Genetic Networks, Adaptation, & the Evolution of Genomic Islands of Divergence

Presented in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy

with a Major in

Bioinformatics and Computational Biology

in the

College of Graduate Studies

University of Idaho

by

Tyler Duncan Hether

Major Professor
Paul Hohenlohe, Ph.D.

Committee
Christine Parent, Ph.D.
Paul Joyce, Ph.D.
James Foster, Ph.D.

Department Administrator
Eva Top, Ph.D.

April 2016

Authorization to Submit Dissertation

This dissertation of Tyler Duncan Hether, submitted for the degree of Doctor of Philosophy with a Major in Bioinformatics and Computational Biology and titled "Genetic Networks, Adaptation, & the Evolution of Genomic Islands of Divergence," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: ______________________ ____________
Paul Hohenlohe, Ph.D. Date

Committee Members: ______________________ ____________
Christine Parent, Ph.D. Date

______________________ ____________
Paul Joyce, Ph.D. Date

______________________ ____________
James Foster, Ph.D. Date

Department Administrator: ______________________ ____________
Eva Top, Ph.D. Date

Abstract

Genetics has made great strides in identifying specific genes that affect traits of biomedical and basic biological importance, and modern genomic technology has greatly increased our power to detect even genes with small effects on phenotype (Yang et al., 2010). However, genes underlying important phenotypes don't exist in isolation; rather, they interact in two important ways that are now amenable to direct observation with genomic techniques.

First, genes interact within genetic regulatory networks to produce complex quantitative phenotypes. For example, in many gene regulatory cascades a given gene product may interact with a number of other genes and proteins. The incorporation of a network-level functional view of genetic interactions into models of multivariate phenotypic evolution represents a new synthesis in biology, enabled by the wealth of empirical genomic data (Zhu et al., 2009; O'Malley, 2012). By modeling relative simple gene regulatory networks, I found that the direction of new phenotypic (co)variation that is supplied to a population from new mutation (the M-matrix) depends on a given network topology. Such mutational (co)variation directly contributes to the shape and orientation of additive genetic (co)variation (the G-matrix) which affects how quickly populations can adapt to a new environment. When letting the network topology itself evolve, I found that populations can quickly explore phenotype space and, as such, can get closer to new phenotypic optima than without mutations in the network topology. Moreover, the adaptive trajectories taken later during the adaptive walk directly depend on historical contingencies (i.e., which networks were selected for in the past). Lastly, when network topology evolves, reproductive isolation can evolve too as a result of persistent overdominance.

Second, genes exist in physical locations along chromosomes so that the action of evolutionary forces like mutation and selection on single loci has impacts on patterns of variation at neighboring loci. Meiotic recombination is central in connecting physical genetic elements to population genetic theory as well as quantitative trait loci. As such, I have created an R package that uses a Hidden Markov Model (HMM) approach to identify recombination hotspots, coldspots, crossovers and non-crossover gene conversion tracts from low-coverage whole genome or reduced representation (e.g., RADseq) data. This approach is applicable for any haploid or diploid organisms with a reference genome.

Under a divergence with gene flow scenario, physical gene interactions can also cause autocorrelation in genetic differentiation (e.g., F_{ST}) across the genome of diverging populations, creating "genomic islands of divergence" – gene regions that have significantly greater differentiation than expected under neutrality (Nosil et al., 2012). An appealing aspect of this model is that regions physically linked to selected loci are relatively buffered from the homogenizing effect of migrant alleles so that new mutations that are tightly linked to selected loci have a higher probability of increasing in frequency, facilitating further genomic differentiation between diverging populations. To examine the relative roles of selection, recombination, and gene flow in creating heterogenous genomic differentiation I employed mathematical modeling and experimental evolution of polymorphic populations of the budding yeast *Saccharomyces cerevisiae*. I found that neither migration nor new mutations were necessary to engender island growth. Instead the segregation of existing standing genetic variation can transiently but quickly generate islands when admixed populations evolve in isolation in different selective environments.

Taken together, this dissertation underscores the importance of genetic interactions in generating phenotypic diversity. Moreover, we find that simple models of gene interaction – together with demography and evolutionary factors – can generate complex, but predictable patterns of adaptation. The models, computational tools, and experimental data herein thus expand our knowledge of how genetic interactions shape the nature of suites of traits.

Acknowledgements

I would first like to thank my mentor, Paul Hohenlohe, for his help and guidance throughout my time at the University of Idaho. Bouncing ideas back and forth over the years undoubtedly enriched my understanding of biology and his help was never ending. Indeed, I took advantage of his open door policy more than he probably envisioned a student would.

I would also like to thank my committee members. Paul Joyce greatly influenced my understanding of mathematical genetics, James Foster was instrumental in engendering my interest in computational biology, and Christine Parent helped me bring these tools together to answer interesting biological questions. Though not officially a committee member, I would also like to thank Luke Harmon for his tutelage these past six years. Additionally, Holly Wichman generously offered her time and equipment, both which greatly improved my experimental designs.

My experience at the University of Idaho was greatly enriched both professionally and personally by the graduate students and post docs associated with the Bioinformatics and Computational Biology program. In particular, I spent countless hours talking science and sharing ideas with Matthew Pennell, Daniel Beck, Travis Hagey, CJ Jenkins, Roxana Hickey, Hannah Marx, Daniel Caetano, Kenetta Nunn, and Josef Uyeda. I'd also like to thank Michael France and Janet Williams for making sure I utilized my classes to their full potential and Matthew Singer and Eliot Miller for making sure I spent enough time outside to appreciate nature and all its diversity.

Funding for me and the research in this dissertation came from the National Science Foundation (NSF) BEACON Center for the Study of Evolution in Action under Cooperative Agreement No. DBI-0939454 and NSF DDIG grant DEB-1406762. The Institute for Bioinformatics and Evolutionary Studies (IBEST) at the University of Idaho provided stipend and core equipment. Funding was also possible through grants from the National Institutes of Health National Center for Research Resources (5P20RR016448-10), and the National Institute of General Medical Sciences (8 P20 GM103397-10). This work was also supported by NIH grants P30GM103324 and P20RR016448 (L. Forney, PI).

Of course, many dissertations are written with copious amounts of coffee and this one was no exception. So I would be remiss if I didn't say thanks to the many baristas of One World Cafe in downtown Moscow for making sure I stayed well caffeinated.

I'd like to thank the members of the Hohenlohe lab – past and present – for all their help. Matthieu Delcourt was always willing to answer any and all quantitative genetics questions I had. Tamara Max and Cody Wiench kept the lab afloat and were always up-to-date on the latest laboratory techniques. Early drafts of this thesis as well as most of my talks were greatly improved by Sarah Hendricks, Amanda Stahlke, Brendan Epstein, and Kim Andrews.

Lastly, I'd like to acknowledge Genevieve Metzger for her infinite friendship and support in all things. From the very beginning of my academic career she has been there. Indeed, without her encouragement I probably wouldn't have started my thesis project in the first place. She has been there with me through the ups and downs of graduate school. Though we have yet to publish anything together, our collaborations in life have been the most rewarding. Thank you.

Dedication

I dedicate this work to Genevieve, Hannah, and Torin.

Thank you for showing me what really matters in life.

Table of Contents

List of Tables

List of Figures

CHAPTER 1
General introduction

Understanding the nature of diversity as long been of interests to biologists. How can interbreeding individuals give rise to such phenotypic and taxanomic diversity? Theoretical and empirical research over the past century has fundamentally changed our understanding of the evolutionary forces that generate diversity and recent work has investigated how suites of interconnected populations can adapt to novel environments via various types of genetic interactions. Herein I extend our understanding of how both epistatic and physical interactions can shape multivariate phenotypic (co)variation and heterogeneous genomic differentiation.

1.1 LAYOUT OF CHAPTERS IN THIS DISSERTATION

Chapter 2 — Models of multivariate phenotypic evolution based on quantitative genetics have largely not incorporated a network-based view of genetic variation. By modeling simple two-gene regulatory networks I found that the nature of the matrix of mutational (co)variation (the M-matrix) is strongly affected by network topology. Both standing genetic variation (the G-matrix) and rate of adaptation are constrained by **M**, so that **G** and adaptive trajectories are curved across phenotypic space. Under weak selection the phenotypic mean at migration-selection balance also depends on **M**.

Chapter 3 — I extended the network models of Chapter 2 to allow mutations in the network topology. I found that network changes can be beneficial early when populations are displaced from their phenotypic optimum. These network changes also created historical contingencies such that the trajectory of later adaptation depends heavily on the resulting network structure that evolved. I also found that network architecture itself can result in overdominance and showed that such overdominance can lead to persistent reproductive isolation between populations adapting in parallel. The C++ based simulation program that I created to model network evolution, *NetworkEvolution*, is freely available online.

Chapter 4 — The role of meiotic recombination in adaptation ties Mendelian principles to the evolutionary processes that occur at the population level. Thus, further understanding of physical genetic interactions would benefit from efficient methods for directly measuring rates of recombination across the genome, including crossovers and non-crossover gene conversion events. I created a Hidden Markov Model-based approach for estimating recombination rates, based on genomic sequence data from haploid products of meiosis and diploid populations, both produced by admixture between two genetically characterized parents. I used this method, together with next-generation sequencing, to identify recombination hotspots and cold spots as well as characterize rates and sizes of gene conversion events in the budding yeast *Saccharomyces cerevisiae*. The methods employed here have been implemented in the R package *HMMancestry*.

Chapters 5 & 6 — It is increasingly evident that taxonomic diversity can occur despite ongoing gene flow between interbreeding populations. Now that we are in the genomic era, studies have reported that loci of adaptive divergence often cluster within the genome of divergently evolving populations or sister species. Mathematical models have been put forth to explain how such "genomic islands of divergence" can form as a result of physical gene interactions (i.e., linkage) between established divergently selected loci and *de novo* mutations. Using both modeling (Chapter 5) and experimental (Chapter 6) approaches I found that an alternative method for island formation can arise from standing genetic variation, which is likely to occur at shorter timescales than *de novo* mutations and might be more in line with empirical studies.

CHAPTER 2

Genetic regulatory network motifs constrain adaptation through curvature in the landscape of mutational (co)variance[1]

2.1 SUMMARY

Systems biology is accumulating a wealth of understanding about the structure of genetic regulatory networks, leading to a more complete picture of the complex genotype-phenotype relationship. However, models of multivariate phenotypic evolution based on quantitative genetics have largely not incorporated a network-based view of genetic variation. Here we model a set of two-node, two-phenotype genetic network motifs, covering a full range of regulatory interactions. We find that network interactions result in different patterns of mutational (co)variance at the phenotypic level (the M-matrix), not only across network motifs but also across phenotypic space within single motifs. This effect is due almost entirely to mutational input of additive genetic (co)variance. Variation in M has the effect of stretching and bending phenotypic space with respect to evolvability, analogous to the curvature of space-time under general relativity, and similar mathematical tools may apply in each case. We explored the consequences of curvature in mutational variation by simulating adaptation under divergent selection with gene flow. Both standing genetic variation (the G-matrix) and rate of adaptation are constrained by M, so that G and adaptive trajectories are curved across phenotypic space. Under weak selection the phenotypic mean at migration-selection balance also depends on M.

2.2 INTRODUCTION

Recent years have seen an explosion in the functional understanding of genetic interactions, including mapping of large genetic regulatory and metabolic networks (Dieckmann and Doebeli,

[1] Previously published as: Hether T.D. and Hohenlohe P.A. 2014. Genetic regulatory network motifs constrain adaptation through curvature in the landscape of mutational (co)variance. Evolution 68:950-964. see APPENDIX A for License Agreement.

1999; Stuart, 2003; Huang et al., 2007; Dixon et al., 2009; Costanzo et al., 2010; Zhang et al., 2011). These data have led toward a more comprehensive understanding of complex phenotypes, and emphasize the complexity and non-linearity of the genotype-phenotype relationship (Benfey and Mitchell-Olds, 2008; Mitteroecker, 2009; Tøndel et al., 2011; Travisano and Shaw, 2013). In particular, pleiotropy and functional epistasis are ubiquitous in genetic regulatory networks (Tyler et al., 2009), and this has important consequences for the evolution of complex phenotypes.

However, traditional quantitative genetic models of multivariate adaptation typically assume phenotypic traits to be affected by a large number of loci with largely independent, additive effects (Lande and Arnold, 1983; Turelli, 1984; Arnold et al., 2001, 2008). While pleiotropy and statistical epistasis are sometimes included in these models (e.g. Jones et al., 2003, 2007; Alvarez-Castro and Carlborg, 2007), the effects of specific genetic regulatory network architectures on quantitative genetic predictions of adaptation are not well understood. Incorporation of a network-level functional view of genetic interactions into models of multivariate phenotypic evolution represents a new synthesis in biology, enabled by a new wealth of empirical data (Zhu et al., 2009; O'Malley, 2012).

An initial step toward this synthesis is to explore the consequences of simple network motifs on patterns of dominance, pleiotropy, and epistasis, considering the equilibrium expression level of a gene in the network as the phenotype (Omholt et al., 2000; Gjuvsland et al., 2007a; Aylor and Zeng, 2008). Here we apply a similar modeling approach to multivariate phenotypic space. In multivariate evolution, mutational and genetic correlation among traits can either constrain or facilitate adaptation, depending on the relationship between the direction of selection and genetic correlation (Schluter, 1996; Hansen and Houle, 2008; Agrawal and Stinchcombe, 2009; Walsh and Blows, 2009). Such correlations are expected to result from factors including the pleiotropy and epistasis inherent in genetic networks. Moreover, non-linearity in the genotype-phenotype map resulting from genetic network architecture means that the patterns of mutational correlation may change across phenotypic space, even when the mutational process at the genotypic level remains constant (Mitteroecker, 2009). This variation across phenotypic space could substantially affect both adaptive and neutral evolutionary trajectories (Steppan et al., 2002; Arnold et al., 2008). However, the ways in which genetic regulatory network architecture may induce this variation have not been well quantified.

Here we consider a set of two-node network motif models, covering all basic types of regulatory interactions, in which the phenotypes of interest are the expression levels of the two loci. The mathematical form we use to model regulatory interactions is general to Michaelis-Menten kinetics as well as other modes of gene regulation (Omholt et al., 2000; Gjuvsland et al., 2007a), and we explore the complete set of possible two-node interactions in this form. The two nodes in the network, while described below as single loci, may also be interpreted as well-connected modules in a larger network that interact in relatively simple ways. We model the interactions in these networks with differential equations describing dynamic gene expression, where the phenotypes are equilibrium gene expression levels. We assess whether simple network motifs lead to non-linearity in the genotype-phenotype map that is sufficient to create not only mutational and genetic correlation, but also variation in patterns of that correlation across phenotypic space. Using simulations of adaptive divergence with gene flow between two populations, we test whether the resulting curvature in phenotypic space constrains rates and trajectories of adaptation.

2.3 METHODS

2.3.1 *Modeling Gene Regulatory Networks*

We modeled a set of six two-node gene regulatory networks using systems of ordinary differential equations (ODEs) describing gene expression levels. These two-locus ODEs are analogous to Gjuvsland et al. (2007a)'s three-locus models, and they describe the rate of change of the concentrations of gene products x_1 and x_2 given the genotypic values α_1 and α_2 and the parameters θ and γ. These ODE systems reach stable equilibrium levels of expression, and we use the equilibrium expression levels of gene products x_1 and x_2 as the two phenotypic traits for any instance of a network motif. We do not explicitly model transcription and translation or specify what type of gene product is involved, in order to apply the models to any type of regulatory signal that could lead to interactions between loci or between tightly connected modules in a genetic network. Our model uses diploid individuals but does not contain any dominance, so we define the "genotypic value" α_i at each locus as the sum of allelic effects for the two alleles. Positive or negative gene regulation was modeled as a sigmoid function (Figure 2.1). For example, concentration of gene product x_1 has a positive effect on dynamic expression levels of locus 2 in the second equation of (Figure 2.1A), so locus 1 positively regulates locus 2. The parameter θ represents the amount of regulator needed to get half of the maximum expression rate and γ is the decay rate of expressed gene product (Gjuvsland et al., 2007a). For simplicity in the current study, these two parameters were fixed ($\theta = 300$, $\gamma = 1$).

Setting the ODEs for each motif to zero and solving for x_1 and x_2 yields unique solutions for the gene expression levels at equilibrium as a function of the genotypic values, θ, and γ (see Appendix A). We assume no environmental variation; therefore, for a given genotype in a particular network motif we can calculate both equilibrium expression levels – i.e. the phenotypic trait values – directly. We assessed stability of equilibrium expression levels by calculating the Jacobian matrix linearization of the ODEs at equilibrium points. Equilibrium trait values are stable for all motifs when allelic effects and trait values are positive, conditions that are assumed throughout this study (see Appendix A). For each regulatory motif we also solved for genotypic value (sum of the allelic values at each locus) as a function of equilibrium expression levels. These solutions are unique, so that the genotype-phenotype map is 1:1 at the level of genotypic values for all motifs across positive gene expression levels.

2.3.2 *Estimating **M**, **G**, and epistatic (co)variance*

We estimated the matrix of mutational (co)variance **M** across phenotypic space for each motif using a linear approximation to the genotype-phenotype map as follows. For each motif we calculated the 2x2 Jacobian matrix J_i of the genotype-phenotype map. Then for motif i, $M_i = J_i \ \Sigma \ J_i^T$, where Σ is the matrix of mutational variance introduced per generation at the level of genotypic values. We assumed Σ to have zero covariance (i.e. no correlation in mutation between loci) and equal variance terms 2σ, where $\sigma^2 = 17.3$ is the per-allele mutational variance in allelic value following a continuum of alleles model (Kimura, 1965). This per-allele mutational variance effectively scales the total size of **M**, but does not affect the covariance structure of **M** at all. To validate the linear transformation approximation of **M** against the M-matrix that would occur in a polymorphic population, we also estimated **M** numerically by creating populations centered at nine points in a grid across phenotypic space (at $x_1 = 200, 300, 400$ and $x_2 = 200, 300, 400$) for each motif. We randomly sampled phenotypic values for 10,000 individuals from a bivariate

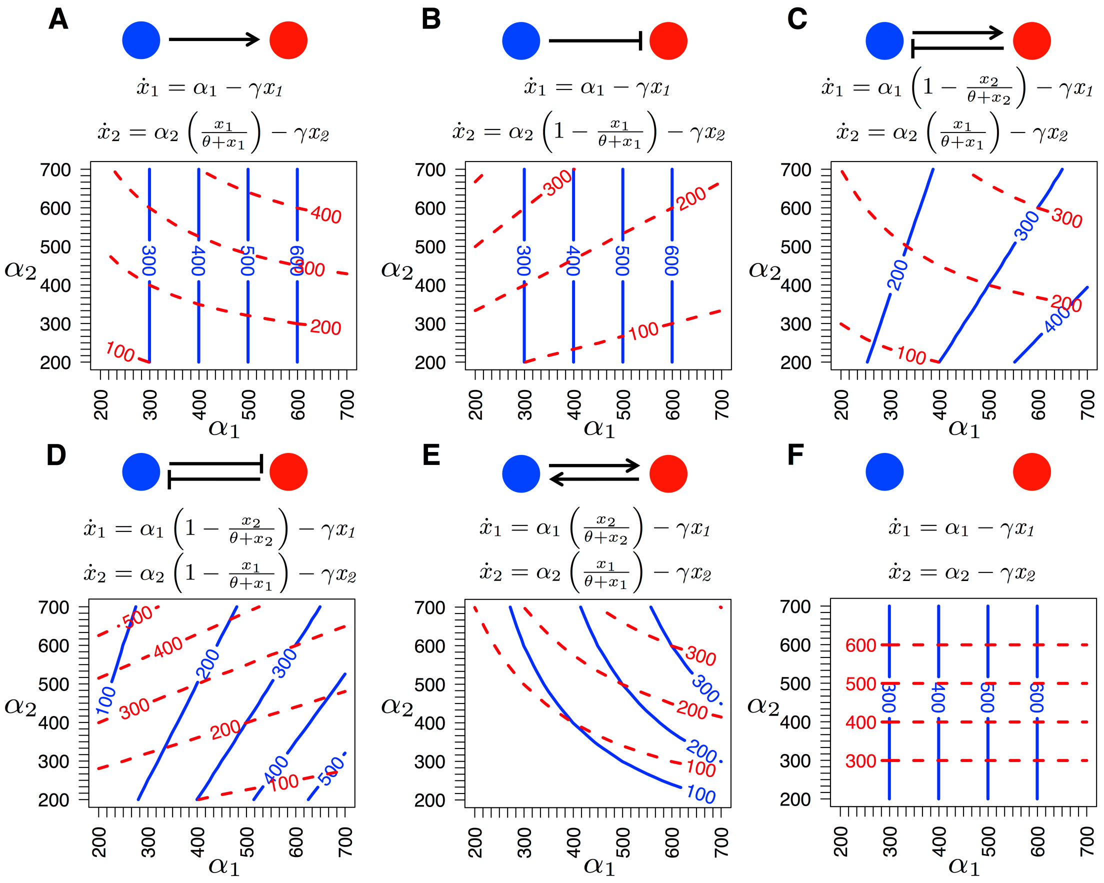


Figure 2.1: Gene regulatory network motifs. Below each motif are the system of ordinary differential equations governing gene expression levels x_i and a graphical depiction of the genotype-phenotype map. Parameters are genotypic value α_i, the sum of allelic values at locus i; θ, concentration of the regulator at which half of the maximum activation level is reached; and γ, gene product decay rate. All motifs reach a single stable equilibrium gene expression level given a pair of genotypic values. Contours represent these phenotypic trait values x_1 (solid blue) and x_2 (dashed red) as a function of genotypic values.

Gaussian distribution with standard deviation of 20 phenotypic units. We mutated each allele in all 10,000 individuals by adding a random deviate, sampled from a Gaussian distribution with variance $\sigma^2 = 17.3$, and calculated M as the (co)variance of phenotypic deviations resulting from allelic mutation. The resulting **M** matrices were indistinguishable from those calculated above, so the linear approximation method was used for all calculations below.

We also estimated the G-matrix of additive (co)variance and the epistatic (co)variance matrix for the nine populations in each motif described above, using the animal model (Kruuk, 2004; Wilson et al., 2010). Each population was evenly split into males and females, and 100 sires were randomly mated to 10 dams each resulting in 1000 offspring, with independent assortment between loci. Using the resulting pedigree information, we obtained breeding values for individuals and population estimates of the G-matrix by fitting a generalized linear mixed model with the R package MCMCglmm (Hadfield, 2010). Because our model includes no dominance (alleles are purely additive within each locus) and no random environmental effects on phenotype, the population-level residual (co)variance matrix includes solely epistatic (co)variance. For the random effects prior, we set the variance component equal to the phenotypic (co)variance and set the parameter "nu" to 2. To the speed up convergence and chain mixing properties we used parameter expanded methods (Liu et al., 1998) with prior means for the working parameter "alpha" set to (0,0) and variances set to 1,000 with zero covariance. For the residual effects prior, we set the variance component of the inverse Wishart distribution to 1,000 along the diagonal with zero covariance and nu to 0.002. We ran the Markov Chain for 12,000 generations following a 1,000-generation burn-in period, sampling every 25 generations to reduce autocorrelation.

Evolvability depends on mutational variation, so phenotypic space can be re-scaled by the mutational distance between phenotypes. To the extent that adaptation is mutation-limited, this re-scaling reflects the "evolutionary distance" traveled during adaptation to a novel phenotype. Mathematically, this distance between phenotypic values is the Mahalanobis distance scaled by the local value of **M**, so that the inverse of **M** is a Riemannian metric tensor (Jost, 2008). We created visualizations of mutation-scaled phenotypic space using an iterative algorithm for deforming a grid of bivariate phenotypes. The algorithm first scaled the grid by mutational variance along single phenotypic axes by multiplying distances from each point to its 4 nearest neighbors by the square root of the corresponding diagonal elements of the inverse of **M**, estimated at each grid point as described above. It then incorporated mutational covariance by sequentially adjusting the position of each point on the grid so that its Euclidean distance to its 8 nearest neighbors (horizontal, vertical, and diagonal) matched as closely as possible to the Mahalonobis distance between phenotypes, scaled by the local M-matrix. Code to perform this deformation was written in R and is available from the authors.

2.3.3 *Simulating divergent selection with gene flow*

We used R to create individual-based simulations to determine the effect of varying gene regulatory network motifs on adaptation under a model of divergence with gene flow. Each simulation replicate included two populations, each of size n = 2,000, exchanging migrants at rate m in an island model (Wright, 1931). To initialize each population, we used the Phenotype-to-Genotype equations (see Appendix a) to obtain the genotypic values α_1 and α_2 that correspond to a phenotype of $x_1 = 300$ and $x_2 = 300$ for each network. We then generated allelic variation by randomly drawing allelic values for each individual using a Gaussian distribution centered at half of the genotypic value and with a variance of 200. We then imposed divergent selection on the two populations by selecting toward two optimum phenotypes. The phenotypic optima for populations 1 and 2 were set to phenotypic points ($x_1 = 150$, $x_2 = 450$) and ($x_1 = 450$, $x_2 = 150$), respectively. Thus divergent selection was imposed along the axis representing negative correlation between the two traits, and selection on the two populations was symmetrical in terms of distance to the optimum and strength of selection.

We used a Gaussian fitness function to calculate individual fitness, w:

$$W = e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x_{opt}})^T\mathbf{\Omega}^{-1}(\mathbf{x}-\mathbf{x_{opt}})} \tag{2.1}$$

where $\mathbf{x}$ is a column vector containing trait values, $\mathbf{x_{opt}}$ is a column vector of phenotypic optima for each trait, and $\mathbf{\Omega}$ is a symmetrical 2x2 matrix describing the landscape of stabilizing selection, analogous to a (co)variance matrix. For simplicity, we assume equal strengths of stabilizing selection for each trait (i.e. the diagonal elements of $\mathbf{\Omega}$ are $\omega = \omega_{11} = \omega_{22}$) and no correlational selection ($\omega_{12} = \omega_{21} = 0$). Individuals were randomly chosen to mate with a probability proportional to their relative fitness ($w\frac{W}{W_{max}}$).

Offspring randomly received one allele per locus from each parent, with independent assortment between loci. This process was continued, sampling with replacement from the parental generation, until the new population's size equaled the parental size, so that generations were non-overlapping. During meiosis there was a probability μ that an allele mutates. In this case the new allelic value was the sum of the original allelic value plus a random value centered at zero with variance $\sigma^2 = 17.3$. For these simulations μ was set to 0.01, so total allelic variance introduced by mutation per generation per allele was 0.173. Note that this represents less total mutational variance, but identical covariance structure, compared to the M-matrices calculated above. Migration between populations followed mating. Individuals (from both populations) were chosen to migrate with probability m, then pooled and redistributed randomly back to one of the populations. To characterize the effects of network motif on adaptation, we simulated 10 replicates of population pairs for each motif for 1,000 generations with parameter values $m = 0.001$ and $\omega = 10,000$. To explore the effects of selection strength and migration rate, we simulated 10 replicates of population pairs for motif C across multiple parameter values (m = 0, 0.0001, 0.001, 0.01; ω = 1000, 10,000, 50,000). We ran these simulations for 20,000 generations to characterize equilibrium levels of adaptation. For reference, at selection strengths of ω = 1000, 10,000, and 50,000, fitness of individuals 10 phenotypic units away from the optimum is 90%, 99%, and 99.8% of the fitness at the optimum, respectively. Initial selection during the simulations was strong; mean-standardized selection gradients per trait (Hansen and Houle, 2008) for the null motif at the initial population mean would be +/- 25, 2.94, and 0.60, respectively.

2.3.4 *Quantifying genetic variation and adaptation*

We estimated the G-matrix of additive genetic (co)variance at generations 50, 100, 500, and 1000 for the shorter simulations, and additionally at generation 20,000 for the longer simulations. Before the mating phase of each of these generations, we conducted a "side experiment" in which 100 sires were mated to 10 dams each to produce 1000 offspring, and we estimated G using MCMCglmm (Hadfield, 2010) as described above. Note that this pedigree data was produced independent of fitness, and these offspring were not those used for the next generation of the simulation.

We used several metrics to quantify adaptation and the structure of **G** during the course of divergence with gene flow. The extent of adaptation was calculated as:

$$A = 1 - \frac{D_{opt,t}}{D_{opt,ini}} \tag{2.2}$$

where $D_{opt,ini}$ is the Euclidean distance between the initial starting position (300, 300) and the phenotypic optimum, and $D_{opt,t}$ is the Euclidean distance between a given population's mean phenotype at time t and its phenotypic optimum. Equation 2.2 represents a ratio where a value of 1 can be interpreted as a population being well-adapted to its respective phenotypic optimum. We quantified aspects of **G** with four metrics (Jones et al., 2003): i) size Σ, calculated as the sum of the eigenvalues, equal to the sum of the variance terms; ii) eccentricity or shape ε, calculated as the smaller eigenvalue divided by the larger eigenvalue; iii) orientation φ, calculated as the angle between the leading eigenvector g_{max} and the axis of x_1, and iv) effective dimensionality n_D, calculated as the total variance divided by the leading eigenvalue (Kirkpatrick, 2009).

2.4 RESULTS

2.4.1 *Genotype-phenotype map under simple network motifs*

We modeled a set of six genetic regulatory network motifs (Figure 2.1). For all motifs, the genotype-phenotype map was 1:1 at the level of genotypic values, although not allelic values (see Appendix A). In the absence of any interlocus interaction (null motif; Figure 2.1F) each phenotype equaled the genotypic value at the corresponding locus. In all other cases, both pleiotropy and epistasis were evident in the genotype-phenotype maps. Pleiotropy resulted from unidirectional (motifs A and B) and bidirectional (motifs C, D, E) regulation between loci, because the genotypic value at a single locus affected the expression levels of both loci. In contrast, the nature of epistasis in allelic effects on phenotypes depended on the type of interaction. Negative regulation led to linear contours on the genotype-phenotype map for the expression level of the downstream gene – but note that where the contours are not parallel, the relationship between multivariate genotypes and phenotypes is still nonlinear (e.g. Figure 2.1B,D; see Appendix A). In contrast, positive regulation led to hyperbolic curved contours in the genotype-phenotype map for the downstream gene (e.g. Figure 2.1A). In both cases, the non-linearity in mapping from genotype to phenotype for one or both traits indicates statistical epistasis; that is, the phenotype resulting from allelic substitutions at both loci differs from the expectation based on the independent additive effects of the alleles considered separately (Phillips, 2008).

2.4.2 *Landscape of mutational variation*

We assessed the landscape of mutational variation using the M-matrix of quantitative genetics, a (co)variance matrix of the phenotypic variation across multiple traits produced by mutation per generation. The motifs produced a wide range of mutational variance in each trait and, with the exception of the null motif (F), correlation between traits (Figure 2.2). Moreover, **M** exhibited striking variation across phenotypic space even when network motif and all other parameters were held constant for all but the null motif. The overall size of **M** – the total amount of phenotypic variance produced by mutation – varied across motifs as well as across phenotypic space within motifs. The magnitude of mutational correlation, and thus the effective dimensionality of **M**, varied across phenotypic space for all but the null motif (Figures 2.1, A.1). The sign of the correlation also shifted under the negative feedback loop motif (Figure 2.1C), leading to the most extreme variation in dimensionality (Figure A.1C).

Although the network motifs exhibit strong functional epistatic interactions between loci and statistical epistasis in the genotype-phenotype map, the patterns of (co)variance in **M** were essentially the result of additive genetic (co)variance with only negligible epistatic (co)variance. We estimated additive genetic and epistatic (co)variance for the two traits across phenotypic space for each motif (Figures A.2, A.3). Matrices of epistatic (co)variance were much smaller in total magnitude than the G-matrix of additive genetic (co)variance, and the pattern of covariance was similar to **G**. Total epistatic variance represented a negligible contribution to total phenotypic variance, such that narrow-sense heritability was greater than 0.99 for both traits in all motifs, for those populations at the center of phenotypic space. Because additive variation contributes most directly to the response to selection, the covariance patterns in **M** are predicted to have a strong effect on adaptation to the extent that evolution is mutation-limited. If this is so, we can get relative estimates of mean evolvability (Hansen and Houle, 2008) from **M**. This also varied widely across phenotypic space (Figure A.4).

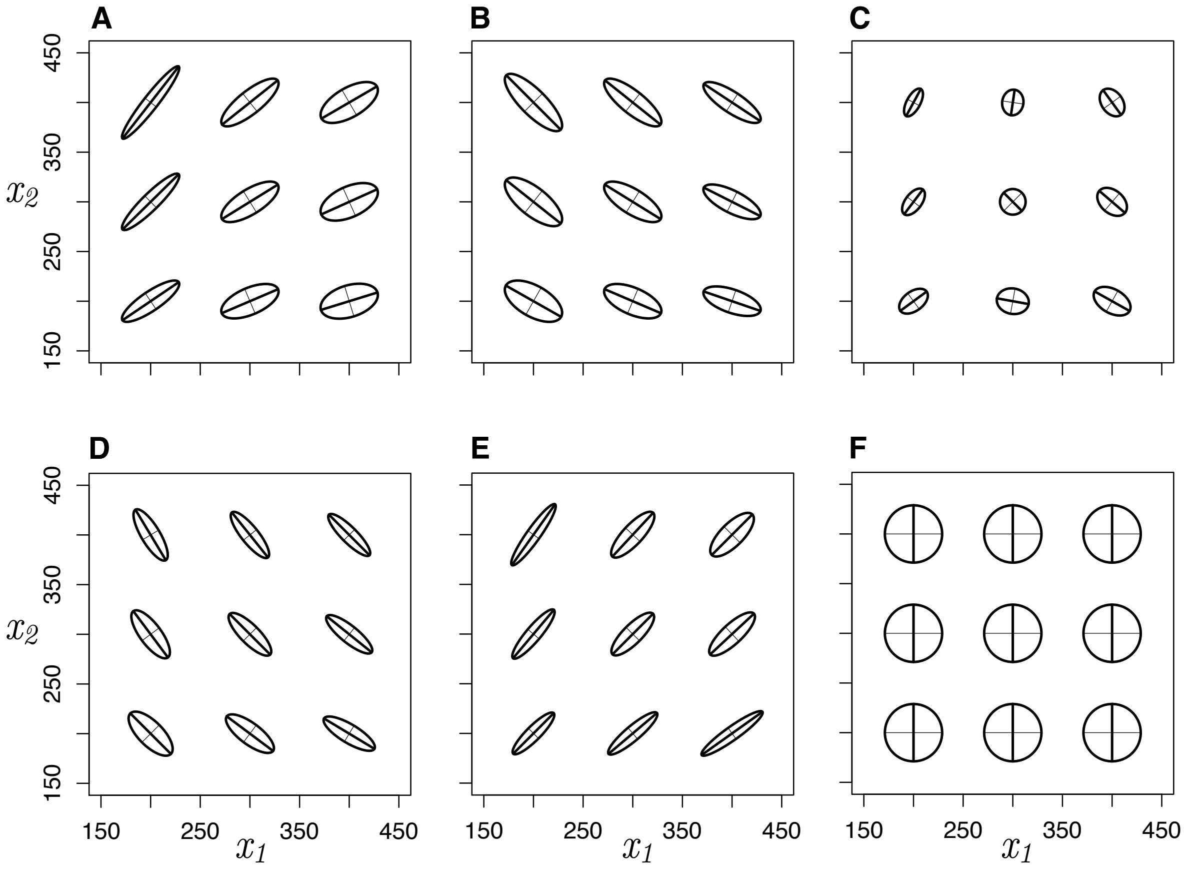


FIGURE 2.2: The mutational (co)variance matrix M across phenotypic space. For each network motif (A-F, labeled as in Figure 2.1), M-matrices for nine populations are plotted as 95% confidence ellipses around mutational variation produced per generation, scaled up by a factor of 2.5 for visualization. Axes within each ellipse represent the first (thick line) and second (thin line) eigenvectors, or principal components, of mutational variation.

Because **M** varied across phenotypic space for all but the null model of network motifs, re-scaling by mutational distance induced curvature in the phenotypic landscape (Figure 2.3). Note that this re-scaled, curved phenotypic landscape may be best represented as an n-dimensional manifold (for n traits) embedded in a higher-dimensional space, but the 2-dimensional projection of this manifold is shown in Figure 2.3. Phenotypic space was generally stretched for all motifs relative to the null. Phenotypic space was also stretched, as expected, in directions of positive correlation between traits in the case of negative gene regulation (Figure 2.3B,D) and directions of negative correlation between traits in the case of positive regulation (Figure 2.3A,E). The extent of deformation varied across network motifs as well as across phenotypic space. Deformation was particularly pronounced in regions of low genotypic value for the upstream gene and high genotypic value for the downstream gene in positive regulation (upper left corners in Figures 2.3A,C, upper left and lower right corners in Figure 2.3E). To the extent that evolution is mutation-limited, these are predicted to be regions of phenotypic space in which adaptation may be particularly constrained.

