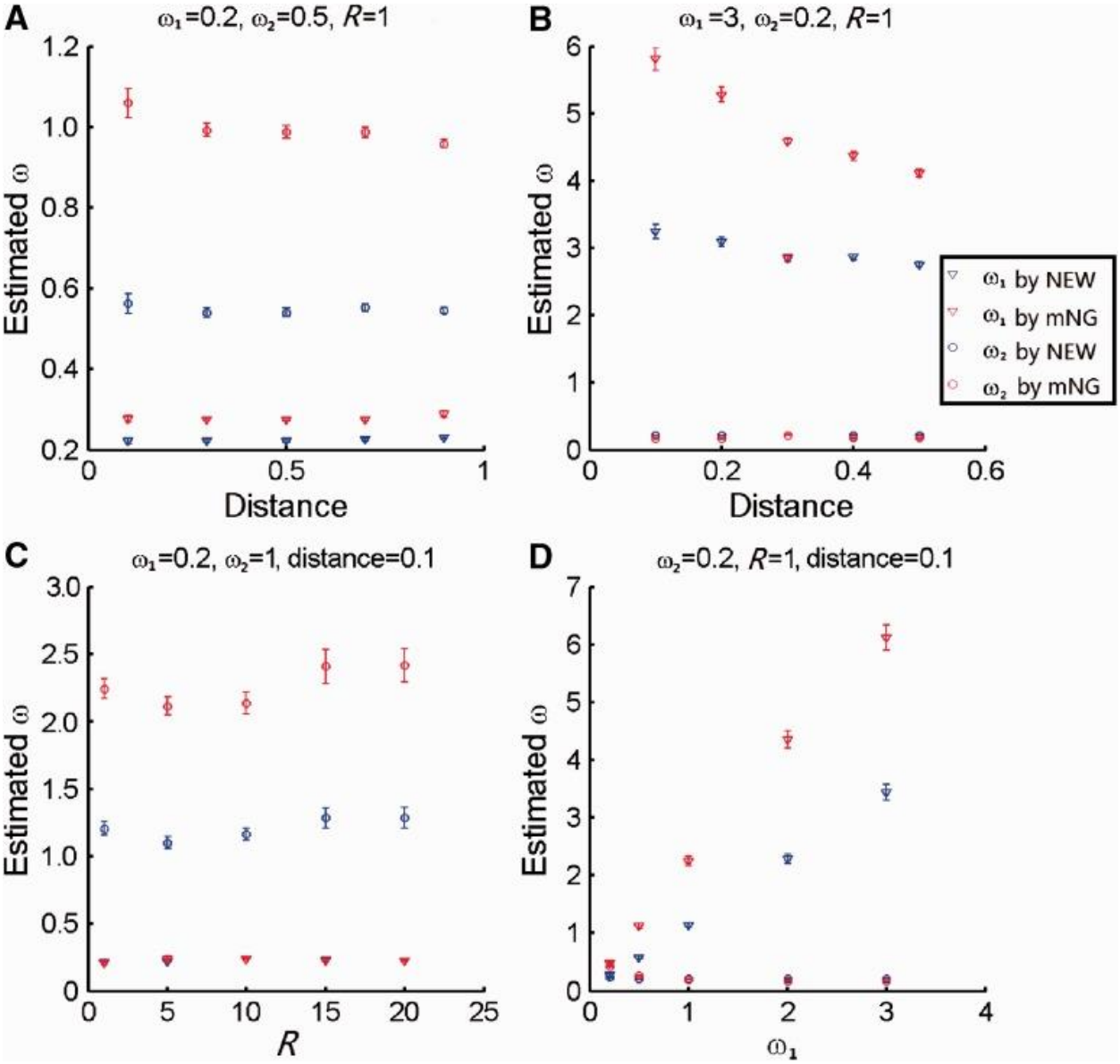


Figure A-3. Performances of the new (NEW) method and modified Nei-Gojobori (mNG) method in estimating the selection intensities (ω_1 and ω_2) on overlapping genes. Shown are results from computer simulations of overlapping genes with the sense-sense overlap. Each symbol represents the mean from 100 replications under a given parameter set, and error bars show the standard error. In each panel, the common parameters are listed above the panel, whereas the varying parameter is shown on the X-axis. Distance is defined as the expected number of nucleotide substitutions per neutral site between the two sequences under comparison.

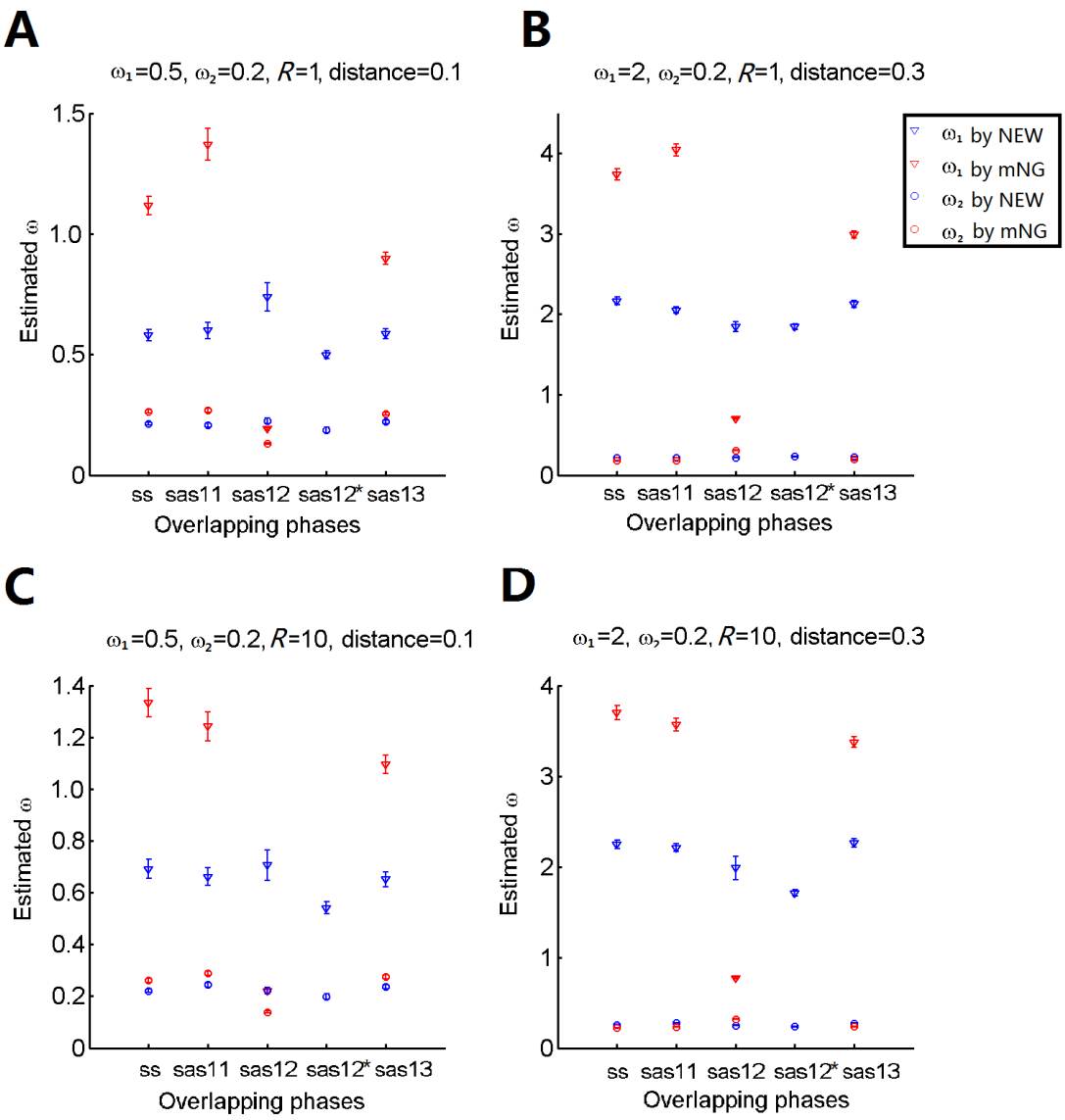


Figure A-4. Performances of the new (NEW) method and modified Nei-Gojobori (mNG) method in estimating the selection intensities (ω_1 and ω_2) on simulated overlapping genes of various phases indicated on the X-axis. Each symbol represents the mean from 100 replications under a given parameter set, and error bars show the standard error. In each panel, the parameters are listed above the panel, whereas different overlapping phases are shown on the X-axis. The results for sas12* are estimates using SS sites (i.e., $\omega_1 = d_{NS}/d_{SS}$ and $\omega_2 = d_{SN}/d_{SS}$) under the sas12 phase.

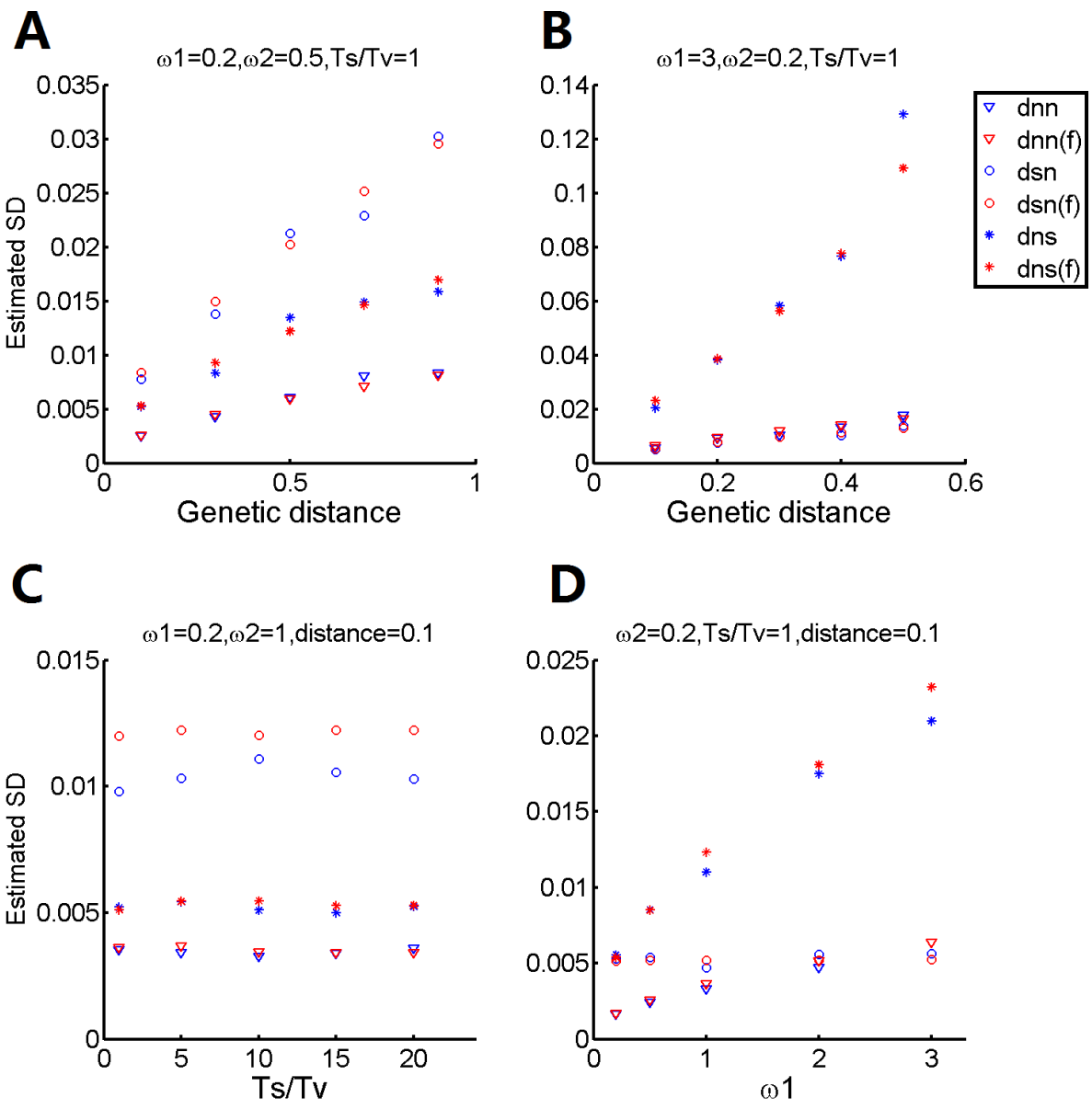
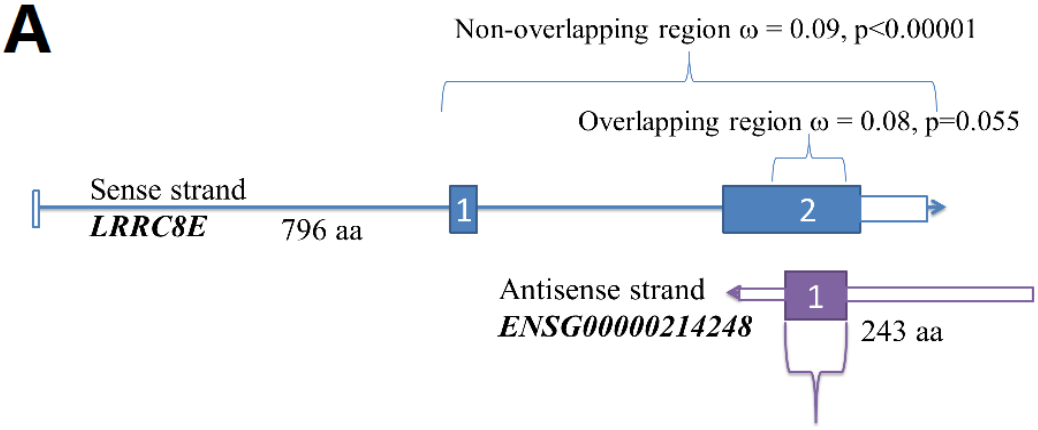
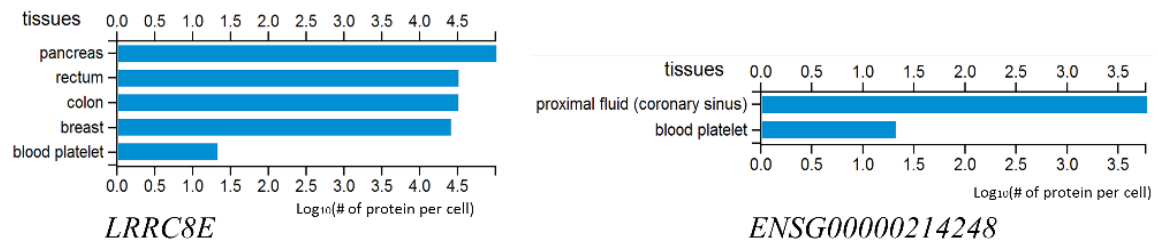


Figure A-5. Performance of the new method in estimating the standard deviation (SD) of d_{NN}, d_{NS}, and d_{SN}. Shown are the results from computer simulations of overlapping genes with the sense-sense overlap. The analytically computed SD, averaged across 100 replications, is shown by red symbols, whereas the actual SD, observed from the 100 simulation replications, is shown in blue. In each panel, the common parameters are listed above the panel, whereas the varying

parameter is shown on the X-axis. Using 400 bootstrap samples of the 100 replicates under each parameter set, we derived a frequency distribution of the observed SD. We found that the mean computed SD is within the central 95% of the frequency distribution of the observed SD under all parameter sets examined.


A
Non-overlapping region ω = 0.09, p<0.00001
Overlapping region ω = 0.08, p=0.055
Sense strand
LRRC8E
796 aa
1
2
Antisense strand
ENSG00000214248
1
243 aa
Overlapping region
ω = 0.20, p=0.0008


B
tissues
pancreas
rectum
colon
breast
blood platelet
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5
Log10(# of protein per cell)
LRRC8E
tissues
proximal fluid (coronary sinus)
blood platelet
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Log10(# of protein per cell)
ENSG00000214248

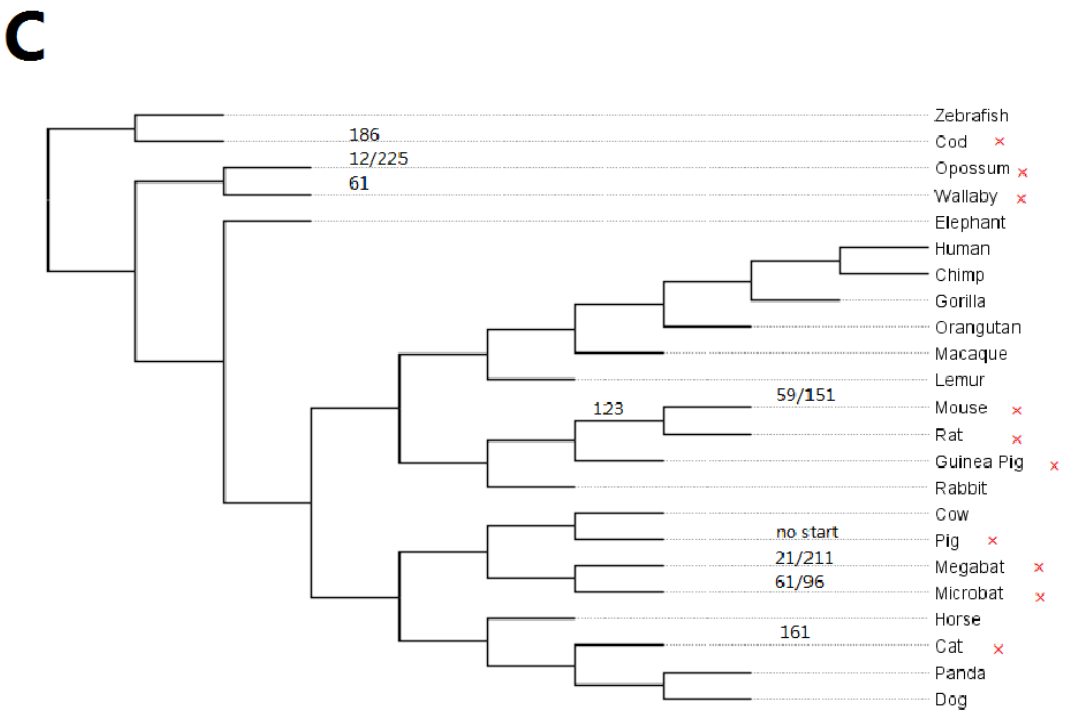

C
186
12/225
61
123
59/151
no start
21/211
61/96
161
Zebrafish
Cod
Opossum
Wallaby
Elephant
Human
Chimp
Gorilla
Orangutan
Macaque
Lemur
Mouse
Rat
Guinea Pig
Rabbit
Cow
Pig
Megabat
Microbat
Horse
Cat
Panda
Dog

Figure A-6. Evolution of the overlapping genes *LRRC8E* and *ENSG00000214248*. (**A**) The structures of the sense-antisense overlapping (sas12) genes of *LRRC8E* and *ENSG00000214248*. The ω values are estimated by comparing the human and macaque orthologs, with *P*-values indicating the probabilities with which the null hypothesis of ω = 1 is true. (**B**) Protein expression levels of *LRRC8E* and *ENSG00000214248*. Median protein intensities from multiple samples, based on ProteomicsDB (Schwanhausser et al. 2011; Wilhelm et al. 2014), are shown for each tissue. (**C**) Evolution of *ENSG00000214248*. Species in which the ORF for *ENSG00000214248* is broken are underlined. Numbers on branches show the amino acid positions of premature stop codons. Branches are not drawn to scale.

