THE UNIVERSITY OF CHICAGO

ON THE MECHANISMS OF GENOMIC AND PHENOTYPIC EVOLUTION

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

COMMITTEE ON GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY
UNJIN LEE

CHICAGO, ILLINOIS
DECEMBER 2021

This work is dedicated to my late wife, Frances Jee Hee Lee. This work would not have been possible without her loving care, dedication, and support.

"It is natural selection that gives direction to changes, orients chance, and slowly, progressively produces more complex structures, new organs, and new species. Novelties come from previously unseen association of old material. To create is to recombine."

— Francois Jacob, 1977

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I'd like to first thank my middle school science teacher for not only tolerating my incessant questions, but also for being knowledgeable and patient enough to provide thoughtful and insightful answers. I would also like to thank Dr. Marsha Rosner and Dr. John Reinitz for their mentorship in guiding me towards the biological sciences as an undergraduate. I would also like to thank my advisor, Dr. Manyuan Long, for giving me the freedom to enjoy science, and my committee, Dr. Martin Kreitman, Dr. Marcelo Nobrega, Dr. Edwin Ferguson, and Dr. John Reinitz for their guidance. I would also like to thank especially Dr. Marcelo Nobrega for allowing me to use his resources. I would also like to thank Dr. Débora Sobreira for her immensely kind and overwhelming generosity and knowledge at the bench. I would also like to thank Dr. Kenneth Barr for helping me make the transition to performing experiments.

I would like to also thank Drs. Linsey Montefiori, Matt Hope, Catherine Wu, Charlene Hoi, and Will Yee for experimental guidance and reagents. I would also like to thank Dr. Ittai Eres for his expertise in Hi-C processing. I would also like to thank Dr. Qi Zhou, Mujahid Ali, Dr. Peter Andolfatto, and Dr. Patrick Reilly for generously sharing their Hi-C data as well as high-quality reference genomes. I would also like to thank the Long lab: Emily Mortola, Daniel Downie, and Drs. Nick VanKuren, Andrea Gschwend, Alex Advani, Jing Yang, Claus Kemkemmer, Jianhai Chen, Li Zhang, Iuri Ventura, Jian Zu, Shengqian Xia, Tansheng Jun, and Deanna Arsala, for fostering a great scientific atmosphere.

I would also like to thank my family and friends for supporting me through difficult times.

ABSTRACT

During the process of adaptation, phenotypic variation present within a population provides the substrate on which selection may act. However, in light of epigenetic processes, the manner in which such variation is produced and the consequences of such variation is of central importance in understanding adaptive processes. In this dissertation, I consider two models of how this variation may arise, either through random, plastic means, or via a rapid recombination of pre-existing regulatory and protein-coding elements within the genome.

While previous models of phenotypic plasticity have generally fallen into two classes of models, either considering adaptive or non-adaptive plasticity, I present a new unified model of phenotypic plasticity using stochastic differential equations and *in silico* selection. Though prior models have suggested that plasticity is deleterious during static conditions, I demonstrate how the relative evolvability of genic or epigenetic control of phenotype can determine whether plasticity will become maintained during stasis.

Additionally, I analyze a model of new gene evolution via enhancer capture, where a new gene may adopt the expression patterns dictated by the regulatory environment into which the gene duplicates, evaluating and comparing it to other major models duplication-based evolution. By comparing the expression patterns of newly evolved essential and non-essential genes, I demonstrate that enhancer capture is likely a significant driver of the evolution of distally duplicated genes via enhancer capture. I then utilize genomic techniques, integrating RNA-seq, ChIP-seq, and Hi-C data, to show that a new essential gene, HP6/Umbrea, is one example of a gene that has evolved in this manner.

Altogether, this dissertation encapsulates the breadth of methods used in evolutionary genetics, providing both a theoretical analysis of a novel model of phenotypic plasticity as well as an experimental validation of an old model of new gene neo-functionalization made possible only through nascent technological development, expanding our understanding of the methods and mechanisms by which phenotypic variation arises.

CHAPTER 1

INTRODUCTION

For millennia, humans have pondered and marveled at the great diversity of life found in nature, seeking to understand our own role within this diversity. Given mankind's compulsive quest for reason and purpose, explanations for this diversity have driven much of man's technological development. Coined in 1937 by Theodosius Dobzhansky [1], the field of evolutionary genetics is mankind's greatest tool in understanding this profound diversity. While it is difficult to truly understand the distant past, evolutionary genetics allows us to take what is to understand what was, using observations of the present to provide a window into the deep, unobservable past. More than being simple comparisons of phenotypes across phyla, the marriage of molecular genetics techniques with the simple-but-powerful mathematical models of population genetics has yielded tremendous insight into the methods and mechanisms by which this diversity arises. As Dobzhansky wrote, "nothing in biology makes sense except in the light of evolution." [2]

The central engine to the advancement of any scientific field is the development of new technology. As science is an inherently predictive process, the advent of new methods allows for the testing and refinement of old models. And while much of the biological sciences is rooted in descriptive science, unlike any other biological discipline, evolutionary genetics remains singular in its predictive power. Indeed, Mendel could never have predicted the direct observation of chromosomes, and the likes of Fisher, Haldane, or Wright could never have predicted the advent of DNA technology. Yet without even a remote hope of the kinds of validation their theories would later receive, their ideas persisted and remained influential not on the merits of overwhelming, definitive proof, but on the explanatory power of their inductive reasoning.

1.1 Plasticity and Evolution

While the Evolutionary Synthesis produced a new paradigm where mutational variation produces selectable phenotypic variation [1, 3, 4], there is little room in this framework for understanding the evolutionary role of epigenetic control of phenotype and phenotypic plasticity. From J. Mark Baldwin's landmark 1896 manuscript describing how the acquisition and learning of new behaviors could lead to the possibility of phenotype-first evolution [5] to C. H. Waddington's 1952 experiments demonstrating the genetic accommodation of an environmentally induced cross-veinless phenotype in a population of flies [6], study of the potential role of non-genetic variation in evolution has been both long and controversial [7, 8], with some contending that such effects are non-existent or inconsequential to adaptation [9, 10].

1.2 Plasticity and Survival

In models of natural selection, an individual's degree of fitness depends upon how well it has adapted to particular external factors, such as the environment in which it lives. Changes in its situation can cause such individuals to become less fit, as phenotypes that are beneficial in one scenario might be ill-suited to another. Nevertheless, unpredictable changes and fluctuations are inevitable, and organisms therefore develop strategies that allow them to cope with such stresses [7, 8]. One mechanism that could help organisms to resist stress and increase fitness in changing surroundings is phenotypic plasticity, the phenomenon by which a single genotype can manifest as a range of phenotypes. In the case of non-adaptive plasticity, this variation among phenotypes is apparent even in the absence of any external stresses, while in the case of adaptive plasticity, variation among phenotypes occurs subsequent to an environmental shift in a manner that increases individual fitness.

1.3 A Need for More

According to the Evolutionary Synthesis, all evolutionarily relevant phenotypic variation is the result of genetic variation. As such, quantitative genetic models of natural selection do not incorporate the existence of non-genetic phenotypic variability [11], instead, portraying one-to-one genotype-phenotype relationships, where a particular genotype (or additive combination of many genetic loci) corresponds to a single phenotype. With the exception of theoretical work directly concerning phenotypic plasticity, most evolutionary models continue to maintain a one-to-one genotype-phenotype relationship. However, recent observations indicate the existence of variable relationships between genotype and phenotype, in which a given genotype may in fact produce a range of phenotypes that can fluctuate dynamically prior to an environmental shift (non-adaptive plasticity), and/or subsequently in response to said shift (adaptive plasticity) [7, 8]. Even within clonal progeny, individuals have been shown to display quantitative differences that may distinguish them from genetically identical individuals. Such differences are important for survival, as in the context of changing environmental conditions, possessing variable phenotypes within even genetically similar populations generates the potential for individuals to be randomly suited to uncertain fluctuations. While prior models have incorporated such effects [11–20], these models remain divided, capturing aspects of either non-adaptive or adaptive plasticity, but not both.

As environmental change plays a key role in understanding the evolution of phenotypic plasticity, currently, the dominating view of plasticity contends that it may only evolve under conditions of environmental fluctuations [21, 22], and that the degree of plasticity in a population should decrease during periods of stasis [11, 13, 23, 24]. However, in light of the abundance of plasticity found in nature, it seems unlikely that plasticity may not evolve under static environmental conditions. In Chapter II I challenge these findings via a unified model of plasticity incorporating both non-adaptive and adaptive plasticity effects, leading

to the following questions: What is the relationship between plasticity (adaptive and non-adaptive) and genic control of phenotype? What are the effects of plasticity on adaptation? Under what conditions may plasticity be favorable and thus evolve in a population?

1.4 An Old Theory of New Gene Evolution

One of the earliest models of new gene evolution was first proposed by Hermann Muller in 1936 [25]. Though he lacked our current molecular understanding of gene regulation, by observation of the double-bar/ultrabar phenotype, he suggested that positional effects on duplicate gene copies may be a source of evolutionary novelty. Indeed, in more recent times, the development of sequencing technology and the assembly of a panel of reference genomes has allowed for a more systematic analysis of new gene origination, while positional effects caused by random integration into the genome has become known as a driving force behind the co-expression of genes with the same chromatin domain.

Using high-quality reference genomes, genes arising from the class of duplication-based mechanisms can now categorically be inferred through synteny- and homology-based searches. Though new genes are systemically understudied in comparison to their older counterparts, new genes originating from duplication-based mechanisms have been relatively well-described and studied, both experimentally and theoretically. However, in studying the evolutionary dynamics of duplication-based origination, a paradox arises: how do functionally redundant copies of the same gene rise to fixation? This paradox has been resolved using various models of duplication-based evolution, including the neofunctionalization model [26], the duplication, divergence, complementation (DDC)/sub-functionalization model [27], the escape from adaptive conflict (EAC) model [28], and the innovation, amplification, and divergence (IAD) model [29, 30].

1.5 Solving a Paradox

Outside of the neofunctionalization model, where a novel function arises in one of the two duplicate copies, the aforementioned models variously invoke alterations of regulatory and/or protein function in resolving this paradox. The DDC/sub-functionalization model allows for complementary degeneration of the functions shared between the paralogous copies. Here, the ability of one gene copy to compensate for loss of function in another allows for the preservation of both gene copies, eventually resulting in the segregation and partitioning of multiple sub-functions to each duplicated copy. Alternatively, while the DDC model merely distributes and sub-functionalizes all gene functions across paralogs, the EAC model allows for increased functionality of multiple original functions in an ancestral gene where simultaneous optimization of all functions is not possible. As such, under the DDC or EAC models, duplication can allow for the relaxation of constraint on the evolution of the ancestral gene, allowing for a selective advantage for both parental and new genes.

Finally, the IAD model begins with an ecological shift allowing for a selective advantage for high copy number. Here, the parental gene has multiple functions, and while its primary function is still maintained, an ecological shift allows for a selective advantage to exist for higher copy number based on selection for an auxiliary function. Importantly, as changes in copy number are more common than point mutations, increased dosage may more rapidly fix than regulatory changes. Following this amplification, subsequent changes are accumulated on the various copies. However, these models do not incorporate the 3-dimensional nature of the eukaryotic genome as well as its role in gene regulation. Crucially, while it is possible to identify both parental and new gene copies within the *D. melanogaster* genome, both parental and new gene copies are indistinguishable in previous models of duplication-based evolution. As such, prior models would predict that essential functionality should partition equally between parent gene and new gene and would not predict an enrichment for non-

essential genes that co-express highly with their neighboring genes. This symmetry and its consequences are considered in detail in Chapter III.

1.6 Co-regulation, Positional Effects, and Genome Organization

The idea that the genome is organized locally into co-regulated chromosomal units, or domains, and that these domains are a fundamental unit of selection has previously been referred to as the domain hypothesis [31]. However, a comprehensive integration of these effects into a single model of new gene evolution has been lacking.

In assessing the degree of co-regulation and positional effects occurring in the *melanogaster* genome, the first comprehensive studies of co-expression and gene order became possible with the advent of micro-array technology. To identify large-scale co-regulation within the *melanogaster* genome, Boutanaev and colleagues systematically assessed the clustering behavior of testes-specific genes on Drosophila chromosomes [32], using publicly available expression data. Here, a cluster is defined as a set of neighboring genes that have testes-specific expression. When compared to a distribution of permuted gene order, an enrichment for testes-specific genes in large clusters (3+ genes per cluster) becomes apparent. Additionally, the number of genes that appear in large clusters is also enriched when compared to a permuted distribution. Overall, the proportion of tissue-specific genes that are clustered is significantly enriched not only for testes-specific genes, but also for head-specific and embryo-specific genes.

One potential bias that has been ignored in the analysis of Boutanaev *et al.* is the effect of tandem duplications. The presence of multiple tandem duplications that result in a local co-regulation of all duplicate copies could inflate the clustering behavior observed by Boutanaev *et al.*, thereby limiting the potential significance of co-regulated gene clusters. To test whether non-homologous co-expressed genes still appear to cluster, Spellman and Rubin utilized a dataset of gene expression across 88 experimental conditions [33]. While

the methodology employed by Spellman and Rubin differs greatly than that of Boutanaev, *et al.*, when compared to random distributions, widespread clustering of co-expression was detected, with 553 genes co-expressing in 46 unique groups ($p < 10^{-4}$). To test whether homologs alone explain the presence of widespread co-expression, the co-expression analysis was repeating after removal of genes with nearby homologs. In this subsequent analysis, 200 genes still remained within 18 co-expressing groups, demonstrating that tandem duplications are not the sole factor explaining co-expression and gene order. Additionally, to test whether genetic order is preserved as a function of biological process, i.e. that genes cluster by pathway, enrichment for Gene Ontology (GO) terms was assessed within each identified co-expression group. 43 GO terms were found to be significantly associated with co-expression groups. However, after neighboring homologs were removed, only 11 GO terms remained significantly enriched with a given group.

1.7 A Model Co-expression Cluster

The combined results of Spellman and Rubin as well as those by Boutanaev *et al.* suggest that large-scale co-regulation may influence gene order within the genome, even after correction for homologous sequences. An in depth look at a model gene cluster by Kalmykova, *et al.* [31] additionally reveals that co-regulation may explain the relative proximity of non-homologous genes that possess similar functions. The model cluster consists of five testes-specific, non-homologous genes (Crtp, Yu, CK2βtes, Pros28.1B, CG13581). Notably, CK2βtes and Pros28.1B were ecotopically duplicated from constituitively expressed genes CK2β and Pros28.1, demonstrating that tissue specificity can potentially arise from duplication alone. The onset of transcription for these genes is highly coordinated and specific, as transcripts can only be found during early spermatogenesis - transcripts are neither present in stem cells, spermatogonia, nor in post-meiotic spermatogenesis. As additional evidence that these genes are co-regulated, the entire 5-gene cluster is inactivated in bag-of-marbles

mutants, while 4 of 5 genes are inactive in always-early mutants. Further DNase sensitivity assays demonstrated that the chromatin profile of this cluster is coordinately regulated across differing tissues, in this case, larval testes, larval brains, and embryos.

The results of Boutanaev, *et al.* and Spellman and Rubin suggest that gene order within the *melanogaster* genome is structured according to function while, the results of Kalmykova, *et al.* provide evidence for such co-regulation. A synthesis of these findings with Muller's ideas regarding the double-bar/ultrabar phenotype leads to a model where positional effects play a central role in the origination of new genes, a hypothesis further developed by Kalmykova, *et al.* [31]. However, crucial evidence guiding this synthesis is lacking and is addressed in Chapter III, leading to the following questions: What forces drive gene duplication? How do we differentiate between evolutionary models of new gene evolution? How can neo-functionalization occur in a new gene copy? What is the role of the 3-D genome in new gene evolution?

1.8 Chapter Overview

In this dissertation, I present both a new theoretical treatment of phenotypic plasticity as well as an empirical validation of an old theory of positional effects and the origination of new genes.

In Chapter II, I articulate a new computational model of phenotypic plasticity using stochastic differential equations. The primary advantage of this method is that it allows for the unification of both non-adaptive and adaptive plasticity effects. In this model, phenotype no longer is directly dictated by genotype, but is the result of both genic and epigenetic effects throughout an individual's life history. As such, any given phenotype may be produced by a large set of genotypes, a key feature in models of non-adaptive plasticity but not adaptive plasticity. Alternatively, the incorporation of an auto-correlation parameter allows for the production of responses well-suited to new environmental conditions, a feature which by

definition is present in models of adaptive plasticity but absent in models of non-adaptive plasticity.

By allowing for the production of any given phenotype through genic and epigenetic means, I demonstrate how when genic means are disallowed during an environmental shift, epigenetic compensation may occur, allowing for the production of individuals that have fully accommodated new environmental conditions without undergoing genetic change. Furthermore, larger degrees of non-adaptive and adaptive plasticity are favorable under these static conditions if control of these effects is allowed to mutate.

When genic parameters controlling phenotype become mutable, the combined effects of plasticity and genic change work hand-in-hand to form a consistent pattern of genetic accommodation that is robust to changes in selective conditions. Alternatively, when plasticity is absent in a population, the rate of accommodation may vary greatly under different selective conditions. When combined together, these results demonstrate how, in agreement with prior models, plasticity may variously accelerate or retard the genetic accommodation of new environmental conditions and how it can be transiently advantageous during adaptation.

Going further, however, the model presented in Chapter II additionally demonstrates how plasticity allows the level of genetic turnover to also be robust to changes in selective conditions at mutation-selection-drift balance. Over long periods of time at mutation-selection-drift balance, the level of plasticity may remain constant or decrease over time, depending on selective conditions. So long as the distribution of phenotypes generated by plasticity falls sufficiently close to new optimal conditions, plasticity may be maintained in a population, contradicting prior predictions of the long-term loss of plasticity under static conditions. Finally, I demonstrate how the relative mutation rate between genic and epigenetic control of phenotype determines the long-term fate of plasticity under stasis, a result that is not possible under prior models of plasticity.

In Chapter III, I use a variety of genomic techniques and novel statistical analyses to

demonstrate how positional effects and enhancer capture may be a driving force for the evolution of new genes originating through ectopic duplication. Prior models of gene duplication-based evolution are symmetric between both new gene and parental gene functions, while being agnostic to the relationship between both new gene and neighboring gene. As such, all functions, including essential functions, are expected to be randomly distributed between both new gene and parental gene copies while expression patterns between new genes and neighboring genes are expected to be random as well. Alternatively, new genes that have evolved via enhancer capture are expected to have essential function remain with the parental gene copy while showing high co-expression with neighboring genes. This difference in expectation was tested using data from new genes evolved in *Drosophila melanogaster*, demonstrating that enhancer capture is a common mechanism for the origination of new genes.

By comparing the co-expression between new gene/parental gene pairs and new gene/ neighboring gene pairs, I identified HP6/Umbrea as a model gene for the enhancer capture model. Prior work on its protein evolution found that its essentiality evolved subsequent to is origination while also showing that protein neo-functionalization was likely not a primary driver of HP6/Umbrea's fixation. Analysis of HP6/Umbrea's expression pattern shows that it is expressed primarily in the imaginal discs and male reproductive tissue. An examination of active enhancer marks revealed the presence of a putative larval enhancer nearby whose activation correlates with the onset of HP6/Umbrea expression. Given that HP6/Umbrea duplicated into a gene-poor region of the genome, HP6/Umbrea remains the likeliest target of this enhancer. Further analysis of the tissue expression pattern of neighboring genes reveals a cluster of 6 putatively co-regulated genes that express in the same tissues as HP6/Umbrea. To test for co-regulation between these different elements, examination of chromosomal conformation capture data in the HP6/Umbrea locus revealed the presence of 3 primary interactions: enhancer-HP6/Umbrea interaction, HP6/Umbrea-6-gene cluster interaction, and

interaction across the entirety of the 6-gene cluster. Higher resolution data for this locus was also generated using 4C-Seq, identifying the locations of 3 putative enhancer candidates.

Central to the enhancer capture model is the ancestral state of the new gene locus prior to duplication. Under the enhancer capture model, the regulatory environment producing co-regulation must pre-date the insertion of the new gene. In the case of HP6/Umbrea, this was tested by comparing the chromosomal conformations of the HP6/Umbrea locus in two in-groups and two out-group species, in this case *D. melanogaster*-*D. yakuba* and *D. pseudoobscura*-*D. miranda* respectively. While the tissue sources of these libraries differed greatly, a comparison of these data sets demonstrated that the 3 primary interactions remained conserved prior to HP6/Umbrea's insertion.

CHAPTER 2

EVOLUTION AND MAINTENANCE OF PHENOTYPIC PLASTICITY

2.1 Abstract

We introduce a novel framework for exploring the evolutionary consequences of phenotypic plasticity (adaptive and non-adaptive) integrating both genic and epigenetic effects on phenotype via stochastic differential equations and *in silico* selection. In accordance with the most significant results derived from prior models, we demonstrate how plasticity is differentially favored when subjected to small vs large environmental shifts, how plasticity is transiently favorable while accommodating a new environment, and how plasticity decreases during epochs where the environment remains stable (canalization). In contrast to these models, however, by allowing the same phenotypic value to be produced via two different paths, i.e. deterministic, genic vs stochastic, epigenetic mechanisms, we demonstrate how when genic contributions alone cannot produce an optimal phenotype, plastic, epigenetic contributions will instead fully accommodate new environments, allowing for both adaptive and non-adaptive plasticity to evolve. Furthermore, we show that while rates of phenotypic accommodation are relatively constant under a wide range of selective conditions, selection will favor the most efficient route to adaptation: deterministic genic response, or stochastic plastic response. As a result, plasticity may evolve or canalization may occur within a given epoch depending on the relative mutation rate of genic and epigenetic contributions to phenotype.

2.2 Introduction

2.2.1 Phenotypic Variability

The modern synthesis requires that all populations of organisms have some appreciable degree of phenotypic variation in order for natural selection to occur. Within this framework, phenotypic variation is a result of genetic variation, whereby selection on phenotypes acts as a feedback mechanism to control how genetic variants flow through a population. However, the extent to which variation in quantitative traits is explained by genetic variation is not fully understood. Under standard models, phenotypic variability, often referred to as V_P, is generally explained by a combination of genetic and non-genetic (environmental) variability, V_G and V_E respectively, such that $V_P = V_G + V_E$ [7, 23]. Note that in the absence of explicit parameterization, gene-by-environment interactions are often also contained within the environmental variability term [34].