2.4.3 *Adaptation under divergent selection*

To test the effect of network-induced curvature in phenotypic space on trajectories of adaptation, we simulated pairs of populations evolving from a common ancestor toward separate phenotypic optima with migration between them, with replicate simulations to minimize stochastic differences (Figure 2.4). Network motifs had a strong influence on both the rate and the trajectories of adaptation. In terms of adaptation rate, most striking is the constraint on adaptation in the direction of negative correlation between traits when gene regulation is positive (Figures 2.4A,C,E). This corresponds to the reduced mutational variation and stretching of evolutionary distance in these regions of phenotypic space illustrated in Figures 2.2 and 2.3. Conversely, adaptation is relatively rapid under negative gene regulation (Figure 2.4D). Network motifs also produced curved trajectories of adaptation through phenotypic space, most notably early in adaptation for motif A and close to the optima for motifs B, C, and D. Curved trajectories represent the tension between the orientation of directional selection and the orientation of additive genetic variation, summarized by the G-matrix (Lande, 1979; Arnold et al., 2008). The effects of mutational variation on adaptation rates and trajectories depend on G (Figure 2.4), and **G** in our simulations

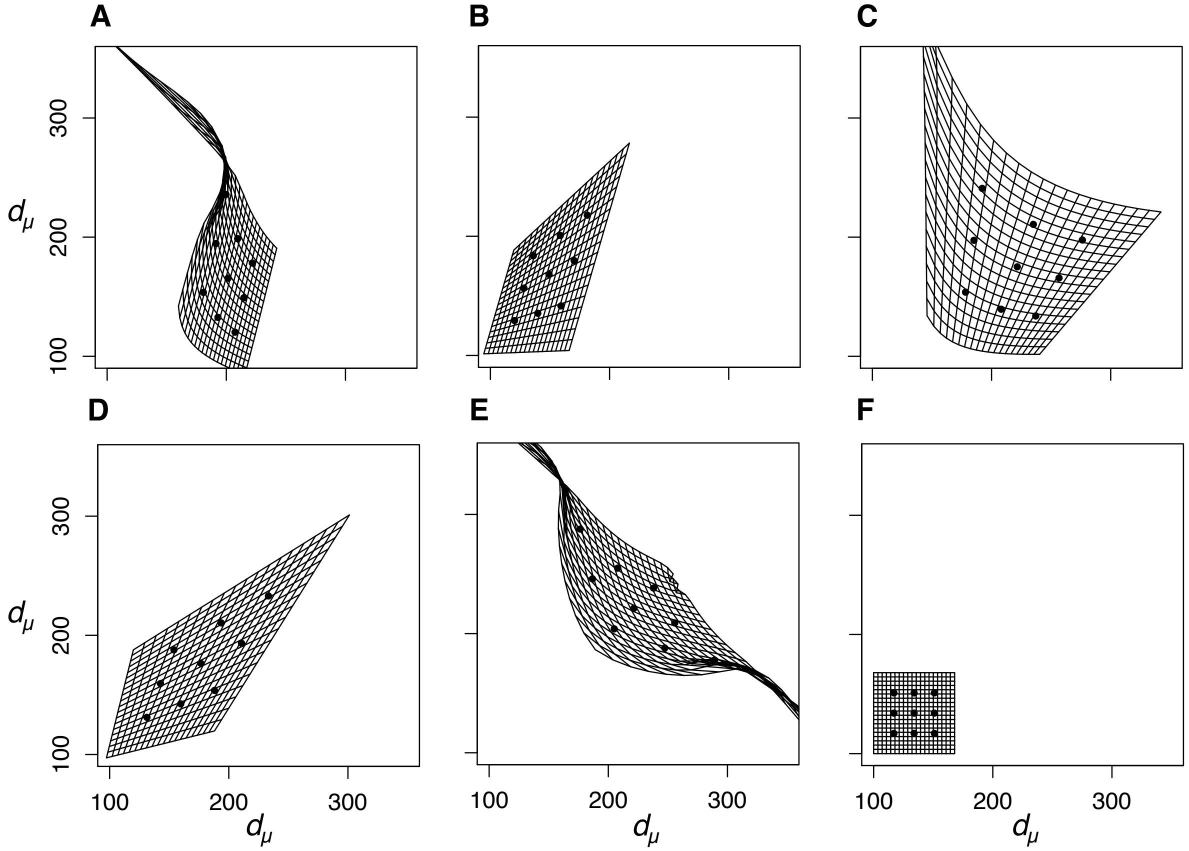


FIGURE 2.3: Phenotypic space re-scaled by mutational (co)variance for each motif. Phenotypic values from 100-500 are represented as a grid that is deformed such that distances between phenotypes, d_μ, in this new depiction represent equal amounts of mutational variation. For visual reference, the locations of the nine populations from Figure 2.2 are plotted as black dots in this new mutational space. Note that this re-scaling may cause the 2-dimensional phenotypic space to curve outward into higher dimensions, but it is represented here as the projection of this curved manifold onto a plane.

was strongly affected by **M**. Adaptation was constrained when the major axis of **M**, and thus the major axis of **G**, is perpendicular to the orientation of directional selection, and adaptation was facilitated when **M** and **G** align with directional selection.

To further explore the interactions among migration, selection, drift, **M**, and **G**, we focused on the negative feedback loop represented in motif C, extending the simulations of divergent selection to reach equilibrium and varying strength of selection and migration rate. Motif C showed striking differences in the degree and direction of mutational correlation across phenotypic space (Figure 2.2C). This is expected to lead to regions of elevated and depressed evolvability (e.g., compare upper left and lower right regions in Figure 2.3C, respectively, and the two populations in Figure 2.4C). As expected, we found that selection strength generally increased and migration generally decreased both the rate and equilibrium extent of adaptation (Figures 2.5 and 2.6; Table 2.1). As seen in the trajectories of adaptation across all motifs (Figure 2.4), the rate of adaptation toward the selective optimum was lower in regions of phenotypic space where mutational variance in the direction of selection was limited for the negative feedback motif, and this effect was consistent across selection strengths and migration rates (Figures 2.5, 2.6, A.5,A.6, A.7, A.8, and A.9). Thus adaptation was slower in the phenotypic region around the optimum of population 1 as opposed to the region around the optimum of population 2.

Re-scaling phenotypic space by mutational distance makes this difference clear: the two populations are seen to travel roughly the same mutational distance over the course of 1000 generations, but the optimum of population 1 is simply farther away from the starting point in mutational distance (Figure 2.7). Accordingly, the distance traveled in phenotypic space by population 1 was much less than that traveled by population 2, despite the entirely symmetrical directional selection, migration, and genetic drift acting on each (Table 2.2). However, the distance traveled by the two populations was much more similar in mutation-scaled space. In fact, population 1 traveled farther in this re-scaled space, as a result of a steeper selection gradient acting during the simulation because population 1 remained farther from its respective optimum than population 2.

With weaker selection, genetic drift had a larger effect, causing higher levels of variation across replicate simulations (Figures A.5, A.6, A.7, A.8, and A.9). Over longer time scales, in the case of weak selection, an equilibrium reflecting drift/migration/selection/mutation balance was

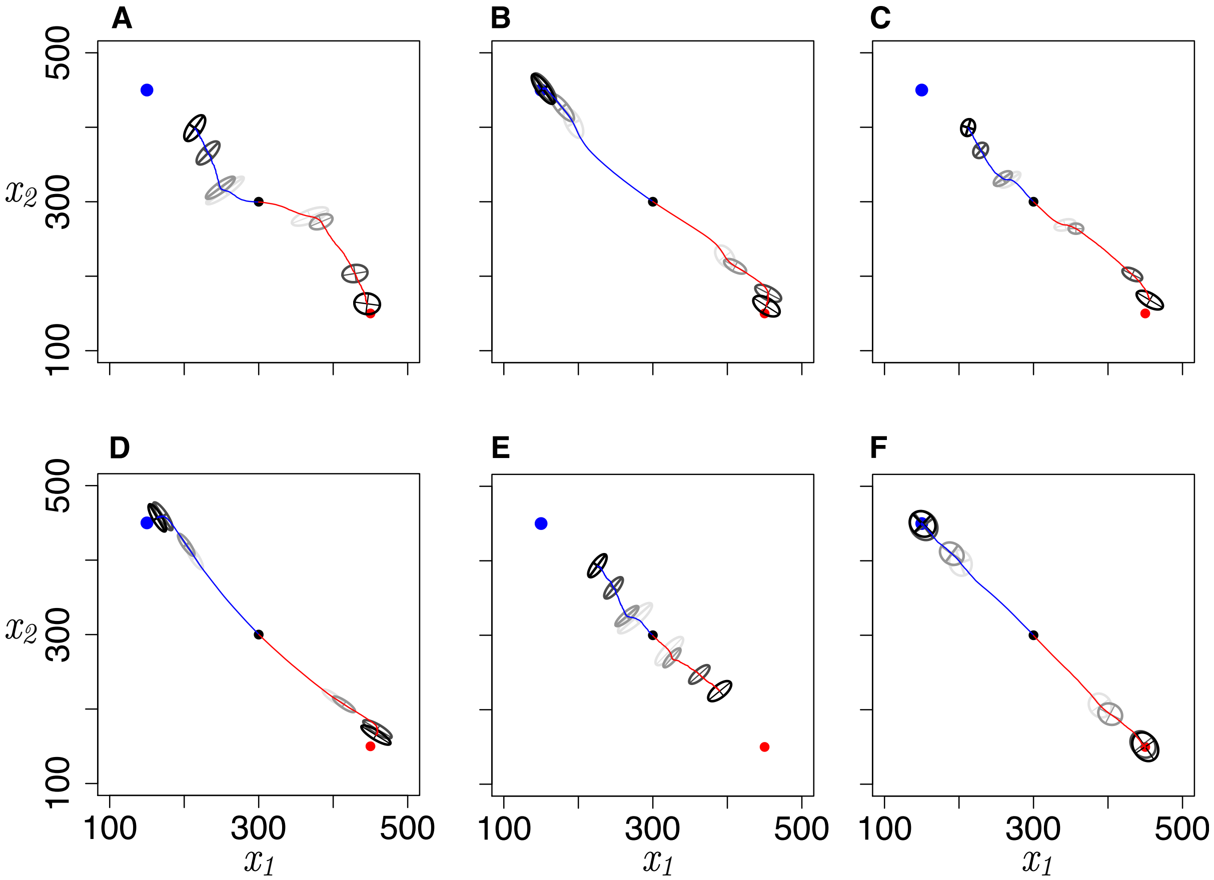


FIGURE 2.4: Evolution during 1,000 generations in response to divergent selection with migration across network motifs. Blue and red lines track the phenotypic means of the two populations evolving toward selective optima at the blue and red points, respectively, averaged across 10 independent replicates for each motif. G-matrices are drawn as 95% confidence ellipses at 4 time points (50, 100, 500, and 1,000 generations; darker ellipses denote more recent G-matrices). Parameter values are m = 0.001, ω = 10,000, μ = 0.01, size of each population = 2,000.

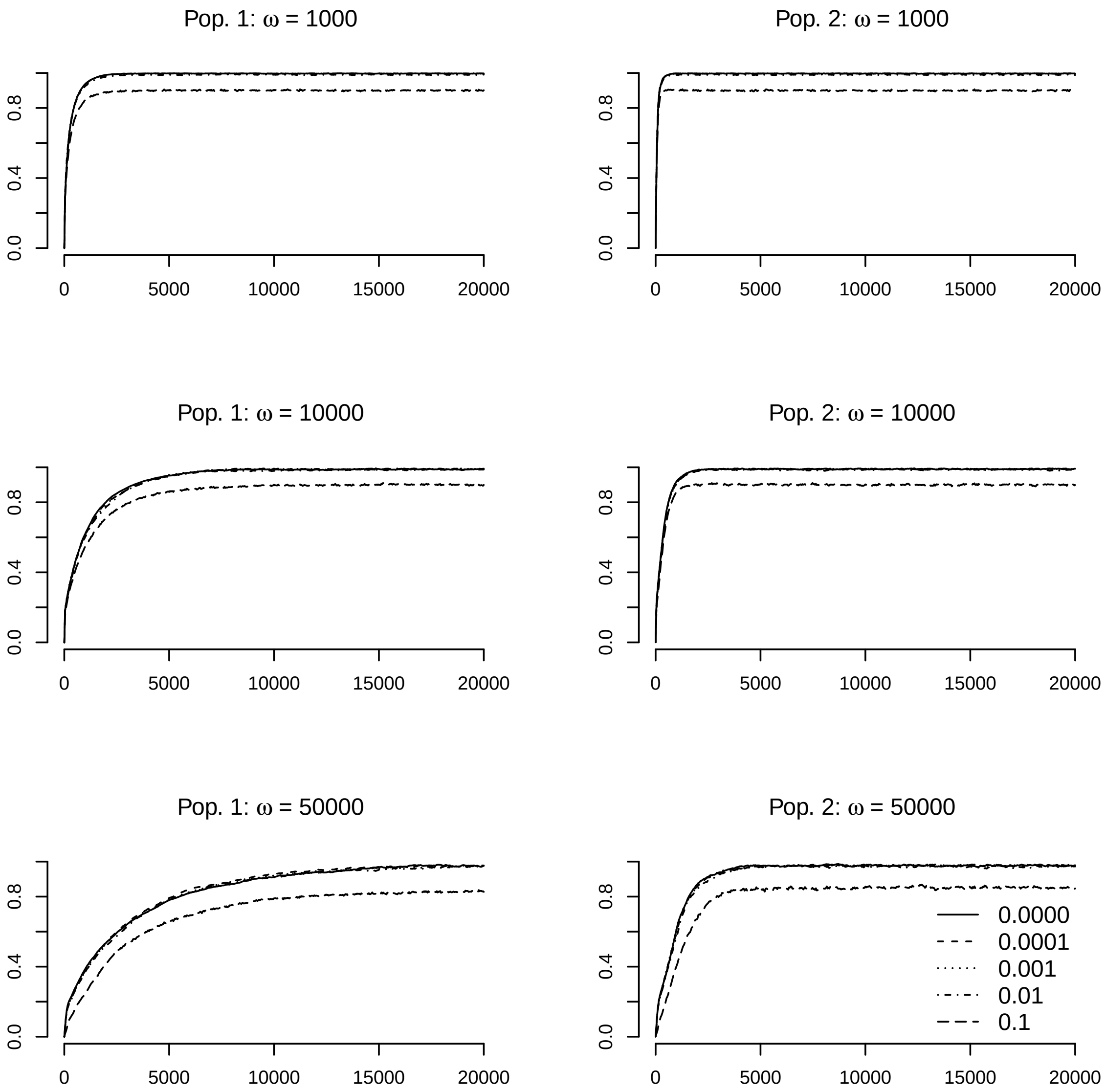


Figure 2.5: Extent of adaptation through time for the negative feedback network (motif C). Rows denote different selection strengths and columns denote either population 1 (left) or 2 (right). Within each panel are 5 different migration rates between the two populations. Plotted is the mean adaptation ratio across 10 simulated replicates for each parameter combination.

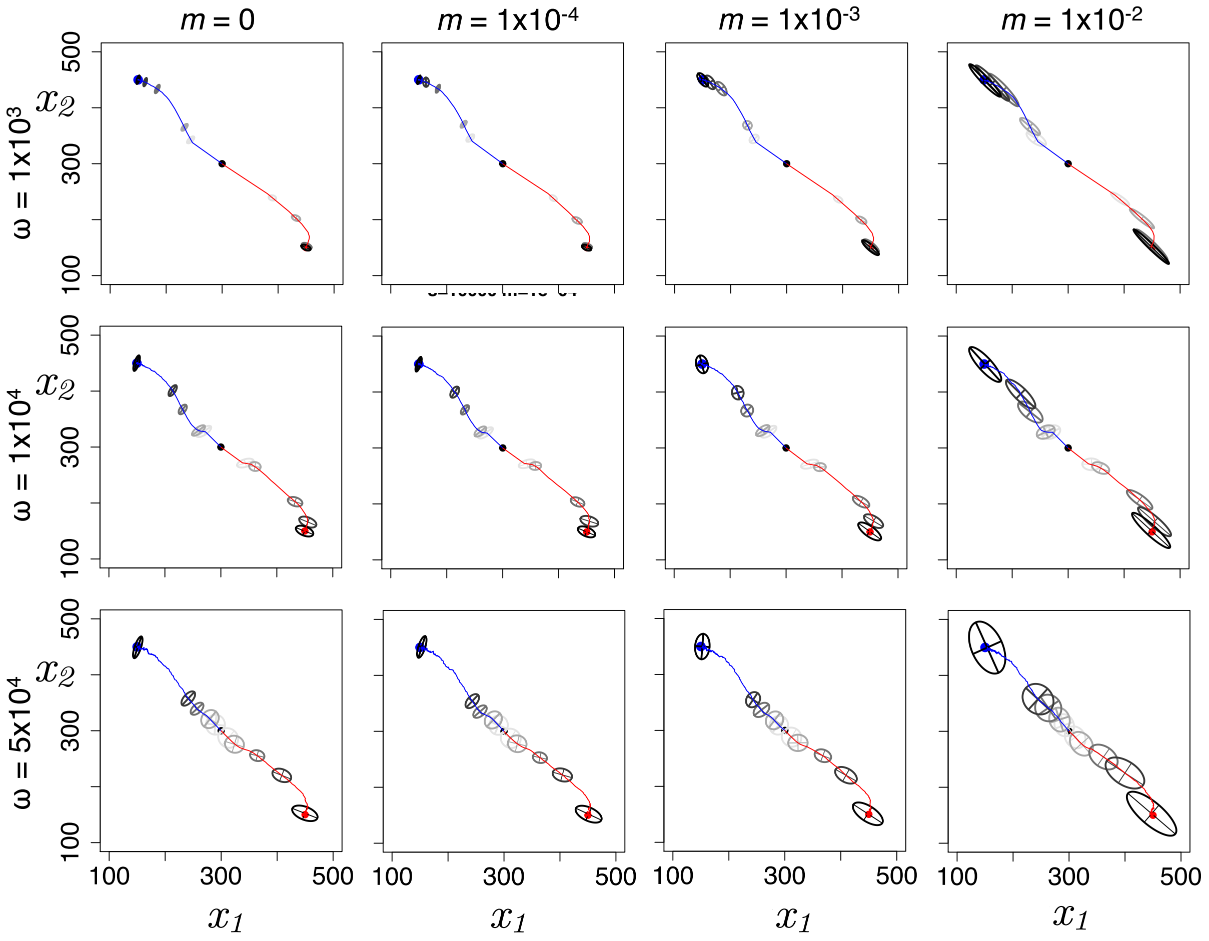


Figure 2.6: Adaptive divergence for the negative feedback network (motif C). Each plot shows the average of 10 replicates of two populations diverging from the initial starting position (black dot) to either the blue or red phenotypic optima for a given combination of strength of stabilizing selection ω and migration m. Blue and red lines track phenotypic means through the course of the 20,000-generation simulation. G-matrices are drawn as 95% confidence ellipses at 5 time points (50, 100, 500, 1,000, and 20,000 generations; darker ellipses denote more recent G-matrices).

Table 2.1: Adaptation and G-matrix summary statistics for motif C at migration-selection balance (20,000 generations; see Figure 2.5). For each combination of migration m and selection strength ω, means and standard errors (in parentheses) are provided for population 1 (first row) and population 2 (second row) averaged over 5 simulated replicates. A, the extent of adaptation (see Equation 4.8); G_{ij}, the ith by jth component of **G**; n_D, dimensionality of **G**; φ, the angle of the leading eigenvector of **G** relative to trait axis x_1; Σ, the sum of the two eigenvalues of **G**; and ε, a measure of the eccentricity of **G** (see Methods).

m	ω	A	G_{11}	G_{12}	G_{22}	n_D	φ	Σ	ε
0	1e3	0.997 (4e-04)	0.9 (0.1)	2.5 (0.2)	8.3 (0.4)	1.01 (0.002)	72.9 (0.5)	9.2 (0.5)	0.01 (0.002)
		0.997 (4e-04)	11.2 (0.4)	-3.9 (0.2)	3.4 (0.2)	1.14 (0.004)	157.7 (1.2)	14.6 (0.5)	0.14 (0.004)
1e-04	1e3	0.997 (5e-04)	1.0 (0.1)	2.7 (0.1)	8.9 (0.3)	1.01 (0.002)	73.0 (0.4)	9.8 (0.4)	0.01 (0.002)
		0.996 (5e-04)	11.7 (0.5)	-3.9 (0.2)	3.6 (0.2)	1.15 (0.005)	158.2 (0.8)	15.3 (0.6)	0.15 (0.005)
1e-03	1e3	0.997 (5e-04)	16.3 (6.5)	-13.2 (6.7)	24.7 (6.7)	1.03 (0.009)	92.1 (7.5)	41.0 (13.2)	0.03 (0.009)
		0.997 (5e-04)	45.1 (12.4)	-37.6 (12.3)	37.4 (12.3)	1.09 (0.016)	148.4 (2.9)	82.5 (24.7)	0.09 (0.016)
1e-02	1e3	0.990 (8e-04)	214.4 (22.4)	-214.2 (22.1)	225.0 (22.7)	1.01 (0.002)	134.2 (0.1)	439.4 (45.1)	0.01 (0.002)
		0.989 (8e-04)	256.0 (16.0)	-245.9 (15.7)	242.2 (15.4)	1.01 (0.002)	135.8 (0.1)	498.2 (31.4)	0.01 (0.002)
1e-01	1e3	0.901 (0.0022)	2291.4 (57.7)	-2301.0 (58.3)	2320.9 (58.7)	1.00 (0.001)	134.8 (0.0)	4612.3 (116.2)	0.00 (0.001)
		0.901 (0.0021)	2469.6 (71.3)	-2424.1 (67.9)	2410.7 (68.0)	1.00 (0.002)	135.3 (0.1)	4880.3 (139.0)	0.00 (0.002)
0	5e4	0.978 (0.0037)	13.7 (1.4)	26.2 (2.8)	80.2 (7.2)	1.05 (0.005)	71.1 (1.0)	93.9 (8.4)	0.05 (0.005)
		0.977 (0.0023)	108.6 (8.3)	-31.9 (4.2)	36.9 (4.2)	1.19 (0.005)	159.9 (1.9)	145.5 (11.7)	0.19 (0.005)
1e-04	5e4	0.978 (0.0024)	11.7 (1.1)	22.3 (2.4)	73.5 (5.1)	1.05 (0.007)	72.2 (1.3)	85.2 (5.9)	0.05 (0.007)
		0.974 (0.0035)	116.9 (6.5)	-40.8 (2.5)	45.2 (2.4)	1.19 (0.009)	155.6 (1.6)	162.1 (7.9)	0.19 (0.009)
1e-03	5e4	0.977 (0.0027)	29.9 (6.2)	7.7 (8.1)	104.6 (8)	1.17 (0.037)	85.2 (4.2)	134.5 (13.5)	0.17 (0.037)
		0.977 (0.0035)	155.4 (13.6)	-79.6 (13.8)	87.5 (13.6)	1.16 (0.013)	150.4 (2.7)	242.9 (26.5)	0.16 (0.013)
1e-02	5e4	0.976 (0.0035)	274.9 (22.0)	-194.5 (28.7)	526.1 (30.6)	1.27 (0.025)	116.2 (2.2)	801.0 (49.4)	0.27 (0.025)
		0.970 (0.0027)	474.2 (19.4)	-353.8 (20)	385.2 (21.6)	1.09 (0.01)	138.9 (0.7)	859.4 (38.2)	0.09 (0.010)
1e-01	5e4	0.829 (0.0059)	2567.8 (57.1)	-2264.7 (51.9)	4815.8 (93.0)	1.19 (0.006)	121.8 (0.6)	7383.6 (130.7)	0.19 (0.006)
		0.853 (0.0050)	3563.5 (81.6)	-3183 (76.2)	3766.6 (109.2)	1.07 (0.004)	134.1 (0.3)	7330.1 (174.9)	0.07 (0.004)

Table 2.2: Alternative distance metrics for quantifying the amount of evolutionary change over the two 1000-generation evolutionary trajectories shown in Figure 2.7. Rescaled space uses the M-matrix as a metric tensor, normalizing phenotypic change by the amount of mutational variance along a trajectory. Populations 1 and 2 are those whose phenotypic optimum is at point (150, 450) and (450, 150) in phenotypic space, respectively.

	Population 1	Population 2
Euclidean distance traveled in phenotypic space	132.7	204.5
Total length of trajectory in phenotypic space	150.5	221.1
Euclidean distance traveled in rescaled space	64.8	57.9
Total length of trajectory in rescaled space	78.6	71.4

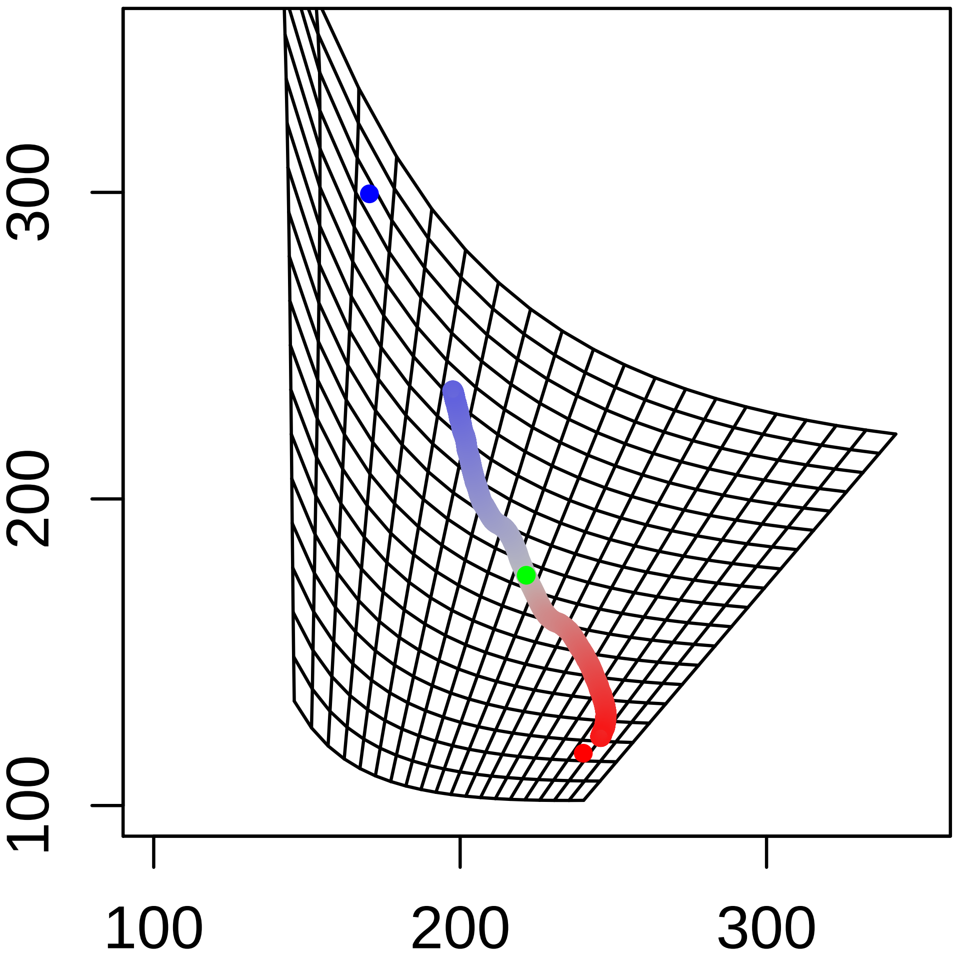


FIGURE 2.7: Single simulation run for motif C, plotted in the re-scaled space shown in Figure 2.3. Parameter values for this run are as in Figure 2.4.

reached farther from the optimum for population 1 compared to population 2, particularly at higher migration rates (Table 2.1).

The G-matrix is affected not only by mutation but also selection and migration. For the negative feedback motif (C), the structure of **G** varied widely across phenotypic space and across simulation parameters (Table 2.1; Figure 2.6). In general, stronger selection produced smaller G matrices (lower overall genetic variance; see Table 2.1), and higher migration rates shifted the pattern of genetic correlation within populations so that the major axis aligned with the direction of divergence between populations (Figure 2.6). **G** also varied strongly between the two populations within simulations, showing the effect of variation in **M** across phenotypic space. Thus the tenuous balance between selection, migration, and mutational variance led to shifts in the sign of genetic correlation across multiple factors: phenotypic space, migration rates, and strength of selection.

2.5 DISCUSSION

2.5.1 *Curvature of the landscape of mutational variation*

The M-matrix of mutational variance and covariance plays a central role in quantitative genetic models of multivariate evolution. **M** provides the ultimate source of additive genetic variation, summarized by the G-matrix, which in turn determines the response to selection (Lande, 1979). However, while increasing attention has focused on both empirically estimating **G** in natural populations and gaining a theoretical understanding of its stability and response to evolutionary forces (Arnold et al., 2008; Björklund et al., 2013), the M-matrix has received relatively less attention in part because of the difficulty of directly measuring it (Houle et al., 1996; Houle, 1998; Houle et al., 2010). One exception is Houle and Fierst (2013), who recently estimated **M** for wing traits in a set of inbred *Drosophila* lines subject to mutation accumulation. They found significant variation in **M** between lines, both in total size of **M** and in mutational covariance structure, although some similarity in **M** was maintained across lines. While the functional genetic basis of these wing traits is unknown, differences in mutation rates between the lines may account for some of the differences in **M**, particularly overall size (Houle and Fierst, 2013).

In the absence of shifts in mutational rates or process at the molecular level, one way in which **M** can evolve and differ across genotypes or populations is through shifts in the architecture of genetic regulatory networks – for example, appearance or disappearance of regulatory connections between genes (Wagner and Altenbery, 1996; Lynch, 2007). An additional way is through changes in allelic values and/or allele frequencies at loci that influence other loci in a regulatory network. In this case, substantial additive genetic variation can be produced by mutation even when genes have strong interactions at the molecular level of genes and their products, termed functional epistasis (Stadler, 2000; Gibson and Dworkin, 2004; Phillips, 2008). Here we explored shifts in the structure of mutational variation caused by functional epistasis with simple but explicit network motif models, holding the network architecture of regulatory connections constant while allowing population variation in allelic values and frequencies. We found striking variation in M at multiple levels, which influenced adaptation under divergent selection in simulation. Our models lead to several conclusions about the effect of genetic regulatory network motifs on mutational and genetic variation and on trajectories of adaptation.

First, we found that positive gene regulation produces more complex patterns of statistical epistasis (Phillips, 2008) than negative regulation, illustrated by the hyperbolic versus linear contours on the genotype-phenotype map (Figure 2.1). This is consistent with the results of Gjuvsland et al. (2007a), who found that positive regulation produces greater and/or more complex patterns of statistical epistasis than negative regulation in a three-locus, one-trait network model. This is also consistent with previous work showing higher mutational robustness resulting from negative feedback (Acar et al., 2010; Paulsen et al., 2011; Denby et al., 2012).

Second, despite the functional epistasis modeled in the network motifs and the statistical epistasis evident in the genotype-phenotype map, epistatic (co)variance at the population level was negligible. Narrow-sense heritability was greater than 0.99 for both traits in all motifs. This is in contrast to the results of (Gjuvsland et al., 2007a), who found moderate levels of epistatic variance across some parameter combinations in their network model. Why the discrepancy between statistical epistasis in the genotype-phenotype map and the lack of epistatic (co)variance at the population level? It appears that the genotype-phenotype map for these models, while curved, is smooth enough that within the phenotypic range of a population it is close to linear. Thus genetic variation within a population is nearly all additive. At this scale, pleiotropy maintains the key role in producing sometimes strong genetic covariance. As populations evolve across phenotypic space in response to directional selection, statistical epistasis then results in shifts in the covariance structure of additive variation, but not a substantial contribution of epistatic (co)variance.

Third, we found that simple network motifs produce striking variation in patterns of mutational variation, even when the mutational process is held constant at the allelic level. The M-matrix exhibits strong correlation as a result of network interactions, as expected. Moreover, the total amount of mutational variation and the sign and degree of mutational correlation depend also on the phenotypic mean, leading to variation in **M** across phenotypic space for a given network. To the extent that evolution depends on genetic variation provided by mutation, variation in patterns of mutational (co)variance effectively bends and stretches phenotypic space. The effect is analogous to the bending of space-time by gravitation under general relativity, so that the inverse of **M** acts as a Riemannian metric tensor that can be used to integrate mutational distance along evolutionary trajectories (Figure 2.7, Table 2.2), analogous to inertial body trajectories in gravitational fields (Jost and Shaw, 2006). Compared to traditional metrics based on phenotypic units, this type of analysis provides an alternative way of quantifying the pace of adaptation. Re-scaling of phenotypic space by mutational distance is straightforward when the genotype-phenotype map is 1:1, as it is for the simple network motifs examined here.

All network interactions that we examined stretched phenotypic distance overall relative to the null model of no interaction. Phenotypic space was especially stretched in directions of low mutational variance (i.e. axes of **M** with small eigenvalues). These axes of low mutational variance correspond to directions in which phenotypic change is relatively small given some amount of mutational input, i.e. axes of mutational robustness. Thus network motifs differ from each other in mutational robustness, but motifs also induce differences in mutational robustness both across phenotypic space and along different axes of phenotypic change from a single initial phenotype. Thus the concept of mutational robustness, like genetic variation (Walsh and Blows, 2009), requires a multivariate view to provide explanatory power for phenotypic evolution.

Fourth, in our model the availability of mutational variation in the direction of selection constrains the speed of adaptation toward a selective optimum, curves the trajectory of adaptation toward the optimum, and shifts the position of the population mean under migration-selection balance. The effect of genetic (co)variance, and by extension mutational (co)variance, on these aspects of adaptation has been previously established (Jones et al., 2003, 2007, 2012). What is new in the current results is the variation in this effect of **M** on adaptation across phenotypic space. While the process of adapting toward a selective optimum can shift the pattern of **G** given constant **M** (Jones et al., 2004), our network model shows that the process of evolving through phenotypic space can also shift **G** because the population experiences different M-matrices. In addition, curvature in trajectories of adaptation caused by mis-alignment of **G** and directional selection is the result of the orientation of **M** across phenotypic space.

Fifth, some theoretical work has predicted that the major axis of the G-matrix in populations experiencing gene flow should align with direction of divergence between them, but this alignment depends on a balance with selection and migration rate (Guillaume and Whitlock, 2007). Our results are consistent with this prediction, with the addition of network-induced changes in **M** across phenotypic space shifting the resulting orientation of **G** as well. It is worth noting that under weak selection, population 1 shows slightly higher rates of adaptation at intermediate migration rates, compared to either higher or lower migration rates. This may be an instance of adaptive introgression; i.e. a low level of migration supplying genetic variation along the axis of divergence between populations, which facilitates the response to selection (Guillaume and Whitlock, 2007; Arnold and Martin, 2009; Abbott et al., 2013). Accordingly, the dimensionality of **G** is highest at intermediate migration rates in this case (Table 2.1). More generally, attention has focused on the question of the stability of **G** over time and among related taxa. Empirically, **G** is observed to change over short time-scales (Björklund et al., 2013), but also retain some aspects of its structure over longer time-scales and among populations (Arnold et al., 2008). Drift, selection, and migration are factors that can de-stabilize **G**, and now we can add network-induced shifts in **M** across phenotypic space to this list.

2.5.2 *Extension of simple network motif models*

The models above are most simply described in terms of two loci that regulate each others' expression level under Michaelis-Menten-like kinetics. However, these network motifs are general enough to apply to pairs of loci with multiple types of gene regulation (reviewed by Gjuvsland et al. (2007a)), and also to two well-defined, interacting modules in a larger regulatory network.

As larger regulatory networks are being empirically mapped, it is possible to abstract features of these networks corresponding to such higher-level motif architecture, and to map these aspects of network architecture to phenotype (Tøndel et al., 2011). This extraction of larger-scale network motifs may suggest general features of mutational and genetic (co)variance that emerge from genetic regulatory networks and that could impact adaptation. It remains to be seen to what extent more complex networks can be approximated by much simpler network models in terms of their influence on mutation and genetic variation, or what degree of network modularity is required for this approximation. The general modeling approach taken here could also be directly extended to larger motifs, using more numerical methods in order to catalog the effects of network architecture on mutational variation and evolutionary constraint.