A.8 Supplementary figures and tables

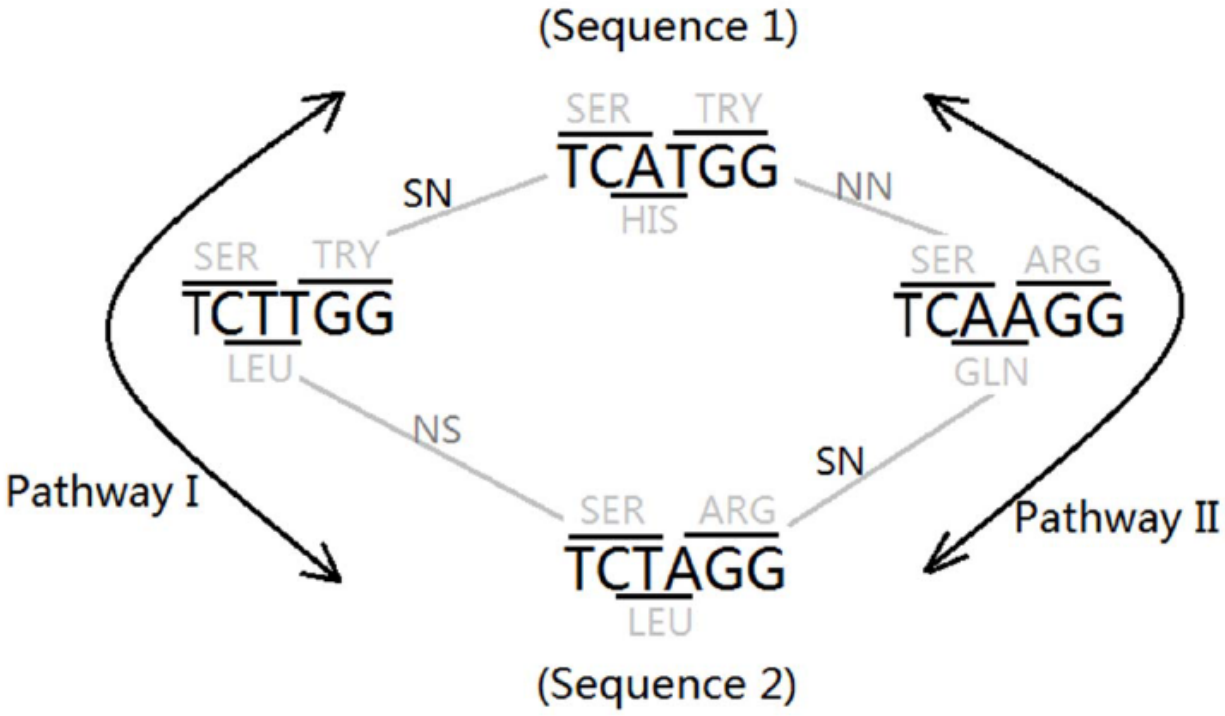


Figure A-S1. An example showing pathways of nucleotide substitutions in a sense-sense overlapping region. Codons and corresponding amino acids in ORF1 are marked with lines above the sequences, whereas those in ORF2 are marked with lines below the sequences. The types of nucleotide substitutions (NN, NS, SN, or SS) are indicated. The two pathways are considered to be equally likely.

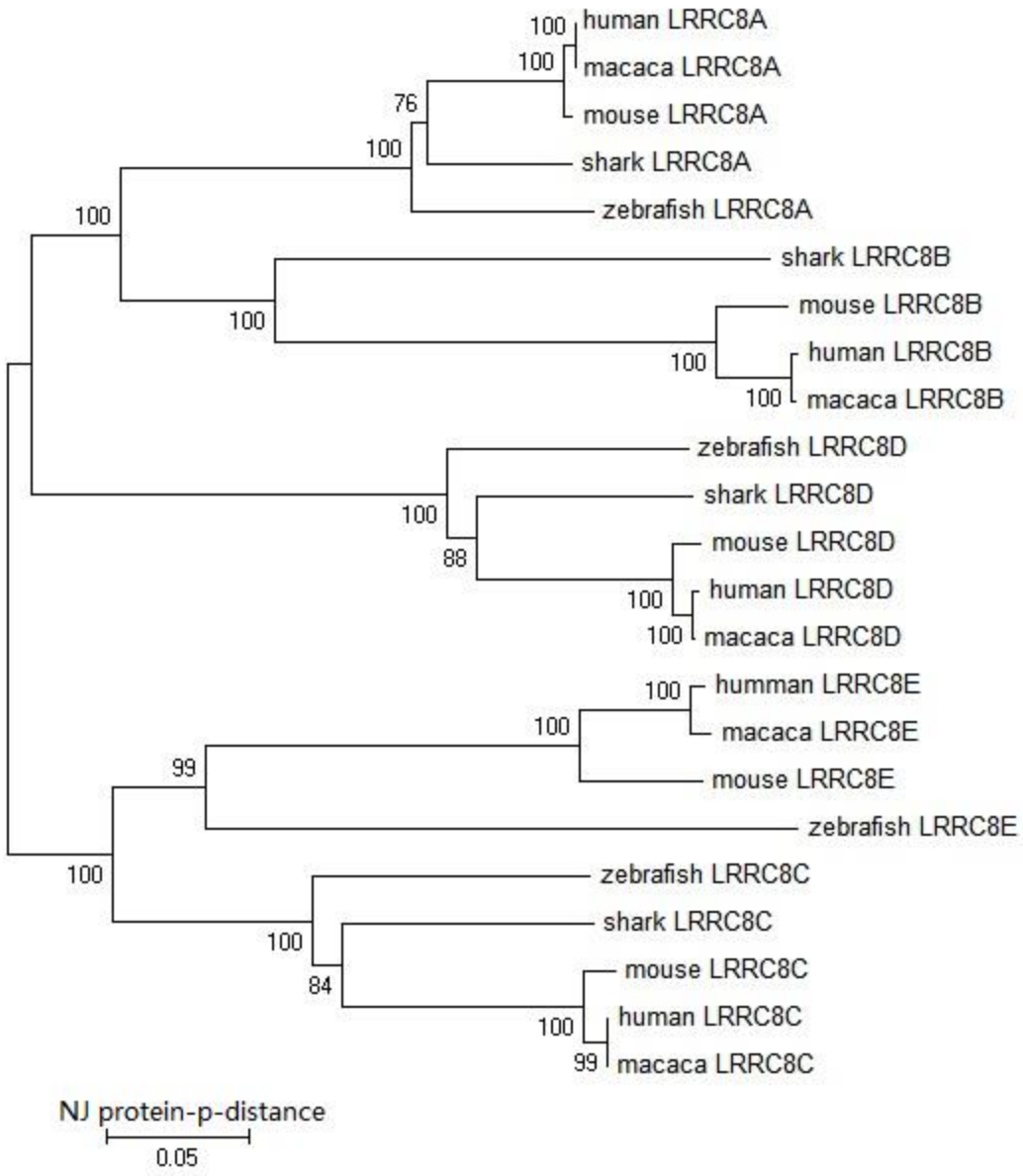


Figure A-S2. The unrooted phylogenetic tree of *LRRC8* genes from human, macaque, mouse, zebrafish, and shark. The tree was reconstructed using the neighboring-joining method with protein *p*-distances. There is no *LRRC8E* homolog in shark and no *LRRC8B* homolog in zebrafish. Bootstrap percentages derived from 1000 replications are shown for each interior branch.

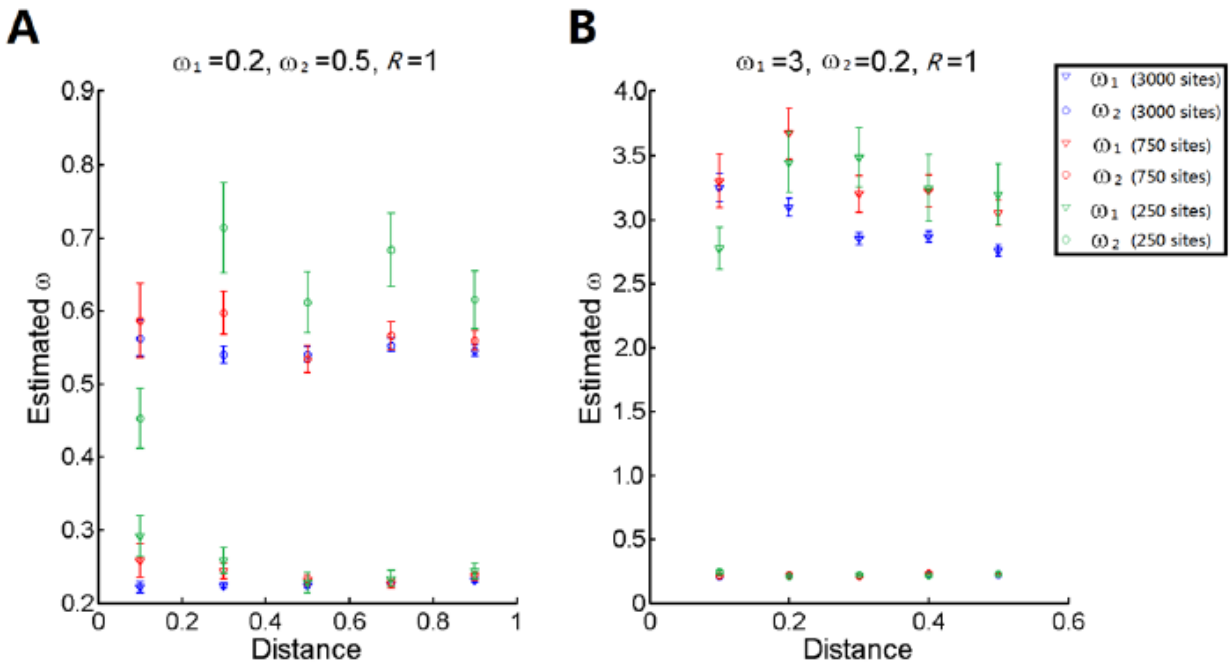


Figure A-S3. Performance of the new method in estimating selection intensities on genes with different overlapping lengths. Shown are results from computer simulations of overlapping genes with the sense-sense overlap. Each symbol represents the mean from 100 replications under a given parameter set, and error bars show the standard error. In each panel, the common parameters are listed above the panel.

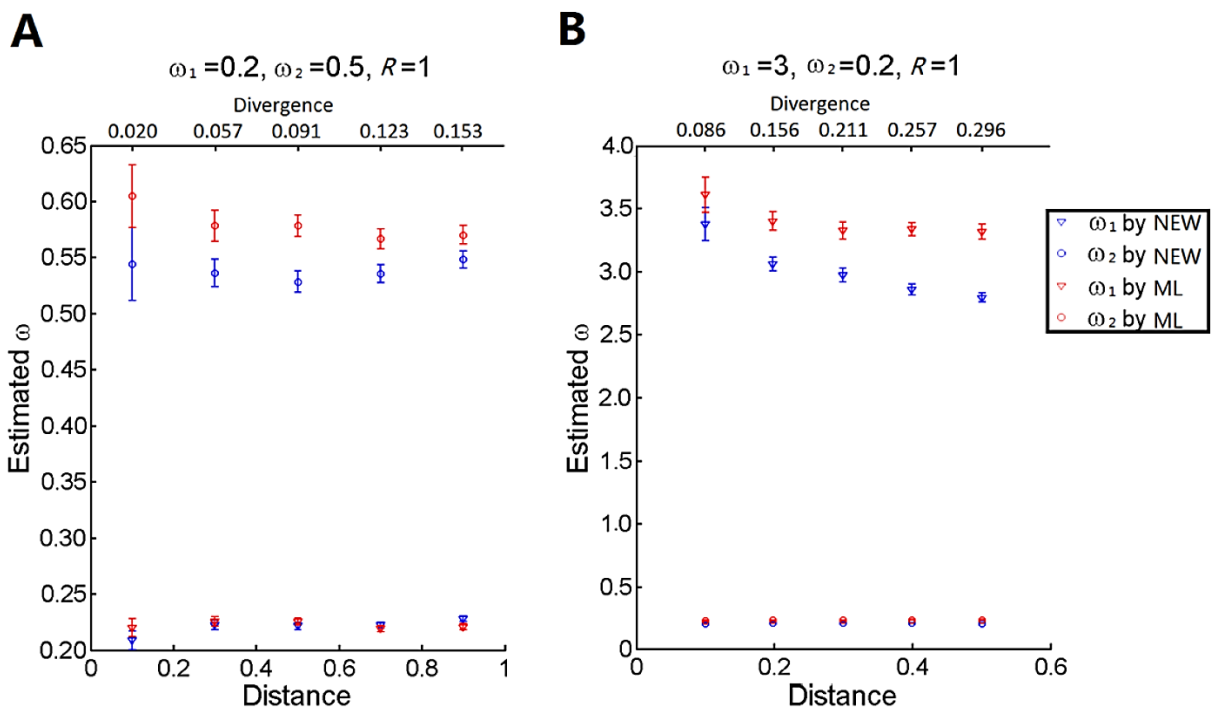


Figure A-S4. Comparison between the new (NEW) method and the maximal likelihood (ML) method. Shown are results from computer simulations of overlapping genes having the sense-sense overlap of 3000 sites. Each symbol represents the mean from 100 replications under a given parameter set, and error bars show the standard error. In each panel, the common parameters are listed above the panel. The average divergence level across all sites between a pair of simulated sequences is shown on the top X-axis.

Table A-S1. Accession numbers of *LRRC8E* sequences in Fig A-6C

Species	Accession number
Zebrafish	ENSDART00000108710
Cod	ENSGMOT00000018936
Opossum	ENSMODT00000019047
Wallaby	ENSMEUT00000010742
Elephant	ENSLAFT00000025823
Human	ENSG00000171017
Chimp	ENSPTRT00000019130
Gorilla	ENSGGOT00000022147
Orangutan	ENSPPYT00000011063
Macaque	ENSMMUG00000005424
Lemur	ENSMICT00000009643
Mouse	ENSMUST00000053035
Rat	ENSRNOT00000035142
Guinea pig	ENSCPOT00000003946
Rabbit	ENSOCUT00000031672
Cow	ENSBTAT00000005636
Pig	ENSSSCT00000014838
Megabat	ENSPVAT00000002294
Microbat	ENSMLUT00000013663
Horse	ENSECAT00000000256
Cat	ENSFCAT00000014946
Panda	ENSAMET00000009429
Dog	ENSCAFT00000029107

Table A-S2. Accession numbers of sequences in Fig A-S2

Species	Gene	Accession number
Human	LRRC8A	NP_001120717.1
	LRRC8B	ENST00000330947
	LRRC8C	NP_115646.2
	LRRC8D	NP_001127951.1
	LRRC8E	NP_001255213.1
Macaque	LRRC8A	NP_001244505.1
	LRRC8B	NP_001248286.1
	LRRC8C	NP_001244532.1
	LRRC8D	NP_001253482.1
	LRRC8E	XP_001093244.1
Mouse	LRRC8A	NP_808393.1
	LRRC8B	NP_001028722.1
	LRRC8C	NP_598658.1
	LRRC8D	NP_848816.3
	LRRC8E	NP_082451.2
Zebrafish	LRRC8A	NP_001025120.1
	LRRC8C	XP_001919381.1
	LRRC8D	ENSDARG00000062113
	LRRC8E	ENSDARG00000078283
Shark	LRRC8A	XP_007900959.1
	LRRC8B	XP_007885367.1
	LRRC8C	XP_007885365.1
	LRRC8D	XP_007885538.1

Appendix B:

Why phenotype robustness promotes phenotype evolvability?

"However, one cannot really argue with a mathematical theorem."

— **Stephen Hawking**

B.1 Abstract

Robustness and evolvability are fundamental characteristics of life whose relationship has intrigued generations of biologists. Studies of several genotype-phenotype maps (GPMs) such as the map between short DNA sequences and their bindings to transcription factors showed that phenotype robustness promotes phenotype evolvability, but the underlying reason is unclear. Here we show mathematically that the expected phenotype evolvability is a monotonically increasing function of the expected phenotype robustness in random GPMs. Population genetic simulations confirm that increasing phenotype robustness raises the probability that a target phenotype appears in a population within a given time, under empirical as well as randomly rewired GPMs. These and other results demonstrate that the positive correlation between phenotype robustness and phenotype evolvability is mathematical rather than biological. Hence, it is unsurprising to observe this correlation in every empirical GPM investigated, although the magnitude of the correlation may vary due to influences of various biological factors.

B.2 Robustness and evolvability

Genetic robustness refers to phenotypic invariance in the face of mutation, and is a widespread phenomenon at multiple levels of biological organization (de Visser et al. 2003; Kitano 2004; Wagner 2005b; Masel and Trotter 2010; Yang et al. 2014; Ho and Zhang 2016). Evolvability is the ability to produce (adaptive) phenotypic variation (Wagner and Altenberg 1996; Kirschner and Gerhart 1998; Wagner 2005b; Masel and Trotter 2010). Although robustness and evolvability are both fundamental characteristics of life, their relationship has been a long-standing controversy (Kitano 2004; Wagner 2005b; Masel and Trotter 2010). On the one hand, they are apparently antagonistic to each other, because the higher the robustness, the lower the probability with which a mutation results in a new phenotype (Ancel and Fontana 2000; Carter et al. 2005). On the other hand, robustness has been suggested to promote evolvability, not least because robustness allows the accumulation in a population of cryptic genetic variations that may be exposed and adaptive in a new environment (Aldana et al. 2007; Elena and Sanjuan 2008; Masel and Trotter 2010). Experimental evolution of RNA enzymes (Hayden et al. 2011), RNA viruses (McBride et al. 2008), and bacteria (Stiffler et al. 2015) showed that robustness can indeed enhance evolvability under certain conditions, but the generality of these findings is unknown.