Historically, such quantitative genetic models do not systematically incorporate the existence of non-genetic phenotypic variability [7, 11, 34]. Instead, such models frequently portray a one-to-one genotype-phenotype relationship, in which a particular genotype corresponds exclusively to a single phenotype, or else is expressed as a summation of the effects contributed across many loci. Any discrepancies between the additive effect of independent loci and phenotypes is often simply modeled as a linear error term, while in practice such discrepancies are dismissed by invoking the relatively poorly understood phenomenon of penetrance. However, as our collective understanding of molecular biology grows, effects previously discarded as environmental error must include an increasingly large number of effects beyond variation in life history, including more complex adaptive and non-adaptive plasticity mechanisms like behavior, epigenetic regulation, diversification, and non-genetic modes of inheritance.

Recent observations indicate the existence of variable relationships between genotype and

phenotype, in which a given genotype may in fact produce a range of phenotypes that can fluctuate dynamically [35–39]. Even within clonal populations, individuals have been shown to display quantitative differences that may distinguish them from genetically identical individuals. Phenotypic plasticity describes the phenomenon by which such individuals within a population may differentiate from genetically similar members by non-genetic means. In the context of rapidly fluctuating environmental conditions, possessing variable phenotypes within even genetically similar populations generates the potential for individuals to be randomly suited to uncertain conditions (i.e. non-adaptive plasticity) [14–20]. Alternatively, in the context of a single environmental shift, survival may increase depending on the ability to produce a phenotype well-suited to new conditions in response to such changes (i.e. adaptive plasticity) [11–13]. While hypotheses concerning the adaptive value of plasticity have been previously described, such variability largely appears to inescapably exist at the very least on a molecular level, regardless of whether or not it may provide an adaptive advantage. In fact, it has been shown that the minimization of such molecular noise requires the evolution of very specific topological constraints [40].

Understanding the potential evolutionary consequences of such variability has often been contentious, with studies suggesting that there is no evidence to show that plasticity influences adaptation [10]. However, various models and observations have delved into the possible mechanisms by which plasticity may have evolved and the role that it may play in shaping biological pathways[11–13, 16, 41–48]. In classical population genetics, standing genetic variation alone produces a wide variety of phenotypes upon which natural selection may act. As such, in the process of adaptation, genotype precedes phenotype under natural selection. In contrast, models of phenotypic plasticity generally assume that, within a single given genotype, an organism's interactions with the environment paired with mechanisms for adaptive and non-adaptive phenotypic plasticity can produce a wide variety of phenotypes. In these models, it is these phenotypes, produced by both plasticity and underlying

genetic variation, that are under selection. As such, phenotypic change may pave the way for genotypic change in the process of adaptive evolution. Such a process has variously been referred to as either genetic assimilation [49] or genetic accommodation [50].

2.2.2 West-Eberhard Model

According to Mary Jane West-Eberhard, the adaptive evolution of plastic traits may follow a four-step process. First, an adapting population must have a degree of phenotypic plasticity. Plastic traits will have the ability to display a range of phenotypes in response to various inputs. These inputs may be simple external variation in the environment, but they could also be novel genetic inputs as a result of mutation. West-Eberhard emphasizes that these plastic phenotypes must have a degree of responsiveness to such inputs; otherwise, environmental or even genetic changes would have no effect on phenotype. This scenario stands in contrast to fully canalized traits with such robust buffering to varying input that no alteration in phenotype would even be possible, resulting in cryptic variation [15]. Second, when presented with a new input, either external or internal, the plastic traits produce novel phenotypes in response. Here, a phenotype may have a wide range of adaptive and non-adaptive responses to an altered input, resulting in phenotypic accommodation. While phenotypes overall may have a broader distribution than in prior conditions, phenotypic accommodation will allow for the production of at least some individuals with an optimal phenotype. Regardless, this novel set of phenotypes now constitutes an altered substrate on which natural selection may act. Third, if some phenotypic response to the new input provide a selective advantage, these phenotypes may increase in frequency in the population given a recurring input. Fourth, If this phenotypic response has a genetic component, genetic accommodation may occur, fixing this new phenotype within a population. Notably, this model departs only slightly from the classic mutation-selection view of adaptation, in that the West-Eberhard model allows for the additional possibility of novelty being generated via

plasticity and not only through mutational processes alone [50].

Computational and theoretical models testing various aspects of plasticity have been been previously published, but few models exist that can directly test the West-Eberhard model (see **Discussion**). Tests of the West-Eberhard model must possess two primary properties. Firstly, individuals in a population should be able to produce a variable, non-genetic response to the environment; and secondly, an optimal genotype should exist. The production of non-genetic responses to the environment is a pre-requisite for both steps 2 and 3 of the West-Eberhard model, while an optimal genotype must exist in order for a phenotype to be capable of becoming genetically assimilated. While unable to fulfill these criteria, models of plasticity and switching are still concerned with understanding the role of plasticity mechanisms in adaptive processes, often addressing specific questions regarding the role of plasticity in surviving uncertain conditions, understanding how plasticity is maintained, and determining whether plasticity accelerates adaptation.

2.2.3 Prior Models

Intuitively, populations with greater phenotypic variation will survive a larger number of stressful conditions than those with less variation, which would impart an advantage to having greater V_E. One interesting class of models that specifically examines this phenomenon are phenotypic switching models [21]. Here, a genotype produces binary phenotypes, which are subject to fluctuating binary environmental conditions. Such models constitute a departure from the notion that one genotype can produce a singular phenotype, particularly given that, under fluctuating environmental conditions, no single phenotype would be able to survive both conditions. One key parameter of genotypes under these models is the phenotypic switching rate. The rate at which phenotypic diversification occurs can differ drastically among populations depending upon how beneficial variability is for survival. When change occurs more quickly than organisms are able to genetically adapt to it—as in the case of

seasonal changes, sudden environmental catastrophe, etc.-other strategies, either those randomly produced prior to or in response to external changes, must then be employed in order to avoid extinction. These models demonstrate how phenotypic switching allow populations to survive disastrous extinction events [18, 41], with the central result of these studies being that the optimal switching rate is equal to the environmental switching rate [16, 21]. Some of these models have been extended to include a spatial dimension with added migration, resulting in an increased risk of disasters over space and therefore favoring an increased switching rate [18]. Certain extensions of these models have considered scenarios in which the phenotypic or environmental switching rates may be auto-correlated. When selective environments are increasingly auto-correlated, slower switching phenotypes are favored, as the likelihood of encountering successive environments to which a phenotype is suited increases [18]. Alternatively, when selective environments are increasingly unpredictable-specifically, when environments fluctuate on random timescales-the phenotypic switching rate will be increasingly auto-correlated [22].

When considering models of phenotypic switching under the broader category of models of phenotypic plasticity, it is important to clarify two key facets. First, the distinction between heritable and non-heritable contributions to phenotype is significant, as under these models, the switching rate is the genetic trait under control while the phenotype is random. Secondly, an understanding of the temporal dynamics of the external environment is required, as the environment is assumed to fluctuate, potentially on differing time scales and possibly with or without a finite auto-correlation time. Such aspects have conceptual consequences in understanding the concept of "genetic adaptation". Regardless, the results of these studies produce key insight into the evolution of non-adaptive plasticity. [16, 18, 19, 22, 41, 51]

Alternate models, including ours, focus on how differences in parameters controlling non-heritable contributions to phenotype affect the fixation times of beneficial alleles [11, 23, 44]. As such, models of binary phenotypes may be insufficient for encapsulating certain aspects of

the dynamics of adaptation. Although many models of phenotypic switching provide insight into why fluctuations are central to the evolution of variability, insight into phenotypes existing on a continuum remains limited. However, these models still can capture some limited aspects of phenotypic plasticity. One such model is the Ancel model, which defines phenotype as a continuous variable [11, 23]. Here, rather than having genetic parameters control a single-valued phenotype alone, these parameters define instead a lower and upper limit for a norm of reaction. During selection, the environmental conditions select for a single optimal phenotypic value. In this model, selection is designed to reward having a norm of reaction that contains the optimal phenotype, but punishes large norms of reaction. Ancel-Meyers finds that in the process of adaptation to an environmental shift, the norms of reaction will undergo two distinct epochs. First, the norm of reaction will expand to accommodate the new environment, greatly increasing the fitness of individuals whose norms contain the new optimal phenotype. Subsequently, the norm of reaction will shrink around the new optimum, increasing fitness even further in comparison to those individuals with a wider norm of reaction, suggesting that in static environmental conditions, plasticity will be deleterious. This model was also used to show that phenotypic plasticity may accelerate adaptation if the difference in optimal and initial phenotypes is sufficiently large. However, if this condition is not met, phenotypic plasticity appears to retard adaption.

A second class of models examines the behavior of a continuous phenotype in the context of binary states of a continuous environmental variable. While the norm of reaction in the Ancel model is defined by two bounds, in this class of models, the norm of reaction is defined as a linear function where a phenotype is strictly a genetic response to environment [13, 24]. In this model, as the environmental value increases, the phenotype variable will also increase, with the degree of increase determined by the slope of the linear environment-phenotype function: the norm of reaction, the only genetically controlled variable in these models. With a high degree of plasticity, individuals may produce ideal phenotypes at both environmental

conditions, with a large slope or gradient in the linear function, while less plastic individuals will be unable to produce significantly different phenotypes when environment is varied, resulting in a small slope or gradient. An intuitive analysis of this model shows that under static conditions, plasticity is deleterious. If the environment fluctuates locally around only one of both environmental conditions, for small perturbations in environment, a smaller norm of reaction produces phenotypes much closer to the ideal phenotype than if individuals had a larger norm of reaction. Specifically, under static conditions, it is highly favorable to be under-responsive to small perturbations. However, this lack of response produces a trade-off when large environmental shifts occur, as individuals should produce a similarly large change in phenotype (thus larger reaction norm) to accommodate such a shift [13].

While the two models of plasticity discussed here differ significantly in interpretation, they both model adaptive phenotypic plasticity and they both reach similar conclusions. They agree that during adaptation, phenotypic response undergoes two epochs where the norm of reaction first increases, then decreases around the new optimal phenotype. Under static conditions, strong responses to small fluctuations are deleterious. However, as the environment shifts, a higher degree of plasticity is favorable, allowing alleles allowing for greater plasticity to increase within the population. As populations continue to adjust to the new environmental regime, such sensitivity to environmental changes is then no longer favorable, instead returning to the original state where plasticity is deleterious. Because of this multi-step process, previous models have demonstrated that plasticity may accelerate natural selection but only in very limited conditions where the distance between initial conditions and optimal conditions is very large.

Though plasticity is deleterious under static conditions, as previously mentioned, an increase in phenotypic variation will allow for potential adaptation to a wider range of selective conditions, potentially allowing populations to cross otherwise uncrossable fitness boundaries. Classically, such variation is often thought to be maintained by genetic variation via

mutation, selection, and epistasis [52]. Graphical models comparing intrinsic trade-offs of fitness between various environmental conditions suggest that the nature of the trade-off may favor various strategies for generating non-genetic variability. Populations that have adapted to certain environmental conditions deal with an intrinsic trade-off between consistent development under the broad range of frequently encountered environmental conditions (in these models, adaptation) and the ability to respond to alterations and new conditions (in these models, switching). The degree of variation among extant phenotypes in a given population depends largely upon the benefit verses the cost of evolutionary trade-offs inherent to phenotypic diversification. Here, weaker fitness trade-offs tend to favor a generalist strategy, while strong trade-offs favor specialist strategies [43]. Additionally, such models have shown that the fitness trade-off inherent to adaptive solutions may predict whether adaptation or switching will be favored [42].

Rather than being a result of fluctuating environments, some have also postulated that such plasticity may accelerate the process of genetic adaptation, and as such may provide a mechanism for maintaining environmental variability [53, 54]. Others have suggested that, overall, plasticity will tend to produce slower genetic adaptation than in the case of no plasticity [46], or that it will only accelerate adaptation when plasticity allows for a quicker encounter with an advantageous phenotype [11]. Regardless, results suggest that potential acceleration of adaptation is insufficient to explain the maintenance of plasticity over long periods [11]. Another model [34] additionally proposes that certain "engineering" costs to precise expression of phenotypes may allow for the maintenance of non-genetic phenotypic variability within a population. A model proposed by Wagner, Booth, and Bagheri-Chaichian additionally shows how pleiotropy or associations between decreases in plasticity and changes in mean phenotype may also allow plasticity to be maintained under static conditions [55].

The models reviewed here all incorporate different aspects of plasticity, such as a separation of phenotype and genotype or random fluctuations in environment or environmental re-

sponse. However, none of these models can fully recapitulate the steps of the West-Eberhard model. Only in models of phenotypic switching are genotype and phenotype fully segregated, allowing phenotypic accommodation to precede the establishment of genetic accommodation of new conditions. However, as the occurrence of a swapped phenotype occurs independently of environmental conditions, such models may only capture certain aspects of non-adaptive plasticity.

Unlike in models of phenotypic switching, previous models of phenotypic plasticity allow for a broad range of phenotypes to exist, rather than simple binary phenotypes. In these models, phenotype is the result of genotypic state (norm of reaction) interacting with the current environmental state. However, as phenotype is a direct result of genotype, and all individuals in a population are subject to the same environmental conditions, phenotype thus has a one-to-one correlation to genotype during the process of adaptation. Due to this, these models of phenotypic plasticity are not able to recapitulate the West-Eberhard model. However, given that new phenotypes are produced as a direct result of an environmental shift, such models may capture certain aspects of adaptive plasticity.

In the following section, we present a single, minimal model that allows for the study of adaptive and non-adaptive plasticity while satisfying the conditions for testing the West-Eberhard model.

2.3 Methods and Materials

2.3.1 Model

The aim of this section is to describe our minimal model of phenotypic plasticity. We compare the results of *in silico* selection between populations that have either plastic or non-plastic phenotypes. In order to model phenotypes with long-term stability and convergence, we model these self-regulating phenotypes as a continuous value that is either the result of two

feedback loops (one positive, one negative) for non-plastic traits, or three feedback loops (one positive, one negative, one negative but noisy) for plastic traits (Figure 2.1). The simplest such model is a logistic growth model.

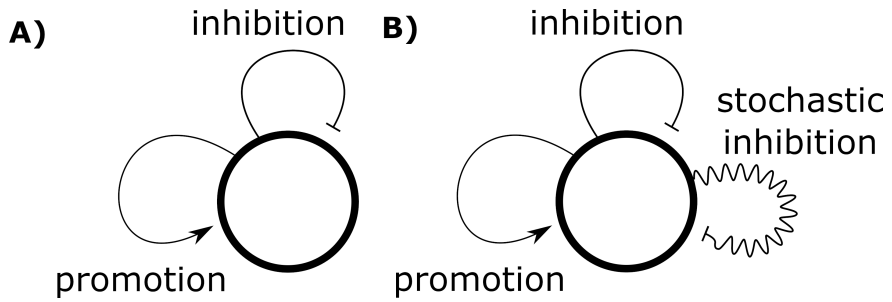

Figure 2.1: **Self-regulating traits.** Phenotypes are modeled as the result of a minimal self-regulating system. **(A)** Non-plastic phenotypes have two feedback loops, one positive and one negative, while **(B)** plastic phenotypes have three feedback loops, one positive and two negative, one deterministic, and one stochastic.

We first model non-plastic phenotypes, where an individual i in a population of size N has a phenotype P_i,

$$\frac{dP_i}{dt} = \gamma_i P_i - \epsilon_i P_i^2. \tag{2.1}$$

which is controlled by a set of genotype parameters $\{\epsilon_i, \gamma_i\}$, representing the total genic (non-epigenetic) contribution to phenotype. Phenotype is represented by an ordinary differential equation, modeled off of a trait that has two feedback loops, and that therefore has the minimum requirements necessary to dynamically maintain a static trait value. There is one positive feedback loop promoting an increase of said trait according to parameter γ_i and a second, negative feedback loop repressing said trait according to parameter ϵ_i, working in concert to maintain the phenotype at a specific value $P_i = \gamma_i/\epsilon_i$ over sufficiently long periods of time. We may assume that this phenotypic value is produced as the result of a population having previously adapted to static environmental conditions, such that the long-term phenotypic value is the optimal phenotype. Notably, this model does not yet have

any plasticity, so that phenotype $P_i = \gamma_i/\epsilon_i$, and, all variability in phenotype (V_P) should be a direct result of population differences in parameters γ_i or ϵ_i. In a population of clones, there should be zero phenotypic variability, as there is no genotypic variability.

As individuals are subjected to a wide variety of non-genetic changes and complex genetic interactions during development, we use a stochastic differential equation with multiplicative noise to model the genotype-phenotype relationship, resulting in log-normal-like distributions of phenotypes [56, 57]. We may change the ordinary differential equation of non-plastic phenotypes represented in Eqn. 2.1 into a stochastic differential equation

$$\frac{dP_i}{dt} = \gamma_i P_i - (\epsilon_i + \xi_i) P_i^2, \tag{2.2}$$

with additional epigenetic parameter ξ_i, representing the total non-genic (epigenetic) contribution to phenotype. ξ_i represents the accumulation of non-genetic fluctuations an individual in a population encounters throughout its lifetime, including environmental, developmental, and behavioral variation, resulting in a plastic, epigenetic response ξ_i. This epigenetic response results in a decoherent feedback response via the second term of the SDE. Importantly, in contrast to the non-plastic model represented in Eqn. 2.1, the accumulation of random epigenetic variation results in a distribution of phenotypes within any given genotype. Similar heavy-tailed distributions have been observed for widely varied phenotypes, such as intracellular protein levels, cell size, or even clutch size [35, 58, 59].

The consequences of this non-genetic variability on population phenotypes are modeled by ξ_i and its genetic control parameters, D_i and τ_i. ξ_i is a random variable with mean 0 resulting from a Wiener process. ξ_i is also auto-correlated with magnitude D_i and time τ_i such that

$$\langle \xi(t_0), \xi(t) \rangle = D_i e^{-|t-t_0|/\tau_i}. \tag{2.3}$$

Notably, the genetic parameters D_i and τ_i control the epigenetic value ξ_i but cannot directly dictate its value, which is still a random variable. As D_i controls simply the magnitude of plasticity, this represents genic control of the non-adaptive plastic response to environment. Similarly, τ_i controls the auto-correlation or memory of plasticity, allowing subsequent generations to have similar, favorable epigenetic responses produced in prior generations. This represents genic control of the adaptive plastic response to environment. Therefore, in contrast to the 2 parameters of the the non-plastic model, individuals controlled by Eqn. 2.2 should have 4 total genetic parameters $\{\gamma_i, \epsilon_i, D_i, \tau_i\}$ (Table 2.1). As an individual's genotype does not have a one-to-one correlation to a given phenotype, phenotypic space is degenerate, where any given phenotype may have been produced by a number of different combinations of genotypes.

Symbol	**Variable**
N	Population size
i	Individual $i \in \{1, 2, ..., N\}$
P_i	Phenotype of individual i
ξ_i	Epigenetic value of individual i
γ_i	Deterministic growth parameter/genotype of individual i
ϵ_i	Deterministic repression parameter/genotype of individual i
D_i	Magnitude of plasticity (non-adaptive) parameter/genotype of individual i
τ_i	Auto-correlation of plasticity (adaptive) parameter/genotype of individual i

Table 2.1: **Summary of variables.**

If the effects of ξ_i are increased (increased D_i), there is greater randomness that is not buffered and is therefore integrated into an individual's phenotype (higher phenotypic variation). Once selective pressures are applied to populations, the auto-correlation term τ_i allows for a more or less consistent response to selection. With a longer auto-correlation time (larger τ_i), offspring will produce similar responses to successive selective events (more auto-correlated), whereas with a shorter auto-correlation time (smaller τ_i), offspring are likely to have a more varied and heterogeneous response (less auto-correlated). As ξ_i has unique auto-correlated properties, ξ_i is generated by using a colored-noise Runge-Kutta method

[60].

This SDE has two key properties: the mean phenotypic value remains γ_i/ϵ_i, and the phenotypic distribution results in a steady-state distribution over sufficiently long periods of time [56, 57]. Under a quantitative genetics framework [23, 52], $V_P = V_G + V_E$, so that individuals within a clonal population may have a non-zero variation of phenotype. Note that the addition of a stochastic term to a simple logistic growth model satisfies the first step of the West-Eberhard model, specifically that a trait will have differential responses to input.

2.3.2 In-silico Selection

Given both plastic and non-plastic models, we then apply *in silico* selection for haploids with Gaussian fitness, resulting in a new selective environment (Figs. 2.2, 2.3). Since selection is applied only to phenotype, and fitness is related simply to the phenotypic distance between the individual's phenotype and the optimal phenotype, there is no direct penalty for plasticity in this model. We note that in these simulations, individuals initially inherit the epigenetic value ξ_i in reproduction, followed by the accumulation of noise, resulting in a new epigenetic value that fits the relationship Eqn. 2.3. As such, this epigenetic inheritance is transient, as the values of ξ_i randomly drift during development and throughout the individual's lifetime, resulting in an inheritance of ξ_i over multiple generations that decreases according to the auto-correlation parameter τ_i [47]. Similar effects have been seen in natural populations [39]. We also note that the random value ξ_i is specific to each individual/lineage, and is not shared across all individuals in the shared environment. We stress, therefore, that ξ_i should not be seen as environmental fluctuations, but instead each individual's unique response to the common, shared environmental conditions defined by the fitness function. The optimal phenotype for the given environment, P_{opt}, does not vary between individuals - it is the same for the entire population.