Traditional quantitative genetics theory deals with epistasis as a source of genetic variation, which is more limited than additive genetic variance in its ability to contribute to adaptive variation (Lande, 1979; Lynch and Walsh, 1998). However, combining epistatic interactions into a single term obscures the wide range of functionally different forms of epistasis. As we found here, detecting little or no epistatic variance using variance decomposition methods may mask relatively strong functional epistatic interactions at the level of gene regulation (Stadler, 2000; Phillips, 2008). Despite the lack of epistatic (co)variance within populations, we showed that functional epistasis can still have an impact on adaptation rates and trajectories. Integrating a regulatory network view into the study of epistatic variance would help to link quantitative genetic theory and models of phenotypic evolution to the emerging wealth of data from systems biology (Gjuvsland et al., 2007a).

As described above, the genotype-phenotype map in this simple model is 1:1. The actual genotype-phenotype map for nearly any quantitative trait is certainly more complex, including dynamic developmental pathways and interactions with environmental inputs, to the extent that some suggest it may not be helpful to consider it as a "map" at all (Pigliucci, 2010; Travisano and Shaw, 2013). Even simple network architecture can limit the ability of quantitative trait locus (QTL) mapping, based on standard assumptions about the distribution of genetic variation, to detect loci underlying a trait (Gjuvsland et al., 2007a). One approach around this issue is to include network parameters directly in the mapping analysis (Wang et al., 2012). On the other hand, it may be that in the case of large genetic regulatory networks with allelic variation at multiple loci, epistatic interactions average out and locus effects are largely additive, so that new approaches to association mapping can indeed account for much of the observed heritability (Allen et al., 2010; Yang et al., 2010).

Given its complexity, one may ask whether the concept of a genotype-phenotype map is obsolete. We argue that it is not. Factors like network motif architecture, developmental processes, and genotype-by-environment interaction certainly add layers of non-linear complexity in the genotype-phenotype relationship. But in both functional studies and predictive models of evolution, approaches can be used to partition these layers. At the phenotypic end, genotype-by-environment interaction can be partitioned out by considering the "phenotype" to be a functional response to environmental inputs – a set of function-valued traits (Kingsolver et al., 2001). Network-based models can also explicitly incorporate phenotypic plasticity into the genotype-phenotype map (Draghi and Whitlock, 2012). At the genotypic end, it may be possible to explain a large portion of the effect of network architecture on relevant evolutionary features, such as **M** and **G**, simply by summarizing complex networks as their canonical motif structure (Tøndel et al., 2011). Explicit models of developmental pathways can also help to focus on particular layers of the genotype-phenotype map (Mitteroecker, 2009; Félix, 2012). These relationships are clearly difficult to unravel, but rapid advances in technology allowing high-throughput empirical measurement at multiple levels (e.g. genomic sequence, genetic and metabolic network architecture), as well as the promise of high-throughput methods at the organismal phenotype level (Houle et al., 2010), may facilitate progress in revealing these connections between genotype, phenotype, and evolutionary trajectories.

2.6 CONCLUDING REMARKS

Our models indicate that the architecture of simple network motifs can potentially have a strong impact on adaptation. Network interactions lead to mutational covariance among traits, and this covariance varies across phenotypic space. Moreover, despite strong patterns of both functional and statistical epistasis, the mutational covariance takes the form of additive genetic variation, so it has a direct impact on the response to selection. The effects of epistasis are observed in changing the covariance structure of mutational and genetic variation as populations adapt toward novel phenotypes. As a result, several evolutionary properties – additive genetic (co)variance (the G-matrix), the rate of adaptation toward a selective optimum, and the trajectory of adaptation – are all essentially stretched and curved across phenotypic space.

CHAPTER 3

Directional selection on a simple genetic network leads to stochastic adaptation, overdominance, and reproductive isolation[2]

3.1 SUMMARY

Evolutionary biology has historically approached the genetics of adaptation from two perspectives: (i) the genetic level, where the focus is on population dynamics and functional roles of single genes, and (ii) the phenotypic level, where quantitative genetics provides a theoretical base and empirical framework. The connection between these perspectives lies in the interaction network among genes that affect a phenotype, but the scale of empirical networks has been a barrier to understanding. Here we start to address two fundamental questions at this interface: 1) How does network architecture affect the ability of complex phenotypes to evolve? and 2) How does network architecture determine the repeatability of evolution? In this study we expand upon previous models of gene regulatory network that connect motif architecture to metrics of phenotypic variation based on classical quantitative genetics theory. Using simulation modeling, we evolved populations to new multivariate phenotypic optima given two classes of mutations: those in the consituitive allelic expressions of coding genes and those in the upstream, *cis*-regulatory region. Mutations in the latter class effectively redraw the genotype-to-phenotype map and so are expected to generate large jumps in phenotypic space. We confirmed that these large regulatory network changes are beneficial early in an adaptive walk but become deleterious when a population is at its optimum and evolving via stabilizing selection alone. We also found that selection can favor "heterozygotes" in the network architecture under certain conditions. This overdominance can persist for millions of generations during which time the population may become reproductively isolated from populations evolving in parallel while exploring their holey adaptive landscape.

[2] Manuscript in preparation for submission to American Naturalist

3.2 INTRODUCTION

The analysis of continuous trait variation has classically fallen into the realm of quantitative genetics. While useful, early assumptions in quantitative genetics has simplified the genetic architecture of complex traits. Namely, there are many alleles affecting trait variation, each with small, additive effects. But identifying the causal loci responsible for phenotypic diversity has been difficult under this paradigm, as evidenced by GWAS studies (for review, see Mckinney et al., 2012). Indeed, genomes are more than just a collection of genetic material. They are highly complex and interactive systems, with expression of genes dependent upon the expression of other genes (Berg and Lässig, 2004).

The interaction between genes, as well as each of the constituent genes' additive effects on phenotype, can be captured under a network theory paradigm (Mckinney et al., 2012). A gene regulatory network (GRN) contains regulatory and signaling genes and the DNA sequences that control their expression (Erwin and Davidson, 2009). These are directed graphs. For example, a transcription factor protein created from one gene transcriptionally regulates a downstream gene (Milo et al., 2002; Babu et al., 2004). The edge connecting two genes (i.e., nodes) therefore gives a graphical representation of the epistatic interactions between them as well as the molecular underpinnings of pleiotropy (Hecker et al., 2009; Phillips, 2008). With GRNs, pleiotropy and epistasis, both of which widely occur in nature, can be explictly modeled to better predict how populations might respond to various selection pressures.

Even simple GRNs, however, can result in complex evolutionary outcomes. For example, Hether and Hohenlohe (2014) recently investigated how different types of two-gene GRNs could create curvature in the genotype-phenotype relationship. Specifically, they modeled 6 different two-gene interaction networks that range from no interaction between genes to negative feedback loops. Nodes in the networks and their interactions affected the continuous expression level of the two genes. They found that network structure predictably influenced the direction of mutational (co)variation and so produced a complex, curved genotype-phenotype relationship. Moreover, when adaptation was mutation-limited, adaptation to a new optimum was fastest when the direction of pleiotropy (i.e., the major axis of mutational covariation) was in line with the direction of directional selection (β).

For the above simulations, Hether and Hohenlohe (2014) disallowed any mutations in the actual *cis*-regulatory modules and hence network architecture was fixed. However, it is reasonable to assume that actual regulatory network structure can evolve over evolutionary time (Erwin and Davidson, 2009). While mutations in the allelic values (i.e., nodes) of a network give incremental and continuous changes in trait values (Hether and Hohenlohe, 2014), mutations in the actual network (i.e., edges) may create evolutionary "leaps" in phenotypic space. This prediction has largely been unexplored empirically but predictions can be gained using Fisher's geometric model (Fisher, 1930; Orr, 2005). Briefly, when populations are adapting to a new, displaced multidimensional phenotypic optimum, selection favors small-effect mutations more often than large effect mutations due to antagonistic pleiotropy. Under an adaptive walk to a fixed optimum the distribution of mutational effect sizes might follow an exponential distribution (Orr, 1998, 2006). Under this scenario, large-effect mutations can quickly move a population closer to its optimum and are so selectively favored early in the adaptive walk. However, these large-effect mutations increasingly come at a cost as the population hones in on its optimum and this cost can come in two forms. First, as mentioned above, mutations in genes can have pleiotropic effects that are antagonistic. Second, assuming a diploid case with no dominance in the network regulation, homozygotes genotypes of the derived regulatory mutation may overshoot the optimum whereas only a single copy of the mutation (i.e., the heterozgyote) was the most fit. When allowing regulatory mutations to co-occur with allelic mutations we would predict that mutations in the former are selectively favored early in the adaptive walk while the contribution of small-effect allelic mutations to be favored over longer timescales.

While Fisher's model is useful in generating predictions of mutational effects fixed during adaptation, it ignores epistatic interactions. Since evolving GRNs are inherently non-additive, adaptation to a new phenotypic optimum might occur along a "rugged" fitness landscape (Wright, 1932). During adaptation in a rugged landscape, the local fitness peak that a given population would reach would be contingent upon past random mutations that fixed early in the its history. Such historical contingencies have been documented in *Escherichia coli* laboratory long term evolution experiments (Blount et al., 2008) and in diversification of three-spined stickleback populations (Taylor and McPhail, 2000).

There are several ways in which independent populations can evolve post-zygotic reproductive isolation. For example, in the Dobzhansky-Muller incompatibility (DMI) model, alternative alleles can fix in 2 or more genes between allopatric populations (Bank et al., 2012; Lowry et al., 2008; Fierst and Hansen, 2010). During a bout of secondary contact between these parental populations hybrids may have reduced fitness since hybrid genotypes have not been tested by natural selection (Dobzhansky and Dobzhansky, 1937; Muller, 1942). Empirical examples of DMIs have been found in some model systems including *Mimulus* (Fishman and Willis, 2001), *Saccharomyces* (Johnson, 2009), and others (Presgraves, 2010).

The main limitation of the DMI model, however, is its dependence of fixation of alleles in each parental population. Indeed, Unckless and Orr (2009) showed that when 2 populations evolve to a similar phenotypic optimum, selection can favor identical alleles in each population, precluding the formation of DMIs. Nevertheless, alternative models of epistatic interactions show that reproductive isolation can readily form without the need for alternative alleles to fix in different populations (Wagner et al., 1994). In one example, Johnson and Porter (2000) showed how reproductive isolation can form as a by-product of independently evolving populations experiencing directional selection to identical optima. Specifically, they modeled quantitative changes in regulatory pathway binding efficacy – not in regulation architecture itself – and found hybrid incompatibility formed under a range of tested parameters when gene regulation contributed to trait values.

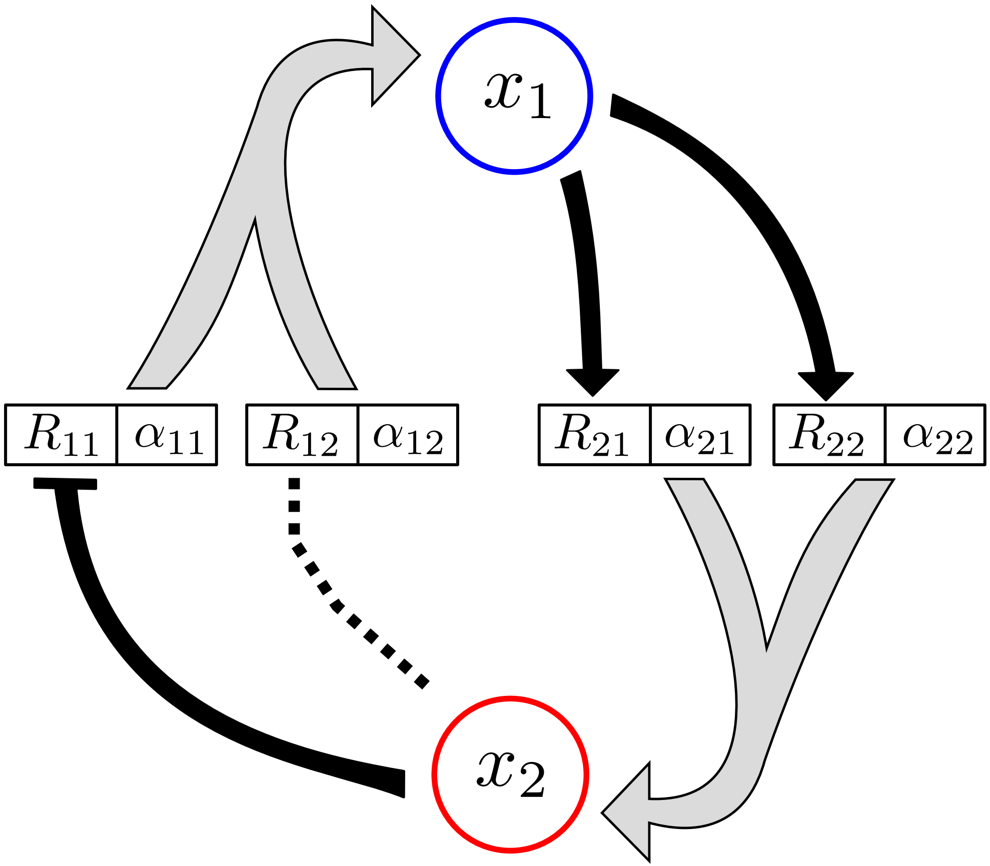
Here we build upon networks models of Gjuvsland et al. (2007b) and Hether and Hohenlohe (2014) to investigate the consequences of phenotypic adaptation when the underlying genetic architecture consists of labile regulatory regions. We first examined the effects of jointly varying allelic and regulatory mutational rates on the adaptive trajectories to a new, multivariate optimum. Second, we asked how the distance to a new phenotypic optimum affects network evolution and reproductive isolation between pairs of population. We found a surprising result over a broad range of parameters where heterozyote network architecture was selectively favored and we further explore the stability of such overdominance in light of holey adaptive landscapes. Third, we identified how the strength of stabilizing selection can maintain reproductive isolation over extended timescales (e.g., on the order of millions of generations).

3.3 METHODS

We used a two-part approach to investigate the evolutionary consequences to evolving networks. First, we modeled quantitative traits by adapting previously described analytical models of gene regulation (Gjuvsland et al., 2007b; Hether and Hohenlohe, 2014). These models translate a given individual's multi-locus genotype to a pair of phenotypes, modulated by its particular genetic architecture. Second we used stochastic simulations to quantify the different adaptive trajectories taken for replicate populations evolving to a new bivariate fitness optimum. Below we provide details of each of these models.

3.3.1 *The network model*

The traits of interest are equilibrium expression rates of proteins. The proteins themselves may interact with one another as they can act like transcription factors. For example, in Figure 3.1 protein x_1, which is transcribed from alleles at gene 1, is a transcription factor that activates the regulation both alleles of gene 2, which in turn make protein x_2. Note that the expression rates of proteins x_1 and x_2 are also influenced by the nodal alleles for each gene (Figure 3.1). For simplicity we ignore environmental variance and so the bivariate phenotypic value only depends on the regulatory architecture of each gene and the allelic values at the nodes. We consider diploid organisms here and the allelic contributions are additive in the sense that the protein expressed both alleles contribute to the final equilibrium expression rate of a given gene's expression rate. We therefore have 4 alleles (from two genes) that can affect the expression of proteins x_1 and x_2 for a given individual. There are $3^4 = 81$ possible regulatory architectures and so there are 81 different systems of nonlinear ordinary differential equations (ODEs) that describe the genotype-phenotype landscape (e.g., Figure 3.1B). Each ODE consists of the 4 allelic values, α_{ij} (the jth allele at the ith gene), the 4 regulatory alleles, and two additional parameters: θ, the concentration of protein at which half of the maximum activation level is reached and γ, the decay rate of the proteins within the cell. For simplicity we held both θ and γ fixed in the current study (Table 3.1). All ODEs produced unique, stable equilibrium values for the expression rate of proteins x_1 and x_2 when allelic values, θ, and γ are non-negative – a condition we fixed in the simulations.



$$\dot{x}_1 = \alpha_{11}\left(1 - \frac{x_2}{\theta + x_2}\right) + \alpha_{12} - \gamma x_1$$

$$\dot{x}_2 = \alpha_{21}\left(\frac{x_1}{\theta + x_1}\right) + \alpha_{22}\left(\frac{x_1}{\theta + x_1}\right) - \gamma x_2$$

FIGURE 3.1: Example network model used in the current study. In this network, circles represent phenotypes and rectangles represent the underlying genes. The phenotypic values depend (gray arrows) on the input from the *cis*-acting regulatory elements (i.e., the R_{ij} values) and their constitutive allelic values ($\alpha_{ij}s$). Single headed, dotted, and blunt ended edges represent positive (+), neutral (0), and negative (-) regulation, respectively. The specific ODE describing this network is given.

Table 3.1: Description of parameters used in the simulation model. For each parameter the "core" value is given and a as well as any range that was investigated.

Parameter	**Core Value**	**Other Values**	**Description**
n_pops	1	2	Number of populations
N	1,000	none	Number of individuals
g	10,000	up to 1e7	Number of generations
μ_C	0.0001	1e-05 - 0	Allelic mutation rate (per allele per generation)
μ_R	0.0001	1e-05 - 0	Regulatory network mutation rate (per allele per generation)
μ_V	10	none	Allelic variance of allelic mutations
$x_1^{(start)}$	1,000	none	Mean x_1 phenotype of the starting populations
$x_2^{(start)}$	1,000	none	Mean x_2 phenotype of the starting populations
$x_1^{(opt)}$	200	200-1,000	Trait x_1 optimum
$x_2^{(opt)}$	200	200-1,000	Trait x_2 optimum
start_network*	"0000"	none	Initial regulatory network architecture
s_V	1,000	10,000	Variance in stabilizing selection of the bivariate Gaussian fitness function
θ	100	none	Concentration of the regulatory at which half of the maximum activation level is reached
γ	1	none	Decay rate of protein products

*Initial regulatory network architecture for each of the four alleles $ijkl$ where i and j are the alleles at the first gene and k and l are the alleles at the second gene; -, 0, & + code for negative, neutral, and positive regulation, respectively.

3.3.2 *The evolution model*

Parameters in the simulation model and their default "core" values are given in Table 3.1 and are described in detail below. At the start of the simulation, each population was seeded with $N = 10^3$ individuals, each with no repression nor activation in the regulatory alleles (i.e., $r_{11} = r_{12} = r_{21} = r_{22} =$ '0' for all individuals). Allelic values for each individual in a population were identical and perfectly adapted to the original population with no standing genetic variation. We allowed each population to evolve under the following life history cycle: viability selection, recombination and mutation of gamete alleles.

For selection, we calculated fitness for each individual using a bivariate Gaussian fitness function (Jones et al., 2003; Hether and Hohenlohe, 2014):

$$W = e^{-\frac{1}{2}(\mathbf{z}-\mathbf{z_{opt}})^T \mathbf{\Omega}^{-1}(\mathbf{z}-\mathbf{z_{opt}})} \tag{3.1}$$

where $\mathbf{z}$ and $\mathbf{z_{opt}}$ are column vectors for the trait values x_1 and x_2 and their optima, respectively. The diagonal components of $\mathbf{\Omega}$ specify the variance in stabilizing selection around the phenotypic optimum and the off-diagonal elements specify the covariance in stabilizing selection (see Table 3.1 for defaults). Thus, there is correlational selection when the off-diagonal components are nonzero. An individual survived if their fitness value was greater than a random number drawn from a uniform distribution between 0 and 1. However, when individuals are far from the optimum, as can occur in the beginning of the a simulation run in which the new phenotypic optimum is greatly displaced from the original, all individuals may have prohibitively low absolute fitness values and this can cause populations to go extinct. We therefore invoke viability selection by selecting on relative fitness w, ($w = \frac{W}{W_{max}}$).

Mating occurs at random amongst the surviving individuals within each population until the number of offspring reaches the carrying capacity, 10^3. During mating, parental gametes form by taking into account recombination at rate r and allowing mutation of both allelic and regulatory alleles at rates μ_C and μ_R, respectively. The regulatory region of an allele is completely linked to its coding region.

3.3.3 *Calculating hybrid incompatibility*

Periodically throughout the simulation we perform a “side experiment” where we assess hybrid fitness between a pairs of replicate populations. The hybrids do not introgress in the main simulation but are used only to assess hybrid incompatibility. Hybrids are formed by first randomly mating parents from different populations of origin with one another and then by recombining and mutating gametes, using the same procedure described above. Absolute fitness of the diploid hybrids is calculated from Equation 3.1. In cases where we looked at F_2 hybrids, we randomly mated individuals of the F_1 generation using the same procedure above. Hybrid incompatibility, I, is calculated by modification to equation 6 of Palmer and Feldman (2009):

$$I = \left\{ \begin{array}{ll} 1 - \frac{\bar{W}_{hybrids}}{\bar{W}_{parents}}, & \text{for } \bar{W}_{hybrids} \leq \bar{W}_{parents} \\ \frac{\bar{W}_{parents}}{\bar{W}_{hybrids}} - 1, & \text{for } \bar{W}_{parents} < \bar{W}_{hybrids} \end{array} \right\} \tag{3.2}$$

Thus, our metric of reproductive isolation, I, ranges from -1 (hybrids much fitter than parents) to 1 (parents much fitter than hybrids).

3.3.4 *Three simulation scenarios*

We investigate 3 specific evolutionary scenarios. In the first scenario, we evolved replicate populations from point $x_1 = 1000,\ x_2 = 1000$ in phenotypic space to point $x_1 = 200,\ x_2 = 200$ and we varied the rates of both types of mutations: regulatory and allelic. We were specifically interested in how these two rates jointly affect adaptive trajectories, defined here as the population mean Euclidean distance from the optimum in phenotypic space. Second, we investigated a particular pair of mutation rates and asked if the adaptive trajectories were sensitive to the location of the new phenotypic optimum. Here we were also concerned with the extent of reproductive isolation, if any, that occurred and how distance to a new optimum affected the likelihood of hybrid incompatibility. Third, we investigated the long-term evolutionary dynamics of adaptation when the phenotypic optimum was an intermediate distance away (at point $x_1 = 600$, $x_2 = 600$) and we kept the mutation rates the same as in scenario 2. The software developed for this simulation is freely available at https://github.com/tylerhether/NetworkEvolution.

3.4 RESULTS

3.4.1 *Phenotypic trajectories taken during adaptation to a distant optimum*

To evaluate the phenotypic trajectories taken during adaptation to a distant optimum we simulated populations under a variety of mutational parameters. Overall we found that the rate of adaptation to a distant peak depended on the relative contributions of regulatory and allelic mutations (Figure 3.2). As expected, without regulatory mutations the rate of adaptation to a new optimum was gradual and populations approached the optimum more quickly as the allelic mutation rate increased (Figure 3.2). With regulatory mutations allowed, adaptation was characterized by large jumps in phenotypic space and that these jumps occurred early during the adaptive walk. Under this scenario, all replicate populations stochastically settled on network architectures that differed from the unconstrained, starting architecture (i.e., "0000"). Following these early large jumps allelic mutations continued to increase population mean fitness at longer time scales by incrementally adjusting the equilibrium expression levels (i.e., selecting for favorable mutations in the α_{ij}s).

When regulatory elements evolved the rate of adaptation at longer timescales was contingent upon the dominant regulatory network that evolved earlier in the population's history. For example, replicates that jumped closest to the optimum in the case of $\mu_C = 10^{-4}$ and $\mu_R = 10^{-4}$ (middle panel of Figure 3.2) did so evolving a negative, double dependency network (i.e., "- - - -", Figure 3.3A) which quickly moved the population close to its optimum. However, there was a noticeable slow down in adaptation for these replicates relative to other network architectures (Figure 3.3B-D).

3.4.2 *Adaptation to other optima*

We investigated how the distance to a new phenotypic optimum altered adaptive trajectories and reproductive isolation. When the optimum was unchanged network architecture did not evolve (upper right panel of Figure 3.4). In all other optima considered, however, changes in the network architecture were favored. In general, selecting on a single trait resulted in replicates evolving one or two different network types but selecting on both traits yielded greater variability across populations. For both short and distant optima, the homozygote derived allele in the network

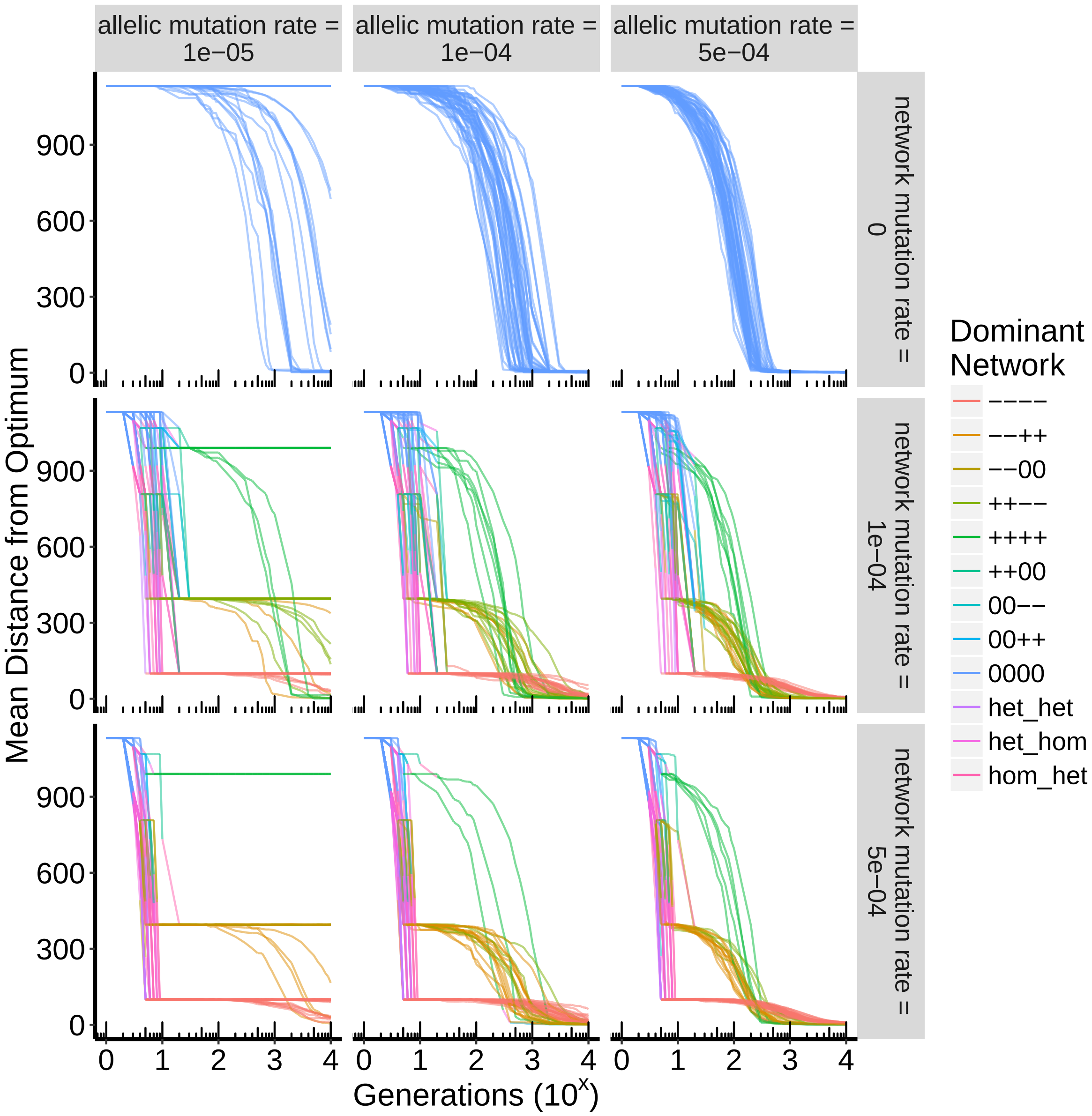


FIGURE 3.2: Summary of adaptive trajectories taken towards optimum $x_1 = 200,\ x_2 = 200$. For each level of allelic mutation rate (columns) and regulatory mutation rate (rows) the population-level mean Euclidean distance from the optimum is mapped for 50 replicate runs. Populations were initialized at point $x_1 = 1000,\ x_2 = 1000$ and evolved towards a new optimum $x_1 = 200,\ x_2 = 200$. Colors denote the "dominant network" type (i.e., the network type that occur most frequent at a given point in time).

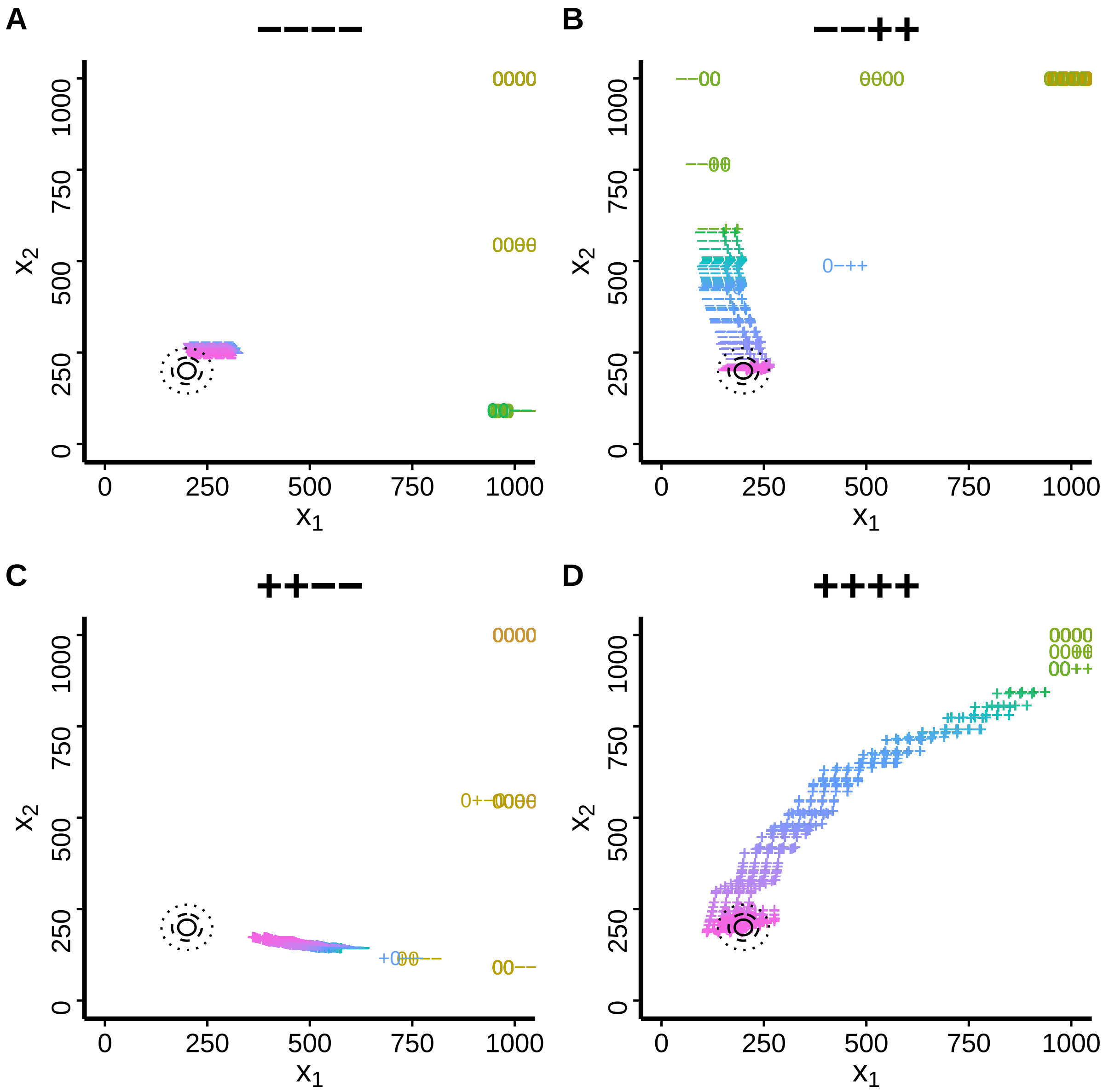


Figure 3.3: Specific examples of adaptive trajectories taken to a new, distant optimum. Shown here are four replicate populations adapting from point $x_1 = x_2 = 1000$ to point $x_1 = x_2 = 200$ with $\mu_C = \mu_R = 0.0001$ (i.e., middle panel of Figure 3.2). At each generation 30 individuals are plotted, indicated by their network architecture. Above each panel the most frequent ("dominant") network architecture at generation 1000 is given. Concentric ellipses show the strength of stabilizing selection (50, 75, & 95% of the (co)variance) and colors show the generation time (orange = generation 1; purple = generation 1000).

was favored and populations steadily adapted to their new optimum with the mean distance from the optimum approaching zero. Interestingly, when the new phenotypic optimum was an intermediate distance away from the original optimum the mean absolute fitness stalls. Such stalling occurs when the highest fit individuals have only a single copy of a derived regulatory allele at one or both loci (i.e., singly or doubly heterozygote advantage; purple shaded lines in Figure 3.4). Moreover, these intermediate distances had a greater propensity to form high reproductive isolation between replicate populations (Figure 3.5).

3.4.3 *Persistent heterozygote advantage evolves as a byproduct of adaptation*

To further investigate the heterozygote advantages seen in Figure 3.4 we evolved replicate populations to an interminable distance ($x_1^{(opt)} = x_2^{(opt)} = 600$) for 10 million generations and under varying selection strengths. We found that heterozygote network architecture persisted throughout these scenarios and adaptive trajectories remained relatively steady for the first million generations (Figure 3.6). Over longer timescales, mean distance from the optimum becomes more variable and we saw that weaker selection resulting in a higher frequency of scenarios in which the heterozygote advantage was replaced with another architecture.

In some populations the heterozygote advantages eventually collapses resulting in an overall decrease in distance from the the optimum (i.e., an increase in mean population fitness). For example, Figure 3.7 shows the allelic value composition of a specific replicate through time. In this example adaptation leads to double heterozygote advantage network architecture. Allelic values then diverge while keeping the mean absolute fitness steady (near $\bar{W}_{ABS} = 0.25$). This divergence is non-linear: small changes that result in increasing smaller neutral allelic values correspond to disproportionally larger changes in allelic values for the heterozygote negatively regulated pair. In this example, the dominant network architecture shifts near 1.5 million generations in gene 2 (Figure 3.7B) and again shortly before 2 million generations for gene 1 (Figure 3.7A), ultimately yielding a '++++' architecture.

An examination of the regulatory allele and genotype frequencies at gene 1 in the above example indicate long periods of stasis supplanted by a large shift in network shortly before 2 million generations (Figure 3.8). A closer examination of these frequency changes at the shift point shows '-/0' heterozygote advantage giving way to a '++' genotype (Figure 3.9).

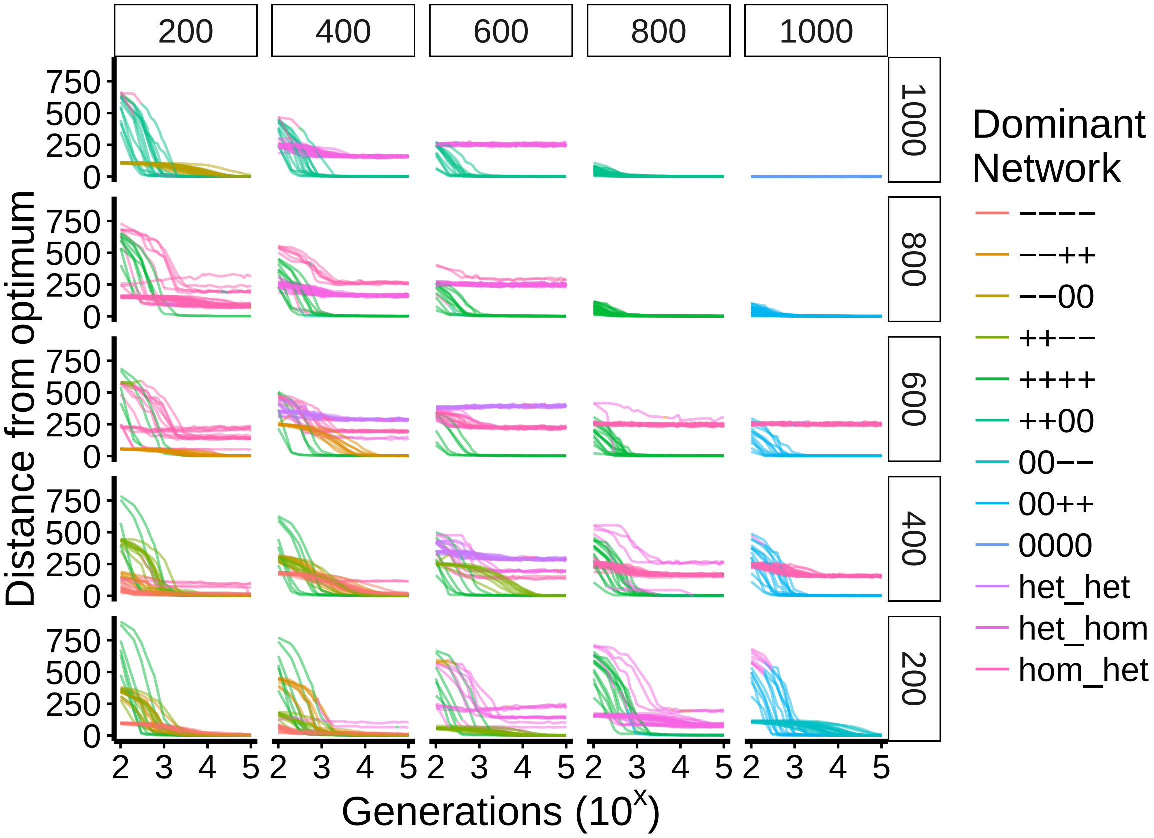


FIGURE 3.4: Adaptative trajectories toward varying optima. For all replicates, populations were initialized at point $x_1 = 1000$, $x_2 = 1000$ (upper right panel). Each sub-panel shows the adaptive trajectories to a new optimum (columns and rows show the location of the new x_1 and x_2 optimum, respectively). Colors denote the "dominant network" type (i.e., the network type that occur most frequent at a given point in time).