Theoretical analysis of the robustness-evolvability relationship is often conducted in the context of a genotype-phenotype map (GPM; **Fig B-1A**), where each node is a genotype, each edge connects two genotypes that differ by one mutation, and nodes are colored based on their phenotypes (Wagner 2012). The set of connected nodes with the same color is commonly referred to as a neutral network (Schuster et al. 1994), because wandering in this network alters the genotype but not the phenotype. Note, however, that phenotypes are defined qualitatively in this context.

A decade ago, Wagner revolutionized the study of the robustness-evolvability relationship by distinguishing between genotype robustness (GR) and phenotype robustness (PR) and between genotype evolvability (GE) and phenotype evolvability (PE) (Wagner 2008). GR is the probability with which a random mutation occurring in a given genotype does not change its phenotype. By contrast, PR is the mean GR of all genotypes exhibiting a given phenotype. GE is the fraction of all phenotypes reachable by one mutation from a given genotype. By contrast, PE is the fraction of all phenotypes reachable by one mutation from any genotype exhibiting a given phenotype. Wagner and colleagues found that, within a GPM, GR and GE are negatively correlated but PR and PE are positively correlated for the phenotypes of RNA structure (Wagner 2008), protein structure (Ferrada and Wagner 2008), and DNA binding to transcription factors (TFs) (Payne and Wagner 2014). However, the broader generality and the underlying cause of the positive PR-PE correlation are unclear.

B.3 PE is expected to increase monotonically with PR in random GPMs

That a positive PR-PE correlation is observed in every GPM investigated (Ferrada and Wagner 2008; Wagner 2008; Payne and Wagner 2014) prompts us to investigate the possibility that this correlation is mathematical rather than biological. To this end, we consider a random GPM between G DNA sequences (genotypes) and their binding to K TFs (phenotypes). Each node represents a genotype of an l-nucleotide DNA sequence, and each phenotype represents the binding of the DNA to a TF. Let the number of genotypes showing phenotype i (i.e., the number of binding sequences of TF_i) be g_i. With a single nucleotide replacement, each genotype can change to one of $m = 3l$ other genotypes, which are collectively called the neighborhood of the

focal genotype. In this random GPM, under the assumption that $1 << g_i << G$ for any i, it can be shown (see Materials and Methods) that the expected phenotype robustness of binding to TF_i is

$$E(PR_i) \approx g_i / G, \quad (1)$$

whereas the corresponding expected phenotype evolvability is

$$E(PE_i) \approx 1 - \sum_{j \neq i} e^{-mg_j g_i / G} / (K-1). \quad (2)$$

Hence,

$$E(PE_i) \approx 1 - \sum_{j \neq i} e^{-mg_j E(PR_i)} / (K-1). \quad (3)$$

Eq. (3) shows that the expected PE_i is a monotonically increasing function of the expected PR_i. In other words, the expected PR and PE are intrinsically positively correlated in random GPMs. Importantly, Eq. (3) does not rely on any specific distribution of g_i.

To evaluate the accuracy of the above formulas that were derived with approximations, we simulated a random GPM with K = 80 TFs that all use 8-mer binding sequences. We chose these parameters because the empirically determined yeast and mouse TF-DNA binding GPMs have 89 and 105 TFs, respectively, and their binding sequences inferred from microarray data all contain 8 nucleotides (see Materials and Methods). To examine the variations of PR and PE in the entire range of possible g_i values, we chose the g_i values to be 15, 25, 35, …, and 805. We repeated the simulation 100 times and calculated the mean empirical PR and PE of binding to each TF. We found that E(PR) (**Fig B-1B**), E(PE) (**Fig B-1C**), and their relationship (**Fig B-1D**) based on the analytical formulas are indistinguishable from the corresponding average values observed from the simulation. This was also the case when g_i follows a normal (**Fig B-S1A-C**), bimodal (**Fig B-S1D-F**), or exponential (**Fig B-S1G-I**) distribution, suggesting that our analytical formulas are sufficiently accurate and general.

B.4 The PR-PE correlation is stronger in empirical than randomly rewired GPMs

We noticed from the analytical and simulation results of random GPMs that PE becomes virtually independent of PR when PR exceeds a certain value (**Fig B-1** and **Fig B-S1**). This phenomenon is much less pronounced in the empirical TF-DNA binding GPMs of mouse (**Fig B-2A-C**) and yeast (**Fig B-S2A-C**). To quantitatively compare empirical with random GPMs, we analytically computed the expected PR and PE for each TF in a randomly rewired mouse GPM, where the number of genotypes for each phenotype is unchanged but the genotype-phenotype relationships are randomized. Relative to a randomly rewired GPM, the actual GPM has higher PR and lower PE values for most TFs (**Fig B-2A**, **B**). This result is similar to that of Payne and Wagner (2014), although they computed PR and PE for a TF by randomly rewiring the binding sequences of the focal TF instead of those of all TFs simultaneously. Furthermore, they did not examine the relationship between PR and PE in any random or randomly rewired GPM. We found that the positive rank correlation between PR and PE is greater in the actual GPM than in each of 100 randomly rewired GPMs (**Fig B-2D**). Similar results were found when the yeast GPM was compared with corresponding randomly rewired GPMs (**Fig B-S2**).

B.5 The increase in the PR-PE correlation is related to large neutral networks

We hypothesize that the differences between the empirical GPMs and their randomly rewired GPMs in PR, PE, and PR-PE correlation are primarily related to the existence of large neutral networks (i.e., genotypes of the same phenotypes tend to be connected) in the former but not the latter. On average, the largest connected network for a mouse (or yeast) TF contains 81% (or 79%) of its binding sequences. This number drops to 1.2% in the randomly rewired GPMs of both species. Based on the definitions of PR and PE, it is obvious that, given g_i values, the

presence of large neutral networks raises PR but reduces PE. As a result, PE increases with g_i in almost the full range of g_i values in the empirical GPMs (**Fig B-2B**; **Fig B-S2B**), but saturates even in the bottom tenth of g_i values in the randomly rewired GPMs (**Fig B-2B**; **Fig B-S2B**).

To further demonstrate that the differences in PR, PE, and PR-PE correlation between empirical GPMs and their randomly rewired GPMs is due primarily to neutral networks instead of other properties of empirical GPMs, we created randomized GPMs with large neutral networks (see Materials and Methods). Indeed, patterns of PR, PE, and PR-PE correlation in these GPMs closely resemble those in empirical GPMs (**Fig B-S3**).

B.6 The biophysics of TF-DNA binding creates large neutral networks

It is interesting to note that, if the binding sequences of a TF were randomly distributed in a GPM, a population starting with a weak binding sequence would have to cross deep binding affinity valleys to reach a strong binding sequence, which is improbable except in very small populations. Thus, the presence of strong TF-DNA binding *per se* implies the existence of large (qualitatively) neutral networks of its binding sequences. But what forces have led to the large neutral networks? It is known that the genotypes for a phenotype tend to form a large neutral network simply by chance when the genotype number is sufficiently large. This phenomenon of percolation is, however, irrelevant here, because the phenotype with the largest number of genotypes contains only 2-3% of all genotypes in the GPMs studied here, much lower than the lower bound required for percolation (6.25%) (Gravner et al. 2007).

TF-DNA binding is known to be primarily determined by specific base-pair recognition (von Hippel and Berg 1986), and at different amino acid binding positions, different base-pairs are preferred due to interaction with hydrogen bonds provided by appropriately positioned amino

acids and peptide functional groups (von Hippel and Berg 1986; Stormo and Fields 1998; Afek et al. 2014). The biophysical property of TF-DNA binding dictates that the binding energy between a TF and a segment of DNA is largely the sum of the interaction energies of individual couples of an amino acid residue and a base pair. Only at 5% of sites does the binding strength deviate from the multiplicative expectation by more than two-fold (Jolma et al. 2013). The scarcity of epistasis means that the one-mutation neighborhood of a strong binding sequence of a TF is likely filled with the binding sequences of the same TF, because a single nucleotide change cannot drastically reduce the TF-DNA binding strength. Indeed, binding sequences with higher binding affinities tend to have higher GR (Payne and Wagner 2014). This property leads to the creation of large neutral networks. A recent extensive analysis of TF-DNA binding affinities generally supports this notion (Aguilar-Rodríguez et al. 2017).

B.7 PR facilitates adaptation in population genetic simulations under randomly rewired GPMs

Because Wagner's definition of PE does not explicitly consider the population genetic process of adaptation, we turn to another, arguably more relevant measure of evolvability–the probability that a target phenotype appears in a population within a given time, which we will refer to as PE'. We start with a haploid adult population with a homogenous genotype corresponding to phenotype *i*, which is optimal in the current environment. All other phenotypes are lethal. In each generation, genetic drift occurs such that *N* offspring are produced and their genotype frequencies may differ from those of the parental population. Each offspring has a probability of μ to become a neighboring genotype due to mutation, and only those with viable phenotypes mature and reproduce (i.e., some of the *N* individuals may not mature). Based on

theory (Nei et al. 1975) and our pilot simulation, we repeat this process for $1/\mu$ generations to allow the population to reach an equilibrium level of genetic diversity. An environmental shift then occurs, which renders phenotype *i* suboptimal, phenotype *j* ($\neq i$) optimal, and all other phenotypes still lethal. We repeat the process of mutation, purifying selection, and drift over many generations until an individual with phenotype *j* appears in the population or the number of generations after the environmental shift reaches a preset limit *T*, whichever occurs first. We examine each and every new phenotype *j* ($\neq i$) and calculate the fraction of phenotypes that can be reached from *i* within time *T*, which is PE'. We repeat the evolutionary simulation 50 times, each starting from a randomly picked genotype of the phenotype *i* and present the average result from these 50 simulations. We consider the first appearance of the adaptive phenotype rather than the first fixation of the adaptive phenotype, because the fixation probability and expected fixation time is the same given N, μ, and selective strength. In all simulations, we use $N = 100$ to speed up the process.

We first conducted the population genetic simulation under the mouse TF-DNA binding GPM using mouse-appropriate $N\mu$. When $T = 10{,}000$ generations is the upper limit in waiting time for the target phenotype, we found a positive correlation between the PR of the starting phenotype and PE' ($\rho = 0.45$, $P < 10^{-5}$; **Fig B-3A**). Similar results were obtained (**Fig B-3B**) when *T* is 1,000 ($\rho = 0.37$, $P < 10^{-4}$), 100,000 ($\rho = 0.46$, $P < 10^{-5}$), or 1,000,000 generations ($\rho = 0.49$, $P < 10^{-6}$). Thus, increasing PR raises the chance of adaptation upon an environmental shift.

We similarly conducted the population genetic simulation under the yeast TF-DNA binding GPM using yeast-appropriate $N\mu$. We again observed that, the higher the PR of the starting phenotype, the higher the probability of appearance of a target phenotype in the population (**Fig B-3B**).

Interestingly, the correlation between PR and PE' becomes even stronger when we conducted simulations under randomly rewired mouse and yeast GPMs, respectively (**Fig B-3C**, **D**). These results indicate that PR promotes PE' and that this property is intrinsic rather than biological.

B.8 Implications

Our mathematical and empirical results showed that (1) the expected PR and PE are intrinsically positively correlated even in random GPMs, (2) compared with the corresponding randomly rewired GPMs, the mouse and yeast TF-DNA binding GPMs show stronger PR-PE correlations, likely because of their large neutral networks, and (3) these large neutral networks are explainable by the biophysical nature of TF-DNA binding. While (1) is a general finding for GPMs of all classes of phenotypes, (2) and (3) are derived from the analysis of TF-DNA binding GPMs. Nonetheless, for any phenotype that can be improved by natural selection, its genotypes must form some neutral networks such that quantitatively better phenotypes are reachable by mutation; otherwise, the phenotype could not be improved by natural selection. Hence, we expect (2) to be true in the GPM for any adaptable phenotype (when adaption occurs primarily via mutation rather than recombination). Note, however, that our finding that the expected PR and PE are positively correlated in random GPMs does not imply that PR and PE cannot have a negative correlation even in hypothetical GPMs. For instance, one could imagine a GPM where the genotypes of some phenotypes form large neutral networks whereas those of other phenotypes are largely unconnected. Compared with the latter group of phenotypes, the former group are expected to have higher PR but lower PE. Consequently, a negative correlation between PR and PE would result when the two groups of phenotypes are analyzed together.

Nevertheless, such GPMs should be the exception rather than the rule. Hence, observing a positive PR-PE correlation in an empirical GPM is expected and does not offer any specific biological insight, as far as Wagner's definitions are concerned.

Our population genetic simulations showed that PR promotes PE' under real and randomly rewired GPMs. PE' is similar to Wagner's definition of PE except that PE' is defined in a population genetic framework and hence is more realistic and more relevant to actual adaptation. Our population genetic simulation differs from a previous treatment of the same subject by Draghi and colleagues (Draghi et al. 2010), who found PR to promote PE' under some but not all circumstances. However, their study contained a number of simplifying assumptions. For instance, they assumed that any genotype has a non-zero probability to show any phenotype by a minimum of one mutation, which is untrue. In addition, no GPM was explicitly modeled and only genotypes of the starting phenotype were assumed to form a neutral network. They also unrealistically assumed that all genotypes of the same phenotype have equal robustness. Furthermore, although the robustness of a phenotype correlates with the number of neighboring phenotypes, they neglected this correlation in their model. Hence, our analysis, based on actual and randomly rewired GPMs, coupled with more realistic assumptions, is biologically more relevant than theirs. Note that, Draghi et al. observed a decrease in PE when PR is very high, which we did not observe in our study. Because such high PR values are not observed in our data, our analysis cannot confirm or invalidate their finding. Together, our findings on the impacts of PR on PE and PE' demonstrate that observing a positive correlation between phenotype robustness and evolvability in an empirical GPM requires no biological explanation. This said, the magnitude of the positive correlation is certainly impacted by some biological factors, as in the TF-DNA binding GPMs studied here.

Compared with phenotypes without large neutral networks, those with large neutral networks (but the same numbers of genotypes) have two apparent benefits. First, mutations are less likely to alter these phenotypes qualitatively. Second, they are more selectable, meaning that mutations could lead to quantitatively fitter but qualitatively unchanged phenotypes. One drawback is that they have a reduced evolvability. Nevertheless, it is clear by comparing the mouse (or yeast) TF-DNA binding GPM with its randomly rewired GPM that the PE and PE' reduction in the empirical GPM is moderate while the PR increase is substantial (**Fig B-2**; **Fig B-S2**; **Fig 3**).

Kitano contended that there are architectural requirements for complex systems to be evolvable and that such requirements also give rise to robustness (Kitano 2004). If his "evolvable" meant "selectable", our results strongly support his hypothesis, because having a large neutral network given the number of genotypes is necessary for a phenotype to be selectable and is also the reason behind its high robustness. If his "evolvable" is in the sense of PE or PE', our findings refute his hypothesis, because the architecture that confers high evolvability—a lack of neutral networks (given the number of genotypes)—reduces robustness.

In the case of TF-DNA binding GPMs, large neutral networks arise naturally from the biophysics of TF-DNA binding. It seems likely that, in other systems such as RNA secondary structures or protein structures, large neutral networks can also result from physical and/or chemical properties of the systems. If this conjecture proves to be generally true, it would mean that simple physical and chemical laws not only permit the origin of life but also provide life with robustness and selectability while allowing reasonably high evolvability. This intriguing possibility is worth exploration in the future.

B.9 Materials And Methods

B.9.1 Expected PR and PE in a random GPM

Let us consider a random GPM, where each node represents a genotype of l nucleotides and the GPM contains $G = (4^l - 4^{0.5l})/2 + 4^{0.5l} = (4^l + 4^{0.5l})/2$ unique genotypes and K phenotypes. The above formula of G was derived by considering that each sequence is equivalent to its reverse complement and that there are $4^{0.5l}$ palindromic l-mers (when l is an even number) (van Helden et al. 1998). Because palindromic sequences constitute a tiny fraction ($< 0.5^{l+1}$) of all genotypes, we ignored their palindromic effects in the following modelling. As shown in the numerical examples (**Fig B-1**; **Fig B-S1**), this approximation is acceptable. Let the number of unique binding sequences of TF_i be g_i. With a single nucleotide replacement, each genotype can change to one of $m = 3l$ other genotypes, which are collectively called the one-step neighborhood of the focal genotype. We assume that $1 << g_i << G$ for any i. The expected GR of a binding sequence of TF_i is the expected number of other binding sequences of TF_i that fall in the one-step neighborhood of the focal binding sequence, divided by m. Because the number of other binding sequences of TF_i is g_i-1 and the probability for any one of them to fall in the one-step neighborhood of the focal binding sequence is $m/(G\text{-}1)$, the expected GR is $E[GR] = [(g_i\text{-}1)m/(G\text{-}1)]/m = (g_i\text{-}1)/(G\text{-}1) \approx g_i/G$. Because PR is the mean GR of all binding sequences of TF_i, the expected PR is $E[PR] = E[\text{mean GR}] = E[PR] \approx g_i/G$.

Now let us consider another TF (TF_j), which has g_j binding sequences. The probability that a particular binding sequence of TF_i is in the one-step neighborhood of a particular binding sequence of TF_j is approximately m/G. Hence, the probability that a particular binding sequence of TF_i is in the neighborhood of any binding sequence of TF_j (or more precisely the expected number of edges between a particular binding sequence of TF_i and all binding sequences of TF_j)

is approximately mg_j/G. The expected number of edges between all binding sequences of TF_i and all binding sequences of TF_j is approximately mg_ig_j/G. Because the number of edges between two phenotypes follows a binomial distribution (with g_ig_j trials each having a success rate of m/G), the probability that the phenotype of TF_j binding is reachable from the phenotype of TF_i binding by one mutation from at least one binding sequence of TF_i equals q_{ij} =1 - $(1 - m/G)^{g_ig_j} \approx 1 - e^{-\frac{mg_ig_j}{G}}$. Thus, PE_i, the fraction of all phenotypes reachable from the phenotype of TF_i binding by one mutation, is expected to be $\sum_{j\neq i} q_{ij}/(K-1) = \sum_{j\neq i}(1 - e^{-mg_jg_i/G})/(K-1) = 1 - \sum_{j\neq i} e^{-mg_jg_i/G}/(K-1)$. One can substitute g_i/G in the above formula by $E(PR_i)$ to obtain $E(PE_i) = 1 - \sum_{j\neq i} e^{-mg_j E(PR_i)}/(K-1)$, which indicates that E(PE) is an increasing function of E(PR).