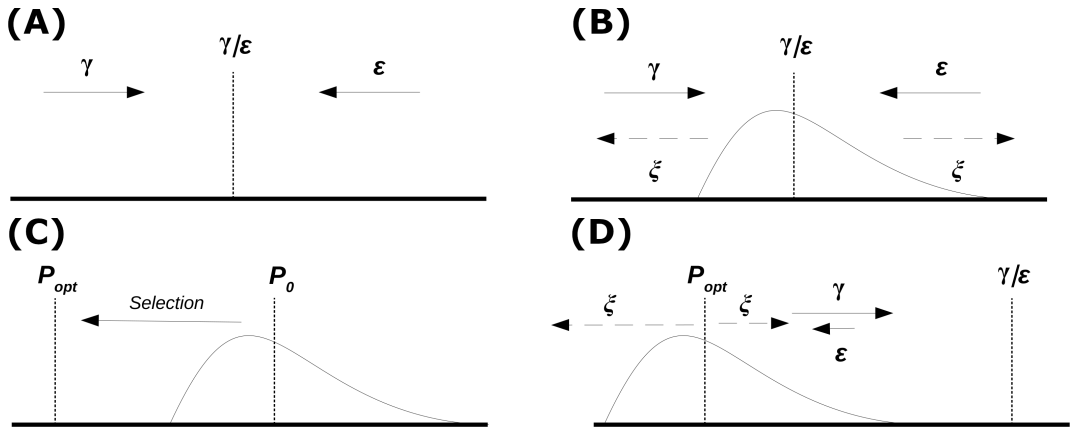

Figure 2.2: **Differential equation illustration. (A) Non-plastic phenotypes.** For non-plastic phenotypes, phenotype is strictly defined by genic control parameters, γ and ϵ, where the trait value is the equilibrium between the "forces" of growth (ϵ) and repression (γ). **(B) Plastic phenotypes.** For plastic phenotypes, phenotype is defined by genic control parameters, γ and ϵ, as well as the random epigenetic parameter ξ, which is controlled by D (magnitude of plasticity) and τ (auto-correlation). With plasticity, trait value is random, but converges on a steady-state distribution around γ/ϵ. **(C) Adaptation.** Populations are challenged with a new environmental condition, favoring an optimal phenotype far from the initial conditions. **(D) Phenotypic accommodation.** By disallowing mutation of genic control parameters, when challenged with a new environmental condition, populations adapt to the new environmental conditions through epigenetic means alone. As a whole, the genic control parameters pull populations towards a distribution around γ/ϵ. To maintain populations around the optimum, epigenetic parameter ξ compensates by favoring larger, non-zero values of ξ, while increasing the degree of plasticity and auto-correlation (c.f. Figure 2.4).

Genetic algorithms (*in silico* selection) generally consist of the following steps: mutation, selection, and repopulation. These steps are iteratively processed over a population of candidate solutions, with each generation yielding a new population that has been genetically influenced by the previous generation. Beginning with an initial seed population, each candidate solution within the population is scored based upon a pre-determined objective function. Forward simulation steps, if necessary, were solved for a set generation time T. Subsequent generations were repopulated using Wright-Fisher/multinomial sampling for monoploids using fitness w.

Objective functions are often a representation of a difference between the phenotypic value

of each candidate solution and the optimal phenotype provided. The objective function here is defined by fitness w, which is defined by a Gaussian function with mean phenotype μ and standard deviation σ, where $w_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left((P_i-\mu)/\sigma\right)^2}$. Here, stringent and relaxed directional selection was applied using $\mu = 1/90$, $\sigma_{stringent} = 10^{-4}$ and $\mu = 1/90$, $\sigma_{relaxed} = 10^{-3}$ respectively. After scoring, a new population is generated based upon a multinomial sampling of the previous generation's population, with the relative probability of generating progeny being proportional to the objective function's score. Individuals with higher fitness values will have higher relative probability of generating progeny, while less fit individuals should have a lower probability. Here, the new generation is populated until a pre-determined population limit has been reached (i.e. fixed population size).

After repopulation, mutations are applied using a move-generation algorithm, sampling neighboring states using a fixed mutation rate. Mutations were applied as changes in parameters $\{\gamma_i, \epsilon_i, D_i, \tau_i\}$ with a mutation rate u of 0.05 mutations per individual per generation. Unless otherwise noted, only the parameter ϵ_i was varied with fixed parameters $\{\gamma_i, D_i, \tau_i\} = \{1, 100, 0.01\}$, with initial conditions $\epsilon_0 = 67$. Note that genetic parameters are integer valued, except in the case of tau_i, which is valued at integer multiples of 0.0005). The magnitude of each mutation was drawn from a geometric distribution with probability $p = 0.5$ (e.g. for γ, $Pr(|\Delta\gamma| = k) = (1-p)^k(p) = (0.5)^k(0.5)$ for $k = 0, 1, 2, ...$) unless otherwise noted. The direction of the mutation was also drawn randomly, with an increase or decrease chosen with probability 0.5. If a forward-simulation step is required, such as in the case of solving a differential equation, the related steps are applied after mutation and before selection. Selection then is again applied over the entire population, yielding the next generation.

Two main population sizes were used in this study, $N = 100$ and $N = 1,000$. Population simulations were performed in replicates of 100 simulations. These were chosen to allow for comparisons between replicate populations, while reducing computational costs. Addition-

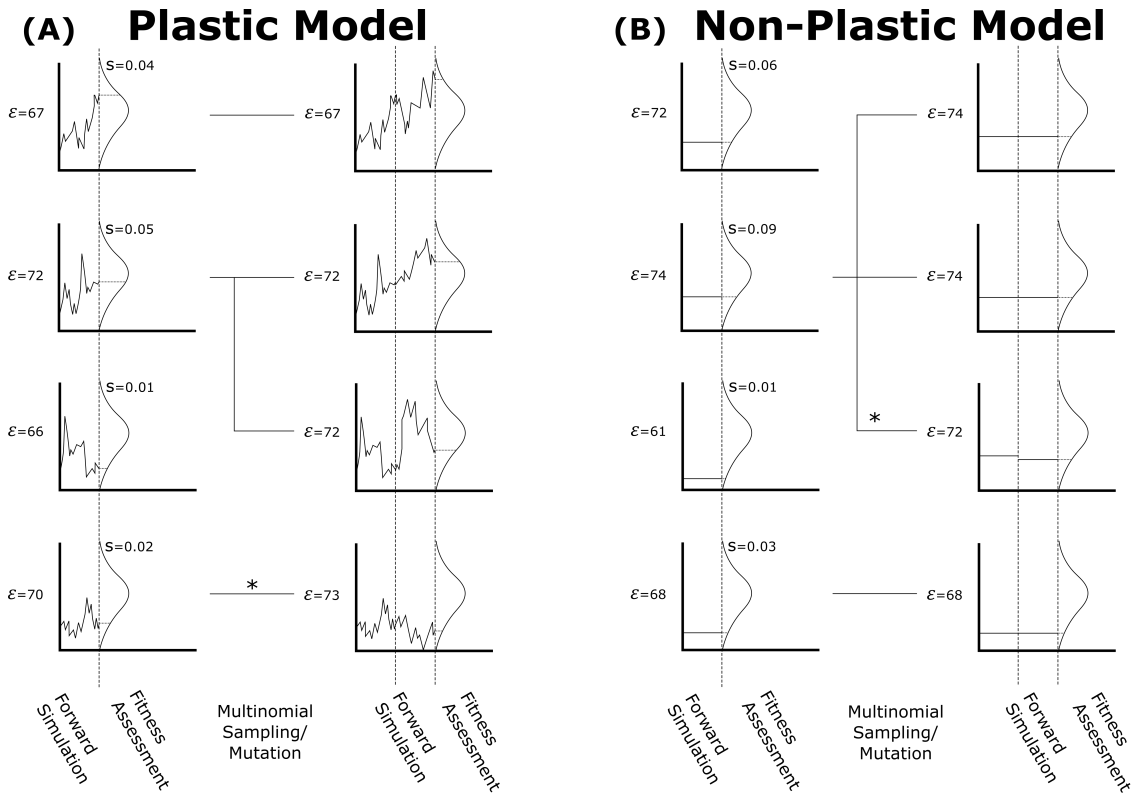

Figure 2.3: **Differential equation illustration. (A) Plastic phenotypes.** A sample population with four individuals is drawn, with vertical axes representing phenotypic value and horizontal axes representing time. Individual phenotypes undergo fluctuations over a pre-determined period of time until selection is applied, using multinomial sampling for monoploids with selection coefficients assigned by a Gaussian fitness function. **(B) Non-plastic phenotypes.** A similar sample population to **(A)**, but phenotypes do not have any plasticity and phenotypic values are determined directly by the genotypic values. Selection is still applied using multinomial sampling for monoploids and Gaussian fitness.

ally, small population sizes allowed for a greater influence of stochastic effects, as well as exaggerating the influence of beneficial phenotypes resulting in multiple bottlenecks.

2.4 Results

2.4.1 Phenotypic Accommodation

Given that the stochastic differential model fulfils the first step of the West-Eberhard model, the second step of the model is the development of phenotypic accommodation. Specifically,

in a shifting environment, certain individuals will be able to produce an adaptive phenotype in response to altered conditions. To determine whether our simulations would be able to produce phenotypic accommodation, we set the mutation rate for genic parameters (i.e. γ_i and ϵ_i) to 0, while allowing epigenetic parameters (i.e. D_i and τ_i) to evolve in populations of size $N = 1000$ and performed 100 replicate simulations. We challenged a population of individuals with stringent or relaxed environmental conditions then allowed individuals to adapt. Given these conditions, epigenetic parameter ξ_i alone could produce fully phenotypically adapted individuals (Figure 2.4), even when the mutation rate for genic parameters (γ_i, ϵ_i) is zero. This accommodation is rapid, with the population fully adapting within approximately fifty generations. As the genic parameters were not allowed to mutate, full adaptation was compensated for by a response in epigenetic variable ξ_i alone (Figure 2.4). Note that, in the initial steps of adaptation, ξ_i produces a strong response out of equilibrium before eventually settling on an equilibrium value far from 0. As ξ_i is the result of a Wiener process, ξ_i should have a mean of 0, but the system is pushed far from that equilibrium value. By allowing the magnitude of plasticity D_i and the auto-correlation/memory parameter τ_i to mutate, these simulations produce a response where increased plasticity and increased memory are favorable. In this scenario, we can consider the shift in selective environment as a recurrent selection input, and due to this recurrence and the inability of the genic parameters to mutate, greater plasticity and memory is advantageous. While we have disallowed genic parameters to mutate at all in this case, a similar response of increased plasticity and memory may also be favorable in conditions where genetic mutation is significantly slower than mutation in epigenetic parameters as well (c.f. **Establishment and Maintenance of Plasticity**). To understand the detailed dynamic of phenotypic accommodation and epigenetic compensation under these developmental conditions, we may consider the equilibrium phenotypic value produced by this population's genotype. In this case, the genic values γ_i and ϵ_i act in concert to pull phenotypic values of the population back to the predicted

steady-state distribution around γ_i/ϵ_i. However, due to selection for a phenotypic optimum that is departed from γ_i/ϵ_i, phenotype must be compensated for by the epigenetic parameter ξ_i (Figure 2.2).

Though ξ_i should accumulate randomness, thereby returning to mean 0 with variability D_i, there are two demands on ξ_i preventing this. The first is the demand that epigenetic parameter ξ_i is sufficiently large so that at least some phenotypes that may be near the new optimum. Overall, plasticity magnitude alone, which is D_i, increases the overall variability in P_i in an unbiased, non-adaptive fashion. To increase the random chance of producing offspring with a sufficiently large epigenetic parameter ξ_i, D_i increases. The second demand is that epigenetic parameter ξ_i continues to remain sufficiently large over time, thus avoiding being distributed around 0. That is to say that, given a parent with an epigenetic parameter ξ_i producing fit parents, the offspring will now be likelier to have a similarly adapted ξ_i as its parents. This results in an increase in the auto-correlation parameter τ_i. Together, these steps represent the second step of the West-Eberhard model, phenotypic accommodation.

2.4.2 Genetic Accommodation

Given that our model can produce phenotypic accommodation, we then allowed only the genic parameter ϵ_i to mutate in our model, excluding mutations in epigenetic parameters D_i and τ_i. To test for genetic accommodation, we performed 100 replicate simulations of populations of 100 individuals under stringent and relaxed directional selection. This population size was chosen to exaggerate certain features of genetic adaptation in plastic regimes (i.e. transient fixation events due to successive bottlenecks, see below & Figure 2.5). Allowing only genetic parameter ϵ_i to mutate, simulations were performed for populations with both plastic (Eqn. 2.2) and non-plastic (Eqn. 2.1) phenotypes.

Figure 2.5 provides a representation of mean genotypic (i.e. ϵ_i) changes for replicate populations with plastic and non-plastic phenotypes under two selection regimes: stringent and

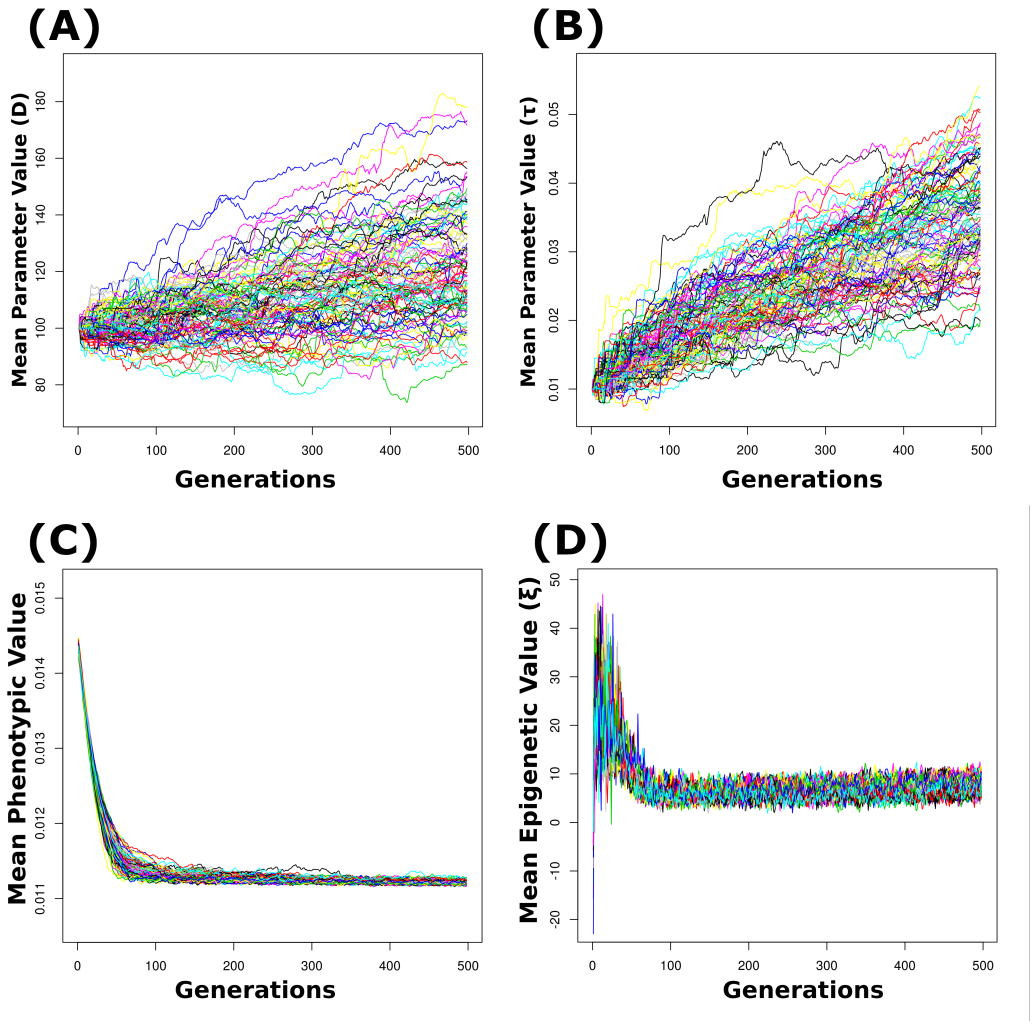

Figure 2.4: **Phenotypic accommodation favors increased plasticity and autocorrelation** 100 replicate populations of 1000 plastic individuals ($T = 2\tau$) were subjected to stringent ($\mu = 1/90, \sigma = 10^{-4}$) selection, with mean parameter values for each replicate population shown above for **(A)** D_i and **(B)** τ_i, one line per replicate population. Though genic parameters were not allowed to mutate, **(C)** population phenotypes fully adapted to new environmental conditions through **(D)** epigenetic compensation. The epigenetic parameter (ξ_i) compensates for a non-optimal genic configuration (γ_i/ϵ_i far from P_{opt}), consistently remaining far the expected value of 0. Under such conditions, increased D_i (non-adaptive plasticity) and τ_i (adaptive plasticity) are favored.

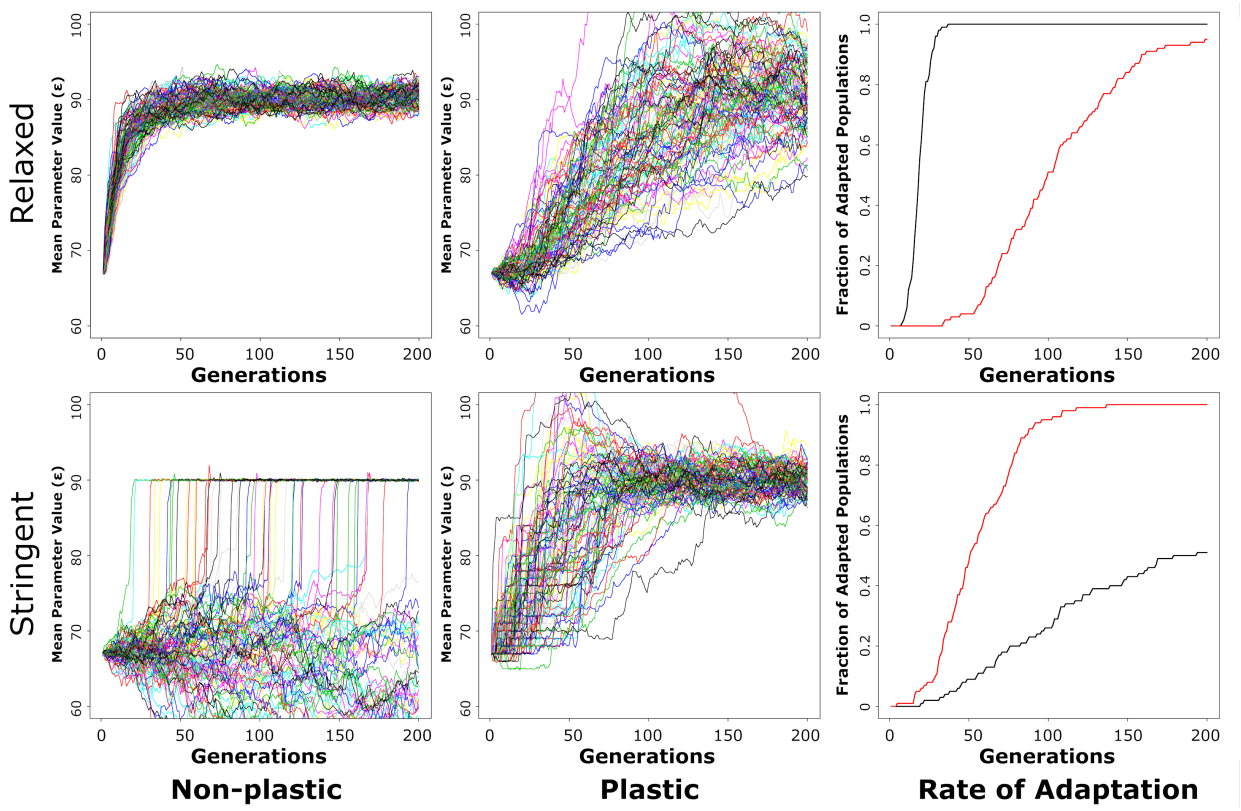

Figure 2.5: **Plastic populations are robust to changes in selective forces.** 100 replicate populations of 100 individuals were subjected to stringent ($\mu = 1/90, \sigma = 10^{-4}$) and relaxed ($\mu = 1/90, \sigma = 10^{-3}$) selection regimes for non-plastic or plastic traits. The mean genotypic values for ϵ_i are shown, one line per replicate population, demonstrating how plasticity allows populations to have more robust responses to changes in selection pressures. When relaxed selection is applied, non-plastic populations converge more rapidly and smoothly to the optimal genotype than plastic populations (top row). However, when stringent selection is applied, non-plastic populations undergo drift until an advantageous genotype is found, followed by rapid fixation, while plastic populations converge smoothly and rapidly on the optimal genotype (bottom row). The cumulative fraction of genetically adapted replicate populations ($\pm 5\%$ of ϵ_{opt}) within non-plastic simulations are significantly more sensitive to differences in selection than plastic simulations. In conditions of relaxed selection, non-plastic populations (black) adapt faster than plastic populations (red), however in conditions of stringent selection, plastic adapt first.

relaxed. By allowing only for mutations in ϵ_i alone, we can see that changes in genic parameters allow for the eventual genetic accommodation of new environmental conditions - the last steps of the West-Eberhard model. In early generations, ξ_i helps to produce selectively advantageous phenotypes (c.f. **Phenotypic Accommodation**). However, as mutations in genic parameters allow for the production of optimal phenotypes through genetic means alone, the epigenetic response is no longer needed to produce fit individuals. Due to this, the epigenetic parameter ξ_i returns back to 0, in contrast to in the case of phenotypic accommodation (Figure 2.4). Overall, in the process of adapting to altered environmental conditions, plasticity helps to ameliorate the negative consequences of an unexpected environmental shift.

For the non-plastic model, the genotypic responses under relaxed selection form a relatively smooth curve with minimal noise on the upward trajectory, revealing a consistent progression of all populations towards the optimal genotype. Because relaxed selection allows for a wider range of advantageous genotypes, adapting populations successively gain increasingly beneficial mutations and follow a selective gradient until reaching an optimum. The genotypes of these adapting populations rapidly converge near the genetic optimum of 90, with small deviations resulting from genetic drift. Additionally, these non-plastic populations rapidly reach an optimal state within approximately fifty generations. By contrast, plastic populations under relaxed selection show a slower pattern of convergence on the optimal genotype within approximately 200 generations. Rather than progressing steadily and consistently towards an optimum, the plastic populations show a very large degree of variation in genotype, sometimes severely over- or undershooting the ideal. Such an effect occurs due to selection acting on phenotype, rather than genotype. However, phenotypic convergence on the optimum remains smooth (Figure 2.6). As the range of possible phenotypes becomes larger due to plasticity, adaptation/accommodation in plastic populations is less smooth and directed than in the case of populations with less plastic genotype-phenotype

relationships under conditions of broad selection. When combined with relatively relaxed selection, this phenomenon causes the genotypes of plastic populations to converge slowly on the optimum phenotype. In contrast to the non-plastic populations, they do not steadily progress as directly towards the ideal and therefore do not adapt as efficiently.