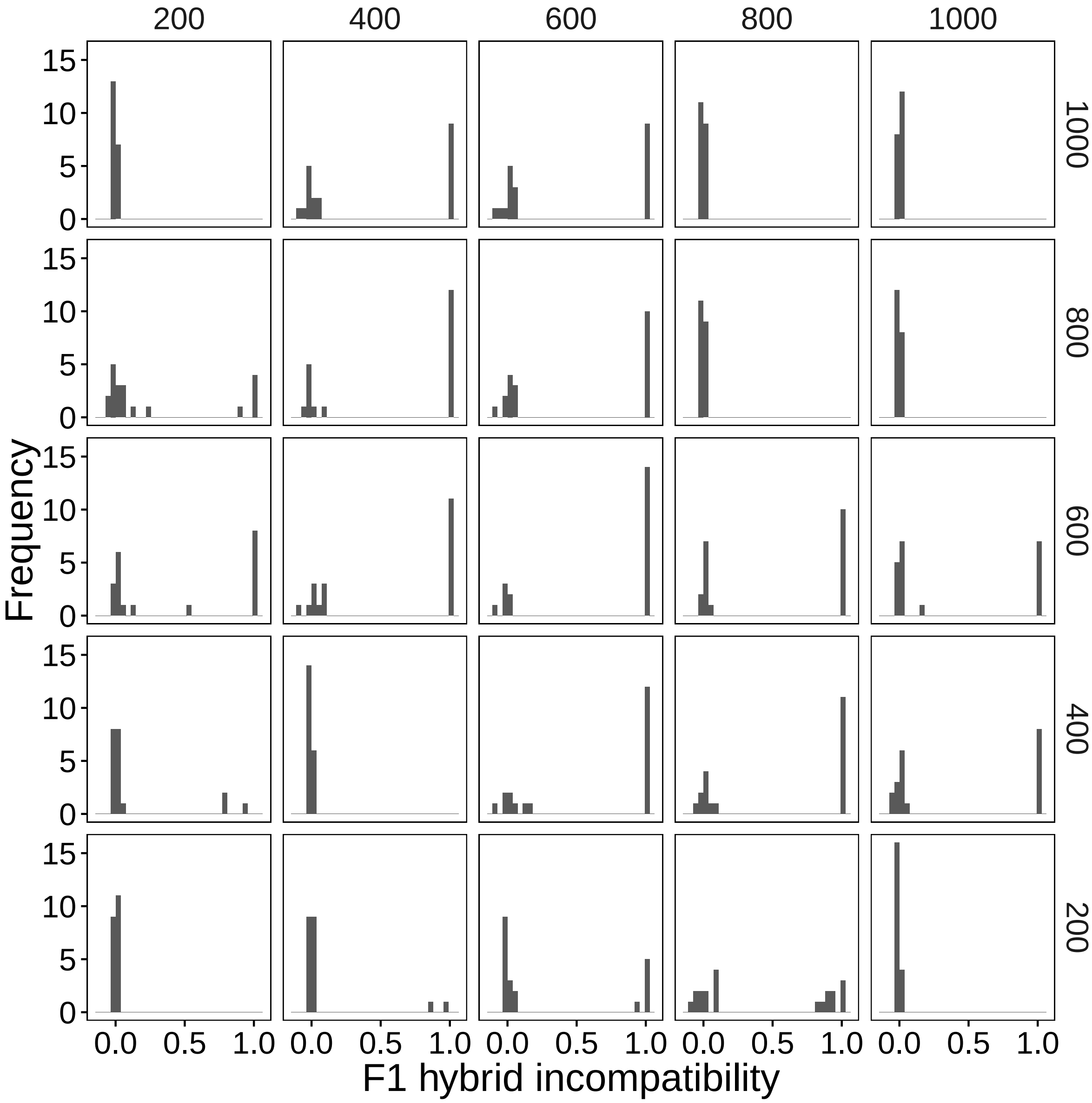


Figure 3.5: F_1 hybrid incompatibility between population pairs. At the beginning of the simulation each population was initialized at point x_1 = 1000, x_2 = 1000 (upper right panel). Each sub-panel shows frequency counts of F_1 incompatibility for the final 100,000th generation (20 replicate population pairs for each optimum). Columns and rows show the location of the new x_1 and x_2 optimum, respectively).

In other populations the mean distance from the optimum actually increases (Figure 3.6) over time. Figure 3.10 shows a specific example. Here, the negative and neutral alleles in gene 1 again diverge but in the opposite direction than in the heterozygote collapsing example above (Figure 3.7). In this reinforcement example its worth noting that the second gene also exhibited a heterozygote advantage; however, the allelic values for the second gene remained unchanged over the same time scale.

3.5 DISCUSSION

Networks are ubiquitous in nature and arise organically in all levels of biological organization. While there is some overlap in specific network interactions between any given pair of taxa, the amount of phenotypic diversity seen across the tree of life is evidence that network topology itself changes over time. For example, sequence differences in the *cis*-regulatory module of the *yellow* gene across *Drosophila* species is partially responsible for wing color pigmentation differences and posterior abdominal coloring in males (Gompel et al., 2005; Wittkopp et al., 2002; Jeong et al., 2006; Erwin and Davidson, 2009). In the present study we modeled adaptation in multivariate phenotypic space when the underlying phenotypic variation was governed by GRNs and our results highlight some of the evolutionary consequences of evolving network architecture.

Fisher's model was biologically grounded based on the observation that organisms are more or less adapted to the environment in which they occur (Orr, 2005). In that scenario, most mutations are deleterious and so large-effect mutations are not expected to be favored. In our model we have two distinct classes of mutations – regulatory and allelic – and mutational effects sizes in the former are larger than those in the latter. When the optimum is unchanged we see no change in network topology (upper right panel of Figure 3.4), consistent with Fisher's model. This model nevertheless misses some biological reality (indeed, all models, by definition, do). Mainly, it underplays the importance of epistatic interactions and so the fitness landscape can be thought of as a smooth surface (Orr, 1998, 2005). Our results instead show that when a population is displaced from its optimum, as can occur via a sudden change in the environment, short term adaptation is aided by these large effect mutations in network structure (Figure 3.2), consistent with a rugged fitness landscape.

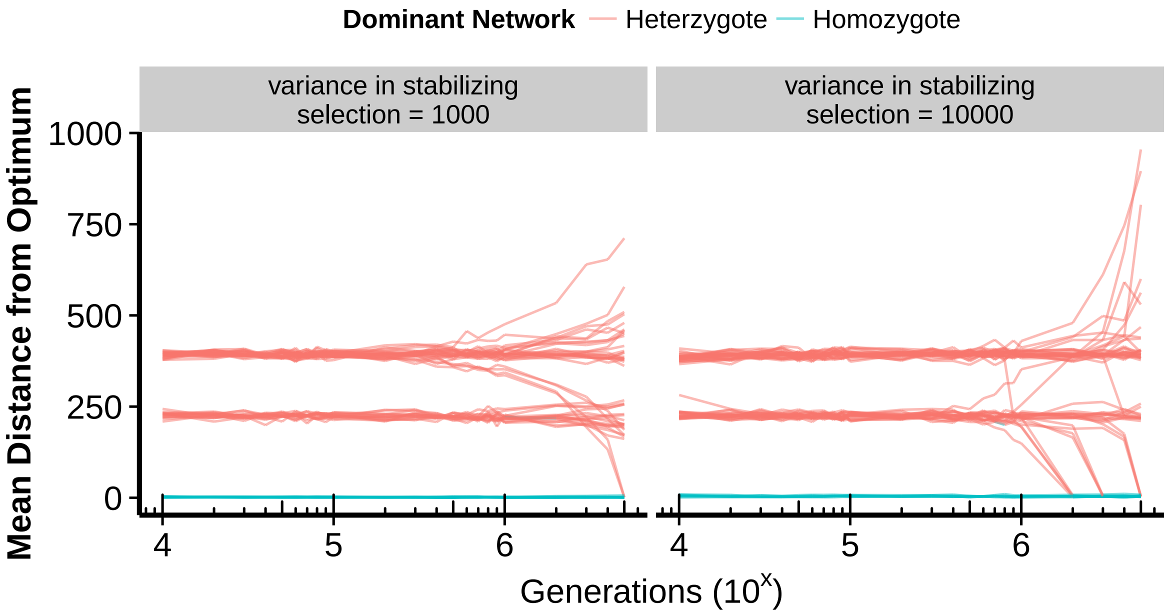


FIGURE 3.6: Long term adaptive trajectories toward an intermediate optimum ($x_1 = 600$, $x_2 = 600$) under varying selection strengths. For all replicates, populations were initialized at point ($x_1^{(start)} = 1000$, $x_2^{(start)} = 1000$). Colors denote whether the “dominant network” type was heterozygous or homozygous.

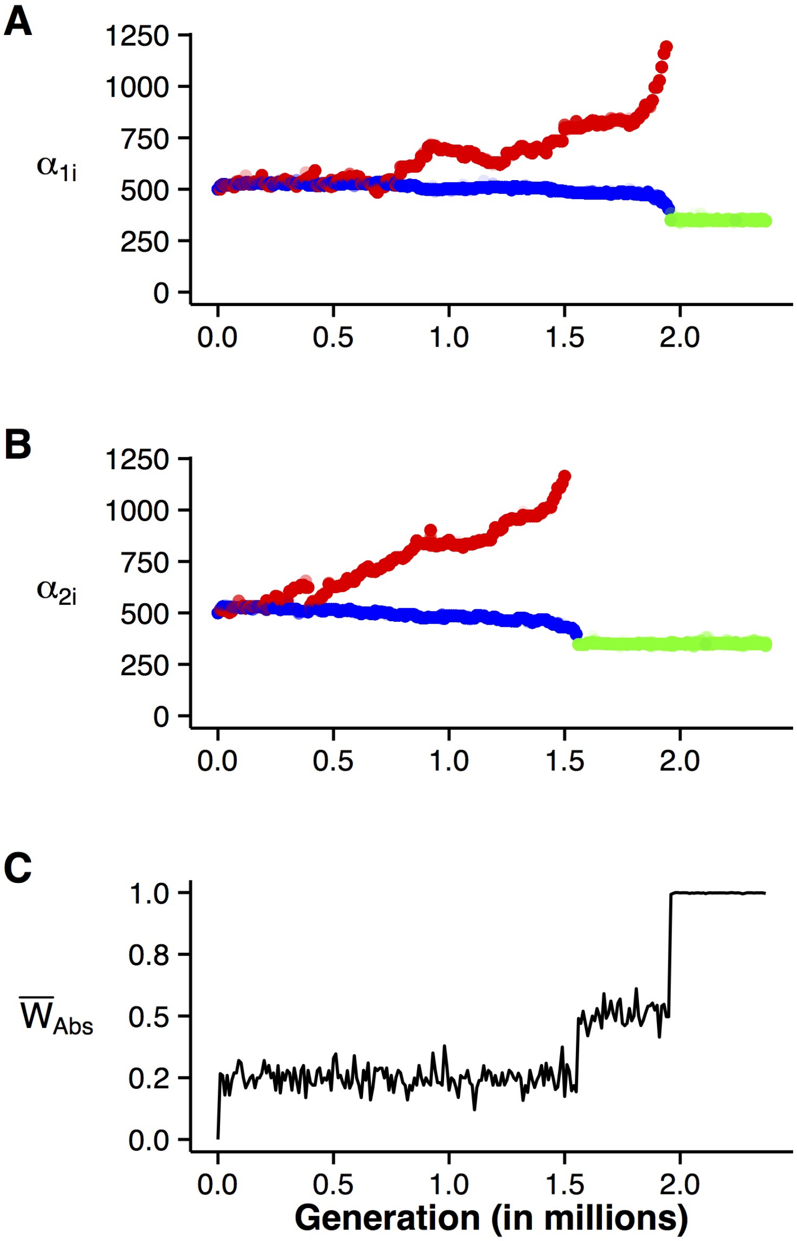


FIGURE 3.7: A specific example of a collapsing heterozygote advantage over time. A) The allelic values for node 1 for 100 random individuals at each time point are plot. Colors denote a given node's type regulation, R_{1i} (red, blue, and green show negative, neutral, and positive regulation, respectively). B) The second gene's allelic values. In both panel A and B allelic values were started at 500, which corresponds to the initial genetic and phenotypic values (see Table 3.1). C) The mean absolute fitness plotted through time.

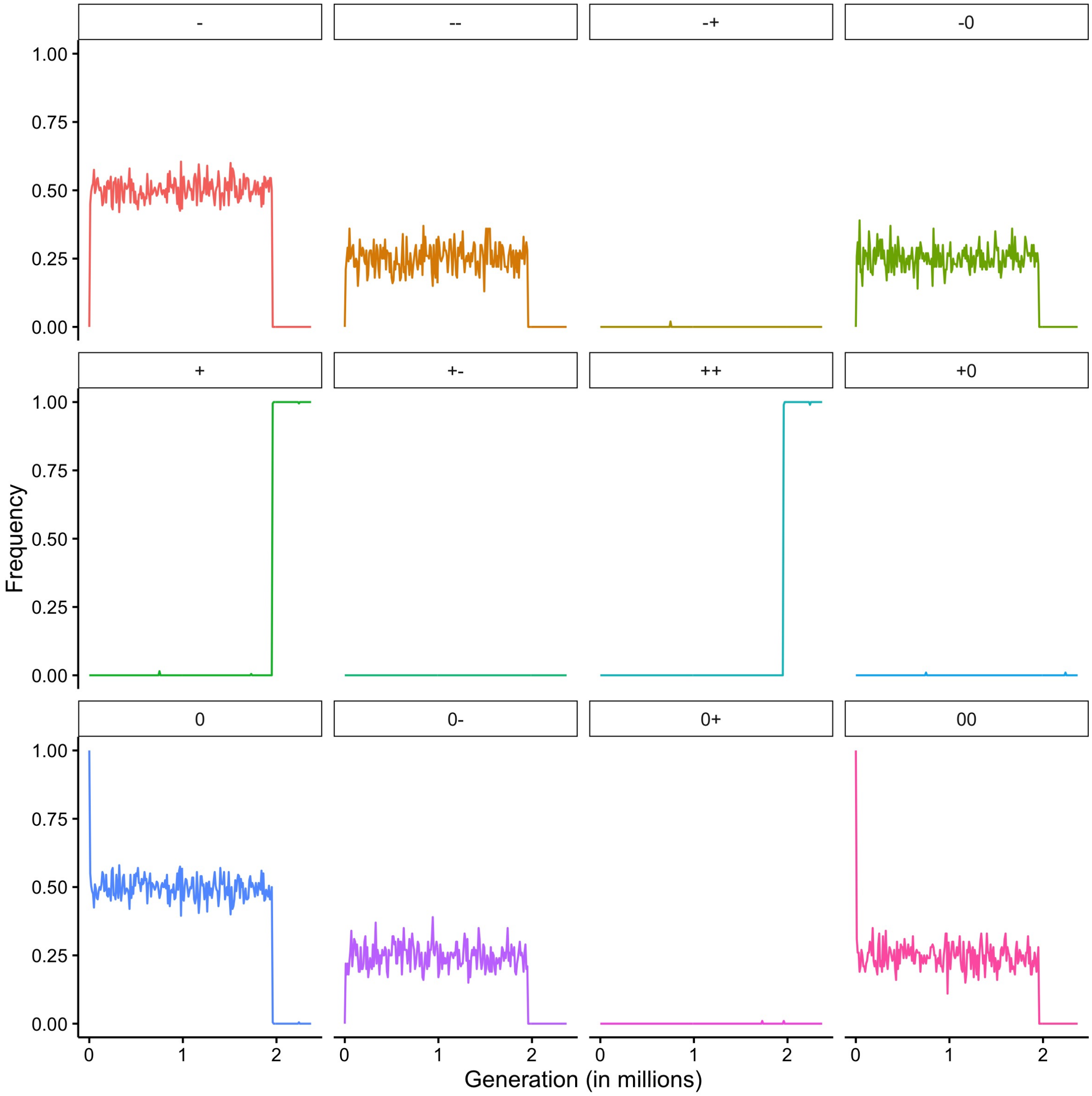


Figure 3.8: Broad scale view of the network shift seen in gene 1 of Figure 3.7. Shown here are the allele (leftmost column) and genotype frequencies (all other columns) for gene 1 for every 10,000 generations. The population was initially fixed for the neutral network.

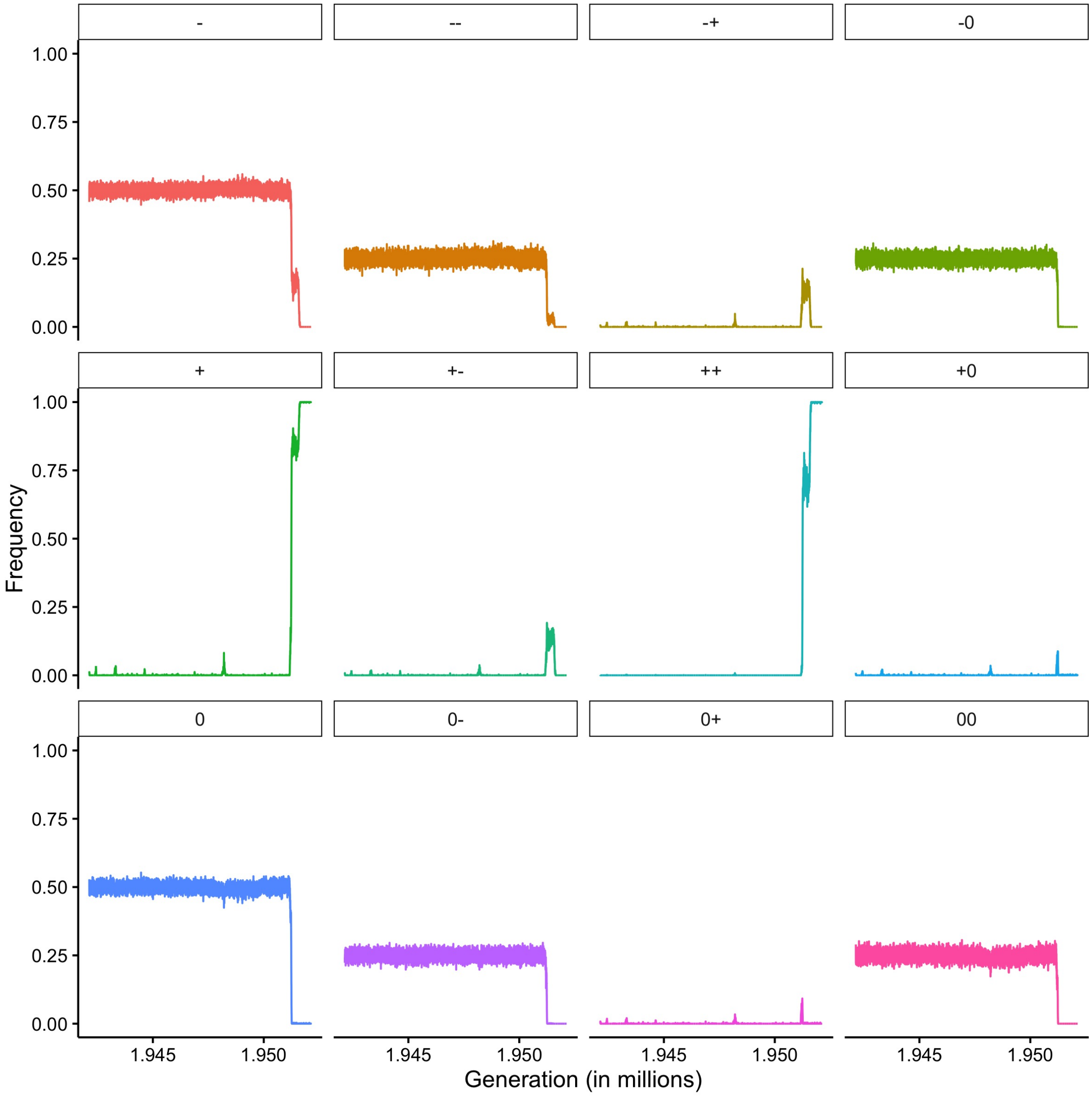


Figure 3.9: Details of the network shift seen in gene 1 of Figure 3.7. Shown here are the allele (leftmost column) and genotype frequencies (all other columns) for gene 1 for every generation, spanning the shift in dominant network type. A deterministic model showing the heterozygous equilibria for the two time points (dashed and dotted vertical lines) are presented in Figure B.1.

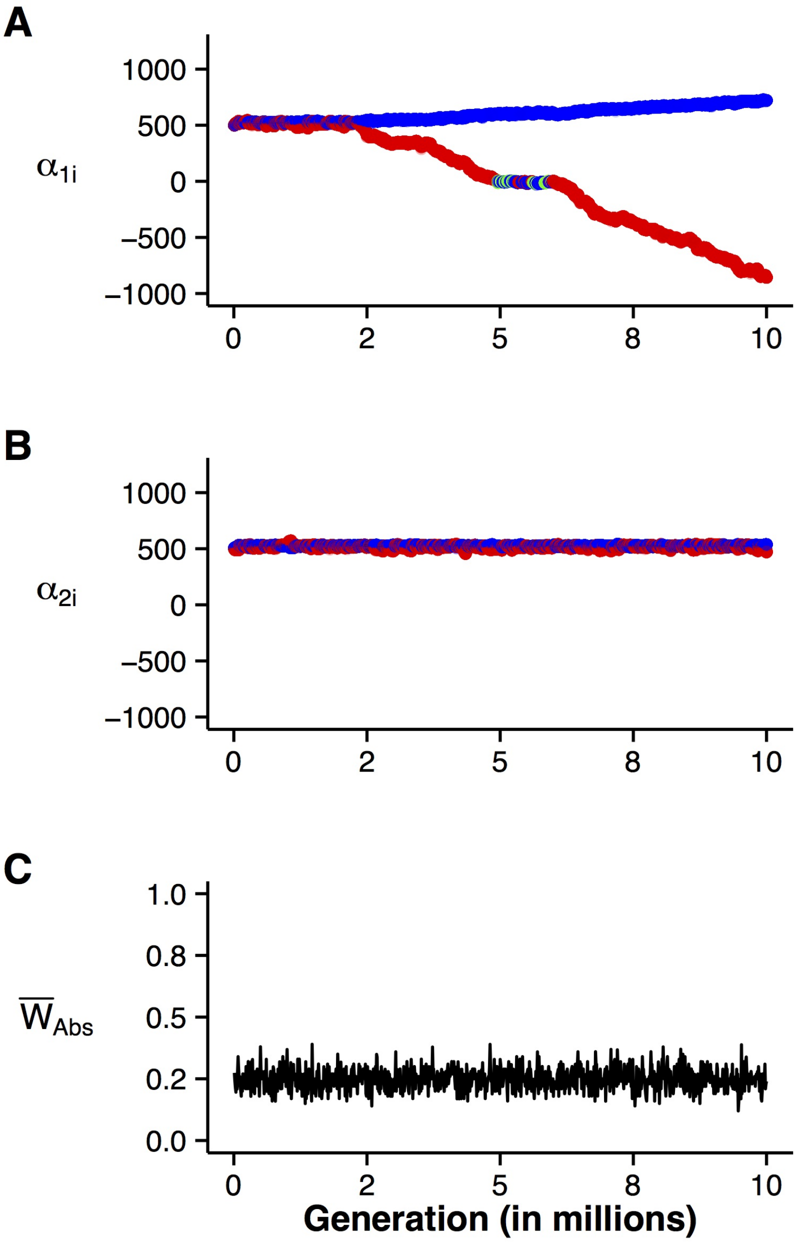


Figure 3.10: A specific example of reinforcing a heterozygote advantage over time. Colors and panel descriptions are identical as Figure 3.7. Note that this example is also a double heterozygote advantage, as indicated by both negative (red) and neutral (blue) alleles co-occurring through time at both nodes.

Here we modeled regulatory mutations on the same orders of magnitude as allelic mutations but this parametrization may seem unreasonably high. Indeed, genes that vary widely in their epistatic effects are shown to evolve more slowly than smaller effect genes in yeast (Fierst and Phillips, 2012). There are several different avenues for which *cis*-regulatory mutations can occur, including singe nucleotide mutations, indels, regulatory cassette shuffling by transposable elements, and large scale genomic rearrangements (reviewed in Erwin and Davidson, 2009). We argue that it is possible that regulatory mutations are occurring at a higher rate than realized since they are selectively favored mostly under cases of strong directional selection. When populations are well fit to their environment these mutations might be so deleterious that an organism would be nonviable early in development.

3.5.1 *Historical contingencies*

Looking across replicate populations we identified a strong pattern of historical contingencies under directional selection, which is a hallmark of a non-Fisherian, rugged landscape (Gavrilets, 2000). Specifically, during adaptation to a new distant optimum populations quickly "settled" into different network architectures. Further adaptation was facilated by allelic mutations, incrementally changing the mean phenotypic values over time. Interestingly, we observed that the longer term rate of adaptation (e.g., greater than 1000 generations) depended on which network a given population discovered along the adaptive trajectory. Relative to no network evolution (top row of Figure 3.2) some adaptive trajectories were ultimately slower even though initial adaptation was very quick. This tortoise-hare pattern makes sense in light of the direction of mutational (co)variation. Mutations from a neutral network ("0000") to a purely negative one ("- - - -") resulted in a large jump in phenotypic space, which was selectively favored early. However, this negative network results in a negative mutational covariance (Hether and Hohenlohe, 2014) which is roughly orthogonal to the direction of selection. In other words, most mutations that occur in this network are not favored. Thus, even though the regulatory mutation to a "- - - -" architecture was selectively favored it slowed down the long term rate of adaptation (e.g., Figure 3.3A). On the other hand, regulatory mutations that that resulted in a "++++" topology have a positive mutational covariation (Hether and Hohenlohe, 2014) and so more steadily adapted compared to the "- - - -" dominated populations (e.g., Figure 3.3D). It should be noted that if the direction of selection was some other pattern (e.g., β in the positive direction for both traits) the "tortoise" and "hare" roles may be reversed.

The degree of historical contingency depends on the the magnitude of directional selection experienced following the environmental change (Figure 3.4). We saw the greatest variation in distance from the optimum across replicates that evolved towards an intermediate optimum. Our GRNs incorporated input from two alleles at each locus. Therefore, a given network mutation, if favored, will produce offspring homozygote for the derived regulatory allele. In the case that the phenotypic optimum is far relative to the regulatory mutational effect size, these derived homozygotes will be favored and the new regulatory allele will sweep to high frequency in the population. However, in cases where a single copy of the large-effect mutation is favored a pattern of overdominance can form. Because there are more heterozygote genotypes than homozyogote genotypes in our GRNs it follows that greater variability across replicates should form in phenotypic regions that promote overdominance.

3.5.2 *Resolution of heterozygote advantages*

There are many examples of overdominance in nature (Allison, 1954; Hollick and Chandler, 1998; Freking et al., 2002; Gemmell and Slate, 2006) and we see this pattern form as a byproduct to adaptation in our simulations. Interestingly this pattern was not stable (Figure 3.6). The collapsing of the heterozygote advantage that we see was always associated with an increase in mean population absolute fitness. Since predicting how a homozygous optimum can be reached when there exist high fitness heterozygous intermediates is not straightforward (Wagner et al., 1994), in this section we discuss how such instability is selectively favored.

Consider a scenario such as the one in Figure 3.11A. Here a population begins far from its optimum in allelic space for a single trait. Without regulatory mutations the population can adapt through allelic mutations (e.g., top row of Figure 3.2) but this adaptation is relatively slow, at least early on in the adaptive walk. A quicker way in which the population can reach high fitness is by regulatory mutation. Such a mutation can be seen between Figure 3.11A and Figure 3.11B. This transition is another way of looking at the benefit of the heterozygote advantage. Note that in this example the allelic values did not change, which is a reasonable assumption given the short time scale at which they are selectively favored (Figure 3.2). Thus, at the level of the allelic values the fitness landscape itself changed.

The allelic values in Figure 3.11B can evolve semi-independently along the nearly neutral ridge of high fitness. While this is theoretically possible in the original genotype (Figure 3.11A) selection and Mendelian segregation ensure that alleles are identical ($\alpha_{11} = \alpha_{12}$) when the regulatory architecture is homozygous and the population is at peak fitness. In Figure 3.11B, instead mutations in α_{11} are nearly neutral so long as there exists compensatory mutations in α_{12}. We see evidence of these compensatory mutations occurring in simulation (Figure 3.7A-B, Figure 3.10A). Therefore, during the long stasis of overdominance, linked allelic values are walking the nearly neutral fitness ridge.

If the population drifts to the left along this ridge the heterozygote advantage is reinforced (e.g., Figure 3.10A). On the other hand, if a population drifts rightwards along the ridge in Figure 3.11B enough the heterozygote advantage will collapse (e.g., Figure 3.7A-B). Within a given network type the ridges are nearly neutral; however, the regulatory mutation effects on fitness need not be the same across the ridge. This is because allelic values linked to their *cis*-regulatory allele can further modify the expression rate. With enough allelic change one allele might overtake the other.

Interestingly, in our example (Figure 3.7A) the regulatory allele that eventually dominated ("+") was not one of the alleles involved in the heterozygote advantage ("0" or "-"; Figure 3.8, Figure 3.9). How is this possible? Panels B and C in Figure 3.11 have similar locations of ridges but differ by a the regulatory allele R_{12}. In this example, the "+" regulatory allele arose several times, failing to overtake until the last time (Figure 3.9). Once it did, there was a brief stable period of a second ("-/+") heterozygote advantage followed by fixation of the "+" regulatory allele. Indeed, this brief '-/+' heterozygote advantage in Figure 3.9 occurs near its expected equilibrium frequency assuming a simple, deterministic model of selection with 3 alleles (Figure B.1; see Appendix B for derivation). Thus, whereas the fitness of both homozygotes in the "-/0" scenario were low ($W_{ABS}^{(00)} \approx W_{ABS}^{(--)} \approx 0.07$) the fitness of the "++" homozygote was higher ($W_{ABS}^{(++)} \approx 0.53$). Thus, even though $W_{ABS}^{(-0)} \approx 0.99 > W_{ABS}^{(-+)} \approx 0.78$ genetic drift eventually transitioned the population to a different fitness ridge (i.e., one more akin to Figure 3.11C). Finally, allelic values α_{12} in the Figure 3.11C network intermediate evolved such that a "++" architecture was favored (i.e.,

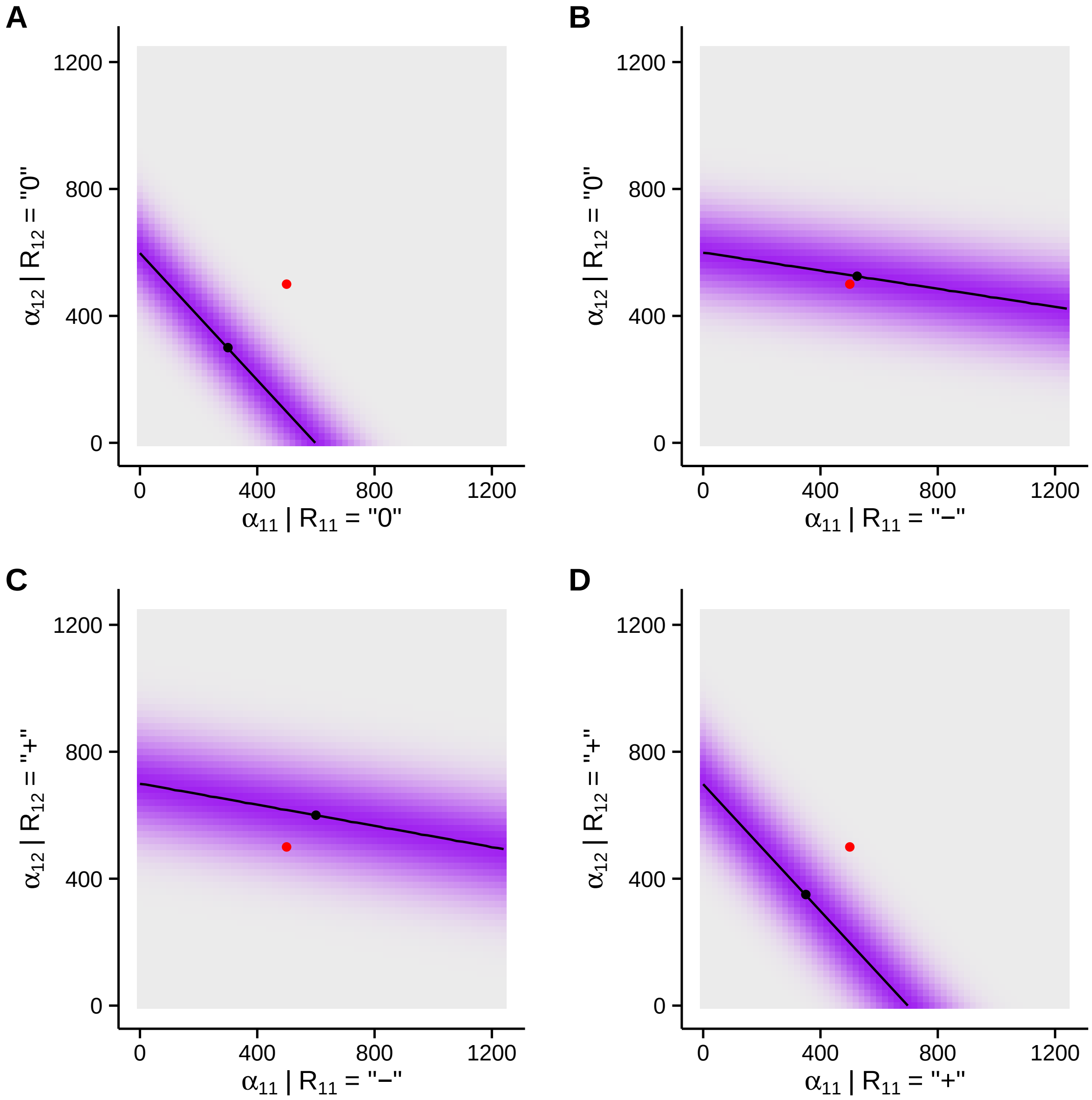


FIGURE 3.11: Fitness ridges for four types of networks used during the resolution of a specific heterozygote advantage seen in simulation. For this example, the second gene's regulatory values were fixed for two "+" alleles. In each panel, the fitness is shown for a given pair of allelic values at gene 1 given its *cis*-regulatory allele. Purple shading shows areas of high fitness. The red dot shows the location of the initial population in allelic space. The black lines show combination of α_{11} and α_{12} values that yield maximum fitness and the black dots shows were $\alpha_{11} = \alpha_{12}$ along the black line. For clarity, we set x_2 to its optimum ($x_2^{(opt)} = 600$) for each network type so that we projecting a higher dimensional space into 2 dimensions.

$W_{ABS}^{(++)}$ increased to 1). Therefore, breakdown of heterozygote advantage was due to populations exploring the nearly neutral, “holey” fitness landscape (Gavrilets and Vose, 2005). Had selection been weaker (e.g., right hand side of Figure 3.6) the nearly neutral ridges would be wider allowing for a greater degree of both ridge walking and genetic drift.

3.5.3 *Reproductive isolation*

In agreement with what Johnson and Porter (2000) found, we did not observe any appreciable F_1 or F_2 hybrid incompatibility under purely stabilizing selection (top right panels of Figure 3.5, Figures B.2, B.3). We instead found that reproductive isolation evolves under directional selection (Figure 3.5). Johnson and Porter (2000) also found that gradually directional selection facilitated reproductive isolation because different populations took different evolutionary routes towards the changing optimum. In the current study we imposed a sudden shift in phenotypic optima. This shift allowed for large effect, regulatory mutations to be favored early in adaptation which contributed to reproductive isolation. Had the optimum moved slowly in our model, these large effect mutations would not be stochastically fixed in different populations and so reproductive isolation would be unlikely.