B.9.2 Microarray data

The TF-DNA binding microarray data for mouse and yeast were downloaded from UniPROBE (http://the_brain.bwh.harvard.edu/uniprobe/downloads.php) (Newburger and Bulyk 2009). We defined binding sequences using the same data and enrichment score (E-score) cutoff (0.35) as in Payne and Wagner (2014); this cutoff corresponds to a low false discovery rate (Payne and Wagner 2014).

B.9.3 PR and PE calculation

We considered only single nucleotide substitutions in computing PR and PE. This is slightly different from a previous study (Payne and Wagner 2014), in which insertions and deletions (indels) were also considered. While considering indels should in theory make the analysis better, Payne and Wagner (2014) assumed that indels are one nucleotide long and are

restricted to the two ends of a binding sequence, which are unrealistic. Contemplating the complication of indels and the problem with the assumption, we decided not to consider indels. Note that our mathematical model has a variable *m* that measures the number of one-step neighbors per node that in theory takes into account all kinds of mutations. Hence, ignoring indels in the empirical analysis does not impact our mathematical analysis. Unlike the previous study (Payne and Wagner 2014), we considered all binding sequences of a TF rather than only those belonging to the largest neutral network (giant component). Because sequences that do not belong to the giant component can also bind to its TF and has potentials to evolve to a binding sequence of other TFs, including all binding sequences makes our analysis more complete. This change in methodology does not qualitatively affect the results on empirical (**Fig B-2A-C** and **Fig B-S2A-C**) or randomly rewired GPMs (**Fig B-S4**). A binding sequence of TF_i can be zero mutational steps away from a binding sequence of TF_j if they share the same binding sequence.

B.9.4 Generation of randomly rewired GPMs

Given the g_i values of all TFs, we randomly picked genotypes from the 8-mer genotype space (with replacement) and assigned the genotypes to each TF. This was done with replacement, because both mouse and yeast GPMs contain genotypes that map to multiple phenotypes and because the sum of g_i exceeds G in both mouse and yeast. A genotype can map to multiple phenotypes but it cannot occur twice for the same phenotype.

B.9.5 PR, PE, and PR-PE correlation in random GPMs with large neutral networks

The ensemble of all binding sequences of a TF is often represented by a position weight matrix (PWM), which shows the frequencies of A, T, G, and C at each nucleotide position of all

binding sequences of the TF. Because potential epistasis is ignored in constructing PWMs from microarray-based TF-DNA binding data, when PWMs are used, all binding sequences of a TF are connected to form one large neutral network in the GPM. We downloaded PWMs for mouse and yeast from UniPROBE (http://the_brain.bwh.harvard.edu/uniprobe/downloads.php) (Newburger and Bulyk 2009). For microarray data, we defined binding sequences using the same data and same enrichment score (E-score) cutoff (0.35) as previously used (Payne and Wagner 2014); this cutoff corresponds to a low false discovery rate (Payne and Wagner 2014). To convert PWMs back to binding sequences, we calculated the probability of each genotype for each TF, and used the cutoff of 0.0000469 in yeast and 0.00023885 in mouse to define binding sequences. Using these cutoffs led to similar total numbers of binding sequences as in the microarray data. We considered all binding sequences passing our cutoff to have equal binding affinities to the TF of concern.

We then constructed a random GPM with large neutral networks. Specially, to remove the evolutionary relationships among the PWMs (and those among their corresponding TFs), we constructed a new set of PWMs by randomly shuffling all nucleotide positions among all existing PWMs of the species. We then used these scrambled PWMs to construct the GPM. In this GPM, large neutral networks are still present (albeit different from those in the empirical GPMs).

B.10 Acknowledgements

We thank members of the Zhang lab and four anonymous reviewers for valuable comments. This work was supported in part by the U.S. National Institutes of Health grant GM103232 to J.Z.

B.11 References

Afek A, Schipper JL, Horton J, Gordan R, Lukatsky DB. 2014. Protein-DNA binding in the absence of specific base-pair recognition. Proc Natl Acad Sci U S A 111:17140-17145.

Aguilar-Rodríguez J, Payne JL, Wagner A. 2017. A thousand empirical adaptive landscapes and their navigability. Nat Ecol & Evol 1:0045.

Aldana M, Balleza E, Kauffman S, Resendiz O. 2007. Robustness and evolvability in genetic regulatory networks. J Theor Biol 245:433-448.

Ancel LW, Fontana W. 2000. Plasticity, evolvability, and modularity in RNA. J Exp Zool 288:242-283.

Carter AJ, Hermisson J, Hansen TF. 2005. The role of epistatic gene interactions in the response to selection and the evolution of evolvability. Theor Popul Biol 68:179-196.

de Visser JA, Hermisson J, Wagner GP, Ancel Meyers L, Bagheri-Chaichian H, Blanchard JL, Chao L, Cheverud JM, Elena SF, Fontana W, et al. 2003. Perspective: Evolution and detection of genetic robustness. Evolution 57:1959-1972.

Draghi JA, Parsons TL, Wagner GP, Plotkin JB. 2010. Mutational robustness can facilitate adaptation. Nature 463:353-355.

Elena SF, Sanjuan R. 2008. The effect of genetic robustness on evolvability in digital organisms. BMC Evol Biol 8:284.

Ferrada E, Wagner A. 2008. Protein robustness promotes evolutionary innovations on large evolutionary time-scales. Proc Biol Sci 275:1595-1602.

Gravner J, Pitman D, Gavrilets S. 2007. Percolation on fitness landscapes: effects of correlation, phenotype, and incompatibilities. J Theor Biol 248:627-645.

Hayden EJ, Ferrada E, Wagner A. 2011. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. Nature 474:92-95.

Ho WC, Zhang J. 2016. Adaptive genetic robustness of Escherichia coli metabolic fluxes. Mol Biol Evol 33:1164-1176.

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. 2013. DNA-binding specificities of human transcription factors. Cell 152:327-339.

Kirschner M, Gerhart J. 1998. Evolvability. Proc Natl Acad Sci U S A 95:8420-8427.

Kitano H. 2004. Biological robustness. Nat Rev Genet 5:826-837.

Masel J, Trotter MV. 2010. Robustness and evolvability. Trends Genet 26:406-414.

McBride RC, Ogbunugafor CB, Turner PE. 2008. Robustness promotes evolvability of thermotolerance in an RNA virus. BMC Evol Biol 8:231.

Nei M, Maruyama T, Chakraborty R. 1975. The bottleneck effect and genetic variability in populations. Evolution:1-10.

Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. Nucleic Acids Res 37:D77-D82.

Payne JL, Wagner A. 2014. The robustness and evolvability of transcription factor binding sequences. Science 343:875-877.

Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Pialek J, Tucker PK, Nachman MW. 2012. Adaptive evolution and effective population size in wild house mice. Mol Biol Evol 29:2949-2955.

Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994. From sequences to shapes and back: a case study in RNA secondary structures. Proc Biol Sci 255:279-284.

Stiffler MA, Hekstra DR, Ranganathan R. 2015. Evolvability as a function of purifying selection in TEM-1 beta-lactamase. Cell 160:882-892.
Stormo GD, Fields DS. 1998. Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci 23:109-113.
Uchimura A, Higuchi M, Minakuchi Y, Ohno M, Toyoda A, Fujiyama A, Miura I, Wakana S, Nishino J, Yagi T. 2015. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Res 25:1125-1134.
van Helden J, Andre B, Collado-Vides J. 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281:827-842.
von Hippel PH, Berg OG. 1986. On the specificity of DNA-protein interactions. Proc Natl Acad Sci U S A 83:1608-1612.
Wagner A. 2005a. Energy constraints on the evolution of gene expression. Mol Biol Evol 22:1365-1374.
Wagner A. 2005b. Robustness and Evolvability in Living Systems. Princeton, NJ: Princeton University Press.
Wagner A. 2008. Robustness and evolvability: a paradox resolved. Proc Biol Sci 275:91-100.
Wagner A. 2012. The role of robustness in phenotypic adaptation and innovation. Proc Biol Sci 279:1249-1258.
Wagner GP, Altenberg L. 1996. Perspective: complex adaptations and the evolution of evolvability. Evolution:967-976.
Yang JR, Ruan S, Zhang J. 2014. Determinative developmental cell lineages are robust to cell deaths. PLoS Genet 10:e1004501.
Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci U S A 111:E2310-E2318.

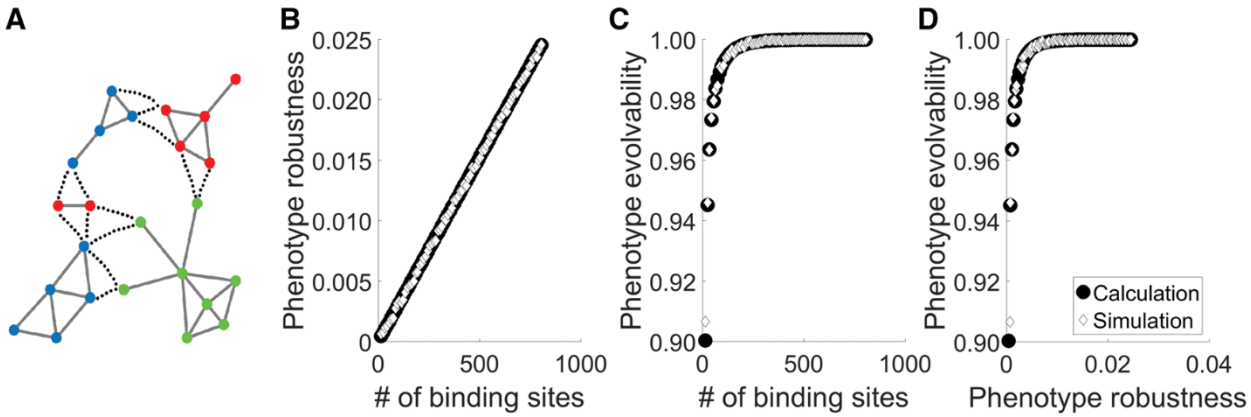


Fig B-1. PR and PE are positively correlated in random GPMs. (**A**) A hypothetical genotype-phenotype map (GPM). Each node represents a genotype, while its color represents its phenotype. Two genotypes that are one mutational step away from each other are connected by an edge, where a solid edge connects genotypes of the same phenotype and a dotted edge connects genotypes of different phenotypes. (**B**) The expected PR increases with the number of binding sequences in random TF-DNA binding GPMs. Each symbol represents one TF. Solid circles show analytically calculated values while open diamonds show corresponding means observed from 100 simulations of random GPMs. The observed standard deviation of PR (average 0.0016) is not correlated with the number of binding sequences. See main text for the parameters of the GPMs used. (**C**) The expected PR increases with the number of binding sequences in these random GPMs. The observed standard deviation of PE (maximum 0.0304) is negatively correlated with the number of binding sequences. (**D**) The expected PE is a monotonically increasing function of the expected PR in these random GPMs.

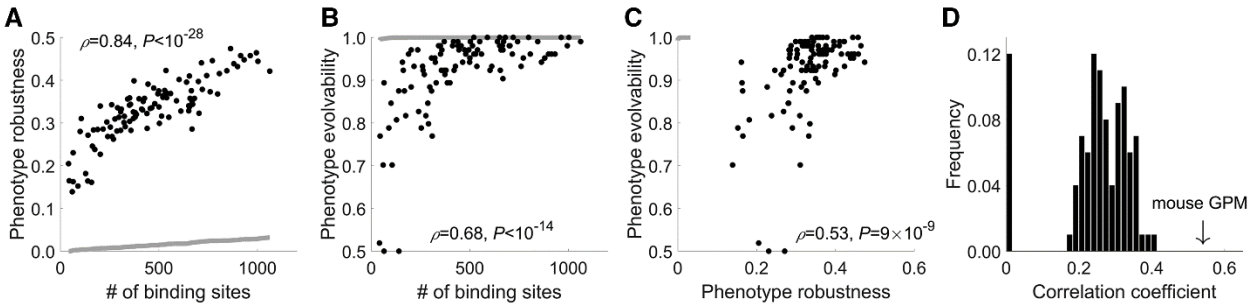


Fig B-2. PR-PE relationships in the mouse TF-DNA binding GPM and corresponding randomly rewired GPMs. (**A**) PR increases with the number of binding sequences in the mouse GPM. Each dot is a TF. (**B**) PE increases with the number of binding sequences in the mouse GPM. (**C**) PE is an increasing function of PR in the mouse GPM. In (A)-(C), the analytically computed results in corresponding random GPMs are presented by the grey curves. (**D**) Frequency distribution of the rank correlation between PR and PE in 100 randomly rewired mouse GPMs. The arrow points to the observed correlation in the mouse GPM.

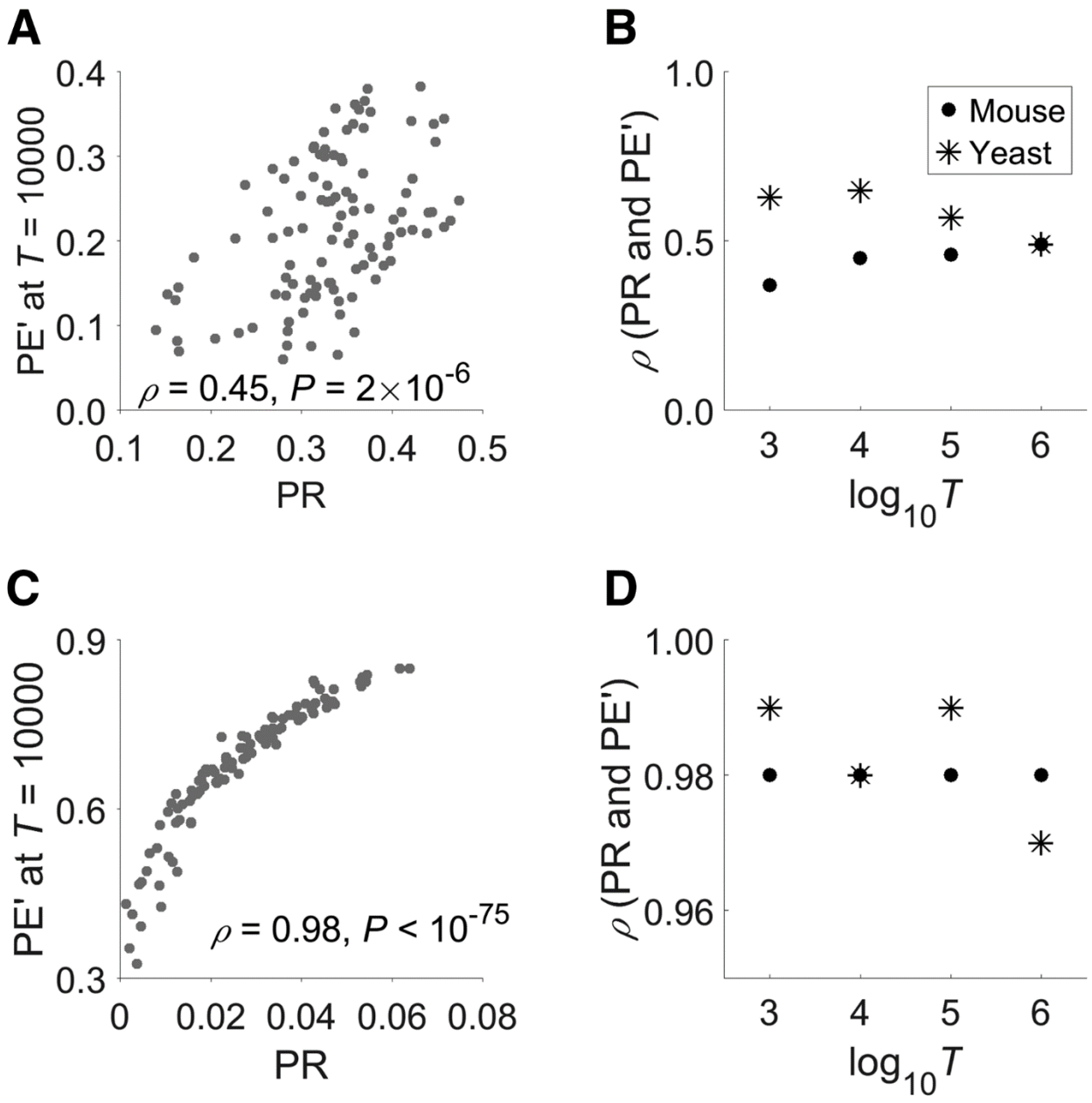


Fig B-3. Population genetic simulations show that PR promotes PE', which is the probability that a target phenotype appears in a population within time *T*. (**A**) Positive correlation between PR and PE' under the mouse GPM when *T* = 10,000 generations. ρ, Spearman's rank correlation coefficient. (**B**) Rank correlation between PR and PE' under mouse (stars) and yeast (dots) GPMs, respectively. (**C**) Positive correlation between PR and PE' under a randomly rewired mouse GPM when *T* = 10,000 generations. (**D**) Rank correlation between PR and PE' under

randomly rewired mouse (stars) and yeast (dots) GPMs, respectively. In panels (B) and (D), all correlations significantly exceed 0 ($P < 10^{-4}$). For mouse, our simulation used $N\mu = 0.004$ per generation per motif, based on the motif length of 8 nucleotides, mutation rate of 5.4×10^{-9} per generation per site (Uchimura et al. 2015), and effective population size of 10^5 (Phifer-Rixey et al. 2012). For yeast, our simulation used $N\mu = 0.016$ per generation per motif, based on its motif length of 8 nucleotides, mutation rate of 2×10^{-10} per generation per site (Zhu et al. 2014), and effective population size of 10^7 (Wagner 2005a).