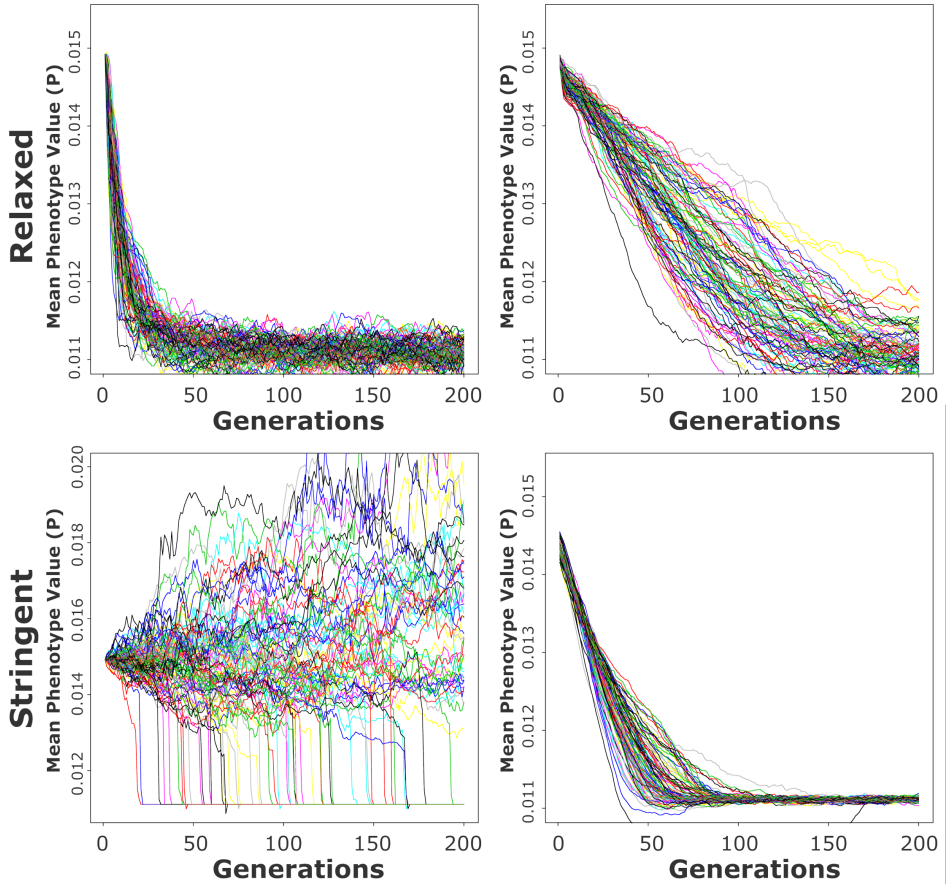

Figure 2.6: **Populations with plastic phenotypes converge smoothly on optimum** Mean phenotypic values for populations shown in Figure 2.5 are plotted, one line per replicate population. The behavior of non-plastic phenotypes match the behavior of population genotypes due to the one-to-one relationship between genotype and phenotype. However, while these phenotypic trajectories for non-plastic populations vary greatly (left), phenotypic trajectories for plastic populations converge smoothly on the pre-defined optimum (right). This effect is more pronounced in conditions of stringent selection (bottom).

Results differ significantly when one compares the effects on genotype under stringent selection for plastic and non-plastic populations. In the latter case, genotypes vary widely among replicate populations as a result of drift. As opposed to conditions of relaxed selec-

tion, non-plastic populations under stringent selection do not progress continuously along a selective gradient towards the optimum, since only phenotypes at or near the ideal confer a survival advantage. Here, phenotypes instead appear to be, in a sense, binary, as they are either largely beneficial to selection or effectively neutral. The populations develop various neutral mutations until hitting upon one at random that provides a large selective advantage. This effect results in a series of rapid fixation events, as each population reaches an optimum and then deviates very little from it.

A measurement of the fraction of genetically adapted/accommodated populations (Figure 2.5b) supports the observation that plastic populations converge on the pre-determined optimum more efficiently under stringent selection, while those that are non-plastic adapt more quickly in response to relaxed selective pressures. Within 150 generations, all non-plastic populations under relaxed selection and all plastic populations under stringent selection have genetically adapted fully to the new selection conditions. In both cases, accommodation occurred along a clearly defined selective gradient. Meanwhile, a number of populations under plastic/relaxed conditions and under non-plastic/stringent conditions have failed to adapt to their respective optima within 200 generations. In particular, non-plastic populations on average adapt far more slowly under stringent selective pressures, as they are limited by the appearance of beneficial alleles through mutations alone.

While our models do not incorporate disasters and total-population extinction due to our fixed-population-size simulations, it should be noted that under conditions of stringent selection, all non-plastic replicate populations would have experienced total extinction in the first few generations of applied stress. Rather than allowing for such scenarios, the model instead allowed populations to undergo genetic drift. Such results indicate that there is a certain degree of plasticity that is required for survival under extreme selective conditions [18, 19]. Regardless, our findings agree as, in general, plasticity does not necessarily expedite adaptation. In the scenario of broad selection, a non-plastic population may more

rapidly adapt to an altered environment when compared to plastic populations. However, in the scenario of stringent selection, plastic populations adapt more rapidly than non-plastic populations.

2.4.3 Genetic Turnover and Mutation-Selection-Drift Balance

To examine how phenotypic plasticity affects mutation-selection-drift (MSD) balance, we determined the degree of genetic turnover for all populations shown in Figure 2.5. We performed an auto-correlation analysis to determine the degree of turnover in mean genotype values (ϵ_i) for plastic and non-plastic populations (Figure 2.7). Specifically, we took replicate populations at this MSD equilibrium, and calculated the auto-correlation (Pearson correlation coefficient) of the mean population genotype values.

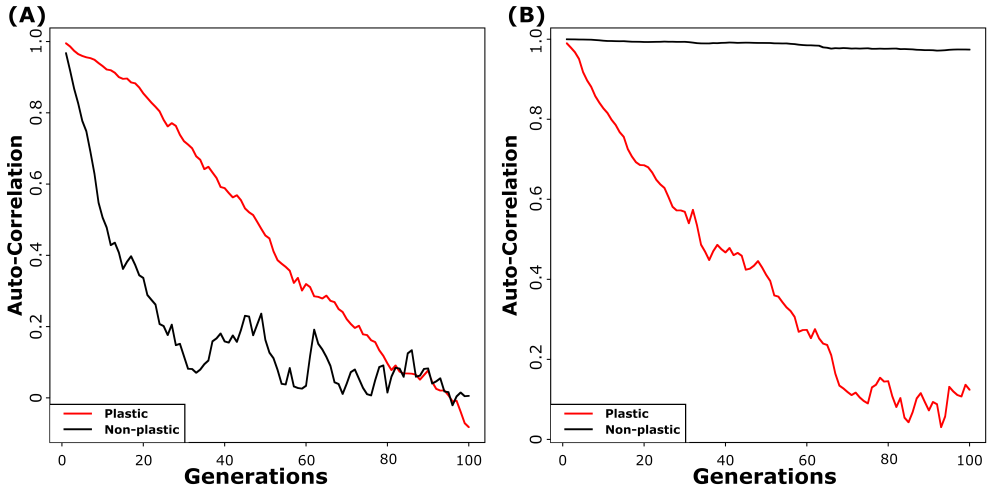

Figure 2.7: **Genetic turnover in plastic populations is robust to varying selection conditions.** At mutation-selection-drift (MSD) balance, turnover of alleles still occurs. Using 100 replicate populations of 1000 individuals at MSD balance, the degree of genetic turnover is represented here by the auto-correlation of the mean genic parameter ϵ_i for the set of 100 replicate populations. **(A)** Under relaxed selection conditions, non-plastic populations turn over more frequently than plastic populations. **(B)** Under stringent selection conditions, plastic populations turn over less frequently than non-plastic populations. Plastic populations appeared to be less responsive to various selection conditions (**(A)** relaxed, **(B)** stringent) than non-plastic populations.

The results of the auto-correlation analyses reveal a more robust degree of genetic turnover

for plastic populations in comparison to non-plastic populations. Regardless of the degree of selection, each generation of plastic individuals gradually diverges from its previous state under plastic conditions. Meanwhile, non-plastic populations do not display this same, robust turnover. Under relaxed selection, the degree of auto-correlation rapidly diminishes for non-plastic populations, while those under stringent selection show an extremely low rate of genetic turnover, with the auto-correlation value varying very little in the examined time frame. Overall, populations display a buffering of the degree of turnover from various selective conditions.

2.4.4 Loss of Variability

The previous sections have demonstrated how plasticity is beneficial under conditions of dramatic selective changes. Additionally, we have shown that plasticity buffers genetic turnover from variations in selection. We now consider whether plasticity is deleterious during static selective conditions. To test whether phenotypic plasticity presents a selective disadvantage at MSD balance, we performed *in silico* selection on plastic and non-plastic populations, making the only mutable parameter the degree of variability (D_i).

In accordance with expectations, populations under stringent selective pressures show a strong downward trend in phenotypic plasticity over many generations (Figure 2.8). Most of these populations show a significant decrease in their degree of variability, with some having their average plasticity decreased by nearly half. Therefore, while plastic populations under stringent selection can reach MSD balance more quickly than non-plastic populations, once the system has come to an equilibrium by reaching an ideal phenotype, the potential benefit for phenotypic variety decreases.

By contrast, populations under relaxed selection do not show a decrease in plasticity after 1000 generations. Unlike in the case of stringent selection, small populations under relaxed selection have a broader range of phenotypes which could be considered beneficial. So long

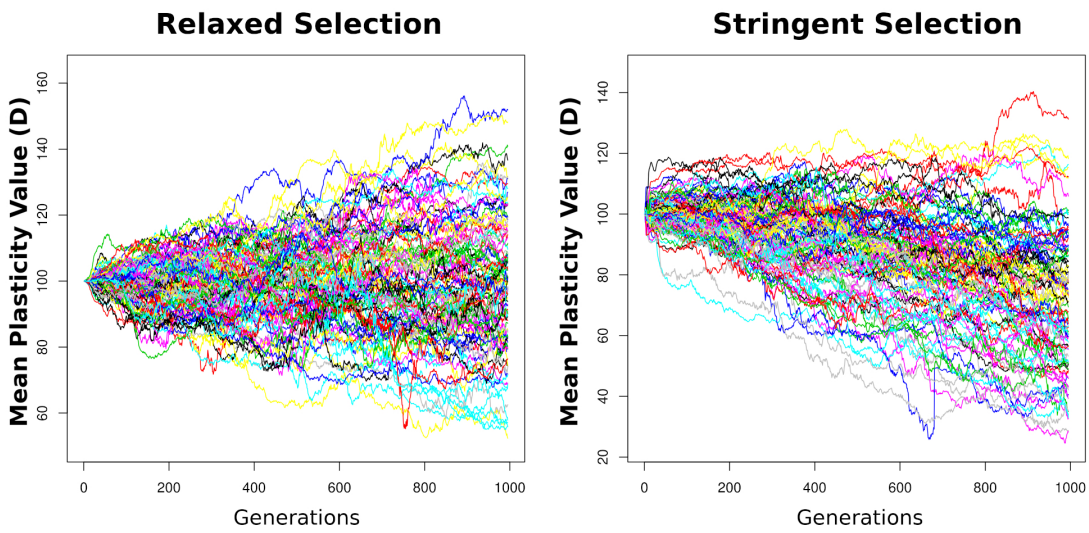

Figure 2.8: **Phenotypic plasticity under static conditions is weakly deleterious.** At mutation-selection-drift balance, plasticity does not to contribute any meaningful selective advantage. Shown are mean D_i values for 100 replicate populations of 1000 individuals at mutation-selection-drift balance, one line per replicate population. So long as the phenotypic variability is sufficiently within the bounds of applied selection, as in the case of relaxed selection, plasticity is near neutral and thus may be maintained in a population for extended periods (left). However, if the stringency of selection is increased, plasticity is weakly deleterious (right).

as the result of phenotypic variability, which includes both V_E and V_G, is sufficiently smaller than the width of the Gaussian fitness function, the degree of plasticity within a population should not be highly detrimental. Under these conditions, the variability originating from plasticity is not as deleterious as it may be for populations under stringent selection.

2.4.5 Establishment and Maintenance of Plasticity

While prior studies have shown how plasticity is often deleterious under constant environmental conditions (c.f. **Prior Models**), at best serving to avoid extinction during dramatic environmental shifts, we have demonstrated that under similar static conditions, plasticity can both be advantageous (Figure 2.4) or deleterious (Figure 2.8). In our model, phenotype is a single-valued variable, produced by a combination of both genic and non-genic/epigenetic

contributions. As multiple mechanisms can produce the same phenotype, any given phenotypic value in our model is highly degenerate, with a broad set of genotypic parameters (genic and non-genic) being able to produce the same outcome. However, while phenotype may be degenerate, thus allowing individuals to survive a single selection event, the degree to which a phenotype is produced through genic pathways vs epigenetic pathway produces trade-offs in subsequent generations. The more a phenotype is dependent on epigenetic compensation, the less likely that subsequent generations will produce a similar response in comparison to producing the same phenotype through mostly genic means. Intrinsically, this effect forces a strong trade-off during the evolution of plasticity - progeny may be better optimized for current conditions but will be less likely to survive further environmental changes.

Consideration of this trade-off leads to the natural conclusion that under constant conditions, quantitative traits should begin to canalize and lose plasticity. However, under changed environmental conditions, natural selection will also favor the fastest route to accommodation of a new environment, whether it be by genic or non-genic means. Indeed, this effect manifests itself as the so-called "Baldwin Expediting Effect" [11], where plasticity appears to be transiently beneficial during adaptation, first increasing, then subsequently being selected against [13]. Similarly, optimization of adaptation rate also explains why the switching rate in models of phenotypic switching [16, 18, 19, 22, 41, 51] will match the rate of environmental fluctuations. In general, such transient increase in variance is a result of minimizing adaptation time [57].

While previous studies suggest phenotypes would begin to canalize under static conditions, we have demonstrated that plasticity may in fact be advantageous under certain extreme conditions, such as when genic solutions are completely inaccessible (Figure 2.4). To further explore this transition, we constrain the relative mutation rates of genic and non-genic control parameters. We define a new variable $r \equiv \mu_{genic}/\mu_{epigenetic}$, where μ_{genic} and $\mu_{epigenetic}$ are the mutation rates for genic and epigenetic control parameters respectively.

We can see now that in the case of full epigenetic compensation, plasticity increases when $r = 0$ (Figure 2.4), and when genic control is mutable, decreases when $r = 1$ (Figure 2.9).

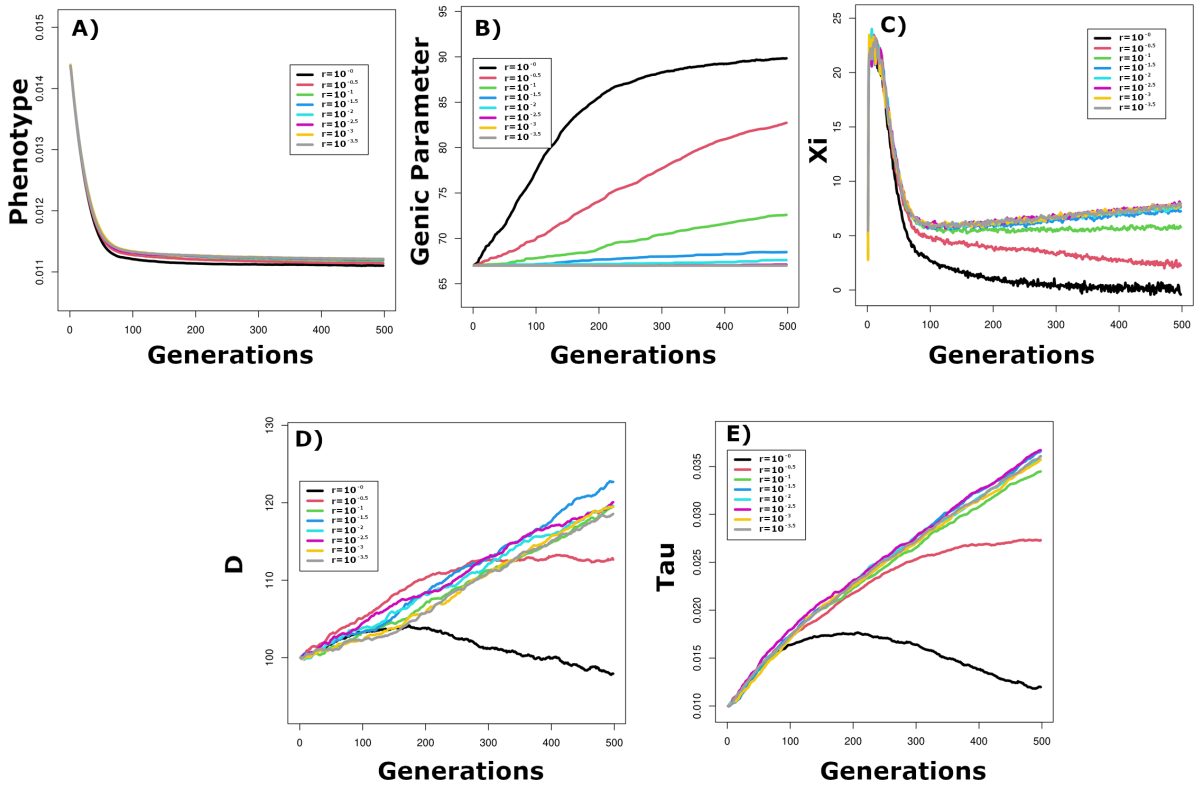

Figure 2.9: **Relative mutation rate determines evolution of plasticity** Populations (100 replicates) consisting of plastic individuals ($N = 1000$) were presented with a new environment while the relative rate of mutation of genic and non-genic parameters (r) was varied. Shown here is the mean population behavior at each r value, averaged over 100 replicate populations, one line per r value. While in all cases, **(A)** phenotypic accommodation of the new environment occurred rapidly, **(B)** the ability of individuals to provide a genic solution (ϵ_i) to new conditions was hindered. **(C)** Where genic solutions were not easily produced, epigenetic compensation occurred, **(D-E)** allowing both non-adaptive and adaptive plasticity to be advantageous within the given epoch.

We applied *in silico* selection (stringent selection) to 100 replicate populations of 1000 individuals while varying the relative mutation rate r (Figure 2.9). As r is varied, the relative rate of phenotypic accommodation remains roughly the same in all cases (Figure 2.9A), though, slightly more rapid accommodation is seen at higher r values. While the mean phenotype in these populations remains similar, the genic contribution to phenotype (Figure 2.9B) varies dramatically as the relative mutation rate for genic parameters is decreased. In

the case of low r, as the new environment is phenotypically accommodated, the epigenetic value ξ_i instead compensates for the inability to produce an optimal phenotype using genic means alone (Figure 2.9C). As seen in previous sections, ξ_i compensates for sub-optimal genotypes initially under all conditions, then within the given epoch, ξ_i is no longer needed for large r, returning to a mean of zero, reflecting the canalization of this phenotype by genic means. However, for small r, ξ_i continues to be important for the production of optimal phenotypes, subsequently increasing again. The genetic control of ξ_i (i.e. D_i and τ_i) similarly stratifies depending on r (Figure 2.9D-E). As r decreases, genic solutions to new environmental conditions become less evolvable and epigenetic compensation becomes an increasingly important mechanism for survival, causing both non-adaptive and adaptive plasticity to be more advantageous within the given epoch. Should a further shift in environment occur after the current epoch, the net result would be that both adaptive and non-adaptive plasticity are overall increased within a population, rather than canalizing. However, as the limit of r approaches zero and genic solutions become completely inaccessible, the amount of time until canalization will approach infinity.

2.5 Discussion

2.5.1 West-Eberhard Model

Mary Jane West-Eberhard has proposed that plasticity plays a central role in the process of adaptation. Under this model, phenotypes that produce differential responses to input encounter a newly recurring input that is both phenotypically and genotypically accommodated. However, previous genetic models have been limited in their capacity to recapitulate this idea. In particular, quantitative genetic models involving explicit and parameterized phenotypes, while able to provide insight into forces acting on overall non-genetic phenotypic variability, typically cannot produce differential responses to a new input and do not

allow for the possibility of phenotypic accommodation [13, 24, 34, 48, 55]. Even models of phenotypic switching–which allow for the possibility of variable responses to a new input, degenerate genotype-phenotype relationships, and phenotypic accommodation [16, 21]–cannot model genetic accommodation beyond adjusting the switching rate to maximize survival to fluctuating environmental conditions.

The model used in this study provides a way of examining the behavior of populations that are phenotypically responsive to novel environmental inputs. As demonstrated by replicate simulations of populations in which only phenotypic parameters were allowed to mutate in response to stringent selection, full adaptation to new environmental conditions occurred within fifty generations (Figure 2.4). All populations reached a new optimum phenotype that was then maintained solely through epigenetic rather than genotypic change, as shown by the fact that the epigenetic parameter ξ_i remained at a quantity above 0. As such, our model not only shows the progression of plastic populations' phenotypic change in response to novel inputs, but our model also has the capacity to distinguish between genic and epigenetic methods of accommodation. It therefore stands in contrast to otherwise similar models, including neural network models of gene regulation [46, 61]. Such models, which can provide differential responses to input, may demonstrate complex behaviors given relatively few assumptions and have even been shown to depict canalization when applied to real data [62, 63]. Such models are able to produce networks that are robust to fluctuating environmental conditions; however, such networks must be evolved genetically/trained to be able to produce such responses [61]. As such, segregation of phenotypic accommodation and genotypic accommodation is difficult in such models.