Reproductive isolation in our model was associated with overdominance. Interestingly, we did not find evidence of reproductive isolation forming between individuals with alternative “pure” network architectures. This seemingly counter intuitive result can be clarified by looking at a specific example. Consider two individuals, *a* and *b*, with the following genotypes: $a = \alpha_{11} = \alpha_{12} = \alpha_{21} = \alpha_{22} = 350$ R = ‘++++’, $b = \alpha_{11} = \alpha_{12} = 300, \alpha_{21} = \alpha_{22} = 350$ R = ‘00++’. Both individuals and their offspring would have the same pair of trait values ($x_1 = x_2 = 600$) and so no reproductive isolation exists. Since the allelic values are identical for each gene at the level of the individual (i.e., $\alpha_{11}^{(a)} = \alpha_{12}^{(a)}$) and the second locus is fixed, only one offspring genotype is possible and the allelic values ‘balance’ out to the same phenotype as the parents. We found that such balancing occurred in the pure networks (e.g., Figure 3.11A,D). On the other hand, with overdominance the allelic values can drift apart lending to more combinations of genotypes that yield identical phenotypic values (e.g., Figure 3.11B,C). Consider two additional individuals (*c* and *d*) that have identical network types but different underlying allelic values: $c = \alpha_{11} = 120, \alpha_{21} = 680, \alpha_{21} = \alpha_{22} = 350$ R = ‘-+++’, $d = \alpha_{11} = 480, \alpha_{21} = 620, \alpha_{21} = \alpha_{22} = 350$ R = ‘-

+++'. In this example, the offspring genotypes have reduced fitness (Table 3.2) even though their parents have the same network topology. Importantly, since the overdominance we observed can be unstable, reproductive isolation is also unstable. Over time, hybrid incomptability might be reinforced via drift or selection on additional genes or it may ultimately collapse (*sensu* the ephemeral speciation model; Rosenblum et al., 2012).

3.6 CONCLUDING REMARKS

Our results highlight how relatively simple models of evolving regulatory network architecture can produce stochasticity in terms of how populations respond to selection. Directional selection preferentially favors large effect mutations early during an adaptive walk which translates to variability in the "evolutionary solutions". Finally, one of the important concepts in network theory is robustness. If a given network is well connected, removal or damage of a single node should not affect the system as a whole because other nodes in the network can compensate for the lost or defective node. In our networks, the coding regions are nodes and the regulatory regions are the edges that connect them. Perturbation of a node via random mutation can cause the multivariate phenotype to shift. However, compensatory mutations in the other node can act to balance the network such that the phenotypic values are shifted back. Over time the phenotypic means of a given population can remain unchanged but the underlying allelic values can evolve via genetic drift and reproductive isolation can transiently evolve as a by-product of adaptation.

Table 3.2: Example hybrid incompatibility between individuals with identical network architecture. Shown here are the hybrid genotypes, phenotypes, and absolute fitness between two parents (*c* and *d*) with identical GRN topology ('-+++'). For all individuals the allelic and regulatory alleles were fixed ($\alpha_{21} = \alpha_{22} = 350, R_{21} = R_{22} =$ '+'). All other parameters were set to their default values (Table 3.1).

Individual	α_{11}	α_{12}	x_1	x_2	W_{ABS}
parent *c*	120	680	600	600	1.00
parent *d*	480	620	600	600	1.00
offspring 1	120	620	547.7	591.9	0.25
offspring 2	480	680	651.7	606.9	0.26

CHAPTER 4

Novel molecular and analytical tools for efficient estimation of rates of meiotic crossovers and non-crossover gene conversions[3]

4.1 SUMMARY

Meiotic recombination plays a central role in structuring genomic diversity. Further understanding would benefit from efficient methods for directly measuring rates of recombination across the genome, including crossovers and non-crossover gene conversion events. Here we describe a Hidden Markov Model (HMM)-based approach to estimating recombination rates, based on genomic sequence data from haploid products of meiosis and diploid populations, both produced by admixture between two genetically characterized parents. We validated this approach on simulated low-coverage sequence data, and we then applied it to an admixed yeast (*Saccharomyces cerevisiae*) line produced by crossing two divergent parental strains. We used two different genomic sequencing techniques. First, we conducted low-coverage whole-genome sequencing of all four spores from single meioses, produced by sporulating diploid F_1 cells. Second, we applied RADseq, a reduced-representation genomic sequencing technique, both to spores from F_1 individuals and to diploid clones from the F_6 generation of an intercross population. Genome-wide rates of crossover (with or without associated gene conversions) and non-crossover were roughly equal, and both displayed a strongly linear relationship with chromosome length. RADseq produced just over one third as many markers as whole-genome sequencing, reducing its ability to detect small-scale non-crossovers, although the two methods performed nearly equally in mapping crossover events. However, RADseq is far more cost-efficient than whole-genome sequencing, particularly in library preparation, allowing many more samples to be genotyped for a given budget. The overall rate of recombination in the F_6 diploids was lower than in the F_1 spore dataset, likely due to strong selection maintaining parental haplotype blocks in our intercross

[3] In review as: Hether T.D., Wiench C.W., and Hohenlohe P.A. Novel molecular and analytical tools for efficient estimation of rates of meiotic crossovers and non-crossover gene conversions. BMC Genomics

line. Our genome-wide estimates for recombination rates largely agree with previous results in yeast, and our methods provide an efficient way of mapping recombination rate heterogeneity specific to any admixed line. Our simulation and analysis software is available as the R package *HMMancestry*.

4.2 INTRODUCTION

Meiotic recombination provides a crucial source of genetic variation by generating new combinations of alleles across loci. Recombination plays a structural role during meiosis by aiding and ensuring correct segregation of homologous chromosomes (Tsai et al., 2010; Anderson et al., 2011; Lichten and De Massy, 2011; Kauppi et al., 2004). Recombination also drives patterns of linkage disequilibrium and haplotype structure across the genome, determining the influence of selected loci on neighboring genetic variants and affecting the power of mapping studies to identify functional genes (Weir et al., 2005; Ott et al., 2015). As a result, much attention has been given to studying how heterogeneity in recombination rate (e.g., hotspots and coldspots) affects population genetic dynamics, phenotypic diversity, and variation in quantitative and disease traits (Price et al., 2009).

While recombination can generate new haplotypes, it can also be associated with a loss of genetic diversity via gene conversion (GC) (Szostak et al., 1983; Chen et al., 2007). Gene conversion as a result of recombination can occur in two ways. First, during a crossover (CO), where there is a reciprocal exchange of DNA between homologous chromosomes (Cole et al., 2012), GC can create small chromosomal "tracts" that lack the typical 2:2 segregation pattern in meiosis (Figure 4.1). Second, these GC tracts can occur without reciprocal exchange – known as non-crossover (NCO) (Yanowitz, 2010). Since COs and NCOs appear to be the result of different double-strand break repair pathways, there has been recent interest in characterizing their abundance, frequency, and location throughout the genome (Pâques and Haber, 1999; Qi et al., 2009; Yanowitz, 2010; Cirulli et al., 2007; Mancera et al., 2008; Anderson et al., 2011).

One approach to measure GC tracts is to sequence or genotype many loci in each of the four products of meiosis. Such an analysis is possible in yeast by sporulating diploid cells and mechanically isolating and sequencing each spore of a tetrad (Sherman, 2002; Anderson et al.,

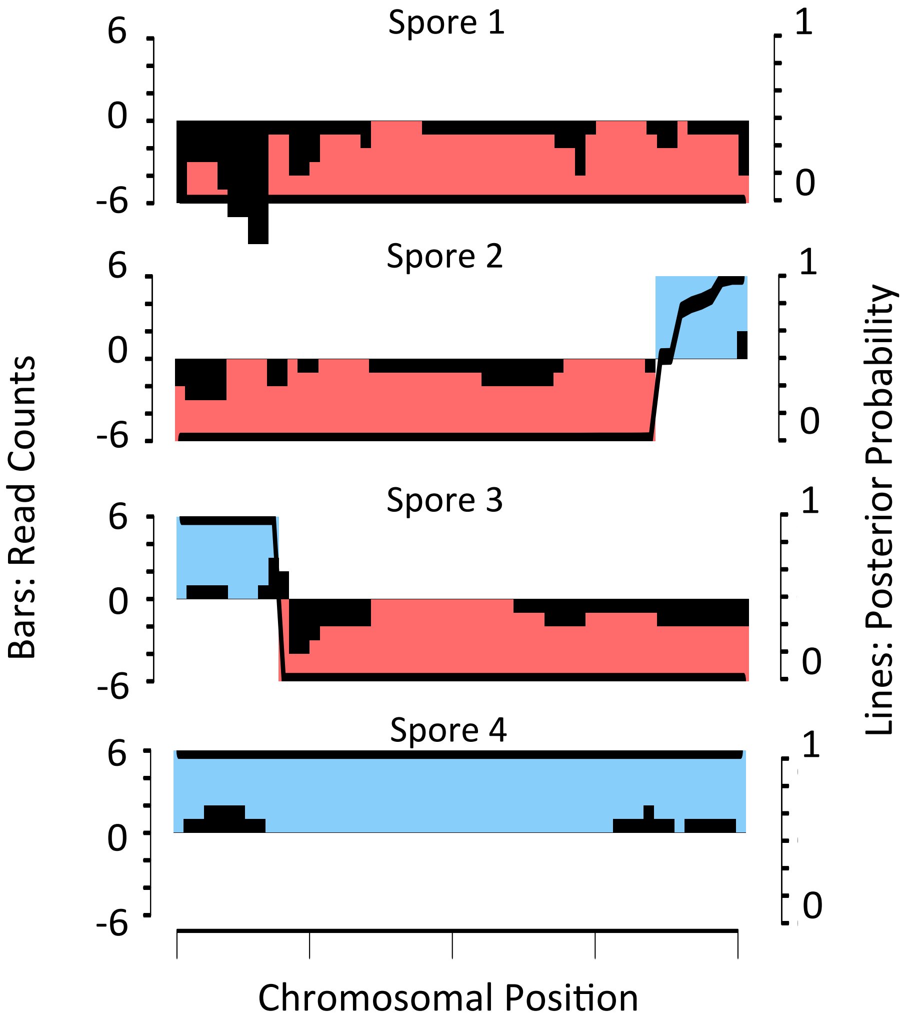


Figure 4.1: Identifying a crossover-associated gene conversion tract from low-coverage sequence data using a Hidden Markov Model. Shown here is a hypothetical chromosomal region across the four products of a single meiosis (e.g., four spores in a yeast tetrad) in an F_1 hybrid produced by crossing two genetically characterized parents. For each spore, vertical black bars (left y-axis) show the number of sequence reads that match either one (positive values) or the other (negative values) parent. Black lines (right y-axis) show the posterior probability of ancestry across genetic markers, and colors represent inferred blocks of ancestry from either the red or blue parent. Note the large gene conversion tract (3:1 red:blue ratio across the middle of the chromosomal region) that is associated with a crossover event between spores 2 and 3.

2011). Despite the recent reductions in sequencing cost, however, this tetrad dissection approach can be costly because 1) high marker density is necessary to obtain precise size estimates of GC tracts and locations of recombination hotspots and 2) libraries for four individual haploids need to be prepared and sequenced for each meiosis event. For crosses between parental genotypes that have a sufficient degree of genetic variation, a reduced representation sequencing approach (e.g., restriction-site associated DNA sequencing, or RADseq; Baird et al., 2008; Davey et al., 2011) could be much more efficient in multiplexing large numbers of haploids, while keeping sequence coverage at marker loci relatively high. A second approach is to reduce the overall sequencing coverage for each individual. The appeal of the latter approach is that one can sequence several times more individual meiosis events to better estimate genome-level recombination rates and – for haploid recombinants – more fully characterize gene conversions.

However, low-coverage sequencing introduces analytical challenges for mapping crossover, non-crossover, and gene conversion events. At any given locus, missing data among the four haploid products of meiosis make it difficult to infer the segregation pattern. In diploids, low coverage can lead to under-estimation of heterozygous genotypes. In addition, other factors can lead to incorrect ancestry assignment at marker loci: ancestral polymorphism, mutation, and sequencing error.

Recently, Hidden Markov Models (HMMs) have been used to probabilistically infer local ancestry (hidden state) along a chromosome from the observed sequencing data in the face of these challenges (for review see Liu et al., 2013). HMMs are computationally efficient and highly accurate in inferring local ancestry from sparse or error-prone data (Figure 4.1). For instance, HMM methods have successfully identified local ancestry tracts in admixed human populations (Price et al., 2009; Hu et al., 2013). One such program, SEQMIX (Hu et al., 2013), takes advantage of low-coverage off-target exome data to refine local ancestry from unlinked SNPs. This program and others focus on diploids, in which NCO and fine-scale GC events are very hard to detect. In order to obtain estimates of fine-scale GC tracts it is necessary to infer local ancestry for haploid gametes.

Here we develop and test an HMM-based inference method for identifying recombination tracts (CO with GC, CO without GC, NCO, and telomeric GC) from low-coverage sequencing of the four haploid products of meiosis in admixed individuals. We validate the method using simulated data. We apply it to two sequence datasets from haploid spores in yeast (*Saccharomyces cerevisiae*): the first produced with low-coverage, whole-genome shotgun sequencing, and the second with a novel modification of the lower-cost, reduced-representation RADseq method (Baird et al., 2008; Davey et al., 2011). We then test whether our estimates of all four types of recombination differ between these two sequencing methods and compare our maps of recombination rates with previously published estimates in yeast. Lastly, we extend our HMM method to map COs from low-coverage sequencing of diploids in an admixed population and apply it to a sample of diploid F_6 advanced intercross line (AIL) isolates. Our results give insight into the frequency and genomic distribution of recombination rates without the need for high coverage, whole genome sequencing. The HMMs, recombination simulator, and the CO/NCO inference algorithm that we describe have been implemented in the R package *HMMancestry*, which is freely available online (https://github.com/tylerhether/HMMancestry).

4.3 RESULTS

4.3.1 *Validation of Ancestry Inference Method Using Simulated Data*

We developed a Hidden Markov Model (HMM) method for inferring chromosomal ancestry and recombination events, called *HMMancestry*, and validated it against simulated low-coverage sequence data. We used the Forward-Backward algorithm (Durand et al., 2008) to assign posterior probabilities of ancestry for each single-nucleotide polymorphism (SNP) locus along a chromosome of an admixed individual. This method infers ancestry at SNP positions that were unobserved due to low sequencing coverage and is robust to missing or misleading genotypic data. We also created a maximum likelihood (ML) method for estimating two global parameters: the genome-wide recombination rate ($\hat{c}$) and the assignment probability ($\hat{p}$). Recombination rate $\hat{c}$ (cM/kb) is multiplied by the physical distance between the flanking and focal SNPs to estimate transition probability from one hidden state (ancestral haplotype) to another. The assignment probability p reflects uncertainty in the assignment of each sequence read to a parental genotype, which can result from ancestral polymorphism, mutations after the admixture event, and sequencing and mapping error.

To examine the performance of our method in accurately genotyping loci, we simulated meiosis events using a range of biologically and methodologically relevant parameters. We were specifically interested in how accuracy, defined as the squared correlation coefficient between the inferred and the true states (Hu et al., 2013), changed with genome-wide recombination rate (c), the assignment probability (p), ploidy, and mean sequencing coverage. Using *HMMancestry*, we simulated several meiosis events from known recombination and assignment estimates (c and p), inferred these parameters ($\hat{c}$ and $\hat{p}$) by ML directly from the simulated data, and inferred local ancestry at all SNP locations.

Overall, the Forward-Backward and ML estimator algorithms performed well in inferring ancestral states across the genome from simulated low-coverage data. Across a broad range of parameters, a sequencing coverage of about 1X per sample optimizes the accuracy of ancestry estimation versus the total sequencing effort (Figure 4.2). This optimum level of coverage matches that for estimating population-level allele frequencies (Buerkle and Gompert, 2013). The median squared correlation coefficient between known and inferred ancestral states for all simulated data combined was 0.997. Model performance for haploids was better than diploids, but this effect was most pronounced under extremely low coverage (e.g., 0.2X) and high recombination rate (1 cM/kb; Figure 4.2). As expected, increasing the assignment probability (i.e. reducing ancestral polymorphism and sequencing and mapping error) also increased model performance.

We also tested the performance of our method to estimate genome-wide rates of recombination and genotyping uncertainty. Our ML estimate of recombination rate, $\hat{c}$, tended to very slightly underestimate the true value under some parameters, including larger values of true c, although this deviation remained 10 orders of magnitude smaller than the true value (Figure C.1). Deviations of the ML-estimated assignment probability, $\hat{p}$, from the true value were lower at higher sequencing coverage and at higher true values of p, and our ML estimate was unbiased (Figure C.2).

4.3.2 *Algorithm to identify recombination events from haploid spore data*

The four products of meiosis in an F_1 individual are expected to have a 2:2 segregation pattern of parental chromosomes. Recombination events (CO with and without GC, and NCO GC) lead to changes in the 2:2 segregation pattern along chromosomes and tracts of non-2:2 segregation

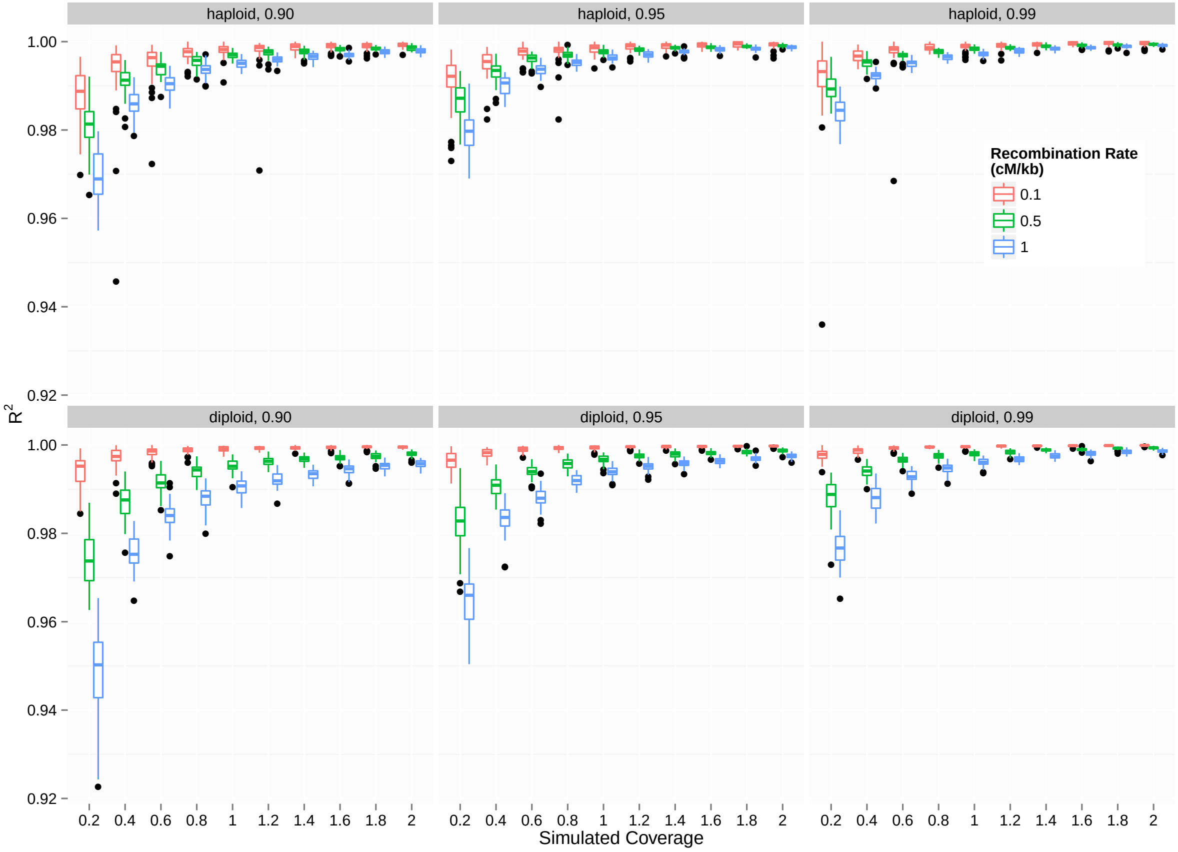


FIGURE 4.2: Performance of *HMMancestry* on simulated data. Shown are squared correlation coefficients between simulated (true) ancestry and inferred ancestry for different levels of ploidy (rows), assignment probability p (columns), recombination rate c, and sequence coverage (x-axis). Box and whisker plots show median (plus first and second quartiles) of 50 simulated replicates for each parameter combination.

(e.g., 3:1 ratio of parental haplotypes). To map these recombination events using the inferred blocks of parental ancestry in the four products of a meiosis event, we created an inference algorithm in *HMMancestry* that uses a three-step classification scheme (Figure 4.3). The first phase moves along a chromosome, identifies the regions with unique segregation patterns, and classifies the 'simple' recombination events. If the focal region is non-2:2 and is located on the end of a chromosome it is classified as a telomeric GC. If the focal region is non-2:2 and is flanked by 2:2 regions with identical or different segregation patterns, it is classified as a NCO or a CO with GC, respectively. Second, the algorithm conducts another sweep along the chromosome to resolve complex GC regions. We define complex regions as regions of non-2:2 segregation (or 2:2 regions of less than 2.5kb) that are themselves flanked by one or more GC tracts. For each of these complex tracts the algorithm identifies whether the flanking 2:2 regions have identical or different segregation patterns and reclassifies the complex tracts as either NCO or CO with GC, respectively. Third, the algorithm screens each chromosome for crossover events that lack a (detected) GC event and classifies them as CO without GC. For each inferred GC tract (CO or NCO) we estimate the size of the tract as the distance between the outermost SNP locations within the tract. For CO events without a detected GC tract, we calculate the size as the distance between the two SNP positions flanking the change from one 2:2 segregation pattern to another; in effect, this distance reflects the maximum size of a GC event that could be associated with the CO but be undetected given the scale of resolution in the marker set.

4.3.3 *Estimation of recombination rates from haploid yeast spores*

First we identified parental SNPs between haploid oak isolate (YPS128) and haploid wine isolate (DBVPG1106) strains of *S. cerevisiae* using whole-genome sequencing. We analyzed 5,674,883 PE250 reads across the two haploid parents before quality filtering, retaining 5,550,596 (97.8%) after quality filtering. We merged overlapping paired ends when applicable. We found that 93% and 87% of YPS128 and DBVPG1106 paired-end reads could be merged, as expected given our targeted insert size (400 bps). We retained 70.9X and 80.9X coverage for YPS128 and DBVPG1106 haploid strains, respectively, and identified 73,581 SNPs between the two parental strains. After removing the 2-micron and mtDNA specific loci, we retained a final count of 73,294 total SNPs that were diagnostic between the two parents.

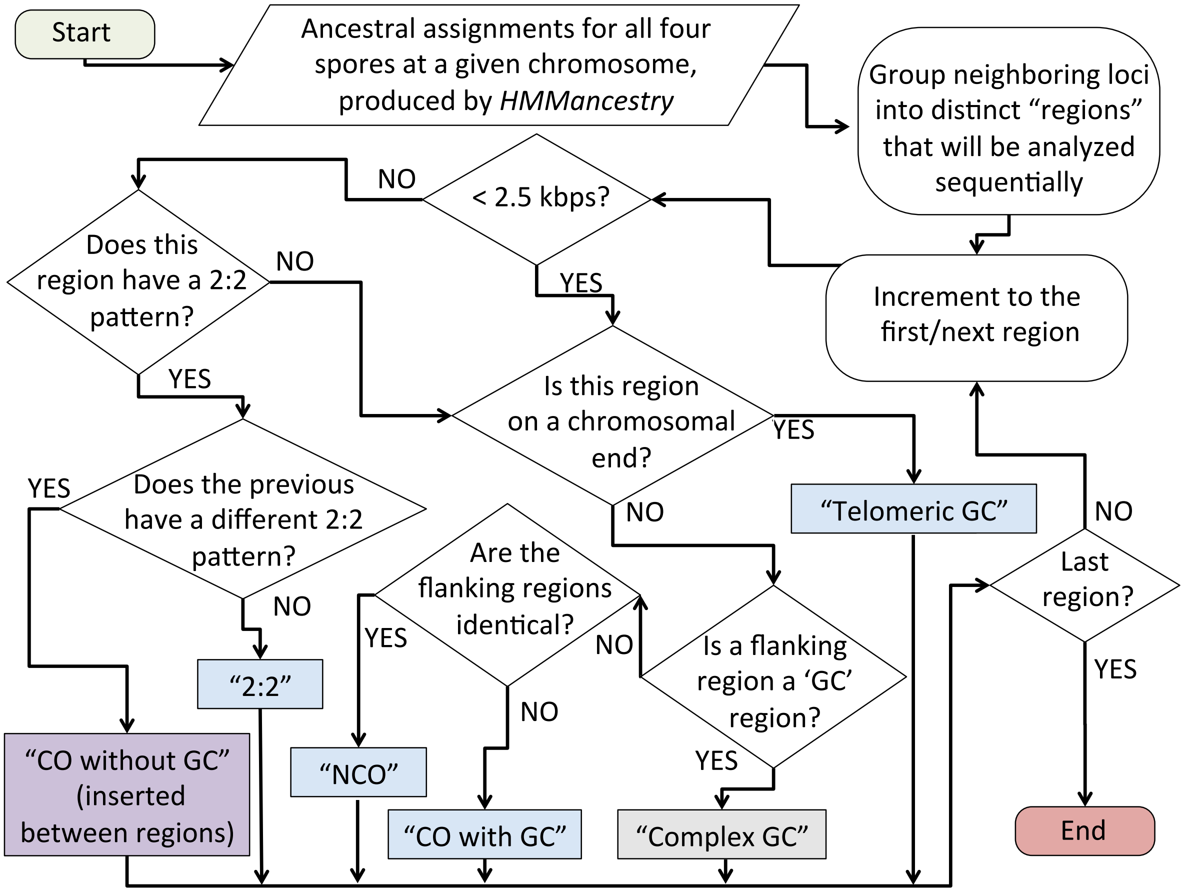

Start
Ancestral assignments for all four spores at a given chromosome, produced by *HMMancestry*
Group neighboring loci into distinct "regions" that will be analyzed sequentially
Increment to the first/next region
< 2.5 kbps?
NO
YES
Does this region have a 2:2 pattern?
NO
YES
Is this region on a chromosomal end?
YES
NO
"Telomeric GC"
Does the previous have a different 2:2 pattern?
YES
NO
"2:2"
"CO without GC" (inserted between regions)
Are the flanking regions identical?
YES
NO
"NCO"
"CO with GC"
Is a flanking region a 'GC' region?
NO
YES
"Complex GC"
Last region?
NO
YES
End

FIGURE 4.3: Inference algorithm used to infer crossovers, non-crossover, and telomeric gene conversion events. Blue, grey, and purple boxes show classifications made in the first, second, and third phases of the algorithm, respectively.

We mated the above parental strains and whole genome shotgun sequenced all 48 meiotic products from 12 independent sporulation events (we refer to this as the WGS dataset). For WGS, we sequenced a total of 4,714,686 PE250 reads. An average 90.6% (SD=2.7%) of reads could be merged into a single read with a minimum overlap of 10 bps. After mapping these reads to our SNP list (see above) we found an average of 62,470.3 informative SNPs (i.e., SNPs that had at least one read mapped to it) per individual (range: 35,069 - 72,329). The mean read coverage for these informative SNPs for WGS samples was 2.8X (SD=0.9X; Table 4.1).

To test a more efficient way of mapping recombination events across a large number of samples, we developed a modification of the Restriction-site-Associated DNA sequencing (RADseq) protocol (Baird et al., 2008; Ali et al., 2015). Our method increases the density of markers across the genome compared to the existing RADseq protocol by digesting DNA with two enzymes in parallel. From the reference genome sequence of yeast strain s288c (Cherry et al., 2012) we estimated that 6,066 RAD loci would be produced by digesting genomic DNA with two enzymes: *nsiI* and *pstI*. We also estimated that 33,983 (46%) of the total SNPs would occur within 600bp of each cut site. Note that our protocol is expected to produce a sequenced locus at nearly every site adjacent to a single recognition site for either of these two enzymes, in contrast to other 2-enzyme RADseq protocols (Peterson et al., 2012; Andrews et al., 2016). Briefly, we split each sample in two and digested the aliquots with either *nsiI* or *pstI*. Performing these digestions separately reduces the bias in shearing efficiency, a source of variance in coverage across loci, by producing larger DNA fragments than would occur in a single 2-enzyme digestion (Davey et al., 2013).

We applied this modified RADseq protocol to all 188 haploid spores from 47 meiosis events (RAD), produced by sporulation of diploid yeast cells from a cross between strains YPS128 and DBVPG1106 as above. We sequenced a total of 9,965,065 PE300 reads. We found an average of 60.8% (SD=2.0%) of read pairs across all 188 samples could be merged. This was unsurprising since RADseq prepared samples had a larger targeted inserted size of 400-600 bps than WGS samples. Mean coverage for RAD was 3.9X (Table 4.1) and we identified 22,304.8 informative SNPs per individual on average (range: 7,918 - 32,193).

We applied *HMMancestry* to the WGS and RAD haploid spore datasets to estimate ancestry at all 73,294 SNPs, identify chromosomal blocks of ancestry, and map recombination tracts. Our ML estimates for assignment probability, $\hat{p}$, were similar and high for both WGS and RAD

Table 4.1: Sequence and SNP information for each dataset used in this study. N=number of individuals; WGS = whole-genome sequencing; RAD = RADseq of haploid spores; DIP = RADseq of diploid F_6 samples. Mean numbers of SNPs, reads, and coverage are per individual sample.

Data	N	Raw read pairs	mean SNPs	range SNPs	mean reads	mean coverage	SD coverage
WGS	48	4,714,686	62,470.3	35,069–72,329	178,583.9	2.8	0.9
RAD	188	9,965,065	22,304.8	7,918–32,193	91,955.6	3.9	1.2
DIP	96	4,855,193	27,076.1	8,478–33,354	93,942.1	3.4	0.7

datasets (Table 4.2). However, WGS contained a 55% higher estimated genome-wide recombination rate $\hat{c}$ than RAD. We used the inference algorithm in *HMMancestry* to map different types of recombination events and found an average of 160.1 recombination events per meiosis event across the two datasets. WGS detected significantly more recombination tracts on average (199.7) than RADseq (151.3; $F_{1,57} = 71.8$; $p < 1e - 11$). This increase in tract count for the WGS dataset is a result of its higher marker density (Table 4.1) compared to RAD, which results in finer-scale detection of recombination events, and accounts for the higher estimate of overall recombination rate $\hat{c}$. Accordingly, the WGS dataset contained significantly smaller tracts for all types of GC (K-S one-tailed test; Table 4.2; Figure 4.4). WGS detected more telomeric GC and NCO tracts than RAD, and it also detected smaller-scale GC events associated with CO than RAD (Table 4.2). Because WGS detected more small-scale NCO events, the WGS data also contained a larger number of separate 2:2 tracts that would have been grouped together with the lower marker density of RAD.

The two sequencing methods detected similar numbers of CO events with and without GC (Table 4.2; Figure C.3). Combining CO events with and without GC, RAD and WGS predicted about 6.2 and 6.4 COs per Mbp, respectively (Table 4.3), which was statistically indistinguishable between the datasets (Pvalue for interaction between the number of COs and data type = 0.55). We found a tight linear relationship between chromosome length and the average number of both CO and NCO events in *S. cerevisiae* haploid spores (Figure 4.5; Table 4.3). On the other hand, we found a striking difference in the number of NCOs per chromosome size between the two datasets (Pvalue $< 2e - 03$), consistent with the failure of the lower marker density in RAD to detect small-scale NCO tracts. For WGS, the rate of NCOs was similar to that of COs (7.1 per Mbp) but much greater than the rate of NCOs found with RAD (4.0 per Mbp).

4.3.4 *Estimation of CO rates in a diploid F_6 population*

We further extended the *HMMancestry* algorithms to infer parental ancestry in diploid individuals, in which chromosomal blocks can be one of three possible states (i.e. homozygous for one or the other parent, or heterozygous). We applied this method to map CO events in the F_6 generation of an AIL between the *S. cerevisiae* strains YPS128 and DBVPG1106. We applied the modified RADseq approach described above to 96 diploid individuals and achieved a similar

Table 4.2: Summary of tract count and size (in kilobases) for RAD and WGS datasets.

Dataset	$\hat{c}$	$\hat{p}$	Tract	mean count	mean size	SD size	median size	K-S (size) P-value
RAD	2.0	0.996	2:2	151.89	72.70	66.73	54.02	1e-14
-	-	-	CO without GC	14.47	1.07	2.22	0.50	0.463
-	-	-	CO with GC	68.19	3.71	12.94	2.10	3e-15
-	-	-	Telomeric GC	15.40	5.65	12.92	1.74	1e-03
-	-	-	NCO	53.23	2.49	8.40	1.41	2e-16
WGS	3.1	0.993	2:2	193.25	57.69	50.45	43.27	-
-	-	-	CO without GC	13.67	0.81	0.90	0.65	-
-	-	-	CO with GC	76.58	1.92	2.21	1.41	-
-	-	-	Telomeric GC	22.42	3.70	4.62	1.52	-
-	-	-	NCO	87.00	1.03	1.86	0.22	-

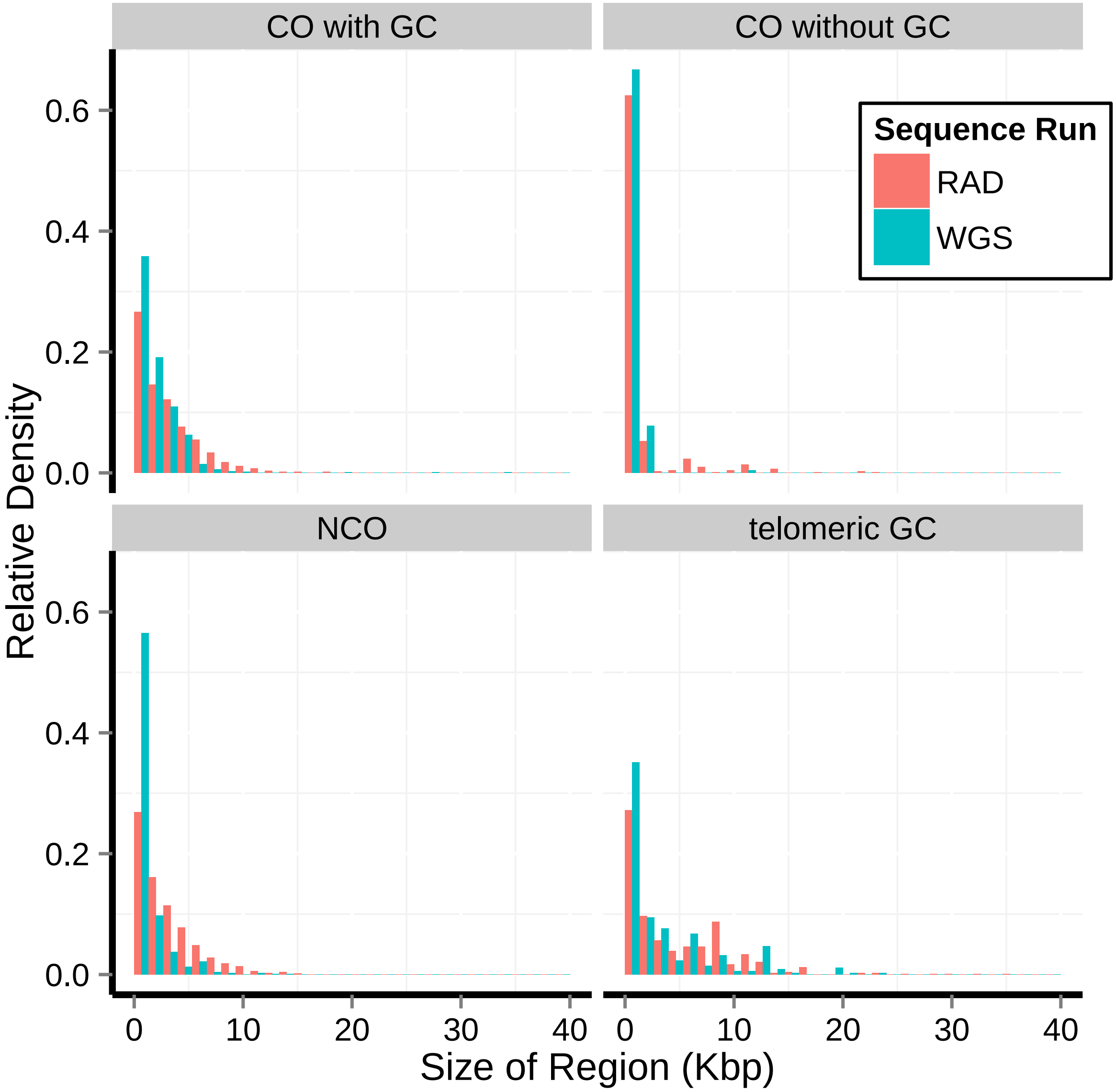


FIGURE 4.4: CO and NCO sizes for WGS and RAD datasets. Bars show the proportion of tracts in each bin, normalized by the total number of each type of tract for each technique; see Table 4.2 for total counts.