B.12 Supplementary figures

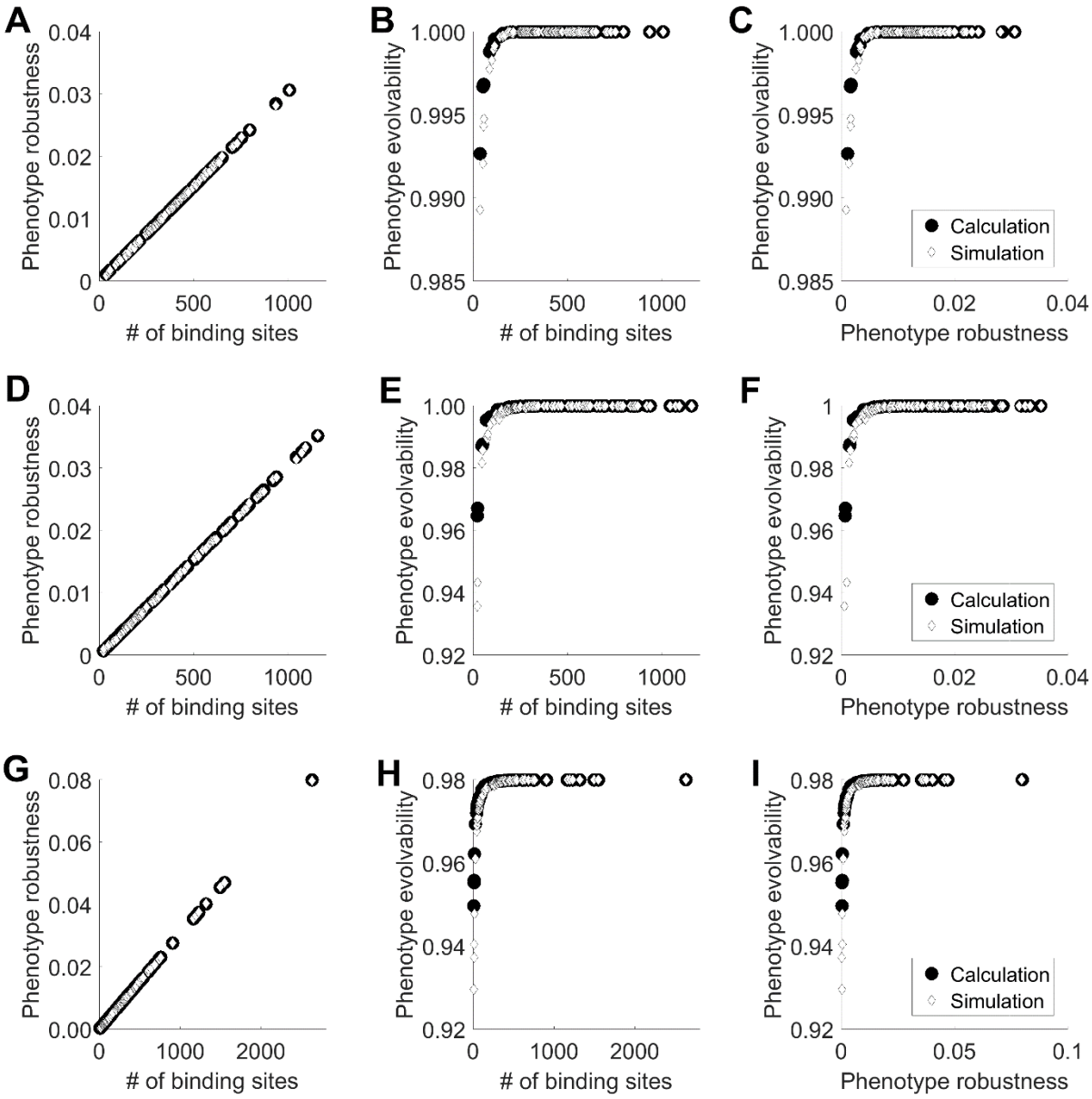


Fig B-S1. Analytical formulas for expected PR and PE in random GPMs are accurate. Each symbol represents one TF. Solid circles show analytically calculated values, whereas open diamonds show the corresponding means from 100 simulations of random GPMs. (**A-C**) Results from using 90 TFs with $g_i > 0$ sampled from the normal distribution of mean = 400 and standard

deviation = 200. (**D-F**) Results from using 90 TFs with $g_i > 0$ sampled from a bimodal distribution. Specifically, 45 g_i values are sampled from the normal distribution with mean = 200 and standard deviation = 100, while the other 45 g_i values are sampled from the normal distribution with mean = 600 and standard deviation = 300. (**G-I**) Results from using 80 TFs with $g_i > 0$ sampled from an exponential distribution with mean = 400.

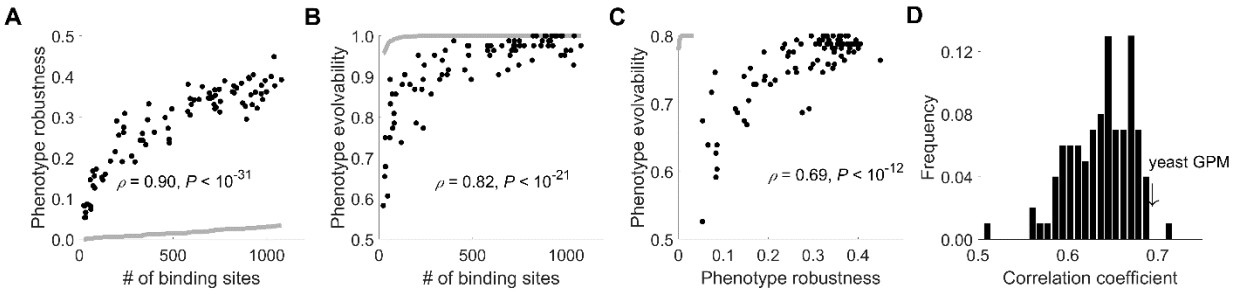


Fig B-S2. PR-PE relationships in the yeast TF-DNA binding GPM and corresponding randomly rewired GPMs. (**A**) PR increases with the number of binding sequences in the yeast GPM. Each dot is a TF. (**B**) PE increases with the number of binding sequences in the yeast GPM. (**C**) PE is an increasing function of PR in the yeast GPM. In (A)-(C), the analytically computed results in corresponding random GPMs are presented by the grey curves. (**D**) Frequency distribution of the rank correlation between PR and PE in 100 randomly rewired yeast GPMs. The arrow points to the observed correlation in the yeast GPM.

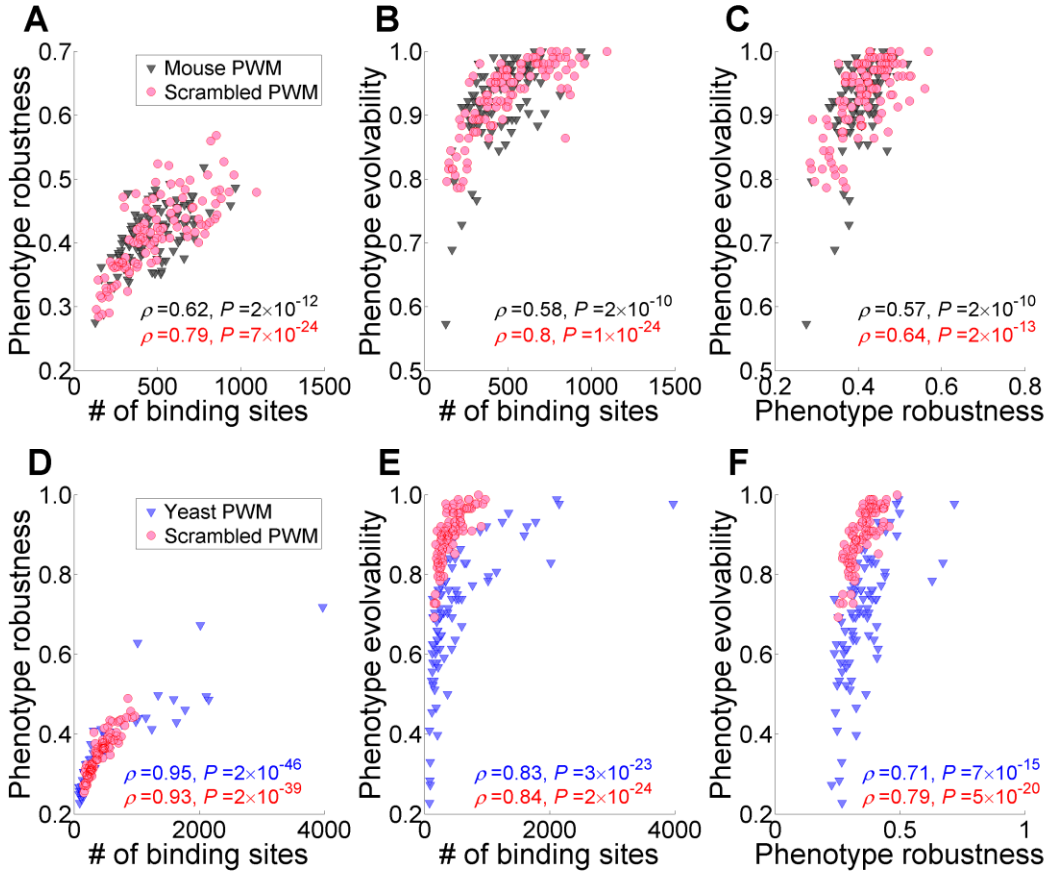


Fig B-S3. PR, PE, and PR-PE correlation based on actual position weight matrices (PWMs) of TF binding sequences and scrambled PWMs. Each triangle or circle represents one TF. ρ, Spearman's rank correlation coefficient. (**A**) In mouse, PR is positively correlated with the number of binding sequences from both actual PWMs and scrambled PWMs. (**B**) In mouse, PE is positively correlated with the number of binding sequences from both actual and scrambled PWMs. (**C**) In mouse, PE is positively correlated with PR for both actual PWMs and scrambled PWMs. (**D**) In yeast, PR is positively correlated with the number of binding sequences from both actual PWMs and scrambled PWMs. (**E**) In yeast, PE is positively correlated with the number of binding sequences from both actual PWMs and scrambled PWMs. (**F**) In yeast, PE is positively correlated with PR for both actual PWMs and scrambled PWMs.

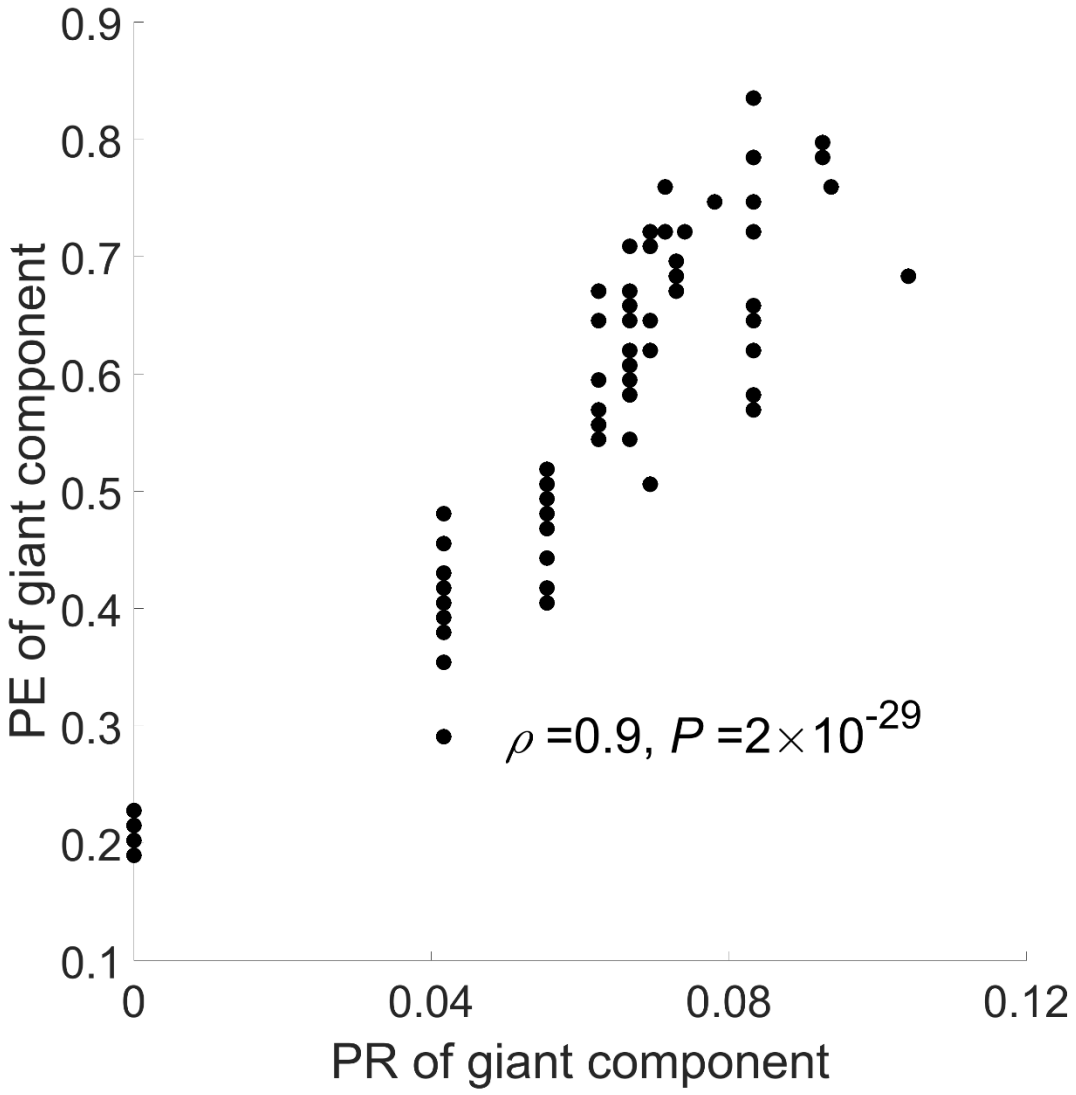


Fig B-S4. PR and PE of the giant components of randomly rewired mouse GPMs are positively correlated. Shown here is the result from one randomly rewired GPM used in Fig B-2D. Each dot represents one TF. ρ, Spearman's rank correlation coefficient. We examined 10 randomly rewired GPMs, and the correlation coefficients are in the range of 0.84-0.91.

Appendix C: Supplementary figures and tables for chapter 2

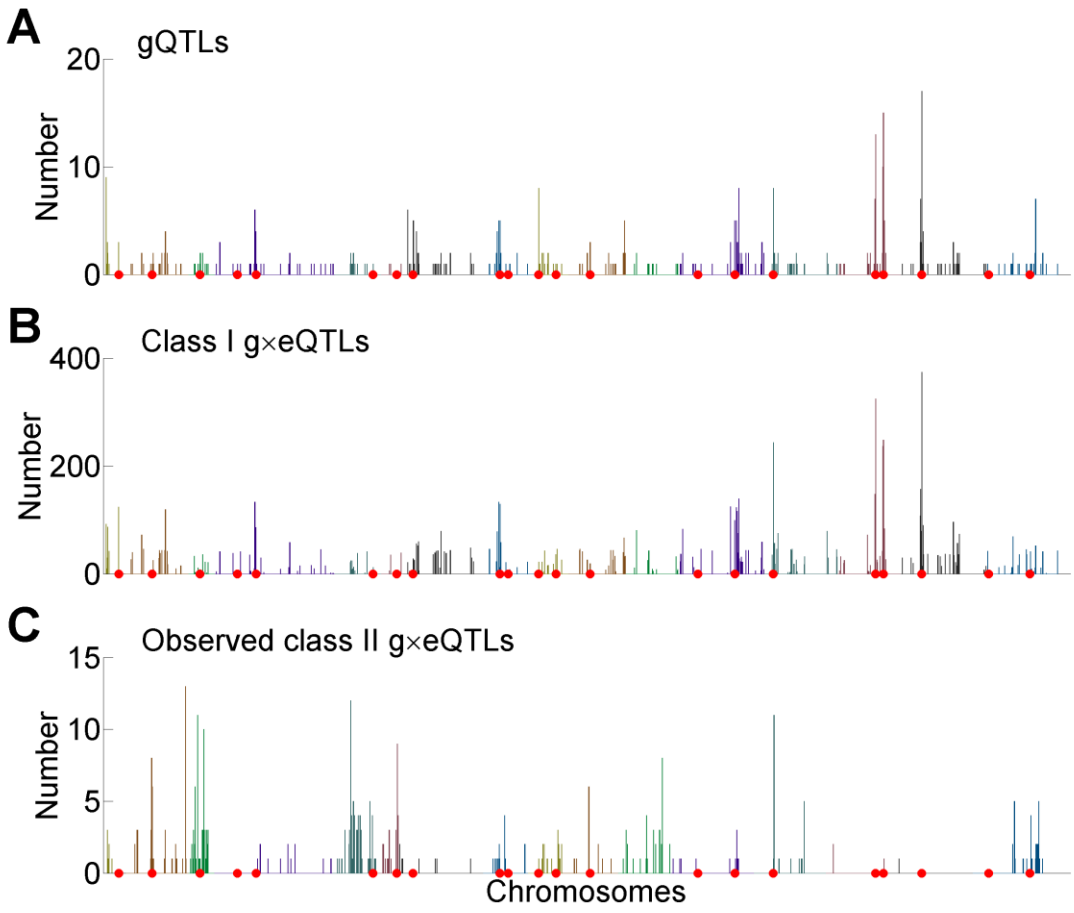


Figure C-1. Genomic locations of mapped gQTLs from the combined data of all 47 environments (red dots) placed against the distributions of (A) all gQTLs, (B) all class I g×eQTLs, and (C) all observed class II g×eQTLs individually mapped in the 47 environments (as in Fig. 2).

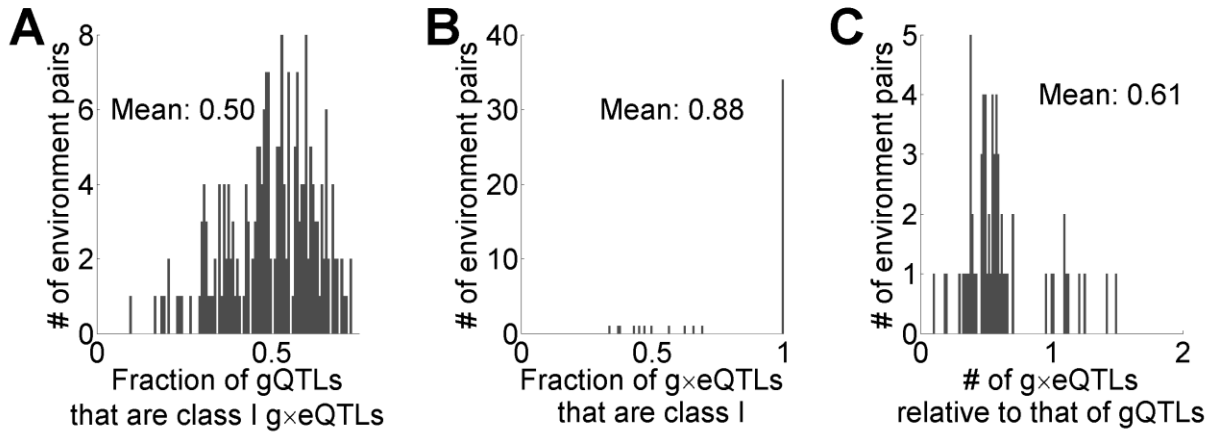


Figure C-2. Relative numbers of g×eQTLs and gQTLs from all pairs of environments mapped using the data from Bloom et al. (2015). (A) Frequency distribution of the fraction of all gQTLs identified from two individual environments that are class I g×eQTLs for the pair of environments. (B) Frequency distribution of the fraction of all g×eQTLs (i.e., class I + extrapolated class II) that are class I. (C) Frequency distribution of the ratio between the number of all g×eQTLs for a pair of environments and the total number of unique gQTLs identified in the two environments.