2.5.2 Changing Environments

Our model shares conceptual similarities to other models of phenotypic plasticity, as well as to learning behaviors [11, 13, 23, 24, 54]. In prior models [11, 13, 23, 24], the survival of

an individual depends upon whether or not the selected phenotype is within an individual's norm of reaction when selection is applied. Similarly, in our model, a genotype is more likely to survive depending on whether or not the optimal phenotype is contained within the phenotypic distribution produced by a given genotype. However, this result is contingent upon the individual with said genotype being in a favorable phenotypic state at the time of selection, a key difference between these prior models and our model. Specifically, the former are unable to distinguish between genetic and epigenetic effects. To wit, in prior models, genotype strictly defines the norm of reaction, where individuals possess elevated fitness only if a new optimal phenotype is contained within that norm of reaction. Such a model is not only unable to recapitulate the West-Eberhard model, but also assumes that plasticity may only be adaptive in response to an environmental shift. This stands in contrast to our model, where an optimal phenotype may be well outside the associated phenotype distribution determined by the genotype but can still be produced regardless via epigenetic means.

Nevertheless, prior models, like the Ancel model, provide insight into how plasticity may ameliorate the effects of an environmental shift [11, 13, 23, 24]. The Ancel model is primarily concerned with understanding whether plasticity/learning may accelerate adaptation by examining the ideas behind the Baldwin expediting effect [54]. Using her model, the author is able to demonstrate that plasticity may only "expedite the search from an initial population distribution to the first encounter with the optimum phenotype" and that this effect is observed for "initial genotype distributions sufficiently distant from the target." [11]. Our model produces similar results in that plasticity does indeed expedite a population's first encounter with a more fit phenotype in the case of both relaxed and stringent selection. However, whether or not plasticity ultimately increases a population's rate of adaptation depends upon the conditions under which selection occurs; in our model, the optimal phenotype remained the same in both stringent and relaxed selection.

Further effects not previously described are also seen with our model. In the case of populations under stringent selection, once both genic and epigenetic parameters are allowed to mutate, plastic populations show a much more consistent and directed progression towards the optimum genotype as opposed to non-plastic simulations (Figure 2.5). Unlike in the case of the non-plastic simulations, where many of the individual plastic populations transiently display several extremely sharp spikes followed by plateaus, plastic populations adopt genotypes progressively nearer to the optimum. The genotypic progression of the plastic populations under stringent selection forms a well-defined curve before reaching a plateau around 90, with deviations caused by genetic drift. Plasticity thus allows these populations to adapt more quickly as they progress along a selective gradient that has become sufficiently smoothed in comparison to the non-plastic populations. This same trend is not seen in plastic populations under relaxed selection, however, which fail to converge on an optimum phenotype in the number of generations that non-plastic populations do in following a selective gradient.

We find that the rate of adaptation, as reflected by the cumulative fraction of genetically adapted populations (Figure 2.5), is dependent on both the manner of selection and the degree of plasticity in a population. This result is in agreement with the Ancel model, which states that plasticity does retard adaptation when both plastic and non-plastic populations readily adapt to a new optimum, as opposed to needing to undergo a large number of changes to adapt. However, in contrast to the Ancel model, genotype in this case may be segregated from phenotype due to epigenetics, and as such, initial genotypic distributions for all simulations in this section were identical with $\epsilon_0 = 67$ in both plastic and non-plastic conditions. Simply taking the difference between initial and optimal phenotypic and genotypic values is not sufficient: the overlap of phenotypic and genotypic distributions with selection gradients must be considered. Our results also suggest that while non-plastic populations are highly sensitive to the conditions of selection that are applied, plasticity

allows populations to be more robust to such variation, a feature that is absent from the Ancel model. Such results may have broader implications for interpreting substitution rates based on sequencing data, as plasticity "smoothens" out the degree of genetic turnover (Figure 2.7).

2.5.3 Maintenance of Plasticity Under Static Conditions

For plasticity to be maintained in a population, the results of previous studies suggest that a constantly and randomly fluctuating environment may be the only method by which plasticity could be maintained (c.f. **Prior Models**). The maintenance of plasticity under static conditions remains an open problem.

Our model has shown how plasticity may variously increase, decrease, or be maintained under different selective conditions. Specifically, we have shown that when mutation of genic parameters is disallowed, increased plasticity is selected for under static conditions (c.f. **Phenotypic Accommodation**). We also have shown that, so long as the range of phenotypes produced both by genotypic and environmental variability is sufficiently within the bounds of directional selection, plasticity may be maintained in a population under static conditions. Alternatively, if the range of phenotypes produced exceeds the bounds set by directional selection, decreased plasticity is selected for (c.f. **Loss of Variability**).

When undergoing change to adapt to unfavorable environments, populations that express a variety of phenotypes are more likely to be able to adapt quickly, since beneficial phenotypes are more likely to be present and selected for. However, this same trait may act as a detriment under steady-state conditions, as the potential for variation causes the phenotypes to deviate from the ideal. Therefore, in agreement with prior results, plasticity within populations may reduce over time, especially if genic solutions are easily evolvable, resulting a loss of variability and canalization (Figure 2.8).

However, our results also demonstrate how plasticity may be favored and subsequently

increase within a population within static environmental conditions (Figure 2.9). We find that in altered conditions, so long as epigenetic solutions are more evolvable than genic solutions, plasticity will be beneficial. Central to this result is variation in the relative mutation rate for genic and non-genic control of phenotype. A number of factors could contribute to r, as pleiotropy, linkage, epistasis, essentiality, or other genetic conflict can prevent the occurrence of certain types of changes. Similarly, the underlying architecture of the genetic networks providing both genic and non-genic control of phenotype may also provide for an overall larger or smaller substrate for mutations by simply having a larger number of mutable positions within the genome.

We also note that the trajectory of genetic control of plasticity is not only dependent on r, but also the epoch of time considered. If conditions remained static for sufficiently large times, all traits would eventually undergo canalization. Similarly, as r approaches 0, the time until canalization increases without bound. The simplest way to define the end of an epoch would be a subsequent shift to a new environment. In this case, plasticity could continue to increase in a population so long as the time before a subsequent environmental shift is shorter than the time required to undergo genetic accommodation of a new environment. Alternatively, the considered epoch may also end when the relative mutation rate r is altered.

2.6 Acknowledgements

UL would like to thank Dr. Marsha R. Rosner for their guidance and help, without which this manuscript would not have been possible. Many of the ideas in this manuscript originated in a workshop on Robustness, Heterogeneity and Adaptation that was organized by Dr. Rosner in cooperation with Dr. John Reinitz. U.L. is funded in part by GM7197-42. M.L. was supported by the National Institutes of Health grant 1R01GM116113-01A1. This manuscript is dedicated in memory of Frances Lee.

2.7 Additional information

All authors declare no competing interests.

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables.

Code to simulate and generate figures may be found at :

'https://github.com/avianalter/sde_evo'

CHAPTER 3

THE ROLE OF THE 3-DIMENSIONAL GENOME IN NEW GENE EVOLUTION

3.1 Abstract

In efforts to explain how duplicate gene copies may rise to fixation in a population, previous models of new gene origination have underappreciated the importance of the 3D genome in this process. We show that positional effects on distally duplicated genes, i.e. enhancer capture, is an efficient mechanism for accommodation of new selective conditions. By performing a co-expression analysis on *D. melanogaster* tissue data and comparing essential to non-essential genes that have newly evolved, we show that enhancer capture is a significant driver of new gene evolution in distally duplicated genes. The new essential gene, HP6/Umbrea, is used as a model for understanding enhancer capture, as it evolved via a full duplication of the parental gene, its subsequent protein evolution is known, and it duplicated into a gene-poor region of the genome. HP6/Umbrea's expression pattern divergence from its parental gene, HP1b, as well as its high co-expression with neighboring genes suggest that it evolved via enhancer capture. ChIP-Seq data shows the presence of active enhancer marks appearing near HP6/Umbrea coinciding with onset of its expression which likely regulates HP6/Umbrea, it's neighboring gene, as well as a distally located 6-gene cluster also found co-express with HP6/Umbrea. To test for co-expression, we find that these three loci, the putative enhancer, HP6/Umbrea, and the 6-gene cluster are in close physical proximity in the 3-D genome of *D. melanogaster*. Finally, we compare Hi-C data from two species with HP6/Umbrea, *D. melanogaster* and *D. yakuba*, to two species pre-dating HP6/Umbrea's insertion, *D. pseudoobscura* and *D. miranda*, showing that co-regulation of these same elements is the ancestral state and thus that HP6/Umbrea evolved via enhancer capture.

3.2 Introduction

Genes arising from the class of duplication-based mechanisms are commonly inferred using synteny- and homology-based searches (Figure 3.1). While new genes are systemically understudied in comparison to their older counterparts, the most well-studied class of new genes are those originating from duplication-based mechanisms. However, in studying the evolutionary dynamics of duplication-based origination, a paradox arises: how do functionally redundant copies of the same gene rise to fixation?

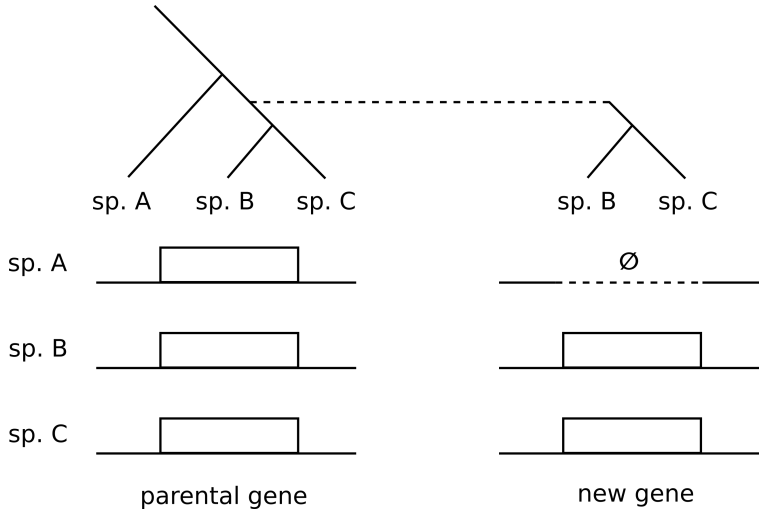

Figure 3.1: **Identification of new genes.** The insertion of a new gene may be inferred by using syntenic alignments of closely related species. Gaps within these alignments may be used to determine the location of a new gene insertion, while reciprocal-best searches may determine whether a gene arose via duplication as well as the identity of the parental gene.

The first models describing new gene evolution proposed that all new genes likely evolve via duplication-based mechanisms [64]. Under such a model, a duplicate copy of a gene is shielded from selective pressures, acquiring new mutations until a neo-functionalized copy of the gene provides sufficient selective force to carry this new gene to fixation. However, until such advantageous function acquired, the new gene copy is subject to genetic drift and is thus unlikely to rise to sufficient prevalence in a population to allow for rare, neo-functionalizing mutations to occur. This problem has been referred to as "Ohno's dilemma." [29].

Various models have been proposed to resolve this problem - the duplication, divergence, complementation (DDC)/sub-functionalization model [27], the escape from adaptive conflict (EAC) model [28], and the innovation, amplification, and divergence (IAD) model [29, 30] as well as its functional equivalent, the Adaptive Radiation model (AR) [65]. However, these models fail to appreciate how the 3-dimensional genome and its corresponding regulatory landscape can drive neo-functionalization of a new gene from the moment of duplication (Figure 3.2).

To address how a duplicate, redundant gene copy may rise to fixation, these models all assume multiple functions for any studied gene. For pleiotropic genes, the DDC/sub-functionalization model allows for complementary non-functionalization of multiple functions that are originally shared between the duplicated copies (Figure 3.2a, b). Given a loss of function in one copy of the gene, the ability of the duplicate copy to compensate for this original loss of function confers a selective advantage to the duplicate copy. Eventually, under the DDC model, increasing divergence allows for the partitioning of multiple sub-functions between gene copies. Alternatively, while the DDC model allows each duplicate copy to posses only a subset of the original functions of the parental gene, the EAC model allows for increased optimization of multiple functions within the ancestral gene as each function partitions to each paralogous copy. Under this model, it may not be possible for a parental gene to simultaneously optimize each of its multiple functions. As such, duplication can allow for the relaxation of constraint on the evolution of the ancestral gene, thus resolving conflict and allowing for a selective advantage in both parental and new genes. While the DDC and EAC models can explain how prior gene functions can be partitioned amongst duplicate copies, these models fail to provide a mechanism for true neo-functionalization.

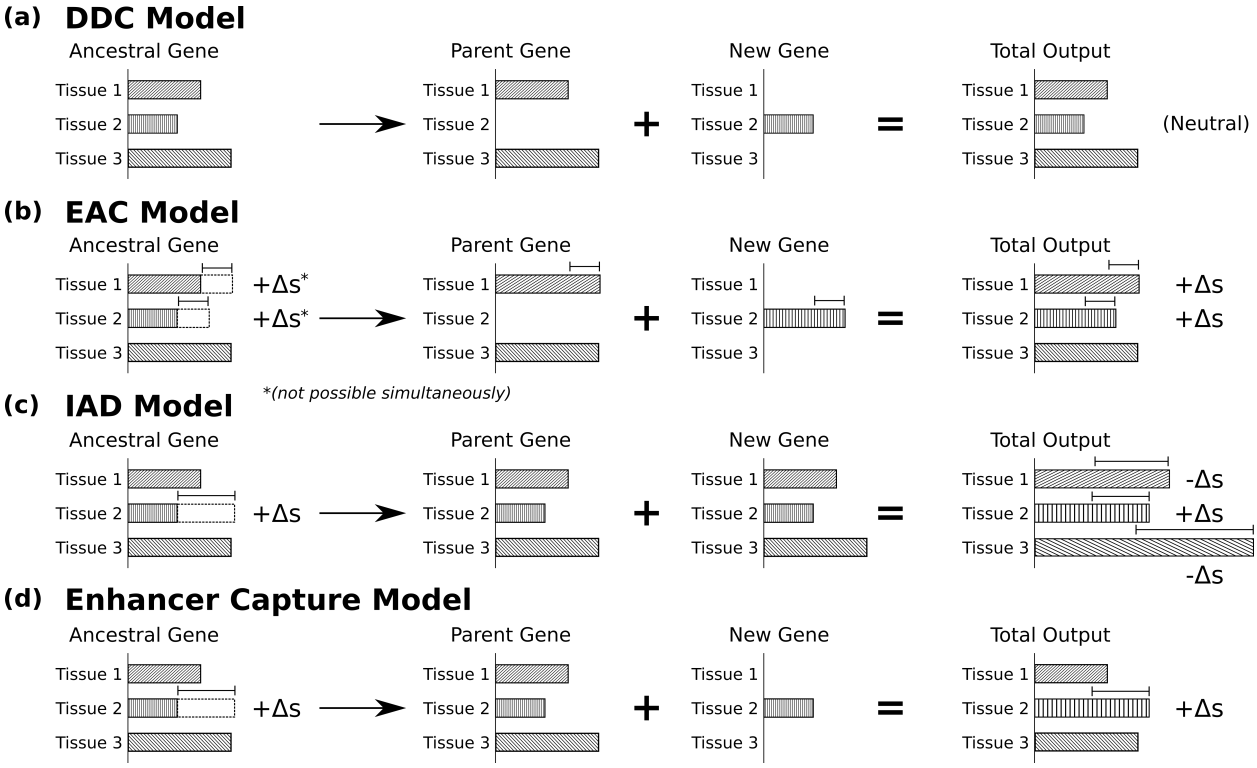

Figure 3.2: **Comparison of extant models.** *Continued on next page...*

One common thread amongst the DDC and EAC models is their conformance to one of Kimura and Ohno's five governing principles of molecular evolution: "Gene duplication must always precede the emergence of a gene having a new function" [66]. While complementary/optimizing mutations may stabilize the appearance of a duplicate gene copy, these mutations may only occur after the duplication of a new gene. The IAD model provides an alternative to this process by allowing for duplication itself to provide neo-functionalization via increased dosage for an auxiliary function of the original gene (Figure 3.2c). Here, the IAD model begins with an ecological shift that favors an auxiliary function of a gene, thus providing a selective advantage for high copy number. Importantly, as events of unequal crossing over are more common than point mutations, gene duplication occurs more frequently than substitutions and can thus fix in a population before regulatory changes evolve. Following this amplification, subsequent changes are accumulated on the various copies, allowing for divergence [29].

Figure 3.2, continued. Various evolutionary models have been proposed to explain how redundant gene copies become fixed in populations ("Ohno's Dilemma"). Presented are illustrations for the **(a)** Duplication-Divergence-Complimentation (DDC), **(b)** Escape-from-Adaptive-Conflict (EAC), **(c)** Innovation-Amplification-Divergence (IAD), and **(d)** enhancer capture models, where the gene regulation of three tissue types are considered and optimal conditions are shown in dotted boxes. Under the DDC model **(a)**, redundancy allows for compensation of any single loss-of-function event, eventually causing the expression pattern of the original gene to be segregated between both parent and new genes. Given that the original protein is produced by both the parent and new genes, the total output is identical to the original gene, and is thus a neutrally evolving process. Under the EAC model **(b)**, two functions cannot be optimized within a single gene copy, and this conflict is resolved via the act of duplication, allowing for simultaneous optimization of both copies. The total output of these two gene copies now has higher fitness than the output of the original gene, rising to fixation. Under the IAD model **(c)**, an environmental shift causes increased selection for an auxiliary function of the original gene. As duplication events (unequal crossing-over) occur more frequently than point mutations, duplication of the original gene provides a more rapid accommodation of the new environmental conditions than regulatory mutations by increasing dosage. However, while this model allows for increased fitness due to increased auxiliary function, one issue in this model is that this increase in fitness must also overcome the penalty imposed by over-activity of all other functions. Over-activity is generally not an issue when environments change sequentially, as is the case of single-celled organisms, but incorrect regulation can be a significant barrier in multi-cellular organisms, e.g. in the case of key transcription factors. Under the enhancer capture model **(d)**, increased expression of a single function provides a selective advantage. A region of the genome contains an enhancer/pre-enhancer that increases fitness once a gene copy duplicates into a region under its control (thus activating it in the case of a pre-enhancer). As the original protein is produced by both parent and new gene, the total output of both parent and new gene increases overall fitness, thus driving both copies to fixation.

While the IAD model provides a reasonable explanation for gene family expansions, particularly in the case of tandem duplications, some serious problems remain with the model, particularly when applied to multi-cellular organisms. While it is assumed that an ecological shift selects for higher copy number, it is not only the auxiliary function that is thus highly expressed, but the original function as well. The selective advantage conferred by increased dosage not only needs to be sufficiently greater than the metabolic costs of excess protein production, but it also needs to exceed potential deleterious effects caused by amplification of the original function. Depending on the spatio-temporal expression of the original gene, duplicate copies of the original gene will likely need to occur in a tissue-specific manner so as not to disrupt processes downstream of the original gene. Such precisely controlled expression is generally not of concern in single-celled organisms, where gene family expansions occur quite frequently. However indiscriminate expression of, for example, transcription factors within multi-cellular organisms will present a large selective barrier that copy number expansion must overcome, particularly if aberrations occur within key developmental processes.

One key factor missing in these models is the effect of chromosomal context on a new gene's regulatory function. A common thread amongst these various models is a separation of the initial establishment of a duplication followed by subsequent changes accumulated by various duplicate copies. Additionally, these models require that genes posses multiple functions. As an alternative to these models, we demonstrate that regulatory innovation via enhancer capture can also be a source of evolutionary novelty, allowing for rapid rewiring of gene regulatory networks in a single neo-functionalization step. During enhancer capture, neo-functionalization arises from the act of duplication itself by recombining pre-existing protein sequences with regulatory sequences, highlighting the importance of the three-dimensional eukaryotic genome in new gene evolution (Figure 3.2d).

3.3 Results

3.3.1 Analysis of Tissue Co-Expression Shows New Genes Evolve by Enhancer Capture

Central to the IAD model is the observation that gene duplication via unequal crossing over is more likely to occur than a point mutation [29, 30]. As previously described, one issue with this model is that there is an implicit assumption that during the environmental shift, the increase in fitness gained by over-activity of the auxiliary function must be greater than the decrease in fitness imparted by over-activity of the gene's original function(s). In the case of single-celled organisms where environments are encountered sequentially, it is reasonable to assume that selection might tolerate over-activity of the gene's original function during the transient environment in which the auxiliary function is favored. However, the decrease in fitness for improper expression or activity is larger in multi-cellular organisms than in single-celled organisms, where a multi-cellular organism's overall phenotype is the cumulative (development) and simultaneous (organ systems) product of many different gene functions.

In the case of multi-cellular organisms, selection may increase for the expression of a gene within a single tissue type (Figure 3.2d). Under the IAD model, a full duplication will drive duplicate gene copies to fixation as it provides the most evolvable solution to new conditions. In contrast, under the enhancer capture model, a copy of the original gene duplicates into a region of the genome containing an active enhancer that increases expression in a tissue-specific manner. Alternatively, the new gene may migrate into a region of the genome containing unbound transcription factor binding sites, thus activating a pre-enhancer region into a new enhancer. Since the total output of the enhancer capture model does not produce over-expression in other tissues like in the case of the IAD model, given sufficiently high population size, enhancer capture will be the more dominant mechanism for gene duplication, particularly with regards to distal/non-tandem duplications. This increase in fitness caused

by the combined output of the new and parental genes thus drives both copies to fixation, providing an alternate resolution to Ohno's Dilemma. While enhancer capture remains the most rapid path to increasing fitness, compensatory mutations in the regulation of the parental gene may also provide a tissue-specific solution to increased selection. Once a compensatory mutation occurs, or even more simply, once the tissue-specific selection is relaxed, the new gene may then begin to diverge, accumulating substitutions.

Each model of gene duplication produces unique relationships between the expression patterns of a new gene vs its parent gene and/or a new gene vs its neighboring genes. As such, we may test whether enhancer capture drives the evolution of new genes evolving via distal/non-tandem duplication by utilizing tissue co-expression data. Specifically, we may predict to what degree a new gene will show tissue co-expression with its parental gene as well as with its neighboring gene depending on if the mechanism driving its evolution falls under the DDC, EAC, or enhancer capture models.