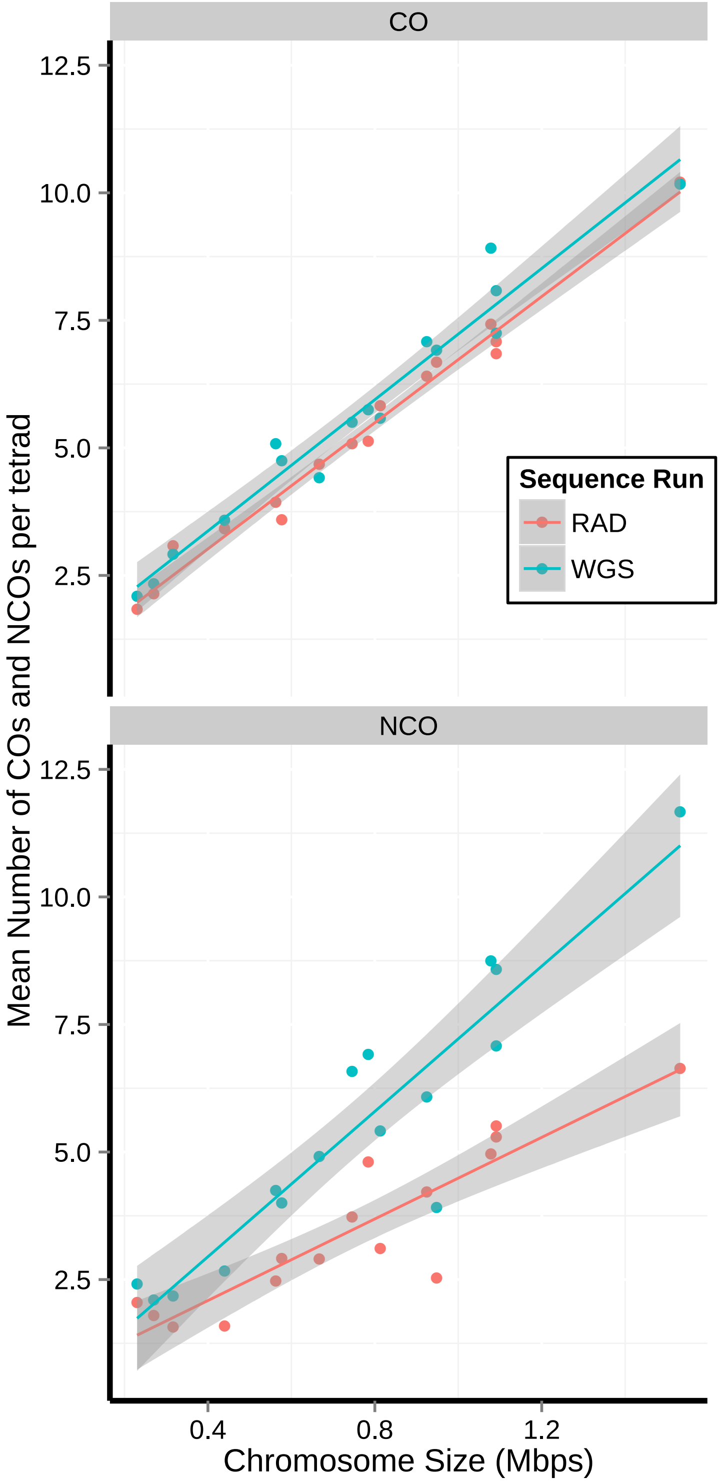


Figure 4.5: Mean number of CO and NCO events per tetrad as a function of chromosome length and data type.

Table 4.3: Regression of the number of COs (with and without GC) and NCOs on chromosome length in the WGS and RAD datasets.

Type	Slope (events per Mbp)	Intercept	R^2
CO | WGS	6.433	0.798	0.955
CO | RAD	6.183	0.5471	0.9822
NCO | WGS	7.118	0.1002	0.8504
NCO | RAD	3.998	0.4895	0.8066

number of SNP markers and coverage level to the haploid RAD dataset (Table 4.1). While NCO and CO-associated GC events cannot be distinguished from these data, we were able to map CO events and estimate genome-wide recombination rates. The maximum likelihood estimates $\hat{p}$ and $\hat{c}$ were 0.999 and 5.6, respectively. Taking into account that the F_6 diploids contained an extra 5 rounds of recombination relative to either WGS or RAD, we observed a noticeable decrease in the genome-wide recombination rate (1.1 cM/kb) relative to the haploid datasets. However, this drop in apparent recombination rate may be the result of selection or assortative mating in the admixed population. At the chromosomal level, large regions (i.e., up to 300 Kb) were dominated by one or the other parental haplotypes with lower than expected heterozygosity (Figure 4.6), in striking contrast to the mean 1:2:1 ratio that is expected to occur throughout the genome with complete admixture and neutrality. In many of these blocks we found that estimates of recombination rates were much lower than in the haploid spores datasets. In some cases, block boundaries coincide with hotspots of recombination identified in both the haploid and diploid datasets, but in other cases regions of elevated recombination rate do not correspond to haplotype boundaries in the F_6 population. Thus, it appears that selection or assortative mating is maintaining large blocks of homozyogous ancestry from one or the other parental strain, and the boundaries of these blocks depend only in part on recombination hotspots.

4.4 DISCUSSION

4.4.1 *Methods for inferring local ancestry with HMMs*

Hidden Markov Models (HMMs) are commonly used to infer local ancestry across the genome in admixed populations, and several related methods have been developed (reviewed in Liu et al. (2013); Liang and Nielsen (2014)). Some of the earlier methods (e.g., ANCESTRYMAP; Patterson et al., 2004), have been combined with unlinked panels of 'ancestry informative markers' to identify genomic regions associated with human diseases (e.g., asthma; Mersha, 2015). HAPMIX (Price et al., 2009) takes advantage of haplotype information to infer local ancestry. With low-coverage whole genome or reduced representation data, however, knowledge of specific haplotypes can be difficult to discern since many loci may be unsequenced across individuals. Andolfatto et al. (2011) developed a diploid HMM-based method for identifying COs from reduced

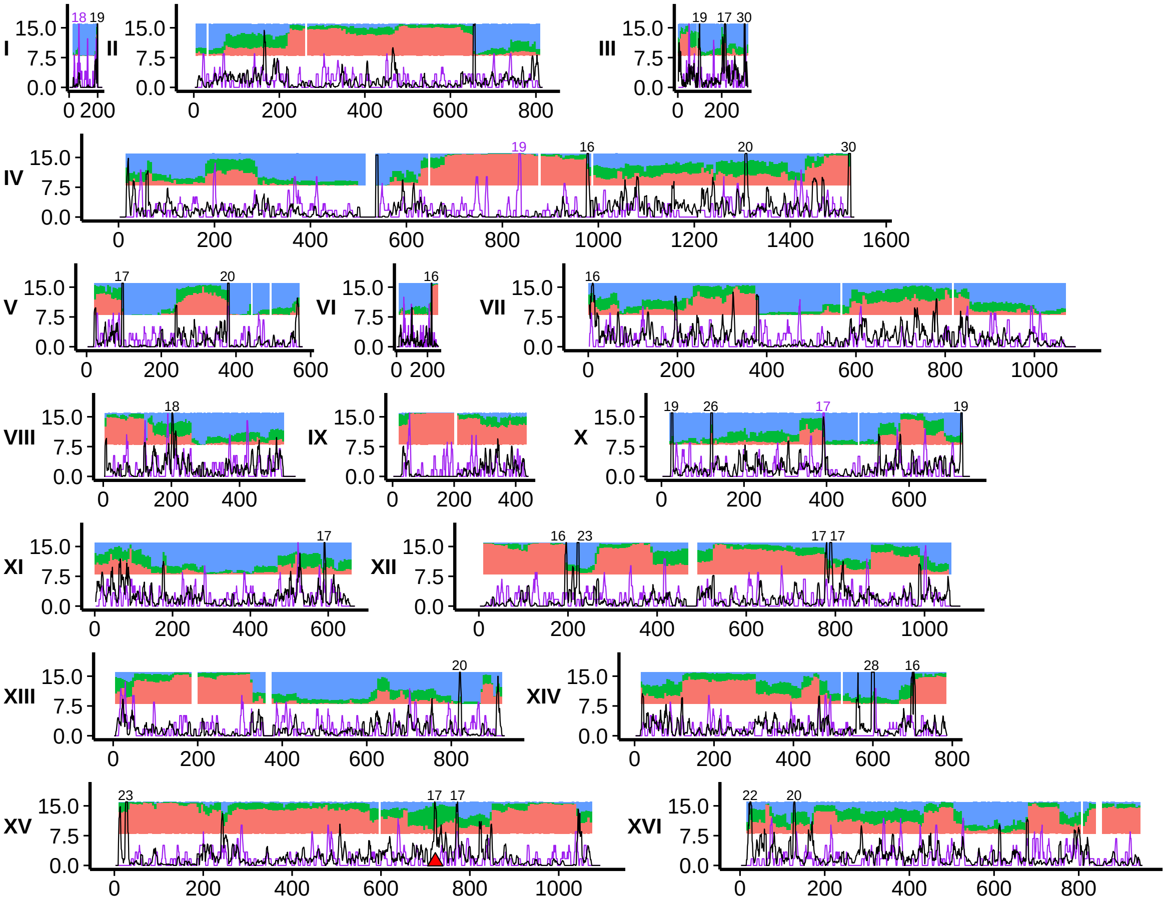


Figure 4.6: Local ancestry and recombination rates in 96 diploid F_6 individuals. Illustrated for each chromosome is the recombination rate (y axis) as a sliding window average of cM per 5 kb window (step size = 1 kb) in diploids (black). For reference, haploid inferred recombination rate is plotted in purple. Values greater than 15 cM are truncated for clarity. Within each plot the frequency of local ancestry is plotted with non-overlapping windows (5 kbp window size). Blue, red, and green show the relative frequencies of YPS128 homozygotes, DBVPG1106 homozygotes, and heterozygotes. The red triangle on chromosome XV shows the location of a homozygote-lethal locus (His3), engineered to ensure that the advanced intercross population remains diploid.

representation sequencing data in *Drosophila simulans* that does not rely on phased data. As with our approach, their method probabilistically assigns local ancestry and is capable of imputing ancestry at loci that lack sequence coverage. Their model had two tuning parameters: γ was used for uncertainty in parental reference sequences and ε was used to incorporate sequencing error based on Phred quality values. Our method incorporates both sequencing error and uncertainty in reference strains into a single parameter, p, which is estimated from the data by maximum likelihood.

Extending *HMMancestry* to other systems carries a few caveats. First, we assume equal admixture between two parental lines, meaning that in the HMM the initial state probabilities are 1:1 for haploids and 1:2:1 for diploids. Unequal admixture would deviate from this expectation, but our method should be robust to this violation. Under the HMM, the posterior probabilities of genotypes at the initial marker depend also on observed read counts at this marker as well as across the chromosome, so that higher sequence coverage and a larger number of markers on each chromosome easily overwhelms the effect of initial state probabilities.

Our method does not strictly account for sex chromosomes which can behave either like a haploid or a diploid chromosome. However, there are two ways to infer ancestry at sex chromosome with our current method. In cases where phenotypic markers to distinguish sexes exists one could use the appropriate HMM (e.g., if the organism is male, use the haploid algorithm). Additionally, one could run both the haploid and diploid algorithms and compare the resulting likelihood values in a odds ratio model comparison framework (Durbin et al., 1998). The latter approach can even be used to determine sex in the absence of other phenotypic markers.

Larger genomes require more computational effort for ancestry estimation, although the HMM approach remains highly efficient. For a single simulated diploid chromosome with 500,000 SNPs, *HMMancestry* estimated posterior probabilities and inferred ancestry in just under a second using a 2.5 GHz Intel Core i7 Macintosh (running OS X 10.10.5). Computationally, we found that both the haploid and diploid Forward-Backward algorithms were 75-100 faster when written in C++ using the R package *Rcpp* (Eddelbuettel and François, 2011) than when written in R alone (tested using *microbenchmark* with default settings; Mersmann, 2011) and thus we have implemented this more efficient version in *HMMancestry*. When scaling up to larger genomes we found a near linear increase in computation time of the forward-backward algorithm with the number of loci. We also make use of R's built in *parallel* package for increased computational efficiency during the ML estimating algorithm.

More complex models of recombination could be incorporated in the simulation functions as well as the ancestry inference in *HMMancestry*. For instance, it would be possible to allow for chiasma interference (Mancera et al., 2008; Zhao and Speed, 1996), or to allow two rates of recombination rather than the single rate that we consider here (c) to account for genomic regions with substantially elevated recombination (hotspots). Nonetheless, we were still able to identify recombination hotspots across the yeast genome (Figure 4.6) even with a single transition rate parameter in our HMM.

4.4.2 *Genomic sequencing approaches*

Our method of estimating recombination rates relies on genomic sequence data across a relative large number of samples, and here we evaluated two sequencing techniques. Low-coverage whole-genome sequencing (WGS) maximizes marker density with the potential to gather data at every polymorphic site across the genome. Accordingly, with WGS we were able to detect not only crossover events, but also small-scale non-crossover gene conversion events. Our modified RADseq technique is a reduced representation technique, leading to a lower marker density, although our two-enzyme approach was successful in increasing the marker density above most other RADseq protocols (Andrews et al., 2016). Nonetheless, RADseq was less able to detect small-scale non-crossover events. The primary trade-off between methods is that roughly three times the number of samples can be multiplexed in a sequencing experiment with RADseq compared with WGS in order to achieve the same mean coverage across marker loci. Depending on the goals of a study, statistical power may be improved by including more samples (i.e. more meiosis events) rather than a denser marker set. A second major difference between the techniques is in cost of library preparation. Whole-genome Illumina shotgun sequencing is typically conducted using proprietary kits in which individual samples are not barcoded (i.e. ligated to adaptors with specific nucleotide sequences that are used to identify individual samples in the sequence data) until near the end of library preparation. The cost of this protocol typically exceeds US$100 per sample. In contrast, RADseq ligates a set of custom barcoded adaptors to each sample early in the protocol, and samples can then be multiplexed during the rest of library preparation

(Andrews et al., 2016; Ali et al., 2015). Adaptor ligation to the single-stranded DNA overhangs left by restriction enzyme digestion in RADseq also tends to be more straightforward than the blunt-end ligation in WGS protocols. As a result, library preparation for RADseq is typically US$5 - 10 per sample, not including the initial one-time purchase of barcoded adaptors that can be used across a large number of sequencing libraries. Thus for an experiment sequencing hundreds of samples, the costs of library preparation for WGS can dwarf the sequencing costs, while a reduced representation method like RADseq is far more cost-effective, and library preparation costs remain a fraction of sequencing costs.

4.4.3 *Recombination rates in S. cerevisiae*

There are two general approaches taken in estimating the recombination landscape by mapping recombination events in a laboratory cross. The first approach, tetrad analysis, has been used extensively for genetic mapping of thousands of genes (Cherry et al., 1997) and genotyping (Mancera et al., 2008) all four individual products of meiosis in *S. cerevisiae*. The main drawback of the tetrad approach is that it only captures a single meiosis event per tetrad and all four spores are required to infer CO rates. We found that this limitation can be overcome either by using a sparser marker density or by lowering the overall coverage (or both) and using a HMM to probabilistically assign ancestry to loci. The second general approach to estimating recombination rates is serially mating advanced filial generations and sequencing a subset of the progeny (Illingworth et al., 2013). This approach can be applied to a wide range of systems in which it is difficult to obtain the complete genotypes for all four meiotic products. Additionally, increasing the number of meiotic generations increases the number of recombination events represented in each sample, thus increasing the probability of detecting recombination hot spots and cold spots. As with tetrad analysis, the use of a HMM can be useful by allowing lower sequencing coverage and inferring blocks of ancestry. The major drawback of this advanced intercross approach is the inability to detect NCO events and other gene conversion information. In addition, multiple generations of an intercross line may be subject to laboratory selection that can decrease the apparent recombination rate, as we observed here in our F_6 population.

Focusing only on CO events, we can compare across studies using different methods to estimate recombination rates in *S. cerevisiae*. Table 4.4 summarizes average rates of COs from a subset of published studies, highlighting the dependence of recombination rate estimates on marker density. The data from Mancera et al. (2008) have been reanalayzed 3 additional times, allowing comparison of statistical methods as well. Our CO/NCO inference algorithm (Figure 4.3) produced similar results as Mancera et al. (2008)'s and Anderson et al. (2011)'s methods, but these rates differed from Illingworth et al. (2013)'s estimate.

The detection of NCO events depends more strongly on marker density, because NCOs are relatively small (about 1kb on average; Table 4.2). Mancera et al. (2008) used microarrays to genotype 51 tetrads, formed from crossing s288c with YJM789, with *ca.* 52,000 markers, intermediate between our WGS and RAD marker sets. Overall WGS marker density was 41% larger than that of Mancera et al. (2008) and we observed 87 NCOs per meiosis in WGS – 1.9 times the amount detecting in Mancera et al. (2008). In our approach, recombination rate can also be estimated from the transition rate parameter in the HMM. At the level of individual chromosomes our ML estimates for transitions were high ($\hat{c}_{RAD}$ = 2.0 cM/kb, $\hat{c}_{WGS}$ = 3.1 cM/kb, $\hat{c}_{DIP}$ = 1.1 cM/kb). These rates, however, correspond to transition rates between states along individual chromosomes and not the CO rate *per se*. Because a CO is often associated with a detectable GC event (Table 4.2) there can be two or more "recombinations" for every CO; similarly, every NCO event has at least two transitions. Thus, it is not surprising that our ML estimates of genome-wide recombination are higher than previously reported rates or when considering the tetrad as a whole.

At the chromosome level we found a tight linear relationship between the number of CO and NCO events and chromosome size (Table 4.3), which is consistent with previous work. Our estimate of CO rate (6.2 per Mb for RAD; 6.4 per Mb for WGS) aligns with that of Mancera et al. (2008) (6.1 per Mb), while our estimates of NCO density (4.0 per Mb for RAD; 7.1 per Mb for WGS) were slightly higher than that of Mancera et al. (2008)'s (3.4 per Mb). We also found that gene conversion tracts were generally longer when associated with a CO event than when associated with a NCO event (Table 4.2).

Analysis of our diploid F_6 AIL dataset (DIP) revealed a surprising result: many large blocks of ancestry that were strongly biased towards one or the other parents (Figure 4.6). Several factors could cause this striking pattern. Segregation distortion, deviation from the expected Mendelian segregation ratio (Zhan and Xu, 2011), is unlikely to be the explanation. High rates of segregation distortion would create frequent non 2:2 segregation at the His3 locus in our tetrad dissections for

Table 4.4: Recombination rate considering only crossovers (cM/kb) per meiosis event in *S. cerevisiae*

Method	Rate	Marker density	Notes	Reference(s)
Tetrad analysis	0.37	>2,600 genes	genetic mapping	Cherry et al. (1997, 2012)
Tetrad analysis	0.45	*ca.* 52,000 markers	Microarray***	Illingworth et al. (2013)
Tetrad analysis	0.75	*ca.* 52,000 markers	Microarray***	Mancera et al. (2008)
Tetrad analysis	0.78	*ca.* 52,000 markers	Microarray***	Anderson et al. (2011)
Tetrad analysis	0.74	*ca.* 52,000 markers	Microarray***	this study
Tetrad analysis	0.68	7,918 – 32,193 SNPs**	RAD	this study
Tetrad analysis	0.75	35,069 – 72,329 SNPs**	WGS	this study
F_{12} AIL*	0.17	52,466 SNPs	2-way cross	Illingworth et al. (2013)
F_{12} AIL*	0.32	82,910 SNPs	4-way cross	Illingworth et al. (2013)
F_6 AIL*	0.88	8,478 – 33,354 SNPs**	DIP	this study

* Advanced Intercross Lines

** All 73,294 SNPs were used to infer COs, though the number of informative SNPs varied across individuals.

*** Reanalysis of data originally published by Mancera et al. Mancera et al. (2008).

the WGS and RAD datasets. Instead, we observed a high frequency of successful dissections in which all four spores formed colonies and proper 2:2 segregation was observed (data not shown).

Could there be a high rate of misclassification that leads to an excess of homozygotes? Simulations show that at low sequencing coverage there is a greater propensity for our HMM to misclassify loci (Figure 4.2). We also found these few, misclassified loci tended to be false homozygotes; however, this effect is not associated with reduced coverage (Figure C.4), and is expected to be dispersed across the genome instead of in large blocks. To confirm this, we performed an *ad hoc* simulation using *HMMancestry* of 96 diploid individuals using the empirically derived estimates of the WGS recombination profile (number of loci = 73,294, displacement between each SNP = 175 bp, $c = 3.1$, $p = 0.993$, coverage = 2.8X, frequency of COs = 0.51, frequency of conversion = 0.85, length of conversion = 1,920 bp). These simulated data showed that 70.7% of genotyping errors were false homozygotes, but that the overall error rate was small (squared correlation = 0.998) and misclassified loci rarely occurred adjacent to each other (Figure C.5). Therefore, the large blocking pattern of pure ancestry that we see in the F_6 AIL dataset is unlikely due to low sequencing coverage or heterogeneous bias detected in the HMM.

It is possible that genetic drift and (or) selection has occurred during the repeated rounds of sporulation. Our sporulation protocol was designed to systematically cull haploid and diploid cells lacking the HIS3/URA3 heterozygosity at the His3 locus (see Materials and Methods). Following each bottleneck, the remaining diploid cells were sporulated to induce sexual reproduction. This process of selecting heterozygote diploids is expected to leave a selective pattern of overdominance at the His3 locus, which is indeed what we observed (Figure 4.6). Despite growing this AIL in rich media, there may be strong selection for particular genotypes as a result of our sporulation protocol that leads to large blocks of homozygosity across the F_6 genomes.

4.5 CONCLUDING REMARKS

We developed and validated a set of methods for efficiently estimating rates of recombination events (CO and NCO GC). Our results demonstrate some heterogeneity in recombination rate across the yeast genome, but overall a consistent pattern of the number of recombination events across chromosomes. Further, we demonstrated that NCO GC events occur at roughly equal frequency to CO events. NCO events have important effects on how meiotic recombination structures genetic variation in diploid populations and on linkage-based mapping approaches (Slatkin, 2008; Lynch et al., 2014). However, they may often be overlooked because they are more difficult to observe empirically. We have shown that NCO tracts tend to be small, requiring a dense marker set, and they typically cannot be detected in advanced intercross lines or diploid samples. However, there are methods for estimating rates of gene conversion from population genetic data in model organisms such as humans, although rate estimates may come with high uncertainty (Gay et al., 2007; Padhukasahasram and Rannala, 2011, 2013). Our results re-emphasize the importance of considering both crossover and non-crossover recombination processes in understanding linkage disequilibrium and the haplotype structure of genetic variation.

4.6 MATERIALS AND METHODS

4.6.1 *The Forward-Backward Algorithm*

We used the Forward-Backward algorithm (Durbin et al., 1998) to probabilistically assign parental ancestry to each locus. This approach has four parts: 1) calculate the emission probabilities 2) calculate the forward probabilities 3) calculate the backward probabilities and 4) combine forward and backward probabilities to infer the most likely ancestral state at each SNP.

For both the haploid and diploid variants of the Forward-Backward algorithm the input consists of a vector of SNP locations along a chromosome and two vectors, k_0 and k_1, containing the read counts of alternative parental alleles at each locus. We calculate the emission probabilities for each locus $i \in (0, 1, ..., I)$. These probabilities correspond to the probability of observing the read counts given an underlying ancestral state, or the likelihood of each ancestral state at the locus given the data.

For the haploid case there are two hidden states corresponding to ancestry from each parent. For the diploid case there are three hidden states: two homozygous states for either parent and one heterozygous state. The emission probability for state j and locus i is calculated by:

$$e_i^{(j)} = \binom{n}{k_j} p^{k_j} (1-p)^{n-k_j} \tag{4.1}$$

where n is the sum of the number of sequence reads from both parents at SNP i. For the haploid case and for homozygous states in the diploid case, p is the assignment probability (see Introduction) and k_j is the count of reads corresponding to parent j. For the heterozygote state, p is set to 0.5. Equation 4.1 assumes that polymorphisms are biallelic (sequence reads that do not correspond to either parental state are discarded) and there is no sequencing bias such that one allele is more likely to be observed, and assignment errors at a heterozygous locus are equal between parental alleles.

The forward probabilities are calculated by starting at the first position of each chromosome with equal probabilities for each state in the haploid case, and a 1:2:1 ratio of probabilities for diploid states. For loci $i = 2$ through $i = I$ of each chromosome we then recursively use the following formula to calculate the forward probabilities. Using matrix notation, the vector of forward probabilities across states at locus i is:

$$\mathbf{f}_i = \mathbf{E}_i \, \mathbf{T}_i \, \mathbf{f}_{i-1} \tag{4.2}$$

where $\mathbf{E}_i$ is a matrix with the emission probabilities (equation 4.1) for each state on the diagonal and zeros on the off-diagonal, $\mathbf{T}_i$ is a 2-by-2 (haploid) or 3-by-3 (diploid) matrix describing the transition (recombination) probabilities between states, and $\mathbf{f}_{i-1}$ is the forward probability at the previous position along the chromosome (5' of position i). To avoid underflow (computational issues with small probabilities) and to calculate the total log likelihood of the data, we rescale the forward probabilities at each SNP by dividing each element of $\mathbf{f}_i$ by the sum of whole vector (Rabiner, 1989).

The transition probabilities in matrix $\mathbf{T}_i$ are calculated for each SNP position and specify the rate of transition from one hidden state to another. For haploids, the transition matrix is:

$$\mathbf{T}_i = \begin{pmatrix} 1-r & r \\ r & 1-r \end{pmatrix} \tag{4.3}$$

and the transition matrix for diploids recombinants is:

$$\mathbf{T}_i = \begin{pmatrix} (1-r)^2 & 2r(1-r) & r^2 \\ r(1-r) & (1-r)^2 + r^2 & r(1-r) \\ r^2 & 2r(1-r) & (1-r)^2 \end{pmatrix} \tag{4.4}$$

where r in equations 4.3 and 4.4 is the probability of getting an odd number of crossovers between loci and is based on Haldane's mapping function (Haldane, 1919; Kosambi, 1943; Gjuvsland et al., 2007a):

$$r = \frac{1 - e^{-2dc}}{2} \tag{4.5}$$

where d is the physical distance (bp) between loci and c is the genome-wide recombination rate (expressed here in terms of Morgans/bp).

The backward probabilities are calculated in a similar manner, but in the reverse (i.e., 3' to 5') direction. We assume that the backward probability at locus I is 1 for each state (i.e., we assume the chromosome can end at any state). At each SNP position $i \in (I - 1, l - 2, \ldots, 0)$ the vector of backwards probabilities are given by:

$$\mathbf{b}_i = \mathbf{T}_i \; \mathbf{E}_{i+1} \, \mathbf{b}_{i+1} \tag{4.6}$$

Lastly, the posterior probability that the hidden state at SNP i is j given the observed sequence of read counts, **X**, is calculated as:

$$P(s_i = j \mid \mathbf{X}) = \frac{\mathbf{f}_i^{(j)} \mathbf{b}_i^{(j)}}{\mathbf{f}_i \mathbf{b}_i} \tag{4.7}$$

HmmAncestry returns these posterior probabilities. This has the benefit of retaining uncertainty in the genotypic calls. For example, one could throw out genotypes where the posterior probability was less than a specified cutoff value. For our analyses, however, we simply called the ancestral state by picking the highest posterior probability. In some cases the denominator in equation 4.7 can be zero. This rarely occurs unless under extremely low signal (i.e., sequence coverage $< 0.1X$). Under those cases we omitted ambiguously assigned ancestry for that SNP.

4.6.2 *Numerically estimating the maximum likelihood parameter values*

There are two parameters in the Forward-Backward algorithm that are specified by the user: the assignment probability, p, and the genome-wide recombination rate, c. These two parameters can also be estimated in *HMMancestry* from the data using an ML approach. The function has a coarse and fine scale method to estimate these parameters. The coarse scale estimates the log likelihood (LnL) of the data across a coarse grid of varying proposed $\hat{p}$ or $\hat{c}$ and the fine scale using a hill-climbing algorithm.

For the coarse scale, we picked the default number of grid points 5-by-5 = 25. The distance between p gridpoints was set to $dx = 10^{-4}$; distance between c gridpoints was $dy = 10^{-5}$. For each parameter combination we ran the above Forward-Backward algorithm and obtained estimates of the LnL calculated by summing up all the scaling factors at each SNP (Rabiner, 1989). We picked the grid point that had the highest LnL.

For the fine-scale step we performed a custom hill-climbing procedure as follows. We used a two-variable Newton-Raphson method to iteratively find better approximations to the maximum LnL parameter estimates $\hat{p}$ and $\hat{c}$. Specifically, we calculated LnL at four points dx and dy distance away in the four cardinal directions, a point $(x_1 - dx, y_1 - dy)$, and a new, proposed point for iteration $n + 1$, calculated from the other points by:

$$(x, y)_{n+1} = (x, y)_n - \mathbf{H}^{-1}\triangledown \tag{4.8}$$

where $(x, y)_n$ is the vector of parameter values at iteration n, $\mathbf{H}$ is the Hessian matrix for the likelihood surface L:

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 L}{\partial^2 x} & \frac{\partial^2 L}{\partial x \partial y} \\ \frac{\partial^2 L}{\partial x \partial y} & \frac{\partial^2 L}{\partial^2 y} \end{pmatrix} \tag{4.9}$$

and $\triangledown$ is the gradient in the x and y directions.

Initially, the distance between points (dx and dy) is identical to that of the coarse scale search. If the proposed point has a higher LnL than all of the six points we accept it; otherwise, we pick the point among the six that had the highest LnL. We repeat the above procedure for a maximum of 25 iterations. For each iteration n we decrease the distance between the six points by dividing the initial distance by the integer n. If the Euclidean distance between points identified at iteration $n - 1$ and n is less than a specified tolerance level (here 10^{-4}) we terminated the search. We bounded the parameter space such that $\hat{p} \in (0.5, 1)$ and $\hat{c} > 0$.

4.6.3 *Simulations*

For each simulation we considered a 250 kb chromosome with 1,000 loci spaced evenly. We were specifically interested in how accuracy changed with ploidy, mean recombination rate, mean sequencing coverage, and the assignment probability. For simulating meiosis events we used *HmmAncestry* to simulate individual tetrads for a given ploidy, c, coverage, and p. We considered haploids or diploids, three levels of recombination rate (0.1, 0.5, & 1 cM/kb, corresponding to a map distance between adjacent SNPs of 0.025, 0.125, and 0.25 cM, respectively), ten levels of coverage spanning 0.2X to 2X, and 3 levels of assignment probability (0.90, 0.95, & 0.99). For each unique combination of parameters we carried out 50 simulation replicates of 50 independent meiosis events.

To simulate meiosis, we crossed two simulated parents. Each parent consisted of a vector of zeros or ones at each locus such that the diploid F_1 was heterozygous at each SNP position. Recombination events between each pair of neighboring loci were randomly sampled using Haldane's mapping function (Haldane, 1919; Kosambi, 1943; Gjuvsland et al., 2007a) to produce a vector of recombination events (0=no recombination, 1=recombination), where the length of the vector was one minus the number of loci. During a simulated meiosis, parental chromatids double creating two pairs of sister chromatids. We allowed recombination to occur between non-sister chromatids. This was done by starting at the beginning of the chromosome and crossing over non-sister chromatids if the corresponding element of the recombination index was 1. For all simulations described herein we ignored any possibility for a non-crossover events or gene conversion events to occur.

The results of the above algorithm give the true parental states at each locus and for each chromatid. To simulate sequence coverage we sampled the total number of reads that map to a given locus following a Poisson random variable with mean equal to the experimentally varied coverage level. Next we simulated sequencing error; this was done by sampling from a binominal distribution: $X \sim Binom(n, p)$ where n equaled the total number of reads and p is the assignment probability. Thus, when p is very high (i.e., close to 1) most of the reads at a given locus map to the correct parental state.

We then applied *HMMancestry* to each simulation replicate. We estimated the genome-wide recombination rate and assignment probability parameters and the parental ancestry at each locus, taken to be the state with the highest posterior probability.

4.6.4 *Yeast strains, crosses, and media*

We crossed two heterothallic and haploid *S. cerevisiae* strains of opposite mating types to create a diploid F_1 progeny. Specifically we crossed YPS128 (mat α, ho△::Hyg, ura3△::KanMX; NCYC# 3632) with DBVPG1106 (mat **a**, ho△::Hyg, ura3△::KanMX; NCYC# 3596). These strains were chosen because they crossed easily in the lab but are genetically distinct (Liti et al., 2009). Disruption of the homothallic switching gene, HO, was necessary for isolating and sequencing of haploid recombinants. Details of these strains' initial construction can be found in Cubillos et al. (2009).

To create the haploid recombinants we sporulated F_1 cells using the following protocol. We harvested 1 mL of F_1 cells grown in standard rich medium (Yeast Peptone Dextrose, YPD; Sherman, 2002) and washed cells in 1 mL of presporulation medium (Yeast Peptone Acetate; Codón et al., 1995). We incubated cells in 1 mL YPA for 12-15 hours at 30C. Following the presporulation incubation, we washed cells in SPO2 sporulation medium (2% KAc, pH 7; Codón et al., 1995). We incubated cells in 1 mL of SPO2 in a roller drum placed in a chilled growth chamber at 22C for 1 to 3 d. This method generally produced high sporulation efficiency (> 95% of cells formed tetrads after 2 days).

Next we killed off unsporulated cells and disrupted the asci of sporulated cells with a combination of heat and enzymatic perturbation. Following the 1 to 3 d incubation in SPO2 we chilled the cells for 15 m at 4C. To kill any unsporulated cells we adapted a protocol from Khare et al. (2011) used to kill unsporulated *Schizosaccharomyces pombe* cells. We placed 1 mL of chilled cells in SPO2 into a 55C water bath for 30 m. To digest the asci we removed the supernatant and resuspended cells in 40 μL of Zymolyase Solution (1 U/μL Zymolyase [MP Biomedicals] in 1 M sorbitol) and incubated them in a 37C water bath for 15-20 m. We added 460 μL of dH_2O to stop the reaction dissected tetrads immediately using standard dissection techniques (Sherman, 2002). Dissected cells were grown on YPD plates for 2 d at 30C and colonies were placed in 15% glycerol and stored at -80C until DNA extraction.

For the haploid whole genome dataset (WGS) we used the above technique to isolate 56 spores (from 14 individual meiosis events). However, two spores from two separate tetrads (Tetrads 3 and 11) appeared to have diploid genotypes (data not shown) and so we report on only 48 haploids. The most likely cause of this error was cross contamination during tetrad dissection. To mitigate this issue we engineered strains using the protocol from Gietz and Schiestl (2008) for the second sequencing run (RAD) so we can check for the presence of a 2:2 segregation pattern at a single locus before sequencing samples. We disrupted His3 in DBVPG1106 (mat **a**, ho△::Hyg, ura3△::KanMX) and replaced it with a functional copy of URA3 from an amplicon from plasmid YEp24 (Struhl et al., 1979). Primers used to amplify URA3 and its promoter region from YEp24 were designed so that the flanking ends had homology to genomic His3. Oligos were long enough so that they only occurred once in the genome. Specifically, the forward 90-mer contained a 34 bp region of 5' homology to genomic His3 and included the start codon (AAATGAGCAGGCAAGATAAACGAAGGCAAAGATG) followed by 18 bp tag (GATGTCCACGAGGTCTCT), a 20 bp barcode (AATTCCGGGCATGCGGCCTT), and ended with a 18 bp region of 5' homology upstream of Ura3 in YEp24 (AGTAACAAAAGAGTGGTA). The downstream primer was constructed similarly with a 34 bp region with 3' homology of genomic His3 including the stop codon (CGTATGCTGCAGCTTTAAATAATCGGTGTCACTA), and 18 bp tag (CGGTGTCGGTCTCGTAGA), a 20 bp barcode (AACCTTGGCCGCTCGGTTCC), and finally a 18 bp region that has 3' homology to YEp24's Ura3 gene (CGATGCGTCCGGCGTAGA). We used the resulting 1798 bp amplicon as the template for recombineering. Successful transformation was confirmed 2 ways. First, we selectively grew transformants on Uracil dropout medium (Sherman, 2002) since the parental strain was prototrophic for Uracil. The transformants grown on Uracil dropout were unable to grow on Histidine dropout indicating that the cells that regain Uracil auxotropy simultaneously became prototrophic for Histidine. Second, we PCR amplified the His3 region of a single transformant colony using forward primer TTCCACCTAGCGGATGACTC and reverse primer TGATGCATTACCTTGTCATCTTC. The native size of the resulting fragment is about 900 bp while the size of a properly placed integrated fragment was about 2 kb. We choose one transformant (DBVPG1106 mat **a**, ho△::Hyg, ura3△::KanMx his3△::URA3) that met the above criteria to mate with YPS128 (mat α, ho△::Hyg, ura3△::KanMX, HIS3+). The resulting F_1 progeny are genetically identical to the F_1 progeny of the WGS except for their heterozygosity at His3.