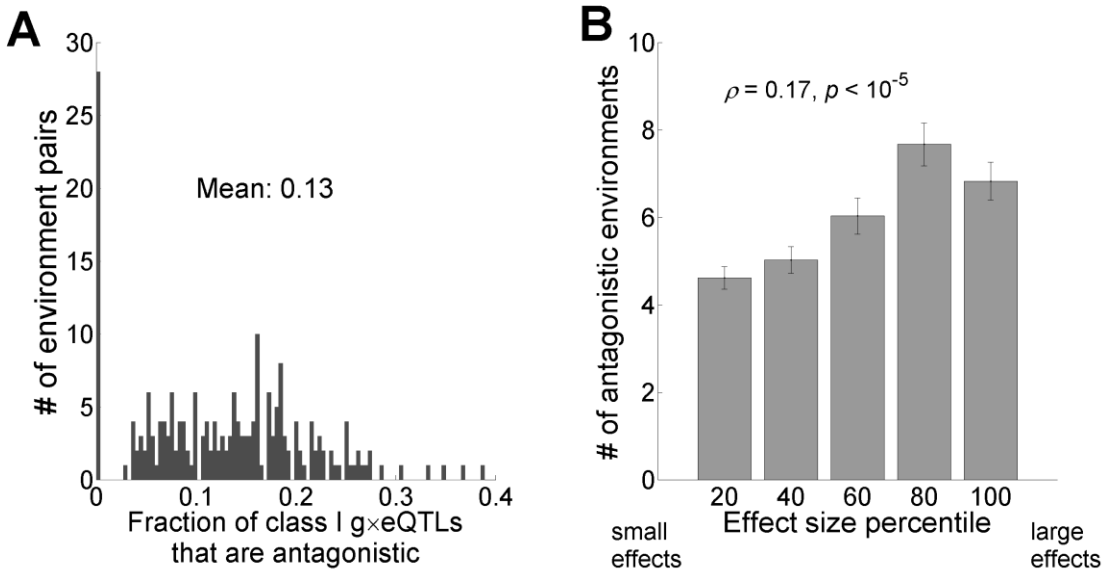


Figure C-3. Patterns of antagonistic G×E based on the data of Bloom et al. (2015). (A) Frequency distribution of the fraction of class I g×eQTLs that are antagonistic. (B) gQTLs with large effects in the environments where they are identified are more likely than small-effect gQTLs to have antagonistic effects in another environment. Error bars indicate one standard error.

Table C-1. Simulation results for g×eQTLs mapping with Q-value=0.005

Method	Narrow-sense heritability	0.75	0.63	0.43
gQTLs	False positive	16.8%	15.6%	14.1%
	False negative	0.03%	0.03%	0.43%
	Percent causal	48.9%	42.2%	31.4%
Class I g×eQTLs	False positive	1.43%	1.43%	1.42%
	False negative	28.5%	29.5%	31.9%
	Percent causal	45.8%	39.43%	29.6%
Direct mapped g×eQTLs	False positive	10%	9.3%	8.6%
	False negative	0.22%	0.73%	3.3%
	Percent causal	34.5%	28.0%	20.0%

False positive is counted when more SNPs are mapped as QTLs than the simulated number on each chromosome; false negative is counted when less SNPs are mapped than simulated number on each chromosome; percent causal is counted if the exact simulated site is identified as QTLs. The results are based on 1000 simulations.

Table C-2 Simulation results for g×eQTLs mapping with different Q-values

Q-value\Narrow-sense heritability	0.75	0.63	0.43
0.05	31:29:40	27:55:18	20:57:23
0.02	36:45:19	28:46:26	22:50:28
0.01	31:30:39	36:46:18	23:44:33
0.005	31:26:43	43:10:47	28:43:29
0.002	37:9:54	46:2:52	30:18:52
0.001	40:5:55	46:6:48	31:12:57

The ratio shows: the number simulations out of 100 that false positive gQTLs are smaller by our method: false positive gQTLs are smaller by method of Bloom et al: number of same mapping result

Table C-3. Distributions of gQTLs and class I g×eQTLs across various genomic regions based on the data from Bloom et al. (2015)

Genomic regions	All SNPs	All gQTLs		Class I g×eQTLs	
	Frequency	Frequency	*P*-value[1]	Frequency	*P*-value[2]
Intronic	0.008	0.006	0.4200	0.012	4.0×10^{-7}
Intergenic	0.331	0.326	0.4083	0.315	0.0313
Coding	0.656	0.666	0.2766	0.671	0.2071
Synonymous	0.558	0.474	2.5×10^{-4}	0.479	0.2431
Nonsynonymous	0.425	0.513	8.4×10^{-5}	0.505	0.1702
Nonsense	0.018	0.014	0.3375	0.016	0.0963

[1]Comparison with all SNPs using a binomial test.

[1]Comparison with all gQTLs using a binomial test.

Table C-4. Distributions of antagonistic and concordant class I g×eQTLs across various genomic regions based on the data from Bloom et al. (2015)

Genomic regions	Antagonistic		Concordant		*P*-value[1]
	Frequency	Occurrences	Frequency	Occurrences	
Intronic	0.0240	20	0.0095	49	2.5×10^{-4}
Intergenic	0.3197	266	0.3142	1625	0.7504
Coding	0.6454	537	0.6752	3492	0.0901
Synonymous	0.5587	300	0.4671	1631	$< 10^{-250}$
Nonsynonymous	0.4227	227	0.5175	1807	$< 10^{-250}$
Nonsense	0.0186	10	0.0155	54	0.4671

[1]Based on a chi-squared test.

File C-1. All gQTLs identified in each of the 47 environments

File C-2. All class I g×eQTLs identified in each of the 1081 environment pairs

File C-3. GO terms significantly enriched or deprived in gQTLs and g×eQTLs

File C-4. GO terms significantly enriched or deprived in gQTLs (tested against genes)

File C-5. Significantly overrepresented or underrepresented GO domains and terms in antagonistic g×eQTLs relative to concordant g×eQTLs

Appendix D: Supplementary figures and tables for chapter 3

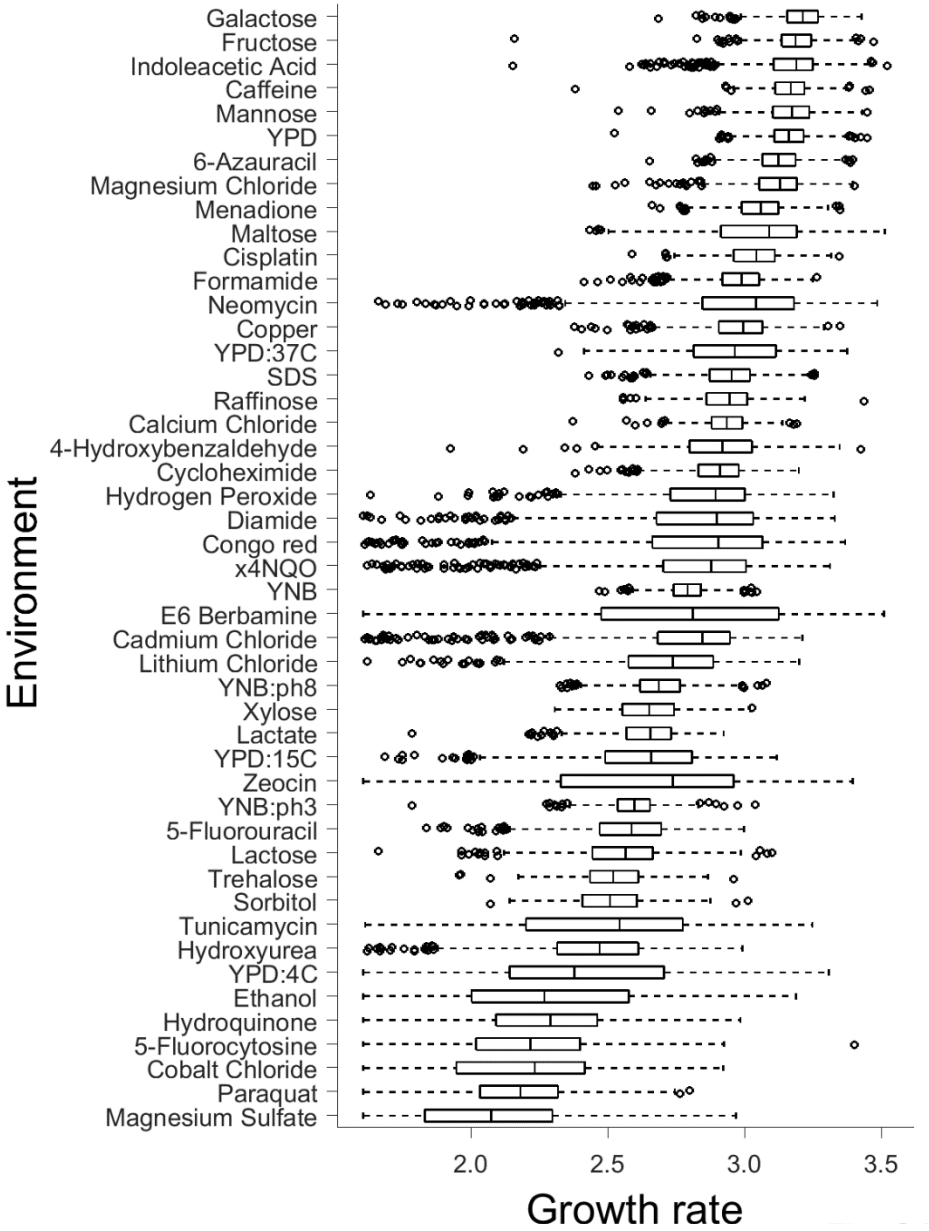


Figure D-1. Box plot of growth rates of segregants in each environment, where the left and right edges of a box represent the first (qu_1) and third (qu_3) quartiles, respectively, the horizontal line inside the box indicates the median (md), the whiskers extend to the most extreme values inside inner fences, qu_1-1.5(qu_3-qu_1) and qu_3+1.5(qu_3-qu_1), and the circles represent values outside the inner fences (outliers). The environments are ordered from low (bottom) to high (top) mean growth rate.

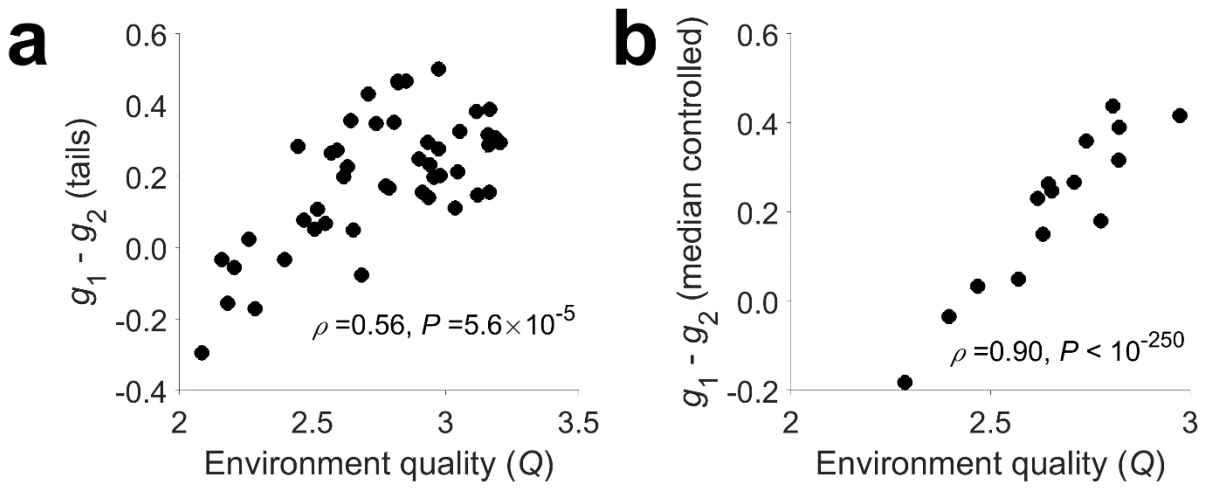


Figure D-2. Widespread narrow-sense diminishing returns among standing genetic variants in yeast. Here, g_1 is the fraction of SNPs showing $s_H < s_L$ and having the same beneficial allele in the slow- and fast-growth segregants, while g_2 is the fraction of SNPs showing $s_L < s_H$ and having the same beneficial allele in the slow- and fast-growth segregants. Diminishing returns epistasis is general if g_1-$g_2 > 0$. (**a**) Estimates of g_1-g_2 in each environment when g_1 and g_2 are estimated using the method in Fig. 1a. We note that g_1-g_2 is positive in 40 of the 47 environments examined ($P < 10^{-6}$, $N = 47$, binomial test). The same is true in 32 of 44 environments when only QTLs are considered ($P = 6.3\times10^{-4}$, $N = 44$, binomial test; three of the 47 environments are not considered either because g_1-$g_2 = 0$ or because no QTL is mapped). (**b**) Estimates of g_1-g_2 in 15 environments that can be studied when g_1 and g_2 are estimated using the method in Fig. 1c. We note that g_1-g_2 is positive in 13 of the 15 environments ($P = 4.9\times10^{-4}$, $N =$ 15, binomial test). The same is true in 11 of the 15 environments when only QTLs are considered ($P = 0.018$, $N = 15$, binomial test). In addition, a strong positive correlation between g_1 and Q is observed, regardless of whether g_1 is estimated using the method of Fig. 1a ($\rho = 0.56$, $P = 5.6\times10^{-4}$) or that of Fig. 1c ($\rho = 0.90$, $P < 10^{-250}$). The corresponding correlations are $\rho = 0.53$ ($P = 1.8\times10^{-4}$) and 0.63 ($P = 0.0091$) when only QTLs are considered.

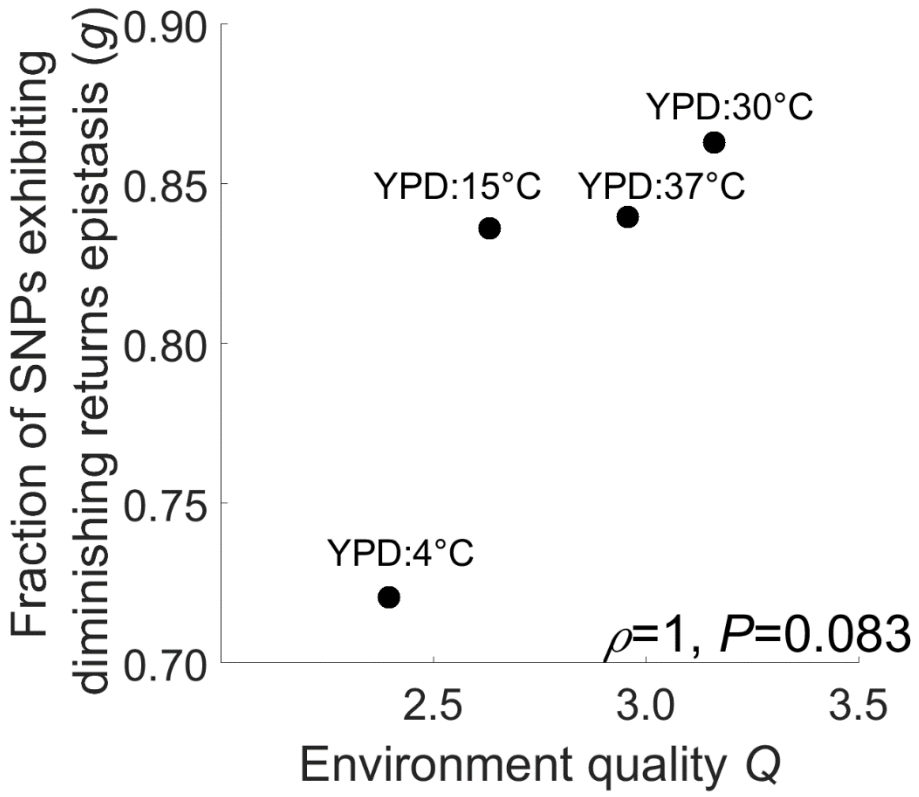


Figure D-3. Fraction of SNPs exhibiting diminishing returns epistasis increases monotonically with environment quality among the four YPD environments that differ in temperature.