Under the DDC or EAC models, the tissue expression patterns of parental and new genes are complimentary, resulting in low co-expression between parental and new gene copies ("parental co-expression"), while the tissue expression patterns of the new gene and its neighboring genes should have no relationship, resulting in random co-expression between the new gene and its neighboring gene co-expression ("neighboring co-expression"). Under enhancer capture, a broadly expressed parental gene acquires increased expression in select tissues by duplicating into a distant region of the genome under the control of an enhancer. Here, parental genes are expected to have broad tissue expression patterns, while new genes have expression patterns with high tissue specificity, resulting in low parental co-expression. On the other hand, since the new gene becomes regulated by the captured enhancer that is already influencing other genes, neighboring co-expression is high.

A tissue expression data set was obtained from FlyBase [67, 68] (c.f. **Methods and Materials**) and co-expression between new/parental and new/neighboring gene pairs was

calculated (Spearman correlation coefficient) for a set of new genes (N=87) which underwent a distal/non-tandem duplication of > 500kb whose essentiality has been validated experimentally [69]. This data contained tissue types extracted from both L3 larvae, pre-pupae, and adult flies, including gut, salivary glands, and imaginal discs from wandering L3 larvae, as well as the head, ovaries, gut, and reproductive organs from adults (c.f. **Methods and Materials**). For tissues that were represented with multiple experimental runs, data from those tissue types were averaged prior to further analyses to avoid representation bias.

The resulting parent/neighbor co-expression plots ("PNC plot") for new essential genes (Figure 3.3**a**), new non-essential genes (Figure 3.3**b**), and both essential and non-essential genes (Figure 3.3**c**) can be used to test whether a significant number of distal/non-tandem duplications evolve via enhancer capture. We may define "low" and "high" co-expression as being below or above the median co-expression value across all distally duplicated new genes respectively. Genes that have evolved via enhancer capture should appear in the lower right quadrant in the PNC plots, as the expression patterns of the new gene diverges from the parental gene while the new gene and neighboring gene share the same expression pattern. Similarly, genes with that have evolved via the DDC or EAC models should appear in the bottom half of the PNC plots, with low parental co-expression resulting from divergent and complimentary expression patterns, and random neighboring co-expression as there is no expected relationship with the new gene and its neighboring genes.

Whiles genes in the lower right quadrant of the PNC plot may have evolved via the DDC/EAC models or enhancer capture, one key distinguishing feature of both models is how essential function is expected to partition between new gene and parental gene. Under the DDC/EAC models, all segregable functions of the original gene are expected to partition randomly between both parent and duplicate gene copies. As such, these models predict that essential gene function should also equally partition between both parent and new genes. The DDC/EAC models thus predict that the ratio of essential:non-essential genes in the entire

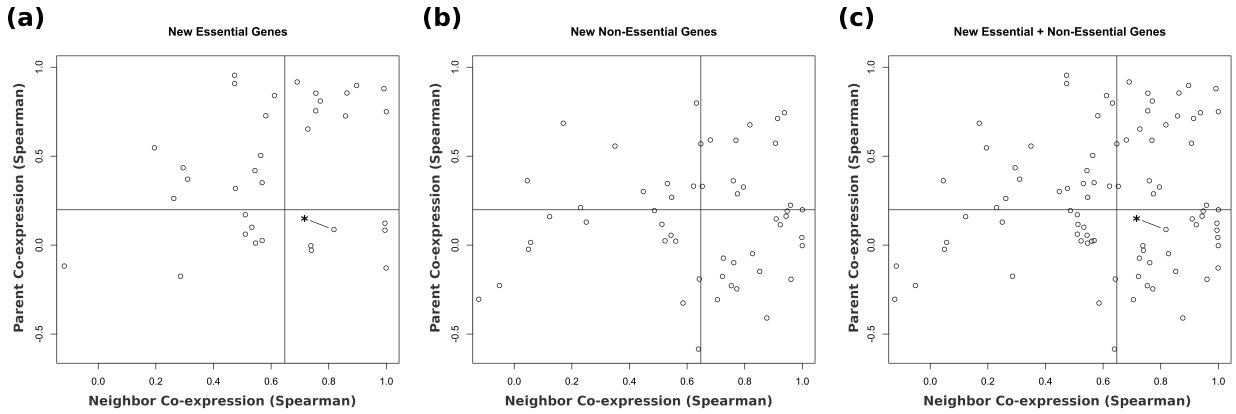

Figure 3.3: **New genes evolve via enhancer capture.** Shown are parent/neighbor tissue co-expression patterns for new genes in *D. melanogaster* which have migrated either more than 500kb away or between chromosomes (new essential genes **(a)**, new non-essential genes **(b)**, and combined essential & non-essential genes **(c)**). Tissue co-expression (Spearman correlation coefficient) between new gene/parental gene pairs is plotted on the vertical axis while maximal tissue co-expression between new gene/neighboring genes pairs is plotted on the horizontal axis. Note, the co-expression between the new gene and four of its neighbors was calculated, two on each side, and the maximal co-expression is reported here. Vertical and horizontal lines indicate median co-expression value of all distally duplicated new genes as in **(c)**. Genes which evolved via enhancer capture are expected to have low parental co-expression and high neighboring co-expression and should thus be present in the lower right quadrant. Genes evolving under the DDC or EAC models should have low parental co-expression due to complimentary expression patterns and random neighboring co-expression. While a new gene's essential function is equally likely to be partitioned between either parent or new gene under the DDC or EAC models, new genes evolving via enhancer capture are unlikely to have essential function, as the expression of the new gene will only augment existing expression of the parental gene, leaving the original essential function intact. Comparing the overall ratio of new essential to new non-essential genes (35:52) to the ratio of new essential to new non-essential genes showing high neighboring/low parental co-expression (6:16) shows that new genes evolve via regulatory capture (Fisher's Exact, p=0.0256). (***** denotes HP6/Umbrea.)

lower half of the PNC plot, including the lower right quadrant, should match the overall ratio of essential:non-essential genes.

Alternatively, the enhancer capture model predicts that most function, including essential gene function, will remain with the parental gene copy, while the tissue-specific expression pattern of the duplicate gene copy serves only to augment the function of the parental gene. Specifically, selection for increased expression in a single tissue will result in elevated tissue-specific expression via the new gene copy, while all other function is retained in the parental copy, including its essential function; the new gene evolving via enhancer capture is expected to be non-essential while the parental gene is expected to be essential. As such, the enhancer capture model predicts that the ratio of new essential:new non-essential genes in the lower right quadrant of the PNC plot should be significantly lower than the overall ratio of new essential:new non-essential genes. Using the parent/neighbor co-expression plots, the ratio of new essential:new non-essential genes in the lower right quadrant (6:16) was found to be significantly lower than the overall ratio of new essential:new non-essential genes (35:52) using Fisher's Exact test (p=0.0256), suggesting that distally duplicated genes in *Drosophila melanogaster* primarily evolved via enhancer capture (Figure 3.3).

3.3.2 HP6/Umbrea as a Model for Enhancer Capture

While new genes categorically remain understudied, the evolution of HP6/Umbrea is a well-suited model system for understanding the enhancer capture model as it is one of the few new genes whose protein evolution has been previously described in the literature (Figure 3.3(*)) [70]. HP1b, a gene located on the X chromosome, duplicated approximately 12-15 mya into an gene-poor intronic region of dumpy, located on chromosome 2L (Figure 3.4). The new gene, HP6/Umbrea, was the result of a full duplication which included HP6/Umbrea's promoter region as well as its three known domains: the chromo domain, the chromo-shadow domain, and the hinge domain connecting the two.

Though HP6/Umbrea was lost ancestrally to multiple speciation events [70], suggesting that the gene was not originally essential for life function, HP6/Umbrea continued to evolve in a step-wise manner. HP6/Umbrea subsequently lost its chromo domain approximately 10-12 mya; this was followed by an accumulation of key substitutions 0-7 mya, resulting in HP6/Umbrea's known essential protein function in *D. melanogaster* [69, 71, 72]. Using these results, protein neo-functionalization may be eliminated as the driving force behind the fixation of HP6/Umbrea given it's step-wise protein evolution. Sub-functionalization and/or subsequent optimization of protein function may also be eliminated for similar reasons.

A simple comparison of HP6/Umbrea's expression pattern to the parental gene HP1b's very broad expression pattern suggests that HP1b is likely under the control of a simple constituitive-on promoter. Alternatively, while HP1b is found in all tissues, HP6/Umbrea is found only in a subset of tissues in which HP1b is found, suggesting that the duplication of HP1b's constituitive-on promoter into a region under control of an enhancer resulted in HP6/Umbrea's tissue expression pattern. This expression pattern is similar not to its neighboring gene, dumpy, but its second neighboring gene, CR44609 (Figure 3.4), expressing primarily in the imaginal discs and male reproductive organs, demonstrating that these genes are likely co-regulated. Given that the tissue expression patterns of HP1b and HP6/Umbrea are not complimentary, sub-functionalization and/or subsequent optimization of regulatory function may also be eliminated as the driving force behind HP6/Umbrea's fixation.

In addition to results excluding other models, publicly available modENCODE ChIP-Seq/ChIP-Chip data [73] provides positive evidence that enhancer capture likely drove the early evolution of HP6/Umbrea. Using the embryonic S2 cell line as a negative control where there is little/no HP6/Umbrea expression, poised (H3K4me1) and primed (H3K27ac) enhancer marks in whole L3 larvae show strong enhancer activity in an intronic, gene-poor region of dumpy (Figure 3.4). Given the absence of other genes in the region (Figure 3.5a), HP6/Umbrea remains the likeliest target of the enhancer based on proximity and expression.

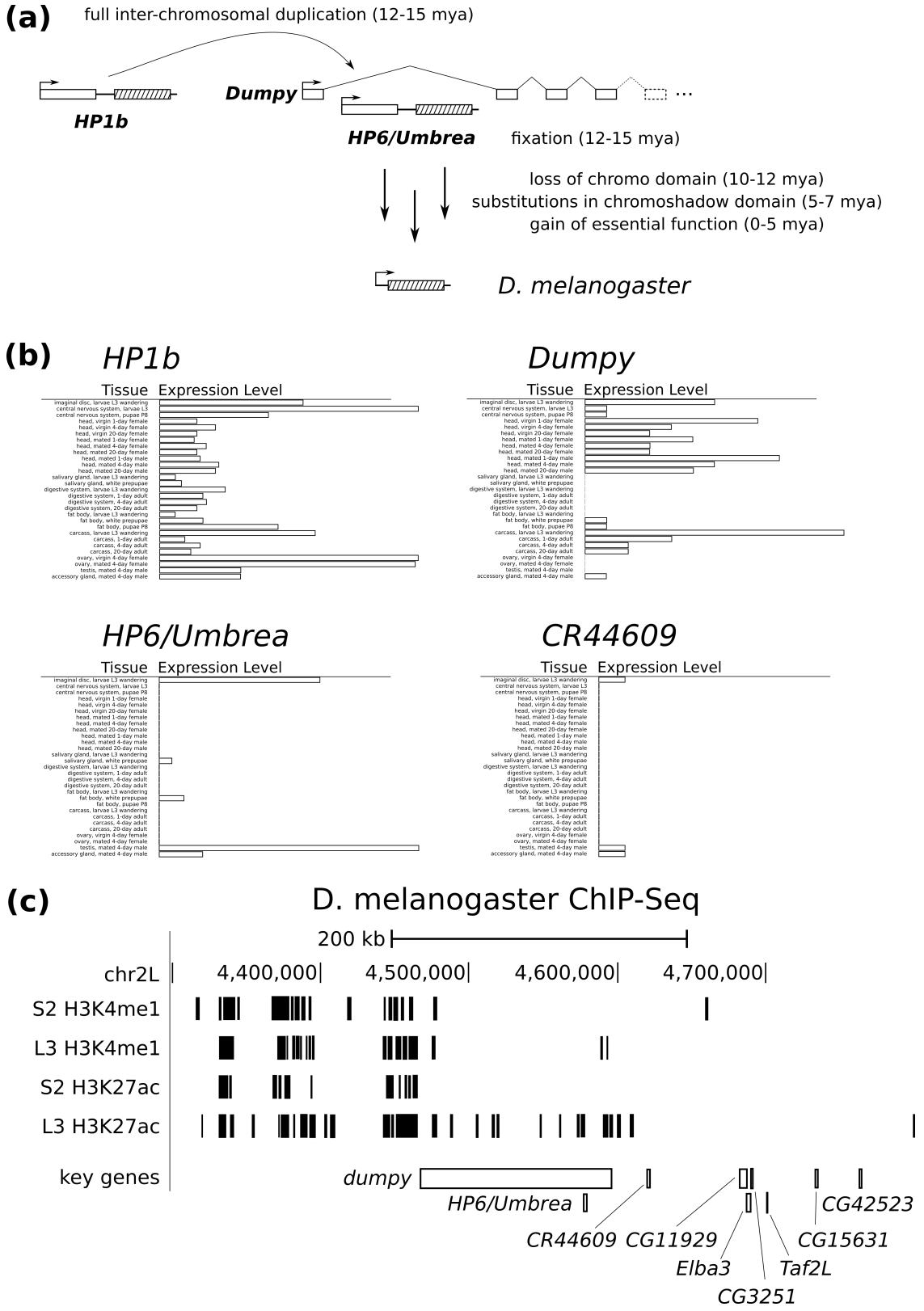

Figure 3.4: **HP6/Umbrea evolved via enhancer capture.** *Continued on next page...*

Figure 3.4, continued. **(a)** HP6/Umbrea is a new essential gene in *D. melanogaster* which arose from a full duplication of HP1b into an intronic region of dumpy, migrating from chromosome X to 2L. HP6/Umbrea's well characterized, step-wise protein evolution suggests that amino-acid substitutions were unlikely to have driven the duplicate gene copy to fixation. **(b)** Unlike the broad expression pattern of HP1b, the tissue expression pattern of HP6/Umbrea is stereotypical of new gene expression patterns, with high tissue specificity, restricted in this case to primarily the imaginal discs and male reproductive organs. This expression pattern is shared with HP6/Umbrea's neighboring gene CR44609. **(c)** A comparison of ChIP-Seq markers for primed (H3K4me1) and active (H3K27ac) enhancers between embryonic S2 (no/low HP6/Umbrea expression) and whole L3 larvae (high HP6/Umbrea expression) tracks shows strong activation of a larval enhancer in a $\tilde{1}$00kb intronic region of dumpy that is, aside from HP6/Umbrea, devoid of protein coding genes.

Given that it appears that HP6/Umbrea duplicated into a region that appears to be under the control of a pre-existing enhancer, we tested for further co-regulation in the region by using tissue expression data (c.f. **Analysis of Tissue Co-Expression Shows New Genes Evolve by enhancer capture**). We then applied a correlational analysis on this tissue expression data set to determine whether HP6/Umbrea is co-regulated with other neighboring genes. We took a 500kb region of the genome centered on the insertion site of HP6/Umbrea and calculated the tissue co-expression of each gene within this region. As enhancers function in a proximity-based manner, we would expect a distance-dependent effect on the co-expression of neighboring genes across the genome. To generate a baseline estimate of this distance-dependent co-expression distribution, we sampled 1000 random genic loci within the *D. melanogaster genome*, calculating the degree of co-regulation expected on proximity alone. Notably, we find that using this distribution, the region of influence of any given regulatory region of the genome appears to be on the order of 25kb, suggesting that this is a characteristic distance for enhancer interaction in *D. melanogaster*. Outside of this region of influence, the likelihood of co-expression relaxes to the genomic average. Therefore, genes found within this region of influence with high tissue co-expression with neighboring genes are likely the result of co-regulation with the focal gene.

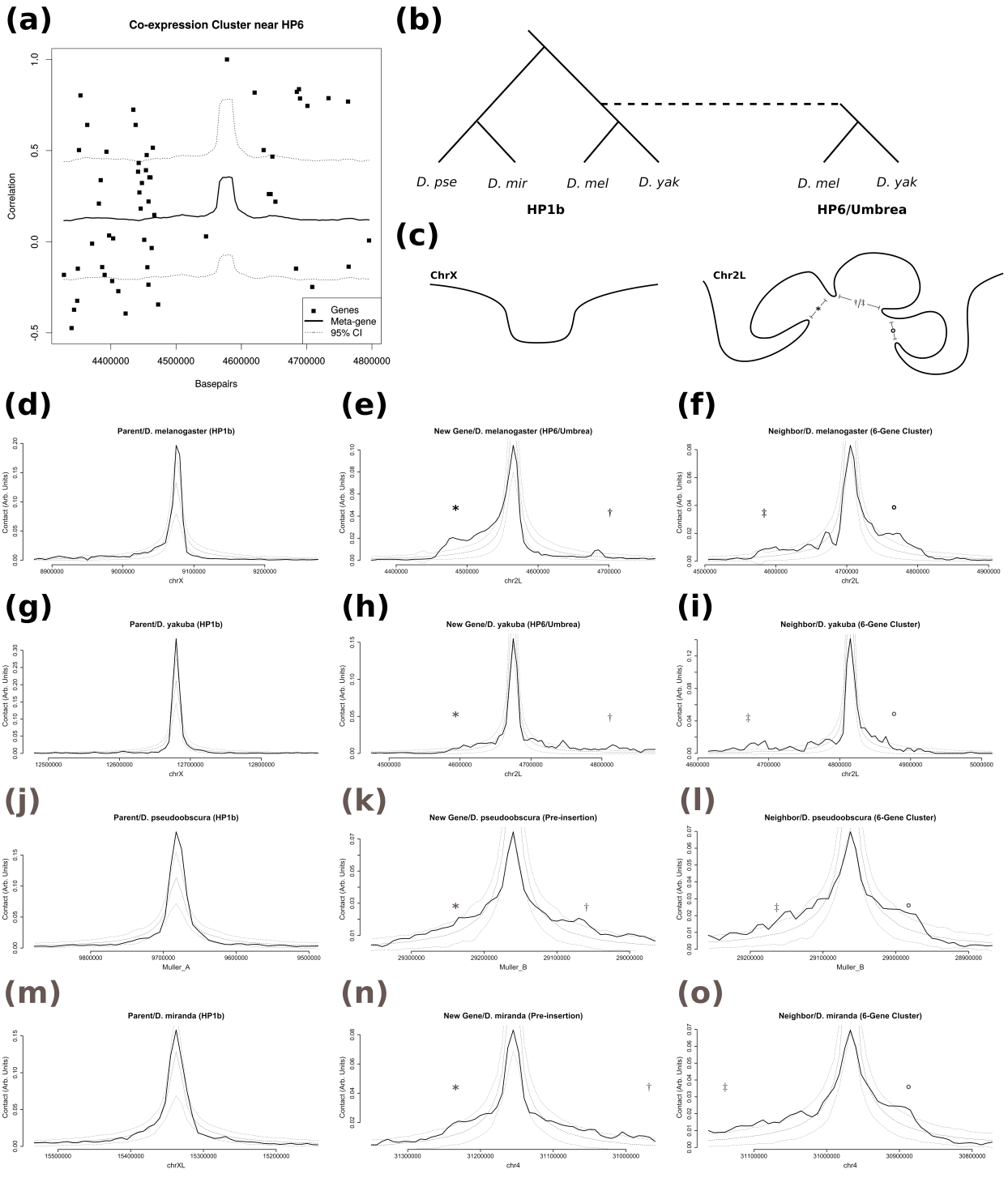

Figure 3.5: **HP6/Umbrea co-expression is associated with conserved chromosomal looping that pre-dates its insertion.** *Continued on next page...*

Figure 3.5, continued. **(b)** Tissue co-expression analysis between HP6/Umbrea and neighboring genes reveals the presence of a co-regulated cluster of 6 neighboring genes. Note absence of other genes within dumpy's intronic regions. **(b)** Two in-group species, *D. melanogaster* and *D. yakuba* (div. $\sim$ 6mya), contain HP6/Umbrea, while two out-group species, *D. pseudoobscura* and *D. miranda* (pse-mir div. $\sim$ 4mya, pse-mel div. $\sim$ 25mya), pre-date HP6/Umbrea's insertion ($\sim$ 12-15mya).**(c)** Cartoon legend illustrating features in **(d)-(o)**. Not drawn to scale. **(d)-(o)** Hi-C data tracks for in-group (*D. mel* **(d)-(f)**, *D. yak* **(g)-(i)**) and out-group (*D. pse* **(j)-(l)**, *D. mir* **(m)-(o)**) species are shown for the parental gene HP1b (left column) HP6/Umbrea's insertion site (middle column) and the co-regulated 6-gene cluster (right column), with a 95% confidence interval generated from genomic sampling plotted in dotted lines. On the vertical axis is contact in arbitrary units, and on the horizontal axis is genomic coordinates centered on the viewpoint location. Conserved feature (*) shows that HP6/Umbrea's insertion site loops with the active larval enhancers contained in dumpy's intronic gene-desert. Conserved features (†) & (‡) show that HP6/Umbrea's insertion site reciprocally loops with the co-regulated 6-gene cluster. Conserved feature (°) shows that the co-regulated gene cluster loops across the entire 6-gene cluster.

By comparing co-expression against this baseline distribution, we may find genes that share the same tissue-specific expression patterns as HP6/Umbrea and are thus likely co-regulated. As expected, we find that the neighboring gene, CR44609, possess the same expression pattern that HP6/Umbrea has. Similarly, we find that a locus of 6 neighboring genes (CG11929, Elba3, CG3251, Taf12L, CG15631, CG42523) located approximately 100kb away from HP6/Umbrea also expresses in the same tissues that HP6/Umbrea does, expressing primarily in the larval imaginal discs and male reproductive organs (Figure 3.5a).