For RAD we dissected 188 spores from 47 meiosis events using the modified diploids described above. For each tetrad we confirmed a 2:2 segregation pattern at His3 gene by selectively growing each spore on Uracil dropout and Histidine dropout. The 2:2 segregation was mutatively exclusive (i.e., the 2 spores that grew on Uracil dropout could not grow on Histidine dropout and *visa versa*).

4.6.5 *Library preparation and sequencing*

Yeast samples were grown up overnight in YPD. After at least 24 h growth, approximately 1.5 mL of overnight culture was harvested and DNA was extracted using the Gentra Puregene kit for yeast (Qiagen #158567). DNA samples were quantified using Quant-It High Sensitivity kit (Q33120) and a subset of samples were haphazardly ran on a 1.5% agarose gel to ensure high quality. For PAR and WGS, samples were submitted to the Institute for Bioinformatics and Evolutionary Studies' Genomics Resources Core Facility at the University of Idaho for whole-genome shotgun library preparation and sequencing (PE250 MiSeq).

For RAD and DIP, we followed the protocol of Ali et al. (2015). A total of 150 ng from each sample were standardized to a concentration of 5 ng/μL, then divided in half, with each half being digested by either 0.7 U of *nsiI* (NEB #R0127S) or 1.5 U of *pstI* (NEB #R3140S) for 1 h at 37C and 20 m at 65C. Each restriction digest included 1X concentration of Cutsmart buffer. After digestion, samples were barcoded, in parallel, with adapters including both a sticky end compatible with pre-existing restriction enzymes and an *sbfI* cutsite to facilitate DNA liberation from streptavidin SPRI beads (see Ali et al., 2015, for BestRAD adapter sequences). Two μmol adapter was ligated onto each sample using 320 U of T4 DNA ligase (NEB #M0202M), 1X NEBuffer 4 (NEB #B7004S), and 0.016 μmol rATP (Thermo Fisher #R0441) and incubated for 1 h at 20C and 20 m at 65C. Approximately 7 ng DNA from both restriction-enzyme treatments from each sample were pooled *en masse*, purified using 1.2X Ampure (Beckman Coulter #A63881) beads, and sonicated for approximately 700 bp fragments using a Covaris M220 ultrasonicator at the Institute for Bioinformatics and Evolutionary Studies. Sheared DNA was incubated with

1X Dynabead M-280 Streptavidin magnetic beads (Invitrogen #11205D) for 20 m at room temperature and washed five times using 1X Binding and Wash buffer (5 mM Tris-HCl; 0.5 mM EDTA; 1 M NaCl). DNA + streptavidin beads were washed once with and resuspended in 1X NEBuffer 4 prior to restriction digestion with *sbfI-HF* (NEB #R3642L) to liberate DNA from the streptavidin beads. After DNA liberation, we used NEBNext Ultra DNA Library Preparation with the following modifications: we used 3.75 μmol of Truseq Adapters for Illumina in lieu of the NEBNext Adapters and omitted the *USER* enzyme step during the adapter ligation. To size select our library we used AMPure beads to select for 500-700 bps. To ensure that the library was in our desired size range and that the TruSeq adapters were ligated on correctly, we did a "test" PCR under the following conditions: 98C for 30 s; (98C for 10 s; 60C for 30 s; 72C for 30 s) x 19 cycles; 72C 5 m. Nineteen cycles ensured enough product to visualize library on a gel to verify it was in the desired size range. Final sequencing libraries had the same PCR profile but with fewer cycles (12). Post PCR, libraries were equimolarly pooled to a final concentration of 10.5 nM and sequenced at the Genomics Core Facility in the Institute for Molecular Biology at the University of Oregon.

4.6.6 *Bioinformatic pipeline*

A major advantage of *HMMancestry* is that the input format, a data frame specifying the allele counts for each parent/population at each SNP, is relatively simple and independent of the choice of bioinformatic pipeline. Here we detail the pipeline used in our analysis. This pipeline is freely available online (https://github.com/tylerhether/Scripts). We used the subprogram *preproc_experiment* in *seqyclean* (Zhbannikov, 2015) to preprocess the 48 recombinant haploids that were whole-genome sequenced (WGS). We deduplicated the raw reads and used the program *flash* (Magoč and Salzberg, 2011) to merge reads that overlapped by at least 10 bps. Merging was necessary since allele counts fed into *HmmAncestry* would be artificially inflated if read pairs overlapped at SNP loci.

We whole-genome sequenced each of the parental haploid strains (PAR) to identify SNPs between them. As with the WGS recombinant data, we used *preproc_experiment* to preprocess the raw reads and separate them into merged (single-end) and unmerged (paired-end) files. To obtain a high quality SNP list we filtered parental reads in *preproc_experiment* with a quality cutoff of 10. Next we aligned each parental set of merged single-end and unmerged paired-end reads to the s288c reference genome (Cherry et al., 2012) and merged the corresponding single- and paired-end BAM alignment files. From these alignment files we created consensus sequences for each parent using samtool's subprograms *mpileup* and *vcfutils* (Li et al., 2009). Fastq files were converted to fasta files using a custom perl script and fed into *nucmer* (Kurtz et al., 2004) where we performed a global alignment without rearrangements, set the minimum alignment length to 10kb, and only retained unambiguous SNPs. We removed multi-base pair indels with a python script from Anderson et al. (2011) to create a list of SNPs between the two parental genomes.

We carried out the following steps to process the RADseq data from 188 individually barcoded haploid recombinants (RAD) and 96 barcoded F_6 diploids (DIP). Because the barcodes from our RAD procedure can be located on either the single end or paired end read, we first ran the raw reads through a custom perl script that flipped any reads in which the barcode was located on the paired-end read. This script was also used to demultiplex the raw reads based on the restriction enzyme and to remove any reads that lacked a restriction site at one end. Second, we used *process_radtags* in the Stacks program (Catchen et al., 2013) separately on the *pstI*-only and *nsiI*-only data to demultiplex samples based on barcode. Third, for each barcode we concatenated the cleaned reads obtained from the *pstI*-only and *nsiI*-only filtering. As with the WGS dataset, we merged any pairs that overlapped with flash and retained the variable length single-end data and the unmerged paired-end data for each sample.

To obtain read counts for each SNP, we first aligned both the merged and unmerged files for WGS, RAD, and DIP datasets using *bowtie2* (Langmead and Salzberg, 2012). For the paired end files we allowed for variable insert size of 400-800 bps, based on the size selection during the library preparation. We then merged the single end and paired end alignment files for each sample using *samtools*. Lastly, we used *vcftools* to include only the SNPs identified between YPS128 and DBVPG1106 and used a custom perl script to parse the VCF files into the number of reads that mapped to each parent at each SNP. These read count data are the input for *HmmAncestry*. Custom scripts for this pipeline are available on github (https://github.com/tylerhether/Scripts).

Using *HmmAncestry*, we estimated the global assignment probability $\hat{p}$ and mean recombination rate $\hat{c}$. We did this procedure separately for each dataset (WGS, RAD, DIP). For each dataset, we used a 10-by-10 grid for the coarse maximum LnL search and ran the fine-scale, hill-climbing search for a maximum of 30 iterations or until the distance between parameters between iterations was less than the tolerance of 10^{-4}. We then estimated the posterior probability of belonging to each hidden state, and took the maximum probability state to be the ancestral state at each SNP locus for each individual.

CHAPTER 5

Uplift and erosion of genomic islands with standing genetic variation[4]

5.1 SUMMARY

Details of the processes that generate biological diversity have long been of interest to evolutionary biologists. A common theme in nature is diversification via divergent selection with gene flow. Empirical studies on this topic find variable genetic differentiation throughout the genome, that genetic differentiation is non-randomly distributed, and that loci of adaptive significance are often found clustered together within "genomic islands of divergence". Theoretical models based on new mutations show how these genomic islands can arise and grow as a result of a complex interaction of various evolutionary and genic processes. In the current study, we ask if such genomic islands can alternatively arise from divergent selection from standing genetic variation and we test this using a simple two locus model of selection. There are numerous ways in which standing genetic variation can be partitioned (e.g., between alleles, between loci, and between populations) and we tested which of these scenarios can give rise to an island pattern compared to no genomic differentiation or complete genomic differentiation. We found that divergent selection, even without reciprocal gene exchange between populations, following a bout of admixture can relatively quickly produce an island pattern. Moreover, we found two pathways in which islands can form from divergence from standing variation: 1) through the build up of islands and 2) through the breakdown of larger, genome-wide differentiation. Lastly, similar to new mutation theory, we found that the frequency of recombination is an important determinant of island formation from standing genetic variation such that mating behavior of a species (e.g., facultative or obligate sexual) can impact the likelihood of island formation.

[4] Manuscript in preparation for submission to Evolution (Brief Communications)

5.2 INTRODUCTION

It is increasingly evident that phenotypic and taxonomic diversity arises despite ongoing gene flow between populations or incipient species (Sullivan et al., 2014). Predicting the genomic response to divergence with gene flow (DGF) in nature is difficult, however, because several interacting evolutionary and genetic factors can occur simultaneously. Moreover, some of these factors can themselves have multiple levels of interaction. For example, divergent selection contributes to genetic divergence both *directly* by its effect on actual selected loci and *indirectly* by 'divergent hitchhiking' (DH) of nearby neutral loci (Via, 2012).

The metaphor of "genomic islands of divergence" has been used recently to integrate the dynamics of migration and divergent selection affecting selected loci and recombination and selection affecting the degree of genetic hitchhiking (Smith and Haigh, 2009; Nosil et al., 2009a). Here, inter-population gene flow homogenizes the neutrally evolving "sea floor" whereas DH creates genomic isolation, reducing the effective migration rate at selected loci as well as loci in tight physical linkage with these selected loci (Via, 2012). Such reduction in effective migration owing to DH can further diverge weakly selected, *de novo* mutations at nearby loci (Yeaman and Whitlock, 2011) that would otherwise be trumped by migration experienced at the sea floor. Thus, over time these divergent islands are hypothesized to grow (widen) with the inverse of the product of migration and recombination whereas height (extent of differentiation) is expected to be proportional to strength of divergent selection.

Mathematical models of GI formation has almost exclusively focused on divergent selection based on new mutations even though many research programs find adaptation from standing genetic variation (SGV; Schluter, 2000; Colosimo, 2005; Carlborg et al., 2006; Michel et al., 2010; Hohenlohe et al., 2012; Nadeau et al., 2012). Adaptation from SGV can lead to faster evolution, fixation of more small-effect alleles, and an increase frequency of beneficial recessive alleles (Orr and Betancourt, 2001) relative to adaptation from new mutations (Hermisson, 2005; Barrett and Schluter, 2008). With regards to GI architecture, however, less is known about the role of SGV in part because such variation can be partitioned several different ways both within and between populations. For example, two populations might be fixed for alternate alleles at all polymorphic loci such that each population is in linkage equilibrium but there is a high degree of cross-population linkage disequilibrium (X-LD) between loci. In such a case all the SGV is partitioned between populations. In other cases, a varying level of polymorphism can occur within one or more populations at one or more loci. It is reasonable to suspect that varying how SGV is partition would likely affect the overall magnitude and localization of genetic differentiation nearby loci under divergent selection.

The arrangement of genetic differentiation that occurs across the genome varies widely in the literature (Nosil et al., 2009a) which makes drawing conclusions on the nature of genomic differentiation difficult. It has been postulated that islands form by DH, with growth of such chromosomal regions possible by further divergent selection occurring at loci that are themselves linked to an already established divergently selected locus (Nosil et al., 2009a; Yeaman and Whitlock, 2011). We hypothesize that another, perhaps more frequently used mechanism for island formation is from the segregation of existing genetic variation between populations experiencing different selection regimes. Herein we modeled genetic divergence from SGV to explore the parameter combinations likely to give rise to islands versus those that generate either genome-wide divergence or no divergence between populations. We considered seven different demographic scenarios that differ in terms of how SGV is partitioned within and between a pair of populations, the mating type, and the migration frequency between diverging populations. Our results highlight how the balance of migration and selection together with meta-population demography can strongly affect short term genome-wide patterns of differentiation.

5.3 METHODS

5.3.1 *Modeling divergence from standing genetic variation*

We were interested in identifying the parameter range likely to give rise to islands (i.e., local differentiation only) from those that give rise to other genomic patterns (i.e., no or genome-wide differentiation). We considered scenarios in which 1) a pair of populations were completely isolated for a period of time that affected the partitioning of genetic variation between populations followed by 2) secondary contact and 3) divergence with gene flow. We were concerned here with SGV only and so we assumed that the genomic response to a given demographic scenario occurs without new mutations or that is on a shorter timescale than is relevant for new mutations. We examined the genome-wide and temporal dynamics of differentiation for 7 specific evolutionary scenarios that vary in how SGV is partitioned within and between populations, the degree of admixture between populations that occurred during secondary contact, the periodicity of migration, and whether individuals are obligate or facultative sexual (Table 5.1). In all scenarios, the initial type of SGV was a parameter of the model and we explore different levels of migration, divergent selection, and recombination.

The general life-history cycle during the divergence with gene flow following secondary contact is as follows. Migration between populations occurs at rate m between populations every m_f generations. For obligate sexual cases, random mating occurs every generation, following migration if applicable. For facultative sexual cases, random mating occurs following migration only, as in the case of the yeast experiment (see Chapter 6). In other words, for the facultative sexual scenarios, cell division occurs asexually and there are m_f rounds of viability selection occurring between migration and random mating. Viability selection within populations occurs at the last step of the life cycle.

In each evolutionary scenario we tracked genetic differentiation between populations at neutral loci linked to a single locus under divergent selection. Locus A is under divergent selection between these two populations and it is linked to a neutral locus B. The dynamics of neutral divergence between populations can be tracked by following haplotype frequencies through time. Because there is only a single locus under selection, we can obtain genomic patterns of differentiation by varying the recombination rate, r, between loci A and B, migration between populations, and the strength of selection at locus A.

Let $g_{ij}^{(k)}$ be the frequency of haplotypes in population k with allele i at selected locus A and allele j at neutral locus B. For convenience we can summarize the gamete frequencies for each population k as a vector, $\mathbf{p}_k$:

$$\mathbf{p}_k = \begin{pmatrix} g_{11}^{(k)} \\ g_{12}^{(k)} \\ g_{21}^{(k)} \\ g_{22}^{(k)} \end{pmatrix} \tag{5.1}$$

TABLE 5.1: Evolutionary models of explored in this study. For each scenario, the starting gamete frequencies, within-population LD, cross-population linkage disequilibrium (X-LD), mating mode, and frequency of migration (m_f) is given. Admixture here refers to the number of random mating that occurred during the period of secondary contact.

Scenario	mating	m_f	Admixture	LD	X-LD	$g_{11}^{(1)}$	$g_{12}^{(1)}$	$g_{21}^{(1)}$	$g_{22}^{(1)}$	$g_{11}^{(2)}$	$g_{12}^{(2)}$	$g_{21}^{(2)}$	$g_{22}^{(2)}$
1	obligate	1	none	0	0.25	1	0	0	0	0	0	0	1
2	obligate	1	none	0	0.125	0.5	0.5	0	0	0	0	0	1
3	obligate	1	none	0	0	0.5	0.5	0	0	0	0	0.5	0.5
4	obligate	1	F_1s	0.25	0.25	0.5	0	0	0.5	0.5	0	0	0.5
5	obligate	1	F_2s	$0.25 - \frac{r}{4}$	$0.25 - \frac{r}{4}$	$0.5 - \frac{r}{4}$	$\frac{r}{4}$	$\frac{r}{4}$	$0.5 - \frac{r}{4}$	$0.5 - \frac{r}{4}$	$\frac{r}{4}$	$\frac{r}{4}$	$0.5 - \frac{r}{4}$
6	obligate	50	F_2s	$0.25 - \frac{r}{4}$	$0.25 - \frac{r}{4}$	$0.5 - \frac{r}{4}$	$\frac{r}{4}$	$\frac{r}{4}$	$0.5 - \frac{r}{4}$	$0.5 - \frac{r}{4}$	$\frac{r}{4}$	$\frac{r}{4}$	$0.5 - \frac{r}{4}$
7	facultative	50	F_2s	$0.25 - \frac{r}{4}$	$0.25 - \frac{r}{4}$	$0.5 - \frac{r}{4}$	$\frac{r}{4}$	$\frac{r}{4}$	$0.5 - \frac{r}{4}$	$0.5 - \frac{r}{4}$	$\frac{r}{4}$	$\frac{r}{4}$	$0.5 - \frac{r}{4}$

Migration and random mating — Migration between the two populations experiencing divergent selection follows a simple two-island model with a migration rate m. For example, the vector of new haplotype frequencies following migration for population 1 is:

$$\mathbf{p}_1^{(new)} = (1 - m)\mathbf{p}_1 + m\mathbf{p}_2 \tag{5.2}$$

Mating is assumed to occur at random amongst the individuals within a given population. The change in haplotype frequency after random mating is:

$$\triangle\, g_{ij} = \pm r D_k \tag{5.3}$$

where r is the recombination rate between loci A and B and D_k is the disequilibrium coefficient ($D_k = g_{11}^{(k)} g_{22}^{(k)} - g_{12}^{(k)} g_{21}^{(k)}$). For the coupling gametes (i.e., $i = j$) the quantity rD_k in Equation 5.3 is subtracted and it is added otherwise.

Viability selection — A matrix describing the fitness values for all zygotes in population 1 is given by the matrix $\mathbf{S}_1$:

$$\mathbf{S}_1 = \begin{pmatrix} 1 & 1 & 1 - sh & 1 - sh \\ 1 & 1 & 1 - sh & 1 - sh \\ 1 - sh & 1 - sh & 1 - s & 1 - s \\ 1 - sh & 1 - sh & 1 - s & 1 - s \end{pmatrix} \tag{5.4}$$

where s and h are the selection and dominance coefficients, respectively. For simplicity in the current model we assume heterozygotes have intermediate fitness between the homozygote genotypes (i.e., $h = 0.5$). In equation 5.4 rows and columns correspond to the elements in $\mathbf{p}_k$. In population 2 the fitness matrix is constructed similarity but the quantity $1 - s$ is replaced with 1 and *vise versa*. The change in haplotype frequencies for each population can be calculated by considering the marginal fitness values for each haplotype. Following Rice (2004), the vector of marginal fitness values is:

$$\mathbf{w}_\mathbf{k}^\star = \mathbf{p}_k^T \mathbf{S}_k \tag{5.5}$$

The change of haplotype frequencies due to viability selection depends on the mean relative fitness of a given population, the current haplotype frequency, and its marginal fitness. The mean relative fitness is the dot product of the haplotype frequencies and their corresponding marginal fitness values:

$$\bar{w}_k = \mathbf{p}_k \cdot \mathbf{w}_\mathbf{k}^{\star\,T} \tag{5.6}$$

Thus, the vector of change of haplotype frequencies after a bout of selection is then:

$$\triangle \mathbf{p}_k = \bar{w}_k^{-1} \left(\mathbf{p}_k \cdot (\mathbf{w}_\mathbf{k}^{\star} - \bar{w}_k)^T \right) \tag{5.7}$$

Numerical methods — Since we were interested in the short-term dynamics of genomic differentiation following secondary contact, we ran each scenario for 500 generations for varying migration rates and strengths of selection and recorded the extent genetic differentiation (F_{ST}, Hartl and Clark, 2007; Hedrick, 2011) at each locus along a simulated chromosome.

5.4 RESULTS

5.4.1 *Genomic differentiation under DGF from secondary contact*

No admixture during secondary contact — Under the evolutionary scenarios in which no admixture during secondary contact occurred (i.e., scenarios 1-3, Table 5.1) we found that the extent of genetic differentiation depended on the relative magnitudes of migration and selection (Figure 5.1). When initial divergence was strong (i.e., scenario #1) the small increases in the migration rate greatly reduced overall differentiation in about 100 generations. Here, under weak to intermediate migration (i.e., $0.001 \geq m \geq 0.01$) and under intermediate to strong selection (i.e., $s \geq 0.05$) genomic differentiation occurred only under tight linkage, consistent with genomic islands. This same general pattern was observed when the initial divergence was weaker (LD=0, X-LD=0.125, scenario #2, Figure D.1) but with less overall differentiation. As expected, when SGV was partitioned completely within populations no differentiation occurred in any migration and selection range (scenario #3, Figure D.2).

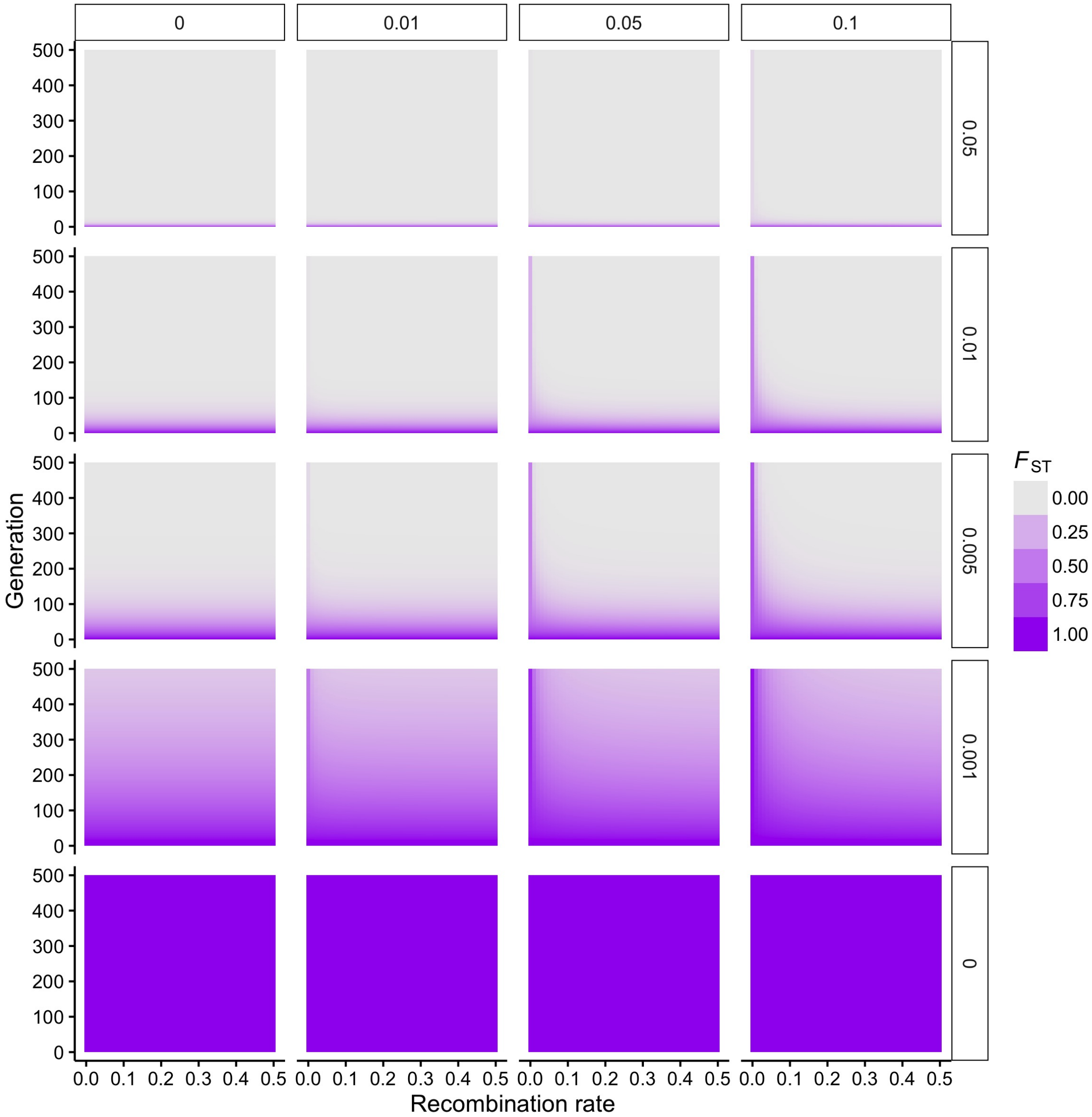


FIGURE 5.1: Dynamics of divergence with gene flow under scenario #1 – obligate sexual, $m_f = 1$, LD=0, X-LD=0.25. For each panel, the extent of divergence (F_{ST}) at neutral loci are given across time. Rows indicate migration rate, m, between diverging populations and columns indicate the strength of divergent selection, s, at selected locus A.

Brief admixture during secondary contact — We found that brief admixture between diverged, locally adapted populations immediately before DGF strongly promoted island formation. Indeed, when DGF was initiated with F_1 individuals – the parents of which were locally adapted to their respective environment – we found that the only divergence that was detected occurred locally within the genome (scenario #4, Figure 5.2). This island pattern was also observed when two rounds of random mating occurred prior to DGF (scenario #5, Figure D.3).

We found a strong effect of mating type on the pattern of genetic differentiation from SGV. As predicted, for obligate sexual mating and when migration occurs periodically (e.g., every 50 generations, scenario #6) selection is relatively strong compared to migration resulting in island formation and persistence even under maximum migration ($m = 0.5$; migration per generation = 0.01). When mating type is facultative, however, the joint contribution of selection and migration can create genome-wide differentiation in addition to islands (Figure 5.4). Here, genome-wide differentiation occurs under strong selection and weak migration.

We identified two pathways in which islands form, depending on the relative strength of selection and migration. First, under strong selection ($s \geq 0.05$) and strong migration ($m \geq 0.2$) islands form from the breakdown of genomic differentiation with time (e.g., upper right panels of Figures 5.3 and 5.4). Second, under weak selection (s=0.01) and weak to moderate migration ($0.01 < m < 0.05$), neutral genetic differentiation began low and increased ("grew") over time (e.g., Figures 5.3 and 5.4). The size (width) of islands differed between the two mating types – with larger islands found in facultative compared to obligate mating types. Interestingly, migration was not need for island growth to occur when admixture occurred during a single bout of secondary contact ($m = 0$, $s = 0.01$, Figures 5.2, D.3, 5.3, and 5.3). This is in stark contrast to scenarios in which no admixture occurred in secondary contact (scenarios 1-3, Figures 5.1, D.1, and 5.1).

5.5 DISCUSSION

5.5.1 *Islands from standing genetic variation*

We found that localized genetic differentiation can readily occur under a wide range of demographic scenarios, depending on the relative strength of migration and divergent selection. Link-

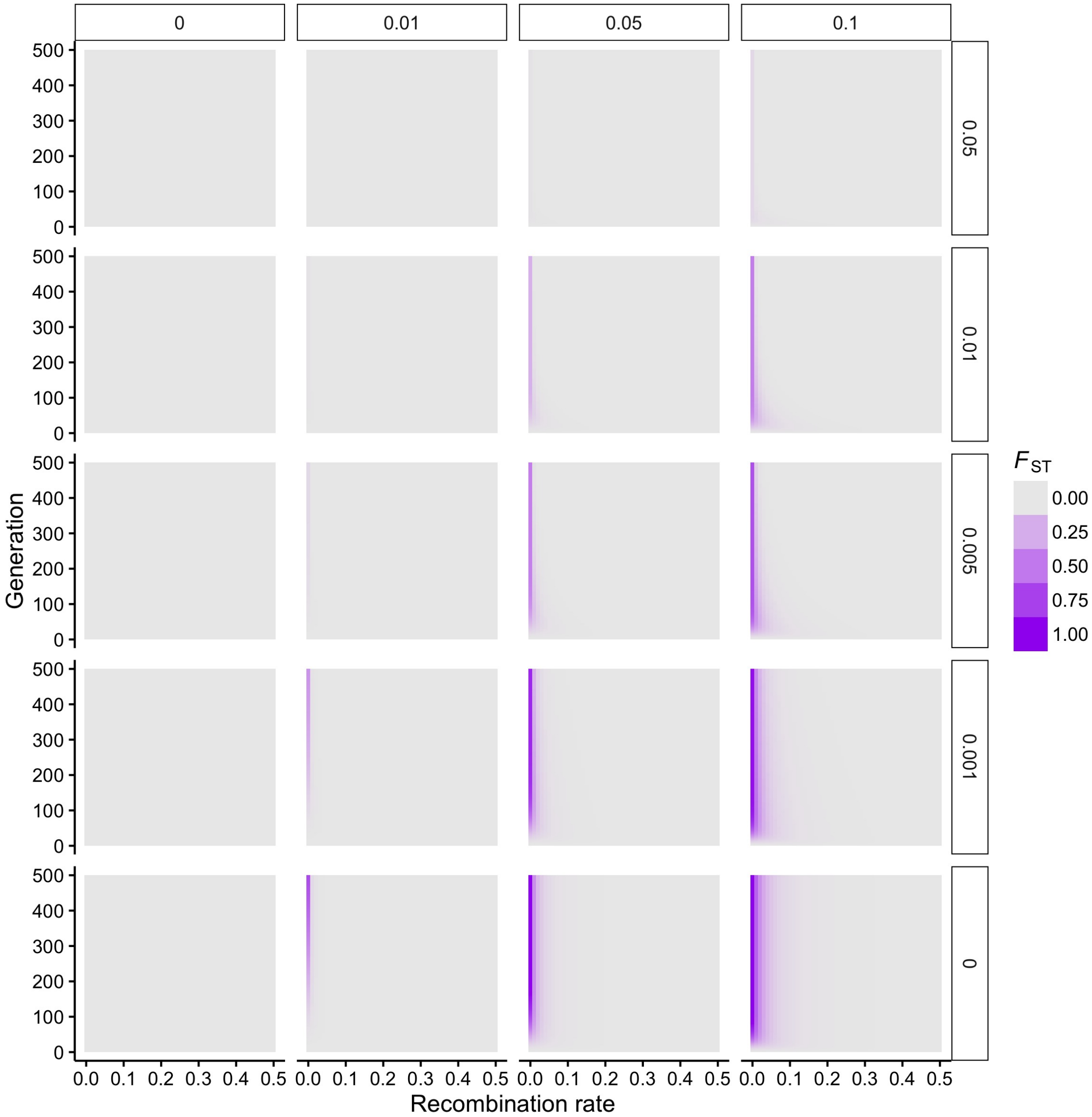


Figure 5.2: Dynamics of divergence with gene flow under scenario #4 – obligate sexual, $m_f = 1$, LD=0.25, X-LD=0.25. For each panel, the extent of divergence (F_{ST}) at neutral loci are given across time. Rows indicate migration rate, m, between diverging populations and columns indicate the strength of divergent selection, s, at selected locus A.

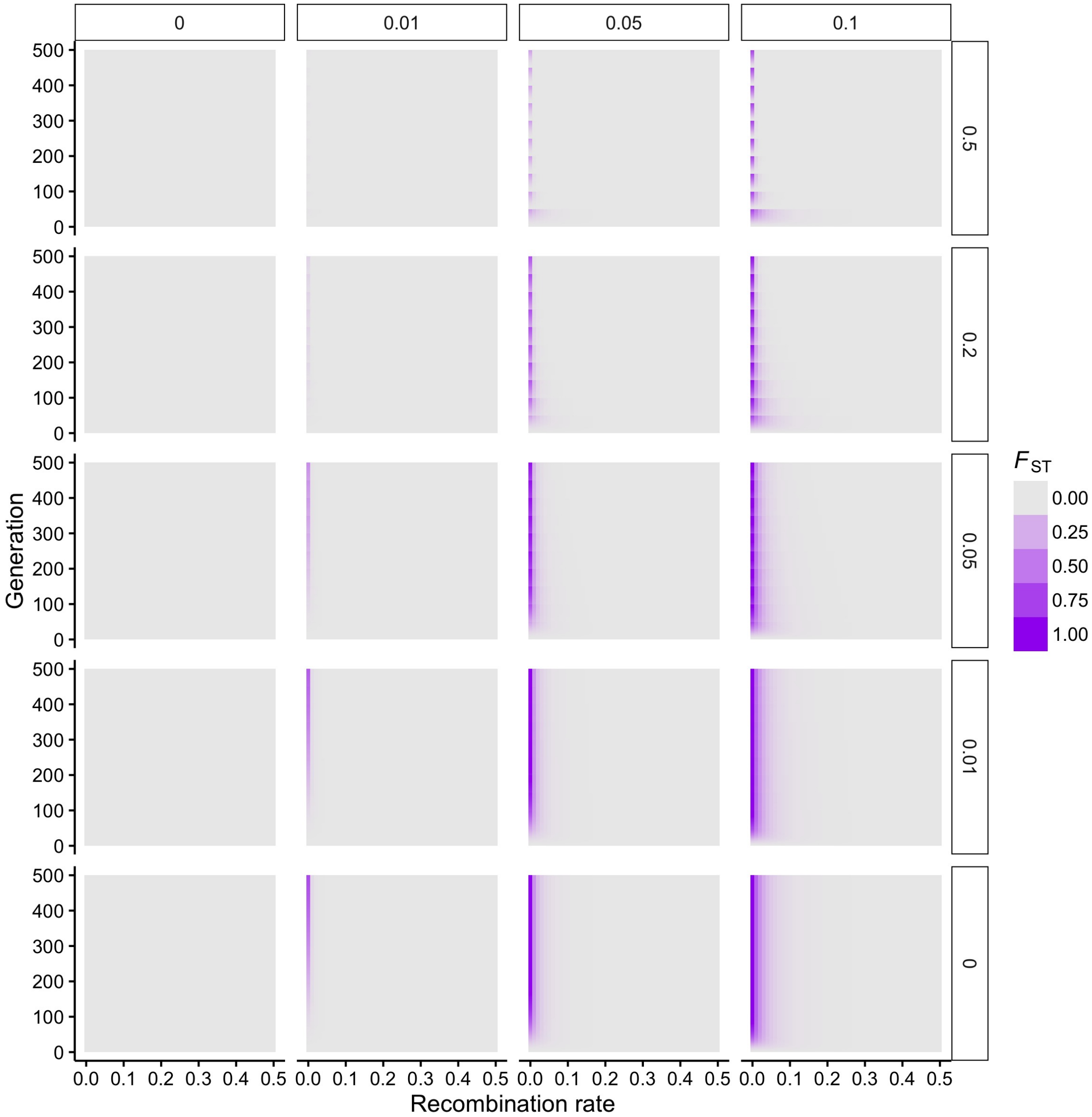


FIGURE 5.3: Dynamics of divergence with gene flow under scenario #6 – obligate sexual, $m_f = 1$, LD=0.25 - 0.25r, X-LD=0.25 - 0.25r. For each panel, the extent of divergence (F_{ST}) at neutral loci are given across time. Rows indicate migration rate, m, between diverging populations and columns indicate the strength of divergent selection, s, at selected locus A.