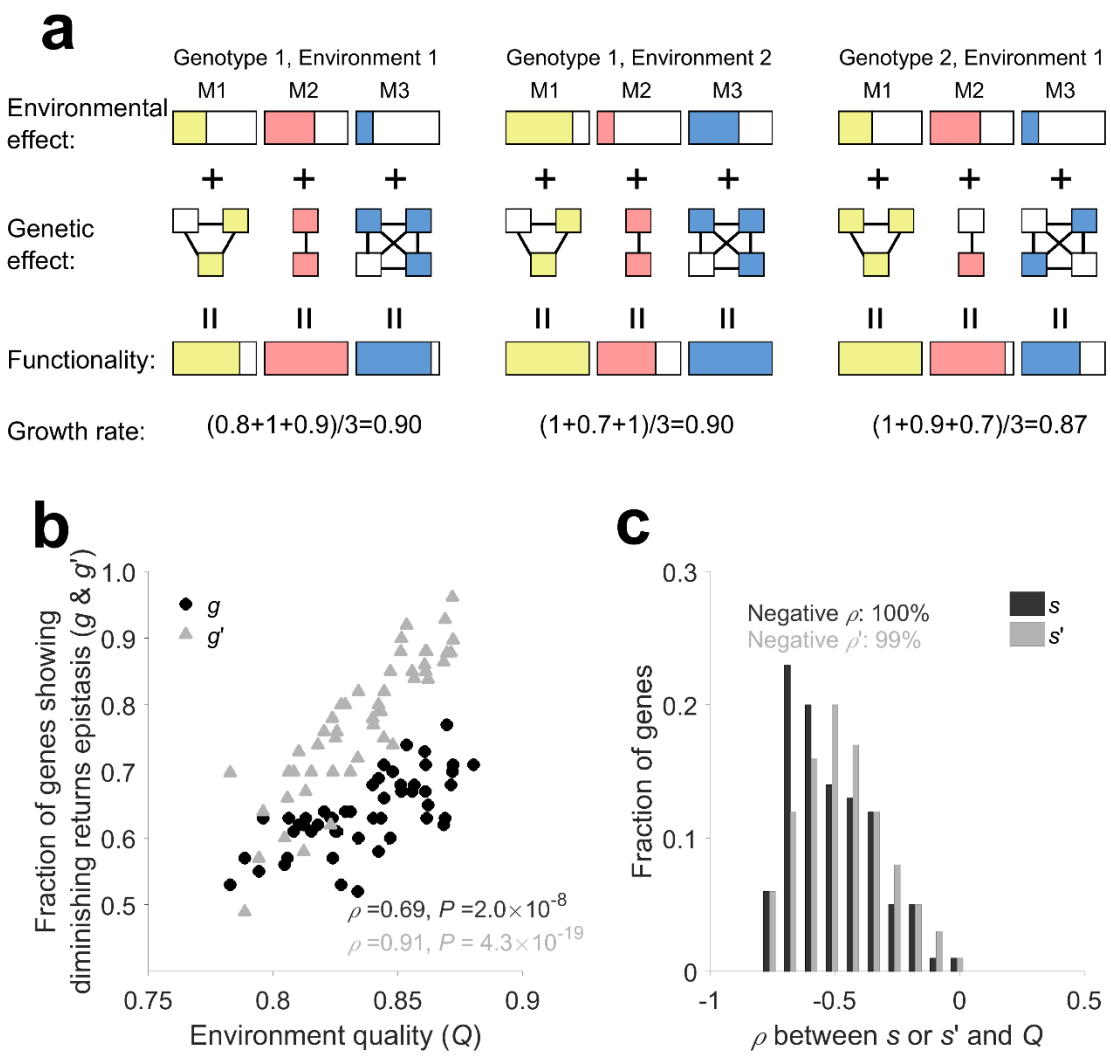


Figure D-4. Simulation of the modular life model in which growth rate equals the arithmetic mean functionality of all modules produces diminishing returns patterns resembling empirical observations. Parameters used in the simulation are the same as those in Fig. 3, except for the following. The maximal contributions of the 10 genes to the functionality of a module are set to be 0.088, 0.096, 0.104, …, and 0.16, respectively. We assume that the functionality contribution of an environment to a module follows a normal distribution with a standard deviation of 0.05. The mean of the normal distribution is 0.2000, 0.2021, 0.2042, ..., and 0.3029, respectively, from the 50 environments. We also added a noise term drawn randomly from the normal distribution

of mean = 0 and standard deviation = 0.008 to the growth rate of each simulated genotype in each environment. (**a**) Simulation scheme. Different modules (M1, M2, and M3) are colored differently. Different environments (Environments 1 and 2) contribute differently to various modules, as illustrated by the three boxes that are filled to different levels. Each module contains a number of genes, each having either a functional allele designated as 1 (filled box) or a null allele designated as 0 (open box). Two genotypes (Genotypes 1and 2) are shown as examples. The functionality of a module equals the sum of environmental and genetic contributions or 1, whichever is smaller. The growth rate of each genotype is computed from the functionalities of the individual modules using the formula indicated. (**b**) Simulation results show that the fraction of genes exhibiting diminishing returns epistasis (g or g') positively correlates with environment quality (Q). Black dots show estimates of g on the basis of the fittest and least fit segregants, whereas grey triangles show estimates of g' from segregants of fixed median growth rates. (**c**) Frequency distribution of the rank correlation (ρ) between Q and the effect of a SNP measured using either all segregants (s; black) or a group of segregants with a fixed median growth rate (s'; grey). The fraction of ρ's that are negative is indicated in black and grey for s and s', respectively. Here, s and s' could be negative if the functional allele is found less fit than the null allele (due to sampling error).

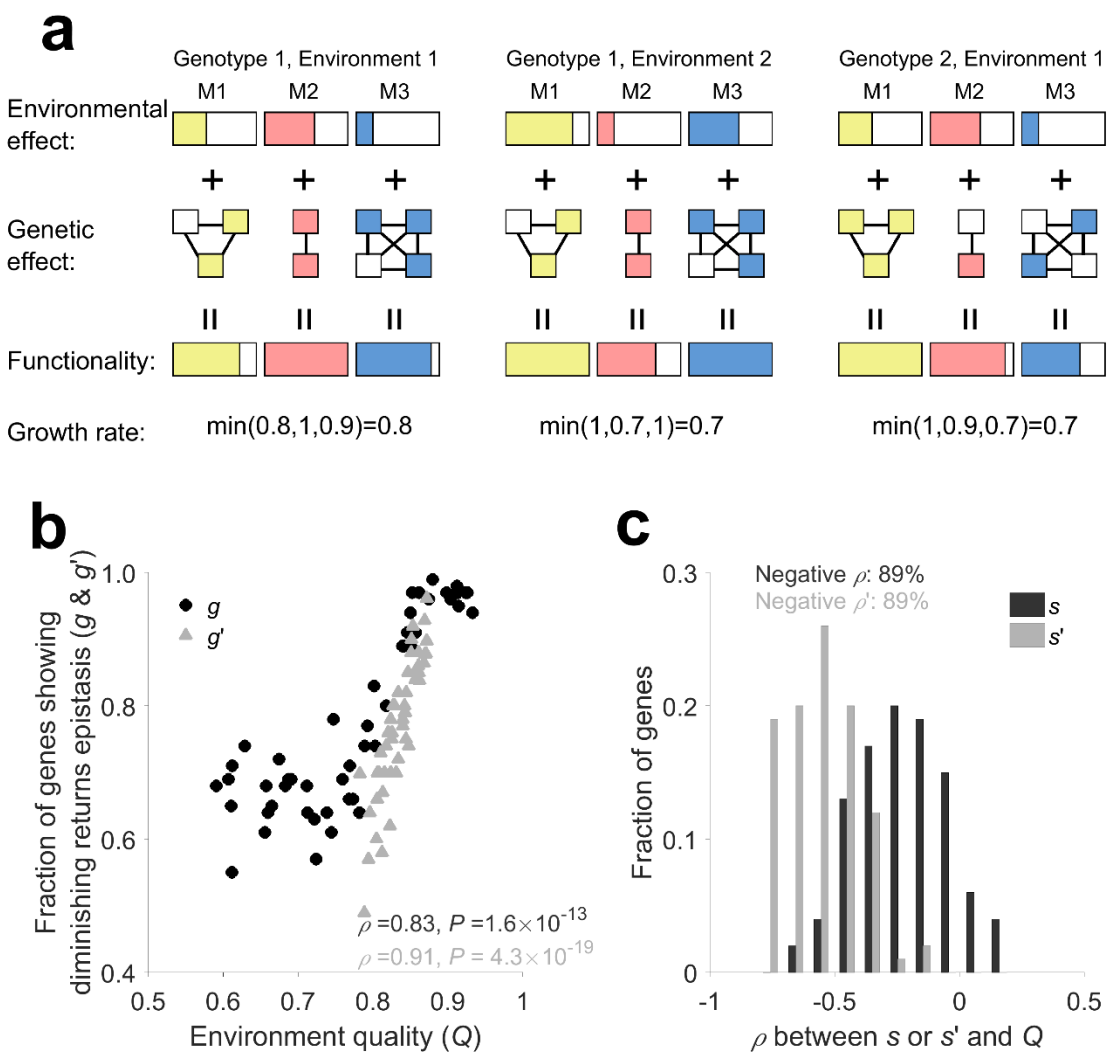


Figure D-5. Simulation of the modular life model in which growth rate equals the lowest functionality of all modules produces diminishing returns patterns resembling empirical observations. Parameters used in the simulation are the same as those in Fig. 3, except for the following. The maximal contributions of the 10 genes to the functionality of a module are set to be 0.088, 0.096, 0.104, …, and 0.16, respectively. The functionality contribution of an environment to a module follows a normal distribution with a standard deviation of 0.05. The mean of the normal distribution is 0.300, 0.307, 0.314, ..., and 0.643, respectively, from the 50 environments. We also added a noise term drawn randomly from the normal distribution of

mean = 0 and standard deviation = 0.008 to the growth rate of each simulated genotype in each environment. (**a**) Simulation scheme. Different modules (M1, M2, and M3) are colored differently. Different environments (Environments 1 and 2) contribute differently to various modules, as illustrated by the three boxes that are filled to different levels. Each module contains a number of genes, each having either a functional allele designated as 1 (filled box) or a null allele designated as 0 (open box). Two genotypes (Genotypes 1 and 2) are shown as examples. The functionality of a module equals the sum of environmental and genetic contributions or 1, whichever is smaller. The growth rate of each genotype is computed from the functionalities of the individual modules using the formula indicated. (**b**) Simulation results show that the fraction of genes exhibiting diminishing returns epistasis (g or g') positively correlates with environment quality (Q). Black dots show estimates of g on the basis of the fittest and least fit segregants, whereas grey triangles show estimates of g' from segregants of fixed median growth rates. (**c**) Frequency distribution of the rank correlation (ρ) between Q and the effect of a SNP measured using either all segregants (s; black) or a group of segregants with a fixed median growth rate (s'; grey). The fraction of ρ's that are negative is indicated in black and grey for s and s', respectively. Here, s and s' could be negative if the functional allele is found less fit than the null allele (due to sampling error).

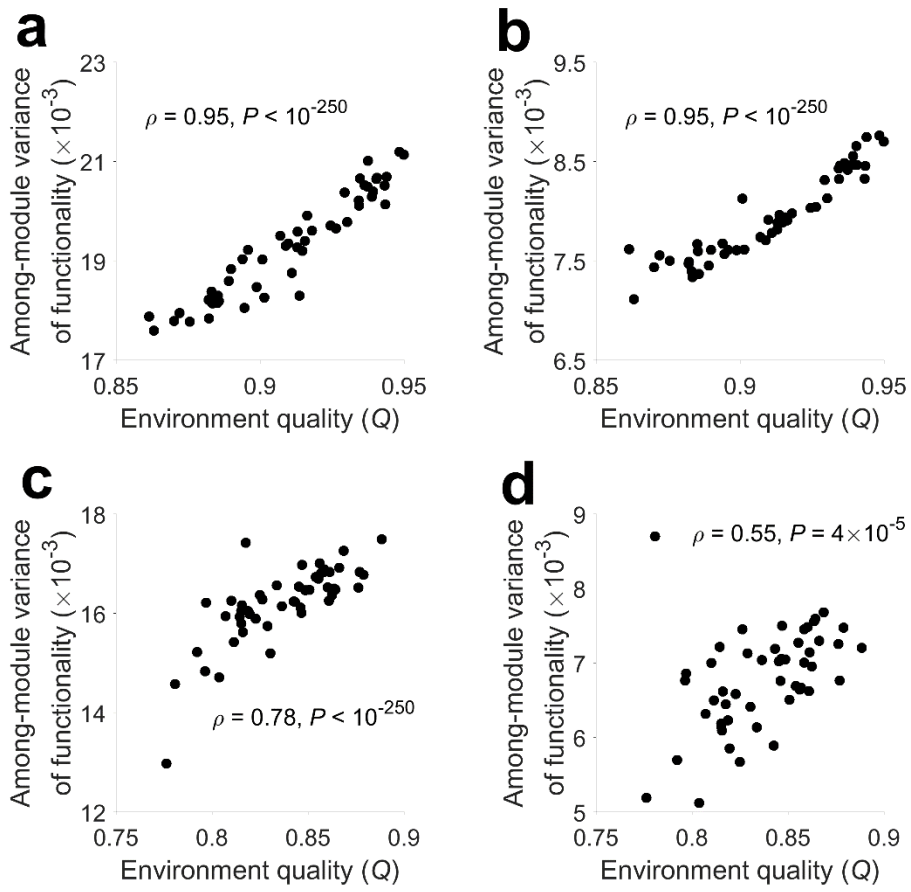


Figure D-6. Among-module variance of functionality in simulated segregants increases with environment quality. (**a**) Among-module variance of functionality increases with environment quality when growth rate is defined by the geometric mean functionality of all modules and is in the range between 0.899 and 0.901. (**b**) Among-module variance of functionality increases with environment quality when growth rate is defined by the geometric mean functionality of all modules and is in the range between 0.949 and 0.951. (**c**) Among-module variance of functionality increases with environment quality when growth rate is defined by the arithmetic mean functionality of all modules and is in the range between 0.899 and 0.901. (**d**) Among-module variance of functionality increases with environment quality when growth rate is defined by the arithmetic mean functionality of all modules and is in the range between 0.949 and 0.951.

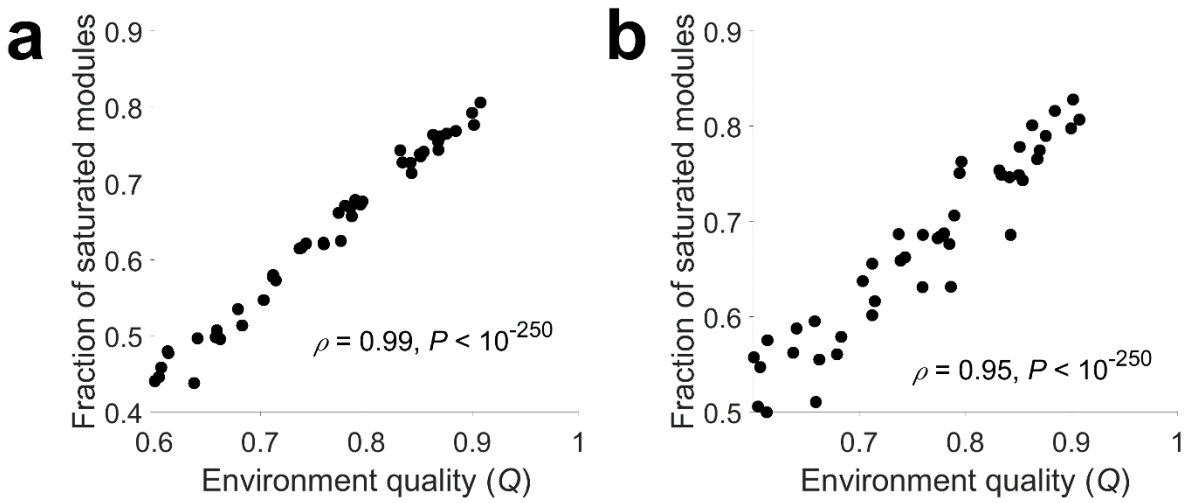


Figure D-7. Fraction of modules with saturated functionality increases with environment quality (Q) when growth rate is defined by the lowest functionality across modules in simulated segregants. (**a**) Fraction of saturated modules increases with Q when growth rate is in the range from 0.799 to 0.801. (**b**) Fraction of saturated modules increases with Q when growth rate is in the range from 0.849 to 0.851.

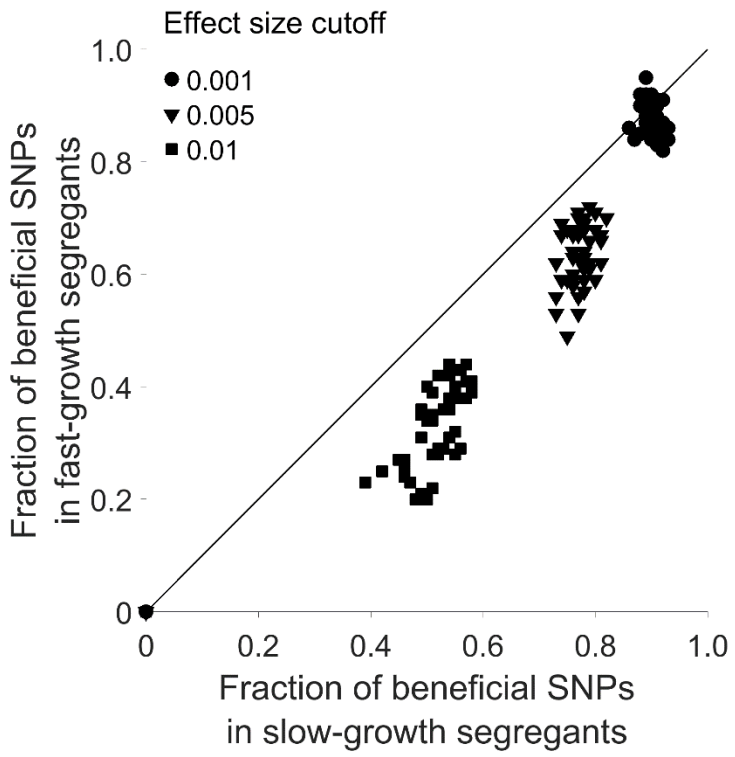


Figure D-8. Fraction of SNPs that can be considered beneficial in the data simulated under the modular life model with geometric mean growth rate. A beneficial mutation must have a growth rate effect greater than the effect size cutoff indicated to be considered beneficial. Three different cutoffs are considered, respectively. Under each cutoff, each symbol represents an environment. The number of symbols below the diagonal is greater than that above the diagonal for each cutoff considered ($P < 3.5\times10^{-5}$ for all cutoffs, $N = 50$, binomial test), demonstrating that the modular life model generates the phenomenon of decreasing supplies of beneficial mutations as the growth rate of the background genotype rises. Under each environment, for each SNP considered, slow-growth segregants refer to the 50 least fit segregants carrying the functional allele and the 50 least fit segregants carrying the null allele; fast-growth segregants refer to the 50 fittest segregants carrying the functional allele and the 50 fittest segregants carrying the null allele.