While the co-expression of HP6/Umbrea's neighboring gene may be explained simply due to its proximity to HP6/Umbrea, the co-expression of the 6-gene cluster is not immediately evident as being a result of co-regulation. However, while this gene cluster is distally located along the chromosome beyond HP6/Umbrea's 25kb region of influence, due to the 3-dimensional nature of the eukaryotic genome, these genes may, in fact, be proximally located near HP6/Umbrea in 3D space and thus be co-regulated. Similarly, while active enhancer marks correlating to the onset of expression appears $\sim$50-100kb away from HP6/Umbrea, it is not immediately clear that these active enhancers are driving HP6/Umbrea expression,

as its distance to HP6/Umbrea exceeds the 25kb region of influence. As the 3-dimensional conformations of the genome may still allow these distal genic elements to interact, we tested whether the putative larval enhancer, HP6/Umbrea, its neighboring gene, and the 6-gene cluster are co-regulated by examining high-resolution Hi-C data for *D. melanogaster* [74] (Figure 3.5e, f). This data was aligned to the *D. melanogaster* genome dm6, and genome-to-genome contact frequencies were estimated using 5kb non-overlapping windows (c.f. **Methods and Material**).

Like co-expression, the frequency at which two genic elements make physical contact is expected to have a baseline, distance-dependent distribution. We may therefore test for co-regulation by predicting significant physical contact between HP6/Umbrea, its larval enhancer, and the cluster of co-expressed neighboring genes. Such an interaction could be detected if contact between these two loci (i.e. HP6/Umbrea with enhancer and HP6/Umbrea with co-expressing genes) exceeds the baseline distance-dependent distribution of contact frequency. We generated an estimate of this baseline contact frequency distribution using 1000 independent loci that were sampled randomly from the genome, where contact data for the flanking regions were used to generate the baseline distance-dependent contact frequency distribution. We then extracted the contact frequency data for the HP6/Umbrea locus alone and compared this to the baseline genome-wide contact frequency distribution (Figure 3.5e, f).

We first note that after self-self interactions are removed, we find that physical interactions in the genome generally remain highly localized, with most interactions lying near the focal locus as expected. Despite this, we find that HP6/Umbrea's complex contact distribution shows significant contact both with the putative larval enhancer as well as the neighboring 6-gene co-expression cluster (Figure 3.5e). Additionally, when the this analysis is repeated for the 6-gene co-expression cluster, we find that this contact is reciprocated, as the 6-gene cluster shows significant contact across the cluster as well as with HP6/Umbrea

(Figure 3.5g). Finally, HP6/Umbrea has enriched contact with the enhancer region that differentially activates at the onset of HP6/Umbrea expression. Combined with the tissue co-expression analysis, these results demonstrate that HP6/Umbrea and these 6 genes are likely co-regulated.

3.3.3 3D Genome Organization Pre-dates HP6/Umbrea Insertion

While we find evidence that HP6/Umbrea, the larval enhancer, and the 6-gene co-expression cluster are co-regulated, it is possible that these interactions evolved subsequent to HP6/ Umbrea's insertion. To determine whether these interactions pre-date HP6/Umbrea's insertion, we examined Hi-C data using a second in-group species, *D. yakuba* (shared by P. Reilly and P. Andolfatto), as well as two out-group species, *D. pseudoobscura* and *D. miranda* (shared by M. Ali and Q. Zhou) (Figure 3.5). While HP6/Umbrea inserted 12-15mya, the divergence between *D. melanogaster* and both outgroup species is 25mya [75]. Within these clades, *D. melanogaster* and *D. yakuba* diverged 6mya, while *D. pseudoobscura* and *D. miranda* diverged 4mya. While *D. melanogaster* Hi-C data was aligned to the standard reference genome (dm6), *D. yakuba*, *D. pseudoobscura* and *D. miranda* were aligned to newer, high-quality reference genomes (*D. yakuba* shared by P. Reilly and P. Andolfatto, *D. miranda* from [76], and *D. pseudoobscura* by M. Ali and Q. Zhou). In comparing the Hi-C contact patterns for both HP6/Umbrea and its neighboring co-expression cluster, we find that key features of the local chromosomal conformation are conserved: contact with larval enhancer, reciprocal contact between HP6/Umbrea and its co-expression cluster, and contact across the entire co-expression cluster (Figure 3.5d-e). The conservation of this chromosomal structure, even despite the subsequent evolution of protein function of HP6/Umbrea, suggests that the neo-functionalization event driving the fixation of the original duplication was likely driven by enhancer capture. Specifically, the 3D structure driving enhancer contacts existed prior to HP6/Umbrea's origination, and by duplicating into this region, HP6/Umbrea immediately

captured this regulatory interaction.

3.3.4 Identification of Enhancer Location

While the resolution of the previous Hi-C data is very high, due to the genome-by-genome nature of Hi-C, this data set's ability to precisely resolve the location of the enhancer elements driving HP6/Umbrea's expression is limited. Furthermore, while HP6/Umbrea's expression is driven primarily in the imaginal discs and testes, the *D. melanogaster* Hi-C data set is derived from embryonic tissue [74]. To identify the enhancer elements driving HP6/Umbrea's expression, we generated a 4C-Seq library using imaginal discs dissected from L3 and pre-pupal larvae, using a DpnII/Csp6I digest. After alignment to the *D. melanogaster* genome, we find that near the single peak identified by Hi-C, the peak has split into three regions, Four-C Located Enhancer Elements 1, 2, and 3 (FLEE1, FLEE2, & FLEE 3) located at Chr2L:4447389-4447781, 4516584-4517321 and 4531137-4532034 respectively (Figure 3.6). Additionally, we see broad agreement with further Hi-C results, such as the large degree of contact between HP6/Umbrea and the neighboring gene and the co-expressing 6-gene cluster.

3.4 Discussion

3.4.1 Enhancer Capture Model

While various evolutionary mechanisms for the origination of new genes have been proposed, these models do not incorporate the 3-dimensional organization of the genome. In the DDC and EAC models, functions are sub-partitioned amongst paralogous copies, resulting in a neutral or adaptive process leading to the fixation of duplicate gene copies. However, in these models, subsequent substitutions in either gene copy are required to explain new gene origination, separating duplication from neo-functionalization. Alternatively, in the IAD

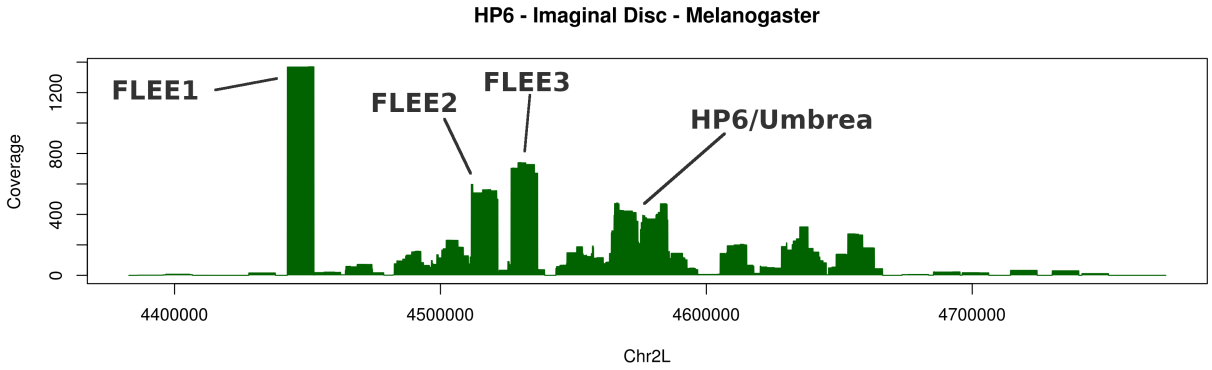

Figure 3.6: **HP6/Umbrea is controlled by 3 putative enhancers** 4C-Seq experiments were performed on larval imaginal disc tissue in *D. melanogaster* as shown above, revealing that the single enhancer peak is associated with 3 different peaks, named Four-C Located Enhancer Elements 1, 2, and 3 (FLEE1, FLEE2, & FLEE3). The vertical axis is sequencing coverage, while the horizontal axis is genomic coordinate. Self-self interactions have been removed.

model, duplication itself provides neo-functionalization by increasing dosage for an auxiliary function. In contrast to these models, we demonstrate how duplication itself may provide neo-functionalization in a tissue-specific manner, a result not predicted by these models. Such neo-functionalization provides a selective advantage in a direct, single-step mechanism without requiring subsequent substitutions as in the case of the DDC, EAC, and IAD models. In addition to producing gene fusions [77] as well as favorable frame-shifts [78], our model highlights the under-appreciated evolutionary value of both the act of duplication itself, and perhaps more importantly, the genomic context in which these duplications occur. While the role of positional effects in gene regulation and evolution has long been appreciated [25, 79], the advent of new chromosomal conformation capture technologies allows us to directly connect the conservation of chromosomal domains [80, 81] and the origination of new genes under a strong conceptual framework.

Under the enhancer capture model, a gene copy duplicates into a pre-existing regulatory context (Figure 3.7a), gaining a new regulatory interaction. Alternatively, the duplication may occur in a region of the genome possessing transcription factor binding sites (pre-

enhancer) but isn't yet acting as an active enhancer due to a paucity of nearby genes to regulate. Regardless of exact mechanism, due to the 3-dimensional looping nature of the eukaryotic genome, duplication recombines genes and enhancers into new combinations, thus resulting in regulatory novelty (Figure 3.7b, c). As such, this model provides an explanation and mechanism for the well-described but poorly-understood phenomenon where new genes often posses highly tissue-specific expression patterns [82–85]. Here, selection for increased expression in a single tissue is most rapidly achieved by acquiring a new tissue-specific expression pattern via distal duplication.

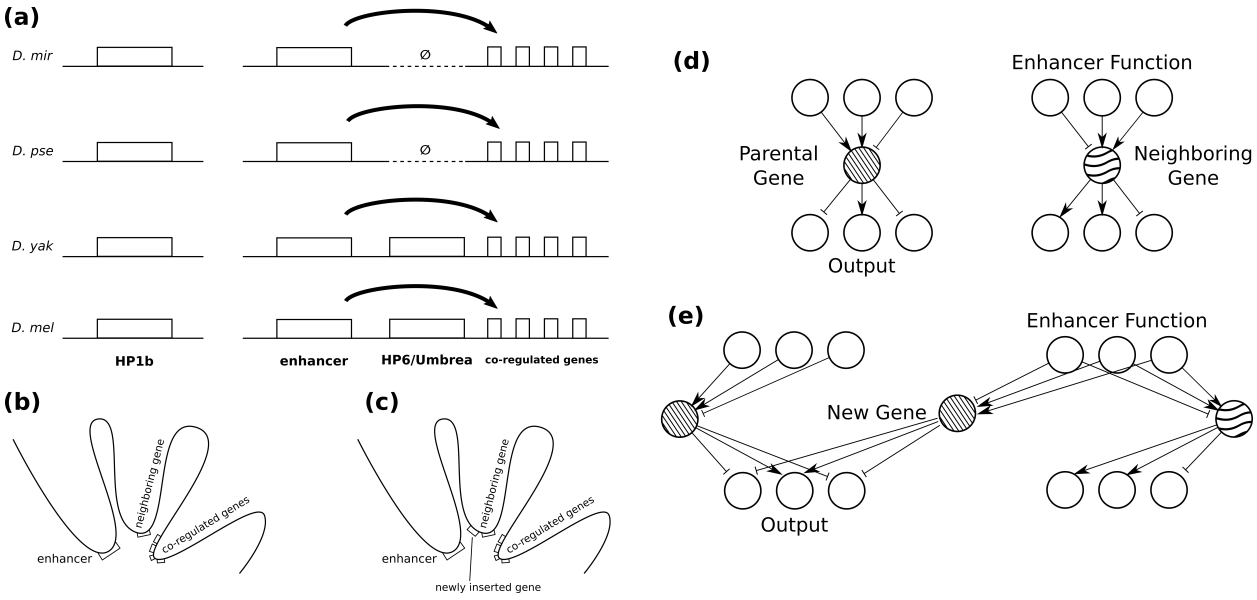

Figure 3.7: **The 3D organization of the genome allows for rapid rearrangement of genetic networks.** Panel **(a)** depicts a cartoon illustration of the action of the larval enhancer on the neighboring cluster of co-regulated genes as well as the future insertion site of HP6/Umbrea. **(b)** Preceding insertion of HP6/Umbrea, the larval enhancer was in contact with both HP6/Umbrea's neighboring gene as well as with the co-regulated 6-gene cluster. **(c)** This looping structure remains conserved following HP6/Umbrea's insertion, allowing for a rapid recombination of elements upstream of HP6/Umbrea's neighboring gene (i.e. larval enhancer) with elements downstream of HP6/Umbrea's parental gene (i.e. HP1b's protein function). A sample gene interaction network, both **(d)** pre- & **(e)** post- duplication, is depicted above. Note that parental gene and neighboring gene's original interactions remain intact, preserving previous function.

The enhancer capture model also provides a mechanistic explanation by which gene in-

teraction networks may rapidly evolve [86]. Under this model, we have two separate gene interaction sub-networks for both parental and neighboring genes (Figure 3.7d). As a new gene duplicates into a region near the neighboring gene, the new gene acquires the upstream regulatory function of the neighboring gene as well as the downstream function of the original parental gene's protein function (Figure 3.7e) while simultaneously preserving the pre-existing interactions from both parental and neighboring genes' sub-networks. As the act of duplication is more likely to occur than a point mutation [29, 30], enhancer capture will therefore be a faster route to generating increased tissue-specific expression of a parental gene (Figure 3.1) than any set of mutations in the parental gene's regulatory sequence. As a consequence, these new genes can fix, allowing for the subsequent accumulation of substitutions. While duplications occur more frequently than substitutions, point mutations altering the regulation of the parental gene will continue to occur. If eventually a compensatory mutation in the parent gene allows for increased tissue-specific expression, this will then allow the new gene to be free from the pressures of natural selection and thus evolve further, resulting either in pseudogenization, e.g. as in the case of HP6/Umbrea's loss in *D. eugracilis*, or the acquisition of further function, e.g. as in the case of HP6/Umbrea's gain of essential function in *D. melanogaster* [70].

One key aspect of the enhancer capture model is the selective advantage imparted by increased tissue-specific expression. While the EAC model describes a very narrow enhancer-based explanation for gene duplication and fixation [28], the resolution of evolutionary conflict, such as sexual antagonism, is a well-known driver of the evolution new genes [87, 88]. While most new genes have highly tissue-specific expression patterns, these often favor either the female or male reproductive organs/germlines in *D. melanogaster* [85]. A close examination of the expression pattern of HP6/Umbrea demonstrates the same - HP6/Umbrea is expressed primarily in the imaginal discs and the male reproductive organs. Similarly, the parental gene HP1b appears to have expression highly skewed towards the female reproduc-

tive organs. As such, it is possible that the selective advantage imparted by HP6/Umbrea's original duplication may have been a result of regulatory sexual antagonism and, given that most new genes show expression specific to reproductive organs, enhancer capture may be a wide-spread mechanism for the resolution of such sexual antagonism, providing a rapid, one-step mechanism for acquiring differential expression between sexes.

Central to both the enhancer capture and IAD models is the rapidity at which novelty is produced. Such rapid evolvability arguments may provide an explanation for the origination of the eukaryotic genome, organized into multiple chromosomal domains that result in a segregation of regulatory enhancer sequences and protein-coding genic sequences. While our model is illustrated with different tissue types, we may easily substitute various environmental conditions for tissue type. Under the context of sequential environmental conditions, the amplification of auxiliary function during transient environments is sensible as described by IAD model [29], as precise spatio-temporal regulation of the original gene is no longer needed, assuming that environmental conditions return back to "normal." Crucially, paralogs become fixed in the IAD model as duplication is the most evolvable solution to altered selective demands. As plasticity arises when permanent genic solutions are not easily evolvable (c.f. **Chapter II**) or when future environmental conditions are completely unpredictable [89], duplications into genomic regions where enhancers already exist may produce precise epigenetic control of a given protein much more rapidly than divergence via accumulated substitutions. Furthermore, while epigenetic mechanisms exist in prokaryotic genomes, these remain simple binary switches as in the case of the lac operon [90]. As the number of environmental conditions increase, the requisite gene-network complexity for such regulation becomes a large barrier for further evolution. Co-regulation of multiple genic units is already an efficient and useful method for dealing with multiple environmental conditions as demonstrated by the lac operon. By developing enhancers that operate in a proximity-based manner, eukaryotic genomes thus provide for the expansion of co-regulation into modular

structures [91] capable of handling greater than two distinct conditions without the need for developing three-way (or larger) switches. Given that enhancer capture can accelerate evolution both through faster-than-substitution alterations as well as modularity, the eukaryotic genome's inherently higher evolvability may suggest that enhancer capture may be one clue in understanding the evolutionary origins of the nucleus.

3.4.2 Revisiting an Old Theory of New Genes

Current models of eukaryotic gene regulation roughly defines two broad classes of genomic sequences: protein-coding sequences and regulatory sequences [92]. Under these models, the precise spatio-temporal control of a protein-coding sequence is provided by genomic enhancer elements where the concerted binding of transcription factors acts to either increase or decrease the activation energy of transcription. Importantly, such control occurs in a three-dimensional, distance-dependent manner - enhancer elements may only control genomic elements that are physically close to these enhancers within the eukaryotic nucleus [92]. Due to this proximity-based effect, the exact conformation of the genome is significantly more important in understanding gene regulation than simple gene order, particularly in gene-dense genomes. Using this proximity-based effect, we show that the chromosomal context into which a gene duplicates, particularly non-tandem/distal duplications, may generate novel enhancer-gene interactions that immediately neo-functionalizes duplicate gene copies.

These positional effects have been well-described since the origins of the field of genetics. The first known positional effect was described in the study of the *bar* gene in *Drosophila melanogaster* in 1925 by Alfred Sturtevant a mere 12 years after he developed the first genetic map [93]. In his original allelomorphic series, Sturtevant surmised that a duplication must have occurred with the *bar* gene, where two copies of the gene were inherited along a single chromosome. Crucially, in comparing the homozygous B/B phenotype to the BB/B^{-} phenotype, Sturtevant found that the *double-bar* or *ultrabar* allelomorph produced a more

extreme phenotype than expected by dosage alone. This *double-bar* or *ultrabar* allelomorph of the classic *bar* gene was found to be the result of a gene duplication event through the examination of polytene chromosomes by Calvin Bridges in 1936 [79]. Dobzhansky recognized this as what he called a positional effect and that it was a result of some kind of chromosomal interaction with neighboring genes [79]. Soon afterwards, Hermann Muller recognized the importance of this observation for the origination of new genes:

> *"We consider the point of chief interest in the Bar case to be its illustration of the manner of origination of extra genes in evolution. Bar had for a long time offered the best case yet known for the idea that genes could arise de novo*. Its interpretation as some sort of duplication met with difficulty, in our ignorance of the real existence of a 'position effect'..."*
>
> -Hermann Muller (**Science**, 1936)

*note "*de novo*" is not used indicate a particular new gene origination mechanism as in [26].

3.5 Methods and Materials

3.5.1 Tissue expression data and analysis

Tissue expression data was retrieved from FlyBase. Pre-computed RPKM data files were downloaded, with RPKM values for each FlyBase transcript being reported for 29 tissues [68]. As many of the tissues types were repetitive, data from head, ovary, carcass, and digestive system were averaged to reduce over-representation bias in further correlational analyses. Gene map data was also obtained from FlyBase to properly identify neighboring genes [67]. Parental/new gene pair information was retrieved from [72]. Spearman correlation coefficients were calculated using the tissue expression data between parental and new gene pairs. Due to intronic structures and variation in gene length, two neighboring genes for each new

gene on each side were assessed using Spearman correlation coefficients and the maximum value of the four neighbors was recorded. Additionally, correlation coefficients for all genes within 500kb of HP6/Umbrea were reported. To generate a baseline distance-dependent genomic estimate of co-expression, 1000 random genic loci were chosen and co-expression values (Spearman) between the randomly selected gene and all neighbors within a 500kb range were calculated. This 500kb region was then divided into 100 non-overlapping windows where mean and variance in correlation coefficients was calculated across all randomly selected loci.

3.5.2 ChIP-Seq data

ChIP-Seq or ChIP-Chip data were obtained for H3K4me1 and H3K27ac for S2 cells as well as whole L3 larvae from modENCODE [73]. H3K4me1 ChIP-Chip data for S2-DRSC cells was obtained using data ID 304 and 3760. H3K27ac ChIP-Chip data for S2-DRSC cells was obtained using data ID 296 and 3757. H3K4me1 ChIP-Seq data for whole Oregon-R L3 larvae was obtained using data ID 4986. H3K27ac ChIP-Seq data for whole Oregon-R L3 larvae was obtained using data ID 5084. For all data sets, data was obtained in .gff3 format and visualized using the UCSC Genome Browser.

3.5.3 Hi-C data

Publicly available Hi-C libraries were obtained from NCBI: *D. melanogaster*, PRJNA393992. *D. yakuba* Hi-C data was shared by Patrick Reilly and Peter Andolfatto, and *D. pseudoobscura* and *D. miranda* data was shared by Mujahid Ali and Qi Zhou. *D. melanogaster* source tissue was S2 cells, *D. yakuba* from adult females, and *D. pseudoobscura* and *D. miranda* were L3 larvae. Hi-C libraries were preprocessed, mapped, and filtered using HiCUP version 0.8.0 [94]. Specifically, reads from fastq files were trimmed at ligation junctions, and subsequently each mate of paired-end sequences were independently mapped to the respec-

tive genomes using bowtie2 version 2.2.9 [95]. Reads were mapped to genomes consisting of canonical chromosomes only (i.e. excluding scaffolds and other unplaced sequences). *D. melanogaster* reference genome was dm6 and obtained from FlyBase [67]. The *D. yakuba* reference genome was shared by Patrick Reilley and can be obtained from NCBI (PRJNA310215). The *D. pseudoobscura* reference genome was obtained directly from Ryan Bracewell (https://www.ryanbracewell.com/data.html) [96] and the *D. miranda* reference genome was obtained from NCBI (PRJNA474939), [76]. HiCUP was used further to remove experimental artifacts based on an *in silico* genome digest as previously described [94]. HiCUP mapped and filtered .sam files were then converted to formats compatible with HOMER version 4.11 [97] and juicer tools version 1.22.01 [98]. To create matrices, HOMER was used to tile the genome into matrices of fixed-size bins, and assign reads to their correct intersecting bins. HOMER was also used to normalize contact counts in these matrices based on known Hi-C biases, as previously described [97]. Juicer tools was used to produce .hic files at resolutions of 5kb for *D. melanogaster* and *D. yakuba* and 7.5kb for *D. pseudoobscura* and *D. miranda*, and to create normalized matrices.

Using Hi-C contact matrices, data rows for HP6/Umbrea and its neighboring cluster were pulled for a 400kb region centered on HP6/Umbrea and self-self interactions were removed. To generate a genome-wide distance-dependent distribution of contact, 1000 random loci were sampled. Contact data for each locus was then normalized with total contact (arb. units) being equal for all loci. The mean and variance for each non-overlapping window was calculated and reported and compared to HP6/Umbrea and the co-expression clusters' data. To generate genomic coordinates for HP6/Umbrea before duplication, *D. melanogaster* sequence flanking HP6/Umbrea's insertion site was aligned to the *D. yakuba*, *D. pseudoobscura* and *D. miranda* reference genomes using blast. Similarly, the promoter region of CG11929 was aligned to *D. yakuba*, *D. pseudoobscura* and *D. miranda* reference genomes to represent the co-expression cluster.

3.5.4 4C-Seq

Tissue from L3 larval imaginal discs was dissected and placed on ice for no longer than 10 minutes. Following dissection, 4C-Seq protocol was followed as per [99] using a DpnII/Csp6I digestion. Sequence data was aligned to reference genome dm6. A virtual digestion of the *D. melanogaster* genome as performed, and reported enhancers are the distance between neighboring virtual fragments.

3.6 Acknowledgements

U.L. is funded in part by GM7197-42. M.L. was supported by the National Institutes of Health grant 1R01GM116113-01A1. This manuscript is dedicated in memory of Frances Lee.

BIBLIOGRAPHY

[1] T. Dobzhansky. *Genetics and the Origin of Species*. Columbia University Press, 1937.

[2] T. Dobzhansky. "Nothing in Biology Makes Sense except in the Light of Evolution". In: *The American Biology Teacher* 35 (3 1973), pp. 125–129.

[3] J. Huxley. *Evolution: The Modern Synthesis*. Allen & Unwin, 1942.

[4] E. Mayr. *Systematics and the Origin of Species*. Columbia University Press, 1942.

[5] J. M. Baldwin. "A new factor in evolution". In: *The American Naturalist* 30 (354 1896), pp. 441–451.

[6] C. H. Waddington. "Genetic assimilation of an acquired character". In: *Evolution* 7 (2 1952), pp. 118–126.

[7] M. Pigliucci, C. J. Murren, and C. D. Schlichting. "Phenotypic plasticity and evolution by genetic assimilation". In: *Journal of Experimental Biology* 209 (Pt 12 2003), pp. 2362–2367.

[8] T. D. Price, A. Qvarnstrom, and D. E. Irwin. "The role of phenotypic plasticity in driving genetic evolution". In: *Proceedings of the Royal Society B* 270 (1523 2003), pp. 1433–1440.

[9] G. C. Williams. *Adaptation and Natural Selection*. Princeton University Press, 1966.

[10] G. de Jong. "Evolution of phenotypic plasticity: patterns of plasticity and the emergence of ecotypes". In: *New Phytologist* 166 (1 2005), pp. 101–117.

[11] L. W. Ancel. "Undermining the Baldwin expediting effect: does phenotypic plasticity accelerate evolution?" In: *Theoretical Population Biology* 58.4 (2000), pp. 307–319.

[12] T. DeWitt, A. Sih, and D. S. Wilson. "Costs and limits of phenotypic plasticity". In: *Trends in Ecology and Evolution* 13.2 (1998), pp. 77–81.

[13] R. Lande. "Adaptation to an extraordinary environment by evolution of phenotypic plasticity and genetic assimilation". In: *Journal of Evolutionary Biology* 22 (2009), pp. 1435–1446.

[14] H. J. E. Beaumont et al. "Experimental evolution of bet hedging". In: *Nature* 462.7269 (2009), pp. 90–93.

[15] M. L. Siegal and J. Y. Leu. "On the Nature and Evolutionary Impact of Phenotypic Robustness Mechanisms". In: *Annual Review of Ecology, Evolution, and Systematics* 45.1 (2014), pp. 495–517.

[16] O. D. King and J. Masel. "The evolution of bet-hedging adaptations to rare scenarios". In: *Theoretical Population Biology* 72.4 (2007), pp. 560–575.

[17] S. F. Levy and M. L. Siegal. "The robustness continuum". In: *Advances in Experimental Medicine and Biology* 751 (2012), pp. 431–452.

[18] W. C. Ratcliff, P. Hawthorne, and E. Libby. "Courting disaster: How diversification rate affects fitness under risk". In: *Evolution; International Journal of Organic Evolution* 69.1 (2015), pp. 126–135.

[19] E. Libby and W. C. Ratcliff. "Shortsighted Evolution Constrains the Efficacy of Long-Term Bet Hedging". In: *The American Naturalist* 193.3 (2019), pp. 409–423.

[20] A. M. Simons. "Fluctuating natural selection accounts for the evolution of diversification bet hedging". In: *Proceedings of the Royal Society B: Biological Sciences* 276.1664 (2009), pp. 1987–1992.

[21] E. Kussell and S. Leibler. "Phenotypic Diversity, Population Growth, and Information in Fluctuating Environments". In: *Science* 309.5743 (2005), pp. 2075–2078.

[22] A. Skanata and E. Kussell. "Evolutionary Phase Transitions in Random Environments". In: *Physical Review Letters* 117.3 (2016), p. 038104.

[23] L. W. Ancel. "A quantitative model of the Simpson-Baldwin Effect". In: *Journal of Theoretical Biology* 196.2 (1999), pp. 197–209.

[24] S. Via and R. Lande. "Genotype-environment interaction and the evolution of phenotypic plasticity". In: *Evolution* 39 (1985), pp. 505–522.
[25] H. J. Muller. "Bar Duplication". In: *Science* 83 (2161 1936), pp. 528–530.
[26] M. Long et al. "New gene evolution: little did we know". In: *Annual Review of Genetics* 47 (2013), pp. 307–333.
[27] A. Force et al. "Preservation of Duplicate Genes by Complementary, Degenerative Mutations". In: *Genetics* 151 (4 1999), pp. 1531–1545.
[28] C. T. Hittinger and S. B. Carroll. "Gene duplication and the adaptive evolution of a classic genetic switch". In: *Nature* 449 (2007), pp. 677–681.
[29] U. Bergthorsson, D. I. Andersson, and J. R. Roth. "Ohno's dilemma: Evolution of new genes under continuous selection". In: *Proc. Natl. Acad. Sci. USA* 104 (43 2007), pp. 17004–17009.
[30] J. Nasvall et al. "Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence". In: *Science* 338 (6105 2012), pp. 384–387.
[31] A. I. Kalmykova et al. "Regulated chromatin domain comprising cluster of co-expressed genes in *Drosophila melanogaster*". In: *Nucleic Acids Research* 33 (5 2005), pp. 1435–1444.
[32] A. M. Boutanaev et al. "Large clusters of co-expressed genes in the *Drosophila* genome". In: *Nature* 420 (6916 2002), pp. 666–669.
[33] P. T. Spellman and G. M. Rubin. "Evidence for large domains of similarly expressed genes in the *Drosophila* genome". In: *Journal of Biology* 1 (1 2002), p. 5.
[34] X. S. Zhang and W. G. Hill. "Evolution of the Environmental Component of the Phenotypic Variance: Stabilizing Selection in Changing Environments and the Cost of Homogeneity". In: *Evolution* 59.6 (2005), pp. 1237–1244.
[35] A. Sigal et al. "Variability and memory of protein levels in human cells". In: *Nature* 444.7119 (2006), pp. 643–646.
[36] H. H. McAdams and A. Arkin. "Stochastic mechanisms in gene expression". In: *Proceedings of the National Academy of Sciences* 94.3 (1997), pp. 814–819.
[37] W. J. Blake et al. "Noise in eukaryotic gene expression". In: *Nature* 422.6932 (2003), pp. 633–637.
[38] S. Bratulic, F. Gerber, and A. Wagner. "Mistranslation drives the evolution of robustness in TEM-1 beta-lactamase". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.41 (2015), pp. 12758–12763.
[39] S. L. Spencer et al. "Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis". In: *Nature* 459.7245 (2009), pp. 428–432.
[40] A. F. Ramos, J. E. M. Hornos, and J. Reinitz. "Gene regulation and noise reduction by coupling of stochastic processes". In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 91.2 (2015), p. 020701.
[41] R. Moxon and E. Kussell. "The impact of bottlenecks on microbial survival, adaptation, and phenotypic switching in host–pathogen interactions". In: *Evolution* 71.12 (2017), pp. 2803–2816.
[42] A. Mayer et al. "Transitions in optimal adaptive strategies for populations in fluctuating environments". In: *Physical Review E* 96.3 (2017).
[43] M. Donaldson-Matasci, M. Lachmann, and C. T. Bergstrom. "Phenotypic Diversity as an Adaptation to Environmental Uncertainty". In: *Evolutionary Ecology Research, v.10, 493-515 (2008)* 10 (2007).
[44] A. C. Tadrowski, M. R. Evans, and B. Waclaw. "Phenotypic Switching Can Speed up Microbial Evolution". In: *Scientific Reports* 8.1 (2018).

[45] C. K. Ghalambor et al. “Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments”. In: *Functional Ecology* 21.3 (2007), pp. 394–407.
[46] Jeremy A. Draghi and Michael C. Whitlock. “Phenotypic Plasticity Facilitates Mutational Variance, Genetic Variance, and Evolvability Along the Major Axis of Environmental Variation”. In: *Evolution* 66.9 (2012), pp. 2891–2902.
[47] D. W. Pfennig and M. R. Servedio. “The role of transgenerational epigenetic inheritance in diversification and speciation”. In: *Non-Genetic Inheritance* 1 (2013), pp. 17–26.
[48] T. D. Price, A. Qvarnström, and D. E. Irwin. “The role of phenotypic plasticity in driving genetic evolution.” In: *Proceedings of the Royal Society B: Biological Sciences* 270.1523 (2003), pp. 1433–1440.
[49] C. H. Waddington. “Canalization of Development and the Inheritance of Acquired Characters”. In: *Nature* 150.3811 (1942), pp. 563–565.
[50] M. J. West-Eberhard. “Developmental plasticity and the origin of species differences”. In: *Proceedings of the National Academy of Sciences* 102 (2005), pp. 6543–6549.
[51] E. Jablonka et al. “The adaptive advantage of phenotypic memory in changing environments”. In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 350.1332 (1995), pp. 133–141.
[52] Nicholas H. Barton and Peter D. Keightley. “Understanding quantitative genetic variation”. In: *Nature Reviews Genetics* 3.1 (2002), pp. 11–21.
[53] G. G. Simpson. “The Baldwin Effect”. In: *Evolution* 7.2 (1953), pp. 110–117.
[54] G.E. Hinton and S.J. Nowlan. “How learning can guide evolution”. In: *Complex systems* 1 (1987), pp. 495–502.
[55] G. P. Wagner, G. Booth, and H. Bagheri-Chaichian. “A Population Genetic Theory of Canalization”. In: *Evolution* 51.2 (1997), pp. 329–347.
[56] U. Lee et al. “Noise-Driven Phenotypic Heterogeneity with Finite Correlation Time in Clonal Populations”. In: *PLoS One* 0132397 (2015).
[57] E. Kim et al. “Geometric structure and geodesic in a solvable model of nonequilibrium process”. In: *Physical Review E* 93.6 (2016), p. 062127.
[58] M. Osella, E. Nugent, and M. C. Lagomarsino. “Concerted control of Escherichia coli cell division”. In: *Proceedings of the National Academy of Sciences* 111.9 (2014), pp. 3431–3435.
[59] W. Jetz, C. H. Sekercioglu, and K. Böhning-Gaese. “The Worldwide Variation in Avian Clutch Size across Species and Space”. In: *PLoS Biology* 6.12 (2008).
[60] R. L. Honeycutt. “Stochastic Runge-Kutta algorithms. II. Colored noise”. In: *Physical Review A* 45.2 (1992), pp. 604–610.
[61] A. Wagner. “Does Evolutionary Plasticity Evolve?” In: *Evolution* 50.3 (1996), pp. 1008–1023.
[62] Manu et al. “Canalization of Gene Expression and Domain Shifts in the Drosophila Blastoderm by Dynamical Attractors”. In: *PLOS Computational Biology* 5.3 (2009), e1000303.
[63] Manu et al. “Canalization of Gene Expression in the Drosophila Blastoderm by Gap Gene Cross Regulation”. In: *PLOS Biology* 7.3 (2009), e1000049.
[64] S. Ohno. *Evolution by Gene Duplication*. New York: Springer, 1970.
[65] M. P. Francino. “An adaptive radiation model for the origin of new gene functions”. In: *Nature Genetics* 37 (6 2005), pp. 573–577.
[66] M. Kimura and T. Ohta. “On Some Principles Governing Molecular Evolution”. In: *Proc. Natl. Acad. Sci. USA* 71 (7 1974), pp. 2848–2852.

[67] A Larkin et al. “FlyBase: updates to the Drosophila melanogaster knowledge base”. In: *Nucleic Acid Research* 49 (D1 2021), pp. D899–D907.
[68] J.B. Brown et al. “Diversity and dynamics of the Drosophila transcriptome”. In: *Nature* 512 (7515 2014), pp. 393–399.
[69] S. Xia et al. “Genomic analyses of new genes and their phenotypic effects reveal rapid evolution of essential functions in Drosophila development”. In: *PLoS Genetics* 17 (7 2021), e1009654.
[70] B. D. Ross et al. “Stepwise evolution of essential centromere function in a Drosophila neogene”. In: *Science* 240 (6137 2013), pp. 1211–1214.
[71] F. Greil et al. “HP1 controls genomic targeting of four novel heterochromatin proteins in Drosophila”. In: *EMBO Journal* 26 (3 2007).
[72] S. Chen, Y. E. Zhang, and M. Long. “New genes in Drosophila quickly become essential”. In: *Science* 330 (6011 2010), pp. 1682–1685.
[73] S. E. Celniker et al. “Unlocking the secrets of the genome”. In: *Nature* 459 (7249 2009), pp. 927–930.
[74] Q. Wang et al. “Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells”. In: *Nature Communications* 9 (1 2018), p. 188.
[75] C. A. Russo, N. Takezaki, and M. Nei. “Molecular phylogeny and divergence times of drosophilid species”. In: *Molecular Biology and Evolution* 12 (3 1995), pp. 391–404.
[76] S. Mahajan et al. “De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture”. In: *PLoS Biology* 16 (7 2018), e2006348.
[77] W. Wang et al. “The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, in Drosophila melanogaster”. In: *Molecular Biology and Evolution* 17 (9 2000), pp. 1294–1301.
[78] W. Wang et al. “Origin and evolution of new exons in rodents”. In: *Genome Research* 15 (9 2005), pp. 1258–1264.
[79] C. B. Bridges. “The Bar ”Gene” a Duplication”. In: *Science* 83 (2148 1936), pp. 210–211.
[80] N. Harmston et al. “Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation”. In: *Nature Communications* 8 (1 2017), p. 441.
[81] J. Krefting, M. A. Andrade-Navarro, and J. Ibn-Salen. “Evolutionary stability of topologically associating domains is associated with conserved gene regulation”. In: *BMC Biology* (1 2018), p. 87.
[82] L. Zhang et al. “Rapid evolution of protein diversity by de novo origination in Oryza”. In: *Nature Ecology and Evolution* 3 (4 2019), pp. 679–690.
[83] H. Dai, T. F. Yoshimatsu, and M. Long. “Retrogene movement within- and between-chromosomes in the evolution of Drosophila genomes”. In: *Gene* 385 (2006), pp. 96–102.
[84] M. D. Vibranovski et al. “Re-analysis of the larval testis data on meiotic sex chromosome inactivation revealed evidence for tissue-specific gene expression related to the drosophila X chromosome”. In: *BMC Biology* 10 (2012), p. 49.
[85] M. Long et al. “New gene evolution: little did we know”. In: *Annual Review of Genetics* 47 (2013), pp. 307–333.
[86] W. Zhang et al. “New genes drive the evolution of gene interaction networks in the human and mouse genomes”. In: *Genome Biology* 16 (2015), p. 202.
[87] L. E. Kursel et al. “Gametic specialization of centromeric histone paralogs in Drosophila virilis”. In: *Life Science Alliance* 4 (7 2021), e202000992.

[88] N. W. VanKuren and M. Long. "Gene Duplicates Resolving Sexual Conflict Rapidly Evolved Essential Gametogenesis Functions". In: *Nature Ecology and Evolution* 2 (4 2018), pp. 705–712.
[89] A. Skanata and E. Kussel. "Evolutionary Phase Transitions in Random Environments". In: *Physical Review Letters* 117 (3 2016), p. 038104.
[90] F. Jacob and J Monod. "Genetic regulatory mechanisms in the synthesis of proteins". In: *Journal of Molecular Biology* 3 (1961), pp. 318–356.
[91] A. Wagner. "Does evolutionary plasticity evolve?" In: *Evolution* 50 (3 1996), pp. 1008–1023.
[92] C-T. Ong and V. G. Corces. "Enhancer function: new insights into the regulation of tissue-specific gene expression". In: *Nature Reviews Genetics* 12 (2011), pp. 283–293.
[93] A. H. Sturtevant. "The Effects of Unequal Crossing over at the Bar Locus in Drosophila". In: *Genetics* 10 (2 1925), pp. 117–147.
[94] S. Wingett et al. "HiCUP: pipeline for mapping and processing Hi-C data". In: *F100Res* 4 (2015), p. 1310.
[95] B. Langmead and S. L. Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9 (2012), pp. 357–359.
[96] R. Bracewell et al. "Dynamic turnover of centromeres drives karyotype evolution in Drosophila". In: *eLife* Sep 16;8:e49002 (2019).
[97] S. Heinz et al. "Transcription Elongation Can Affect Genome 3D Structure". In: *Cell* 174 (6 2018), pp. 1522–1536.
[98] N. C. Durand et al. "Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments". In: *Cell Systems* 3 (1 2016).
[99] D. R. Sobreira et al. "Extensive pleiotropism and allelic heterogeneity mediate metabolic effects of IRX3 and IRX5". In: *Science* 372 (6546 2021), pp. 1085–1091.

APPENDIX A

ADDITIONAL PUBLICATIONS

This appendix contains publications that I co-authored and my personal contributions to these works is highlighted below.

A.0.1 Kim, et al. (2016) Physical Review E

This manuscript arose through a collaboration with Prof. Eun-Jin Kim. With her experience in theoretical physics, we began to collaborate on a model of clonal phenotypic heterogeneity. We then extended this model to examine how a forcing function affects adaptation. We found that minimization of information length produced a geodesic solution that additionally minimized adaptation time. Most importantly, we found that this geodesic solution produces cyclical variation in the variance of the probability distribution functions, explaining how a large degree of plasticity is only transiently beneficial during adaptation.

Kim, E., Lee, U., Heseltine, J., Hollerbach, R. Geometric structure and geodesic in a solvable model of nonequilibrium process. *Phys. Rev. E* **93**, 06127 (2016) doi: 10.1103/PhysRevE.93.062127

A.0.2 Leypunskiy at al. (2017) eLife

This manuscript concerns the relationship of the minimal KaiABC circadian clock and seasonal variations in daytime duration. While circadian rhythms follow 24-hour cycles, the division of day-/night-time fluctuates greatly between terrestrial seasons, ranging from a ratio of 8:18 hours to 18:8 hours of illumination:non-illumination. To understand what mechanisms underlie this robustness, I began with a luciferase reporter assay under the control of the native promoter for the KaiABC system and subjected replicate populations to various day/night ratios and recorded the transcriptional output of the reporter. Though I

did not choose to personally pursue the project, we found that the resulting phase-shifting behavior of the system is linear and well-tuned to respond within physiological limits, offsetting any unexpected shifts in day/night ratios thereby allowing the clock to anticipate future recurrences of daylight onset.

Leypunskiy, E., Lin, J., Yoo, H. Lee, U., Dinner, A., Rust, M. The cyanobacterial circadian clock follows midday in vivo and in vitro. *eLife* **6**:e23539, (2017). doi: 10.7554/eLife.23539

A.0.3 Zu, et al. (2019) Science China

This manuscript is concerned with the network dynamics of new gene evolution. When new essential genes evolve, they must somehow integrate themselves into pre-existing genetic interaction networks. To understand this process, I proposed that we utilize publicly available human tissue expression data to generate a systems-biology level view of the genome. Using this map, we then analyzed the aging process of new genes via topological properties of these interaction networks, finding that new genes are generally highly specialized and become less specialized as they age. These new genes also expand in function and interactions according to a power-law distribution through a "rich-get-richer" mechanism.

Zu, J., Gu, Y., Li, Y., Li, C., Zhang, W., Zhang, Y., Lee, U., Zhang, L., Long, M. Topological evolution of coexpression networks by new gene integration maintains the heirarchical and modular structures in human ancestors. *Science China* **62**, 4 594-608 (2019) doi: 10.1007/s11427-019-9483-6

A.0.4 Xia, et al. (2021) PLoS Genetics

This manuscript is a examination of technical concerns regarding genetic knock-down/knock-out techniques and its implications regarding new gene essentiality. While a prior study claimed to have failed to replicate our prior analysis of new gene essentiality, further examination found that technical issues with RNAi and CRISPR/Cas9 techniques and analyses

were present in the replication study. In this study, we examined replicability issues with their RNAi constructs by using multiple targeting constructs per gene, finding that our original conclusions remained correct. In this work, I contributed statistical analyses as well as a conceptual analysis of incorrectly applied inferences.

Xia, S., VanKuren, N., Chen, C., Zhang, L., Kemkemer, C., Shao, Y., Jia, H., Lee, U., Advani, A., Gschwend, A., Vibranovski, M., Chen, S., Zhang, Y., Long, M. Genomic analyses of new genes and their phenotypic effects reveal rapid evolution of essential functions in *Drosophila* development. *PLoS Genetics* **17**, 7:e1009654 (2021) doi: 10.1371/journal.pgen.1009654