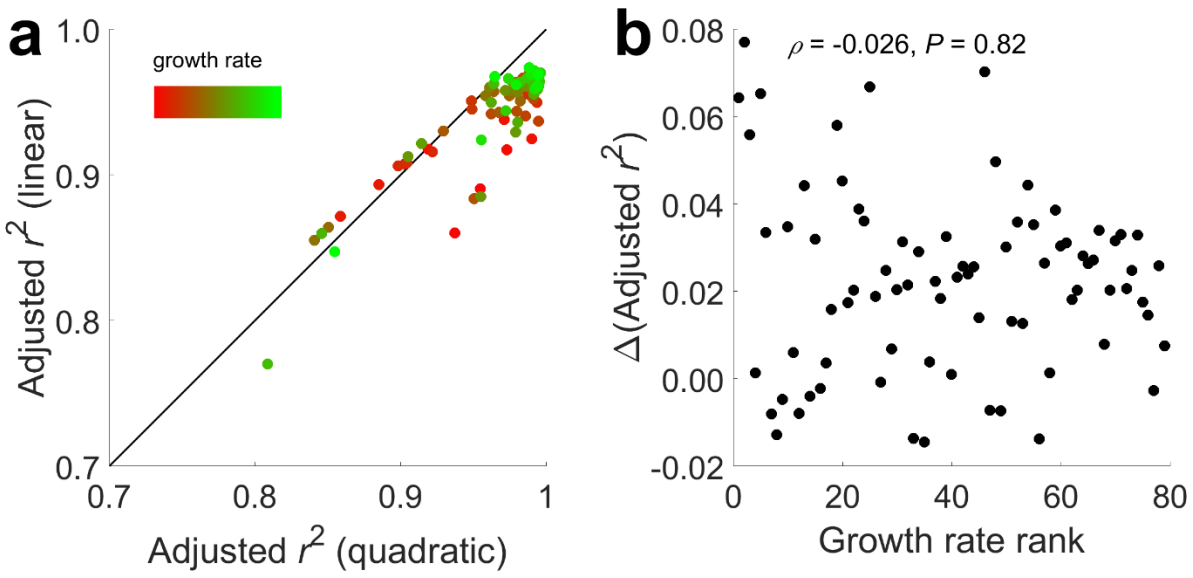


Figure D-9. Assessment of yeast growth saturation and its impact on the analysis of diminishing returns using 79 randomly picked segregants. (**a**) Adjusted r^2 values in linear and quadratic models that respectively describe the relation between growth time and ln(colony radius). Each dot represents one genotype, with the color showing the growth rate rank determined at 48h (faster growth genotypes have larger ranks and are greener). The diagonal line indicates equal adjusted r^2 values of the two models. (**b**) Absence of significant correlation between the difference in adjusted r^2 of the two models and the growth rate rank. Δ(Adjusted r^2) = quadratic adjusted r^2 - linear adjusted r^2.

Table D-1. The correlation between *g'* and *Q* is robust for different F_L and F_H

Low median	High median	# of environments	% of environments with >50% BSDR	Fig.1D ρ	Fig.1D *P*
1.186	1.221	9	100%	0.52	0.16
1.206	1.242	9	100%	0.82	0.01
1.227	1.263	15	100%	0.90	4.9×10^{-4}
1.247	1.284	9	100%	0.82	0.01
1.268	1.306	4	100%	0.80	0.33

Table D-2. Rank correlations between *s'* and *Q* are robust to *F*

Median	#% $\rho<0$
1.201	59.5%
1.232	85.7%
1.263	98.9%

Appendix E: Supplementary figures and tables for chapter 4

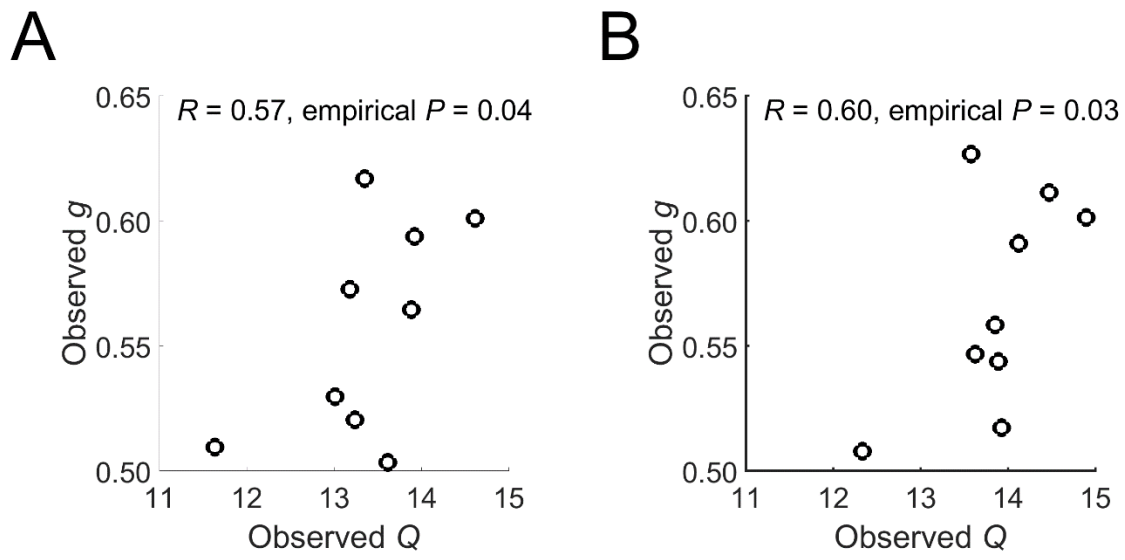


Fig. E-S1. Positive correlation between environmental quality and the fraction of SNPs showing $h < 0.5(g)$. (A) The observed result at 32h between Q and g. (B) The observed result at 48h between Q and g. Each dot represents one environmental condition. Linear correlation coefficient R and empirical P (from 1000 random shuffling of x- and y- axes numbers) are listed.

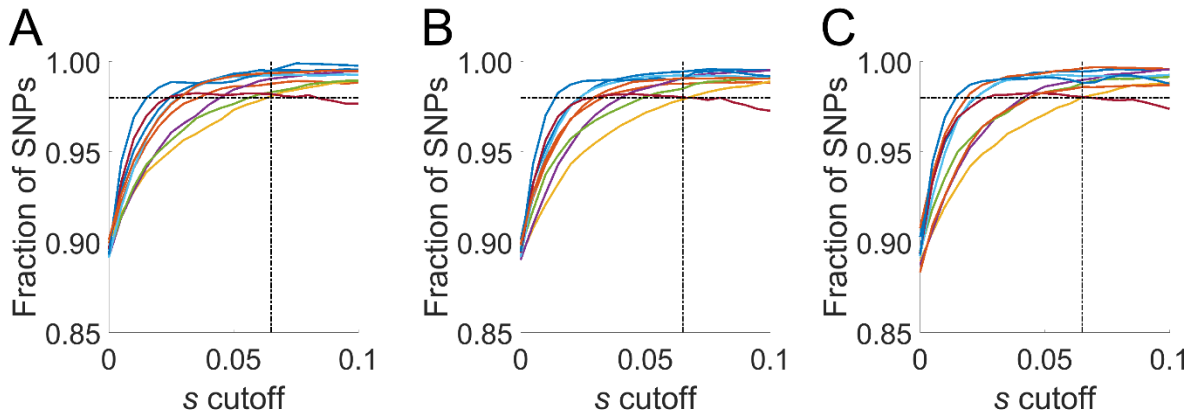


Fig. E-S2. Cutoff for *s* for each time point. (A) The fraction of remaining SNPs satisfying $0 \leq h \leq 1$ at 32h for different cutoffs of *s*. (B) The fraction of remaining SNPs satisfying $0 \leq h \leq 1$ at 40h for different cutoffs of *s*. (C) The fraction of remaining SNPs satisfying $0 \leq h \leq 1$ at 48h for different cutoffs of *s*. Different environments are colored differently. X-axis is the s cutoff used, and y-axis is the fraction of remaining SNPs satisfying $0 \leq h \leq 1$ for each cutoff. The vertical dashed line is x = 0.065. The horizontal dashed line is y = 0.98.

Appendix F: Supplementary figures and tables for chapter 6

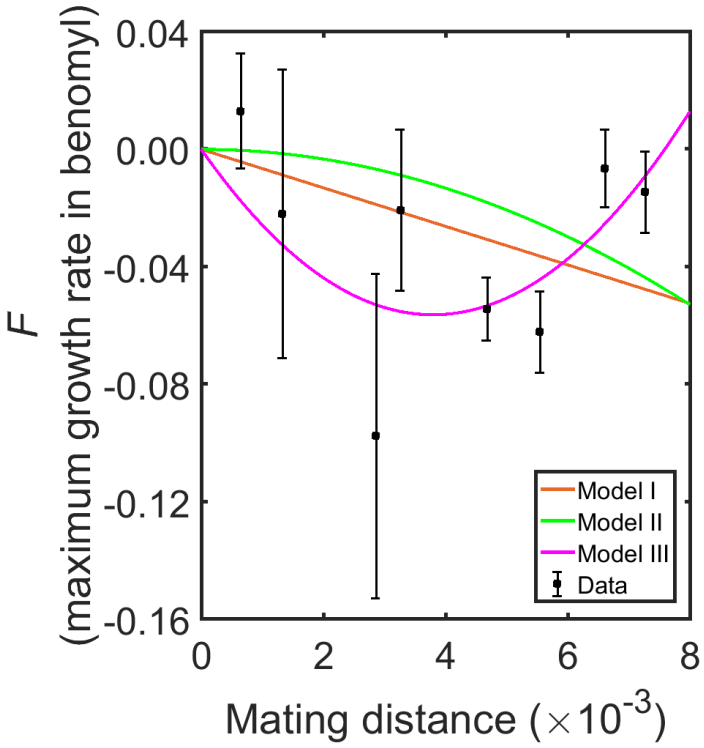


Fig F-1. Hump-shaped relationship between *S. cerevisiae* mating distance (*D*) and hybrid performance (*F*) measured by maximum growth rate in the benomyl medium. The mean and standard error of *F* are respectively shown by black squares and associated error bars. The fitted *D*-*F* curves under different models are shown in different colors.

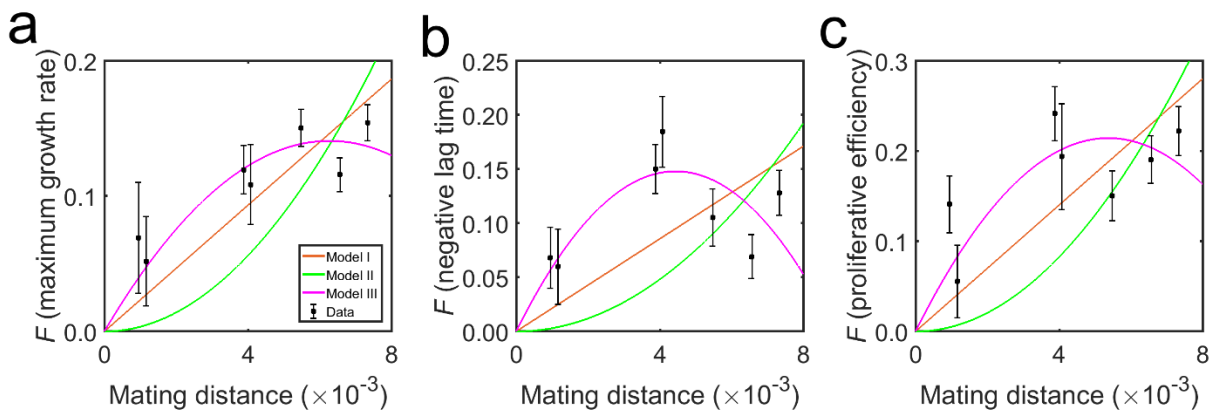


Fig F-2. Hump-shaped relationship between *S. cerevisiae* mating distance (*D*) and hybrid performance (*F*) in (**a**) maximum growth rate, (**b**) negative lag time, and (**c**) proliferative efficiency averaged across 56 environments. The mean and standard error of *F* are respectively shown by black squares and associated error bars. The fitted *D*-*F* curves under different models are shown in different colors.

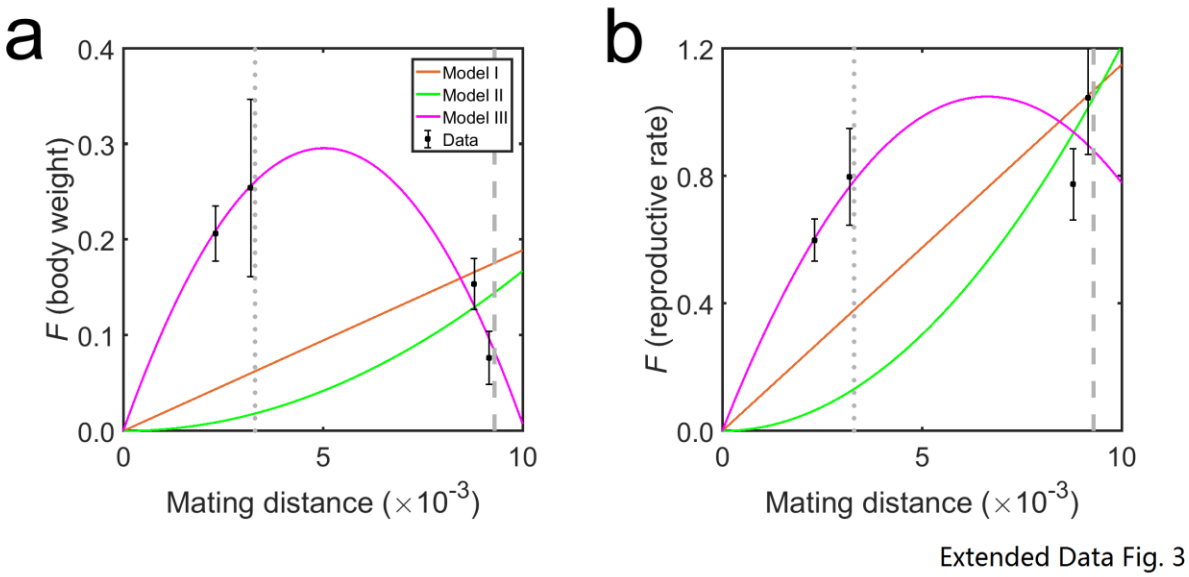


Fig F-3. Hump-shaped relationship between *Mus musculus* mating distance (*D*) and hybrid performance (*F*) in (**a**) body weight and (**b**) reproductive rate. The mean and standard error of *F* are respectively shown by black squares and associated error bars. The fitted *D*-*F* curves under different models are shown in different colors. π and D_{max} are respectively indicated by vertical dotted and dashed lines.

Table F-1. Fitting of the three models to *A. thaliana* data using alternative window sizes

Window size ($\times 10^{-3}$)	Traits	Models	R^2	$2\Delta \ln L$[1]	P-value[2]	OMD [95% CI] ($\times 10^{-3}$)
0.6	Shoot weight					
		I	-3.61	31.6	1.8×10^{-8}	
		II	-9.76	80.8	2.5×10^{-19}	
		III	0.34			5.0 [4.7-6.7]
	Rosset diameter					
		I	-2.38	22.4	2.2×10^{-6}	
		II	-6.66	56.6	5.2×10^{-14}	
		III	0.42			5.0 [4.7-6.8]
	Leaf area					
		I	-3.66	32.8	1.0×10^{-8}	
		II	-9.62	80.4	3.0×10^{-19}	
		III	0.44			4.9 [4.6-6.5]
	Leaf number					
		I	-1.62	16.0	6.2×10^{-5}	
		II	-3.78	33.3	7.9×10^{-9}	
		III	0.39			4.5 [4.4-7.3]
1.0	Shoot weight					
		I	-6.45	35.4	2.7×10^{-9}	
		II	-23.67	121.5	3.0×10^{-28}	
		III	0.63			5.7 [5.0-7.5]
	Rosset diameter					
		I	-3.55	20.3	6.8×10^{-6}	
		II	-11.39	59.4	1.3×10^{-14}	
		III	0.50			5.3 [4.8-6.3]
	Leaf area					
		I	-8.98	48.8	2.9×10^{-12}	
		II	-27.18	139.8	3.0×10^{-32}	
		III	0.78			5.2 [4.7-6.0]
	Leaf number					
		I	-2.38	13.8	2.1×10^{-4}	
		II	-8.55	44.7	2.3×10^{-11}	
		III	0.38			5.5 [4.4-11.2]

[1]Twice the difference in ln(likelihood) between Model III and the model being compared.

[2]*P*-values of likelihood ratio tests are determined using chi-squared tests with 1 degree of freedom

Table F-2. Fitting of the three models to the *S. cerevisiae* data (averaged across 11 environments) using alternative window sizes

Window size($\times 10^{-3}$)	Models	R^2	$2\Delta \ln L$[1]	P-value[2]	OMD [95% CI] ($\times 10^{-3}$)
0.25	Model I	-0.13	11.7	6.3×10^{-4}	
	Model II	-0.40	19.0	1.3×10^{-5}	
	Model III	0.31			3.9 [3.7-4.2]
0.50	Model I	-0.25	11.2	8.0×10^{-4}	
	Model II	-0.51	15.2	9.6×10^{-5}	
	Model III	0.50			3.6 [3.5-3.9]
0.75	Model I	-0.24	8.6	3.3×10^{-3}	
	Model II	-0.41	10.6	1.1×10^{-3}	
	Model III	0.55			3.5 [3.3-3.7]

[1]Twice the difference in ln(likelihood) between Model III and the model being compared.

[2]*P*-values of likelihood ratio tests are determined using chi-squared tests with 1 degree of freedom.

Table F-3. Fitting of the three models to Zorgo et al.'s yeast data

Traits	Models	Adjusted R^2	$2\Delta\ln L$ [1]	*P*-value [2]	OMD [CI 95%] ($\times 10^{-3}$)
Rate					
	Model I	0.17	3.1	0.080	
	Model II	-1.47	12.6	3.8×10^{-4}	
	Model III	0.63			6.3 [4.9-14.5]
Negative lag time					
	Model I	-1.10	8.8	3.0×10^{-3}	
	Model II	-2.70	18.1	2.1×10^{-5}	
	Model III	0.26			4.4 [4.0-5.3]
Efficiency					
	Model I	-0.56	5.0	0.025	
	Model II	-2.29	15.1	1.0×10^{-4}	
	Model III	0.13			5.4 [4.5-8.6]

[1]Twice the difference in ln(likelihood) between Model III and the model being compared.

[2]*P*-values of likelihood ratio tests are determined using chi-squared tests with 1 degree of freedom.

Table F-4. Fitting of the three models to *Mus musculus* data

Traits	Models	R^2	$2\Delta \ln L$[1]	P-value[2]	OMD [95% CI] ($\times 10^{-3}$)
Body weight					
	I	-3.23	16.7	4.4×10^{-5}	
	II	-4.75	22.8	1.8×10^{-6}	
	III	0.95			5.1 [5.0-15.3]
Reproductive rate					
	I	-2.49	12.0	5.4×10^{-4}	
	II	-6.59	28.4	1.0×10^{-7}	
	III	0.51			6.6 [5.2-7.5]

[1]Twice the difference in ln(likelihood) between Model III and the model being compared.

[2]*P*-values of likelihood ratio tests are determined using chi-squared tests with 1 degree of freedom.