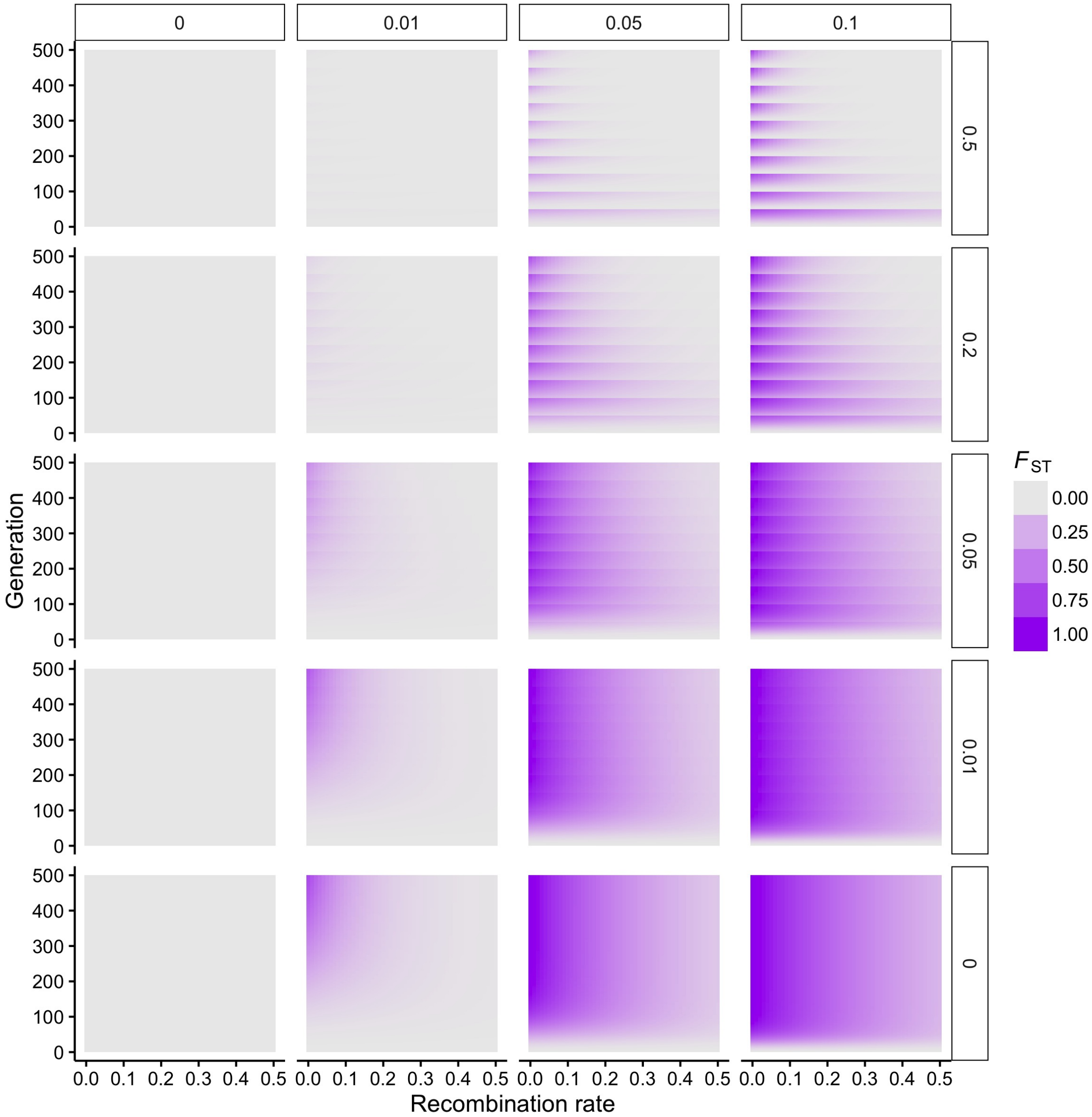


FIGURE 5.4: Dynamics of divergence with gene flow under scenario #7 – facultative sexual, m_f = 50, LD=0.25 - 0.25r, X-LD=0.25 - 0.25r. For each panel, the extent of divergence (F_{ST}) at neutral loci are given across time. Rows indicate migration rate, m, between diverging populations and columns indicate the strength of divergent selection, s, at selected locus A.

age disequilibrium within and between isolated populations can be generated a number of ways prior to the onset of a divergent selection regime. For example, genetic drift can fix alternative alleles between two isolated populations such that there is no LD within but maximum LD between populations. Of course, the fixation of alternative alleles in each isolated population can occur due to preexisting divergent selection on new mutations. In general, the breakdown of linkage disequilibrium under divergence is required for islands to form.

5.5.2 *Islands uplift and islands erode*

Under new mutation theory of island formation, divergent hitchhiking allows for increase establishment probability of new mutations (Yeaman and Otto, 2011) and so islands can "uplift" from the metaphorical sea when seeded with divergently selected loci. We found that such uplifting can also occur from standing genetic variation. An admixture event between genotypically distinct populations creates a high degree of within population LD (Hedrick, 2011). Such a case may occur between hybridizing sister species or through the ephemeral breakdown of a migration barrier. When divergent selection occurs following such an event there are two mechanisms in which islands can form, depending the strength of selection relative to gene flow. During the time in which LD is broken down within a population by random mating, differentiation at both selected and neutral loci increases (though this increase is faster at the selected locus; Figure 5.5E-F). In the case of no migration between populations (e.g., left column of Figure 5.5), neutral differentiation will remain steady since no migration (or mutation) is occurring. Islands can also buildup quickly and erode. For example, under strong divergence with moderate gene flow there is a rapid breakdown of LD early with a slower breakdown of LD later (Figure 5.5D). During the rapid breakdown phase, where the change in haplotype frequencies is dominated by selection and F_{ST} increases with time for both selected and linked neutral sites. During the slow breakdown of LD phase, the change in haplotype frequencies are dominated by migration. Here, differentiation at the selected locus is stable whereas differentiation decreases at the neutral locus (Figure 5.5F) owing to recombination. With tighter (weaker) linkage the decrease of F_{ST} will be slower (higher). Thus, under divergence with gene flow we would expect islands to erode with time.

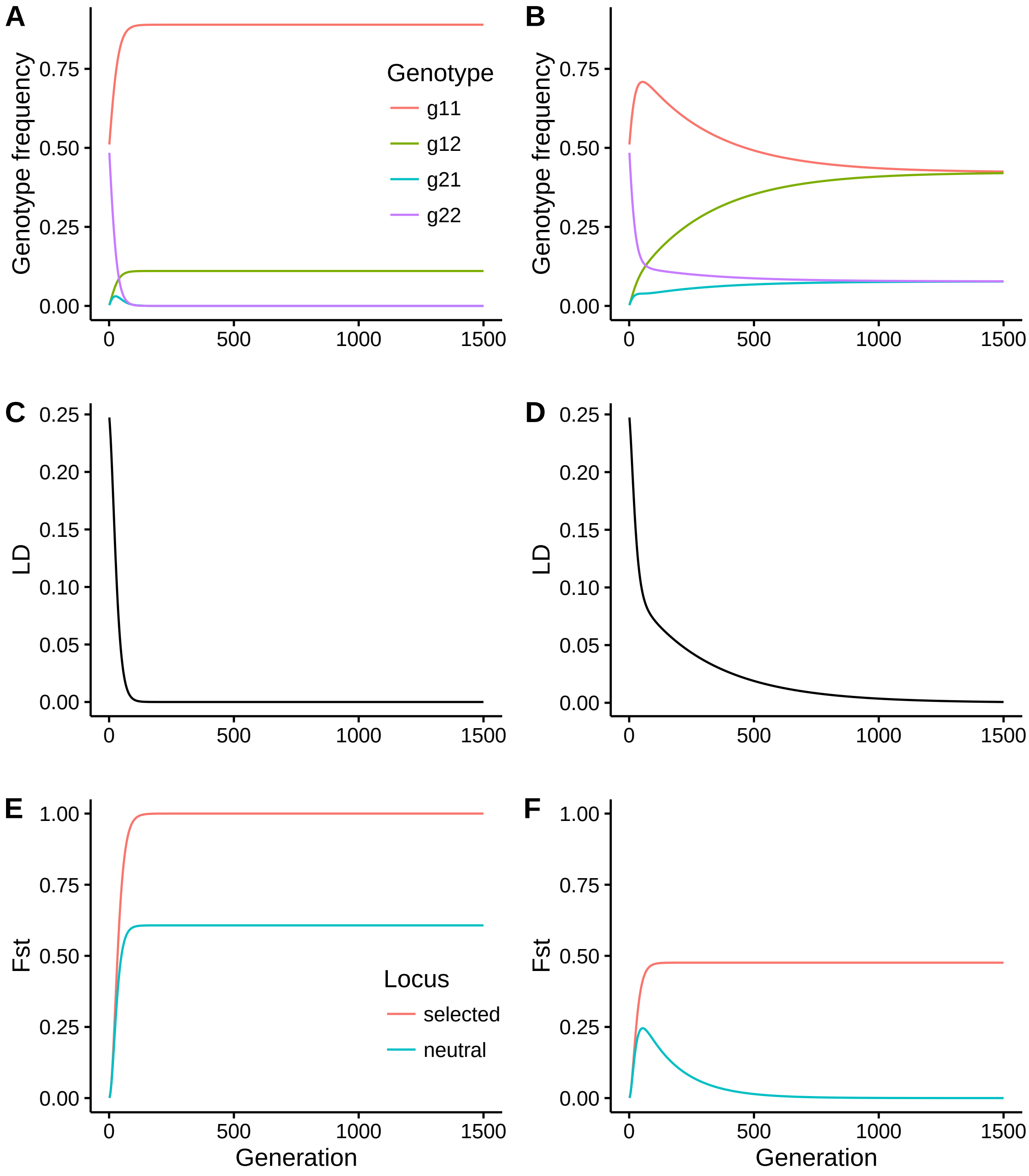


FIGURE 5.5: Temporal dynamics of genotype frequencies, LD, and differentiation at the selected and linked neutral loci (scenario #4). Two specific examples (left and right columns) of island formation are given. For each example we plotted the results for population #1 only so that genotypes g_{11} and g_{12} are favored and g_{22} and g_{21} are disfavored. Left column, migration is absent. Right column, migration is weak ($m = 0.01$). For each condition the selection coefficient was strong ($s = 0.1$) and the recombination rate between the selected and neutral loci was 0.01.

CHAPTER 6

The genomic response to adaptation from standing genetic variation in experimental yeast populations[5]

6.1 SUMMARY

The recent flood of population genomic data has provided exciting new insights and challenged our understanding of how evolution structures genomes. Rather than simply "population genetics with more markers", population genomics promises a transformative shift in our understanding of real-time evolution. For example, markers exhibiting elevated genetic differentiation between populations have traditionally been used to detect loci under divergent selection and genomic data reveal that regions of elevated differentiation often extend across large, physically linked regions of chromosomes. These "genomic islands of divergence" are not simply the predictable result of divergent selection – rather they reflect complex interactions among selection, epistasis, demography, migration, and recombination. Comparative population genomic data exhibit wide diversity in the number, size, nature, and dynamic behavior of islands, presumably reflecting differences in underlying evolutionary processes. Beyond simply detecting such regions, the volume of data produced by next-generation sequencing has the potential to provide the statistical power necessary to test specific hypotheses about how interacting evolutionary forces structure genomic variation. Herein we use experimental evolution to test the overall hypothesis that such genomic islands of divergence can manifest from divergent selection on standing genetic variation. We test this hypothesis by crossing two haploid strains of budding yeast to generate an admixed polymorphic population from their F_2 offspring. We then evolved replicate populations in two different stress environments – sodium dodecyl sulfate and sodium chloride – for 12 days and looked for genomic regions that differentially responded to these environments. We found that genomic islands can readily evolve after an episode of introgression, without further gene flow or *de novo* mutations, which may broadly explain the prevalence of

[5] Manuscript in preparation for submission to Molecular Biology & Evolution

genomic islands seen in nature. Next we tested the efficacy of gene flow in reducing the size and extent of genomic islands. We found that gene flow between locally adapted populations did not correlate with island size. Instead, we found that the genomic response to additional divergence (with gene flow) created a genome-wide and stochastic pattern of divergence.

6.2 INTRODUCTION

Evolutionary biologists have long been interested in identifying genes of adaptive significance, especially in populations experiencing divergent selection pressures. In this spirit several early empirical studies – reviewed in Nosil et al. (2009a) – report loci with 'outlier' status in genome scans (Beaumont and Nichols, 1996). These studies report on the order of 5-10% of the markers surveyed exceed neutral expectations (Via, 2012). More recently, genomic studies have put genetic differentiation in a genomically explicit context and show that such outlier loci can cluster together along chromosomes – creating regions of localized genetic differentiation or "genomic islands of divergence".

Nosil et al. (2009a) summarized the metaphor of genomic islands of divergence. Briefly, when a pair of populations are experiencing divergent selection pressures, gene flow is effectively weaker nearby divergently selected loci. Under these regions of "divergent hitchhiking" *de novo* mutations can further expand the region under divergent selection. Over time these genomic islands can grow and merge with other islands, eventually creating a genome-wide pattern of divergence between pairs of populations or sister species (Wu, 2001).

One question that remains, however, is whether such islands can form from standing genetic variation. All studies that we are aware of look at dynamics and patterns of genomic islands using a retrospective approach (i.e., natural experiments). While this approach has been useful in identifying and cataloging the size, number, and dispersion of genomic islands that occur in nature, it has its drawbacks. Namely, the compounding of evolutionary history over time in these natural experiments can erode the signal of early divergence, making it difficult to genomically reassemble the genetic roots of adaptation to divergent environments (Nadeau et al., 2012). Therefore, while the ultimate source of new genetic variation is new mutations, is it necessary that the heterogeneous nature of genetic differentiation is mainly a result of *de novo* mutations occurring at loci experiencing divergent hitchhiking or can these patterns be explained from standing genetic variation? In Chapter 5 we used a simple diploid, two-locus model and found that genomic islands indeed can assemble quickly from standing variation when recently admixed populations diverge.

Selection can be multifarious, acting across genetically independent traits and among loci across the genome (Nosil et al., 2009b). Empirical studies have cataloged extensive genomic island patterning (Turner et al., 2005; Via and West, 2008; Nadeau et al., 2012; Hohenlohe et al., 2012) as a result of selection acting in concert with migration, drift, and other evolutionary processes. The biggest limitation to these empirical studies, however, is that they necessarily only take snapshots of genomic islands at a point in time after divergent selection has taken place, and so it is difficult to assess whether standing genetic variation or new mutations were involved in island formation. At the same time, our previous mathematical models showed that island formation is possible from standing genetic variation but ignores complex selection and drift.

One approach to test if genomic islands can form from standing genetic variation is to experimentally evolve recently admixed populations in divergent environments. The yeast *Saccharomyces cerevisiae* can be grown in a range of environmental conditions, have a short generation time, and populations can be frozen and later revived for direct comparison to their ancestors or sequenced to create a catalog of genomic differentiation through time. Yeast can reproduce asexually as diploids or haploids, haploids can be crossed to form diploids, and diploids can be induced to sporulate, which means a single diploid cell undergoes meiosis to produce four haploid spores. Genetic manipulation in yeast is common (Scannell et al., 2011). Thus, crossing two genetically diverse strains to create standing genetic variation and maintaining these population as diploids via genetic engineering can be accomplished relatively easily. Moreover, artificial selection combined with high-throughput sequencing has been applied successfully for association and QTL mapping, so the identification of loci responsible for phenotypic divergence has become routine as well (Ehrenreich et al., 2010; Parts et al., 2011). One approach that is particularly useful in identifying loci of adaptive significance in yeast is extreme QTL (X-QTL, Ehrenreich et al., 2010). In X-QTL, parental strains are crossed and aliquots of progeny are selected for a period of time in either a stress medium or in a control medium. Comparing the allele frequency differences between treatment and control allows for the identification of both large and small effect QTL associated with tolerance to the stress. A direct but unexplored extension of this approach is to evolve aliquots in two different stress environments to tease apart genomic regions under directional selection and those under divergent selection.

While the above attributes make yeast an ideal system in which to study the genomic response to divergent adaptation, it is necessary to take their life history into account. *Saccharomyces cerevisiae* is facultatively sexual. Indeed, clonal reproduction occurs about 25,000 - 35,000 times more frequently than sexual reproduction (Magwene et al., 2011), though sexual reproduction can certainly occur more frequently in the laboratory (e.g., Nishant et al., 2010). Thus, while most models of divergence with gene flow consider obligate sexual organisms, the quick generation time of yeast (*ca.* 100 minutes; Herskowitz, 1988) prohibits them from evolving solely in this manner. In our previous models (Chapter 5) we found that the strength of selection is enhanced relative to migration for facultative sexual organisms such as yeast that mate and migrate periodically (Figure 5.4).

The genomic island metaphor holds promise to integrate many evolutionary processes that act in natural populations. However, it is unclear whether new mutations or standing genetic variation drive genomic island growth more commonly in natural populations. Herein we experimentally evolved polymorphic yeast populations to test the hypothesis that genomic islands can quickly form from standing genetic variation. We performed two complementary experiments to test this hypothesis. First, we created standing genetic variation by crossing two diverged strains and evolved replicate F_2 populations in isolation in alternative stress environments: SDS or NaCl supplemented media. We measured the evolutionary response to growth in these alternative environments by quantifying the degree of adaptation and the genomic pattern of differentiation between replicates and 1) their ancestors and 2) replicates grown in alternative stress environments. Second, we tested the efficacy of migration and recombination to reduce the extent and size of genomic islands from locally adapted pairs of populations. Here we tested 4 levels of gene flow. We evolved populations without migration and either with or without mating. Sex in yeast in harsh environments has been shown to increase the rate of adaptation (Goddard et al., 2005) and so we predicted that mating would have a measurable effect on island size. Also, we evolved populations with mating and either intermediate (0.2) or high (0.5) migration between populations grown in alternative environments and tested the prediction that island size decreases with increasing migration rate.

6.3 METHODS

6.3.1 *Yeast strain crosses, media & methods*

Crossing strains — Details of the yeast strain construction and media used have been described in detail in Chapter 4. Briefly, to create an admixed ancestral population we crossed two heterothallic and haploid *S. cerevisiae* strains: YPS128 (mat α, ho△::Hyg, ura3△::KanMX; NCYC# 3632) and DBVPG1106 (mat **a**, ho△::Hyg, ura3△::KanMX; NCYC# 3596). See Cubillos et al. (2009) for initial strain construction. Previous work shows that these two strains have 73,294 high quality SNPs between them (see Chapter 4).

The presence of haploids during divergent selection represents a confounding factor since these haploids may outcompete the diploids under certain conditions (Zeyl, 2006; Otto and Gerstein, 2008) yet not survive the sporulation procedure (see below). Thus, to promote diploid growth (i.e., select for diploids) we disrupted His3 in strain DBVPG1106 and replaced it with a functional copy of URA3 from an amplicon from plasmid YEp24 (Struhl et al., 1979) using the protocol from Gietz and Schiestl (2008). We then mated a single transformant with YPS128 described above. The resulting F_1 progeny are heterozygous at the His3 locus for both HIS3 and URA3 alleles and so can successfully grow on Histidine-Uracil double dropout medium. Proper placement of homologous recombination was confirmed via PCR (see Table 6.1 for recombineering and confirmation oligos). We sporulated a single F_1 individual using the sporulation protocol below and randomly mated recombinant spores together to create an admixed F_2 population. Specifically, we spread 100 μL of washed (YPD) spore solution and mated the spores onto YPD plates for 36-48 h (30C). Using a toothpick, we swabbed cells from each YPD plate and inoculated them in 2 mL of Ura-His dropout medium incubated on a roller drum at 30C. Only the diploid F_2 recombinants are expected to be heterozygous at the His3 locus (URA3/HIS3). This admixed F_2 population was used to seed the initial divergence (see below).

Table 6.1: Oligos used for the construction of the deletion cassette and the confirmation of correct integration into the genomic His3 locus of the target DBVPG1106 strain.

Step	Primer	Template	Sequence (5' → 3')
Deletion cassette	Forward	YEp24	AAATGAGCAGGCAAGATAAACGAAGGCAAAGATGGATGTCCACGAGGTCTCTAATTCCGGGCATGCGGCCTTAGTAACAAAAGAGTGGTA
Deletion cassette	Reverse	YEp24	CGTATGCTGCAGCTTTAAATAATCGGTGTCACTACGGTGTCGGTCTCGTAGAAACCTTGGCCGCTCGGTTCCCGATGCGTCCGGCGTAGA
Confirmation PCR	Forward	gDNA	TTCCACCTAGCGGATGACTC
Confirmation PCR	Reverse	gDNA	TGATGCATTACCTTGTCATCTTC

Sporulation — During the initial crossing of F_2 ancestors and during the course of the evolution experiment we induced meiosis to produce haploid spores. We used the following protocol to sporulate diploid cells. For each population we harvested 1 mL of cells grown in standard rich medium (Yeast Peptone Dextrose, YPD; Sherman, 2002), washed them with 1 mL of presporulation medium (Yeast Peptone Acetate; Codón et al., 1995), and we incubated them in 1 mL YPA for 12-15 hours. Next we washed cells in SPO2 sporulation medium (2% KAc, pH 7; Codón et al., 1995) and incubated cells in 1 mL of SPO2 in a roller drum. Unsporulated cells were killed off using heat and enzymatic perturbation. Asci were removed using 40 μL of Zymolyase Solution (1 U/μL Zymolyase [MP Biomedicals] in 1 M sorbitol). To recover spores from sporulation we allowed them to grow on YPD plates for 2 days. Next we transferred 1 swab of cells to 2 mL of Histidine-Uracil double dropout medium to enrich for diploids.

6.3.2 *Experimental evolution*

We conducted two evolution experiments. In "Experiment I" we allowed admixed (F_2) populations to diverge from one another in isolation. This initial divergence was done to select for alternative alleles between the SDS and NaCl environments. In "Experiment II" we split populations at the end time point of Experiment I into 4 different migration treatments and evolved them under a divergence with gene flow scenario. A schematic of the experimental design is presented in Figure 6.1 and detailed below.

Experiment i, admixture followed by isolation — We aliquoted the F_2 ancestral pool into 6 populations. We subjected three of these populations to YPD containing 0.04% SDS in daily (batch) transfers. The remaining three populations were grown in batch with YPD with 3% NaCl. Each population was grown for 23 hours in 100 μL reactions in a 96-well microtiter plate with continuous shaking. Following incubation, cells were washed twice with water and resuspended with YPD. A total of 5 μL of resuspended cells were added to 95 μL of their respective (matched) stress medium as well as 95 μL of the opposite (unmatched) stress medium. Population growth in the unmatched medium was only used to assess local adaptation (see below) and so these cells were not propagated further. Approximately every 6 minutes during the shaking incubation the optical density (600nm) was automatically measured for all samples,

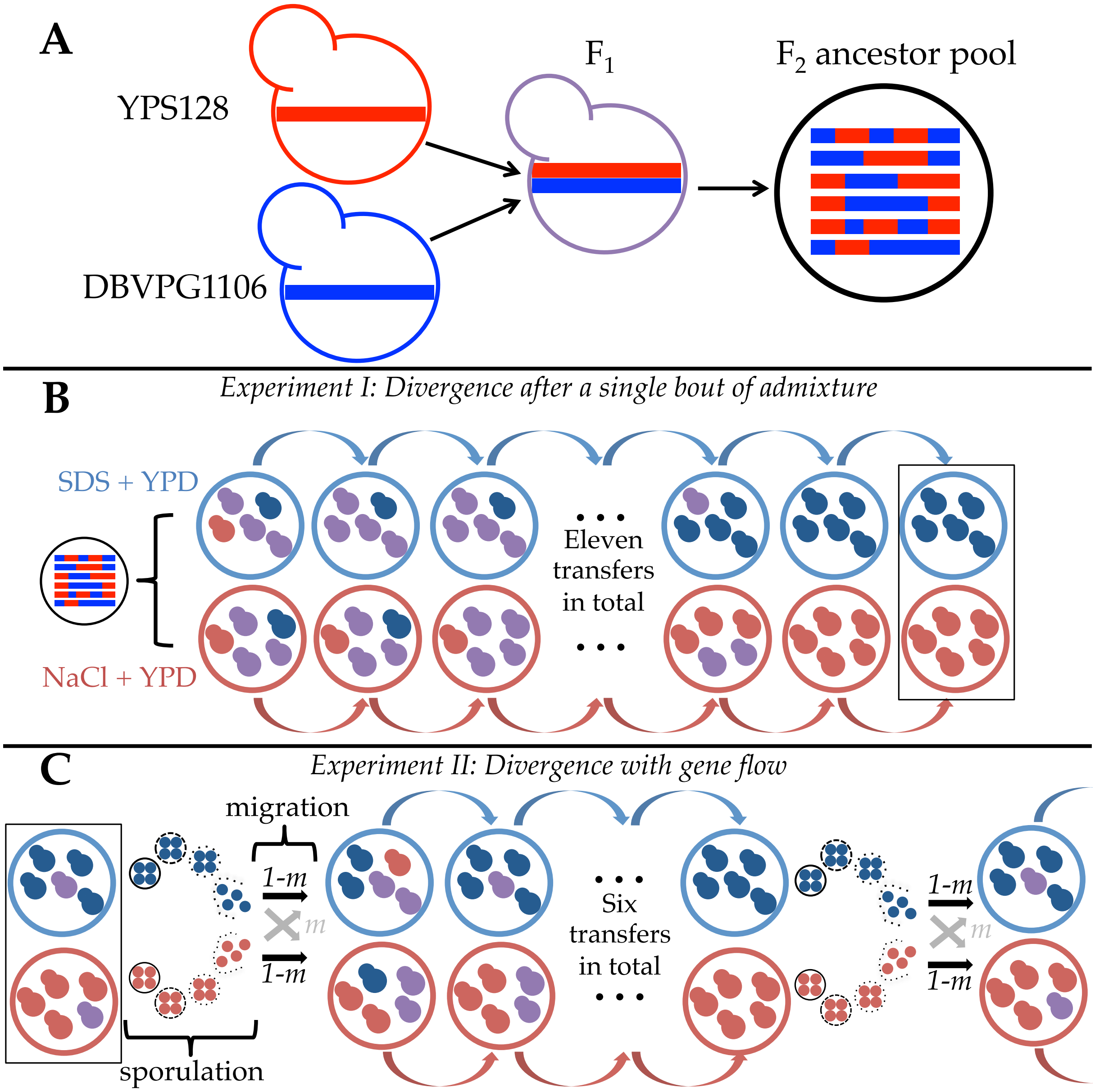


Figure 6.1: Schematic of the experimental design used in the current study. A) The starting F_2 recombinant pool was constructed by mating haploid YPS128 and DBVPG1106 strains and sporulating a single F_1 zygote. B) This F_2 ancestral pool was used to seed Experiment I in which 3 replicate populations (only one of which is diagrammed) were grown in SDS stress (blue circles) and 3 replicates were grown in NaCl stress (red circles) and propagated in batch (curved arrows). This propagation occurred 11 times during Experiment I. C) The last time point of Experiment I – indicated by black rectangle – was used to seed Experiment II. Here we only show one "cycle" of divergence with gene flow from one replicate population pair. Within this cycle populations are sporulated and the resulting spores are migrated at one of three levels (m = 0, 0.2, & 0.5) and allowed to mate (except for the no migration, no sporulation treatment). Resulting diploids are propagated in batch for 7 days. This cycle of divergence, sporulation, and gene flow occurred a total of 4 times.

giving daily growth curves for each population in its local and in its foreign stress environment. We performed 11 transfers (12 days total) of growth in isolation so that local adaptation could evolve.

Experiment II, divergence with gene flow — We also evolved populations under a divergence with gene flow (i.e., multiple bouts of admixture) scenario. For this experiment we continued to evolve populations at the end time points of Experiment I. Specifically, the 6 populations from the end time point of Experiment I were split into 4 different treatments: 1) no migration & no mating 2) no migration with mating (within populations) 3) migration at rate $m = 0.2$ with mating and 4) migration at rate $m = 0.5$ with mating. In each treatment and each replicate (n=3) we grouped one SDS evolved population with one NaCl evolved population, resulting in 3 population pairs in each migration treatment. Based on simulated results (Chapter 5) we tested for genomic island formation under 3 migration levels (m = 0, 0.2, & 0.5). The divergence with gene flow scheme used here consisted of daily batch transfers in selective media with periodic migration. We grew populations asexually in 100 μL of their matched and unmatched media under daily transfers for 7 days (i.e., about 50 generations of asexual growth). As with Experiment I above, the unmatched populations were only used to assess local adaptation. To better estimate growth curves we performed two "technical" replicates of each population and report on the average growth curve for each condition. For the matched populations we sporulated diploid individuals. We standardized all sporulated populations to the same concentration (*ca.* $7.5x10^6$/mL) and mixed spores between population pairs at the experimental varied migration level. We selected for diploids as described above. After growing in Ura-His dropout medium for 24 h we transferred 5 μL of washed (YPD) cells to 95 μL matched and unmatched stress medium. We performed 4 of these selection/gene flow cycles for a total of 28 asexual days (*ca.* 200 asexual generations). Interestingly, newly formed diploid zygotes grew poorly – or not at all – in 0.04% SDS which cause poor estimates of growth curves from well to well and between technical replicates (results not shown). Thus, for Experiment II we reduced the selective concentration of SDS by 25% (0.03% SDS).

6.3.3 *Data analysis*

Measuring initial change in fitness — We tested whether and how populations of F_2 individuals responded to a given stress environment over time during the initial 12 days of batch transfer in isolation (Experiment I). For each day we fit growth curves (optical density, OD, at 600nm) for each of the six populations (3 evolving in SDS and 3 evolving in NaCl) using a generalized logistic (Richards) curve in the R package *grofit* (Kahm et al., 2010). Like the logistic growth equation, the Richards curve increases monotonically with time but the latter is more flexible since the maximum growth rate can occur anywhere between the lower and upper asymptotes (Birch, 1999). This flexibility is accomplished via a shape parameter, v. When v is 1, the curve resembles the logistic and as v increases the maximum growth rate is found closer to the upper asymptote. This latter case more closely resembled our data – especially when populations were tested in the SDS environment.

To measure population growth we fit Richards curves to the data using *grofit* and extracted the four model parameters: lag, rate, shape and carrying capacity (see Figure 6.2A). Prior to fitting the data we zeroed optical densities such that the first measurement was set to zero. Additionally, to measure overall population growth in a given environment we took the integral of the model fit. This area under curve (AUC) metric was used as a single estimate that encompasses all four parameters of the model. For each metric, we averaged the values between the two technical replicates to mitigate well-to-well variation.

To test if these growth curve metrics changed with time or between stress environments during initial divergence we used a repeated measures ANOVA on each of the 4 growth metrics and their composite statistic, AUC.

Measuring local adaptation — At a given time in Experiment I and Experiment II there are two values for each metric and for each environmental patch – one for the local population and one for the foreign population. Our measure of local adaptation follows the "local-foreign" criterion (Kawecki and Ebert, 2004). For example, to estimate local adaptation in AUC we subtracted the AUC value of the foreigners from that of the locals. In this way, values greater than zero indicate local adaptation.

Testing for differences in local adaptation — We tested for differences in local adaptation between stress environments (Experiment I and Experiment II) and between migration treatments (Experiment II only). To test if local adaptation increased over time in Experiment I we used a repeated measures ANOVA for each growth metric as above. Additionally, to test for differences at the final time point of each experiment we fit two separate MANOVAs: one for the last time point of each experiment. The dependent variables were local adaptation metrics for each of the four Richards curve parameter estimates. For Experiment I we fit a one-way MANOVA with the independent variable being environment tested in (2 levels). For Experiment II, we added an additional independent variable: migration with three levels. We also tested for an interaction of environment-by-migration on the multivariate response to local adaptation.

Sequence preparation — To identify the genomic response to adaptation, we sequenced Restriction-site Associated DNA (RAD) markers in 25 populations: the T-11 F_2 ancestor, the 6 populations at the end of time point T0 (i.e., end of Experiment I) and the 18 populations at the end of time point T28 (i.e., end of Experiment II). We used the general RADseq methodology of Chapter 4 to estimate genome-wide differentiation during the final time point of each population.

Bioinformatic pipeline — To process RADseq data we used the general method described in Chapter 4. Broadly, this pipeline preprocesses raw paired-end fasta files, demultiplexes samples (i.e., individual populations), maps each to the reference *S. cerevisiae* genome, and estimates allele frequencies of the previously identified SNP sequences.

We used the following methods to process the raw data. We ran the raw paired-end data through a custom perl script to remove any reads that lacked any sample barcodes. This script also flipped the forward and reverse read if the barcode was detected on the paired-end read, as is expected given our RADseq library preparation (see above). This flipping was necessary to ensure reads were processed correctly in downstream programs that expect the barcode to occur on the forward read. Lastly, this script parsed out *nsiI* and *pstI* RAD sites into separate paired-end fasta files. Since these two restriction enzymes differ by only a single basepair, inclusion of both of them into a single *process_radtags* run resulted in poor quality scores of reads (data not shown).

To call alleles at predetermined SNPs we first used *process_radtags* in the *Stacks* program (Catchen et al., 2013) on each enzyme separately. This procedure was done to demultiplex samples based on barcode. We then concatenated the *process_radtags* results for each sample into a single pair of forward and reverse fasta files and used *flash* (Magoč and Salzberg, 2011) to merge any pairs that overlapped by 10 bp or more. To obtain reliable estimates of change in allele frequencies we targeted high sequencing coverage. The RAD protocol we used randomly shears the end opposite of the restriction site which results in SNPs in these regions having lower sequence coverage compared to regions nearby the restriction site. Thus to obtain higher overall coverage of SNPs we mapped only the forward or flashed reads to the s288c reference genome (Cherry et al., 2012) using *bowtie2* (Langmead and Salzberg, 2012). We then merged and sorted the forward and flashed mapping results with *samtools*. In a previous study we identified a list of >73,000 high quality diagnostic SNPs between haploid strains YPS128 and DBVPG1106 and we filtered the mapping data with this SNP list. Such filter effectively ignores any variation introduced by new mutations and previously fixed positions.

We had three classes of pairwise comparisons for which we estimated per-SNP genetic differentiation. First, we compared allelic change at replicates at the end of Experiment I with their F_2 ancestor (6 comparisons). Second, we compared the three population pairs at the end of Experiment I (3 comparisons). Third, we compared 3 replicate population pairs in each of the migration treatments at the end of Experiment II (12 comparisons). For each of these 21 pairwise comparisons, we used *mpileup* (Li et al., 2009) and *PoPoolation2* (Kofler et al., 2011) to obtain counts of alleles at each of the predetermined SNPs. For each dataset, we further filtered to remove any SNPs with less than 30 alleles in either population using a custom python script. To estimate F_{ST} at each SNP position we used *PoPoolution2* with the following parameters: minimum minor allele frequency, 1%; pool size, 1000; maximum coverage, 5000; window size, 1; step size, 1. We estimated F_{ST} following Hartl and Clark (2007).

The results of the above pipeline yield values of F_{ST} across each of the 21 pairwise comparisons. We used the following methods to characterize the genomic landscape of differentiation in our yeast data and to test the effect of migration on the change of size of islands. First, because F_{ST}

is a relative measure of divergence, two populations can appear differentiated in the presence of directional selection that is in the same direction but with different magnitudes. Such cases might occur when populations are adapting to laboratory conditions that are not associated with SDS or NaCl tolerance (e.g., adaptation to specific YPD, aeration, or temperature). Since we were only interested in divergent loci, we subset the total number of SNPs to include only those that showed evidence of a divergent selective response between the two stress environments. Here we implemented a novel extension of the X-QTL method (Ehrenreich et al., 2010). Specifically, we identified divergent loci from Experiment I by the following criterion. First, all three replicate populations evolving in SDS needed to exhibit identical response to selection (e.g., all show an increase in the YPS128 allele frequency relative to the F_2 source population). Second, and similarly, all three replicates of the NaCl evolved treatment needed to have identical selective response. Loci under parallel directional selection occur when all replicates (in both environments) show identical responses. This might occur for loci responding to general laboratory conditions. Loci under divergent selection can be inferred by opposite and repeatable responses to each treatment. The chances of each condition (directional or divergent) happening by chance is low (*ca.* 3% for each unlinked locus). We made an additional filter such that we discard SNPs in which the average magnitude of differences between treatments was less than 15% unless they occurred within 43.27 kbp of a SNP with a larger difference. This size was chosen based on the median tract length of a non-recombining chromosome (see Table 4.2 in Chapter 4).

Within a given pairwise population comparison we estimated island "clumpiness" by the average (within chromosomes) Moran's I at a 1 Kbp distance lag. When nearby locations of the genome are more similar in F_{ST} than by chance, they are autocorrelated (Moran's $I > 0$) and the size of their correlation is captured by the magnitude of I. Additionally, within each chromosome we estimated the distance at which autocorrelation is expected to be 0; this "neighborhood size" can be an indication of the width of islands when they are present. We used these two statistics to test the hypothesis that increased migration breaks down genomic islands. Finally, to test the similarity of genomic response between replicates with a given treatment or across treatments we estimate Pearson's correlation coefficient for each pairwise population combination.

6.4 RESULTS

6.4.1 *Divergence following admixture in yeast populations*

Comparing individual growth curve parameters across time revealed a difference in how populations initially adapted to the stress environments. In general, SDS evolved populations showed the greatest change over the initial 12 days (Figure 6.2) and these changes were most pronounced in lag, carrying capacity, shape, and AUC. We detected a change in population growth over time in all growth metrics that we analyzed using a two-way repeated measures ANOVA (alpha value = 0.05). Moreover, we detected significant interaction between NaCl and SDS stress environments and time for lag (P<0.0034), carrying capacity (P<0.00026), shape (P<0.0017), and AUC (P<7.6e-05). No environment-by-time interaction was observed for growth rate (P<0.5436; Figure 6.2C) though populations in both stress environments showed an increase in growth rate with time.

We estimated the genomic response to growth in each environment following 11 days of batch transfer by comparing evolved replicates to their F_2 ancestors. We found that our RADseq method yielded a high overall alignment rate (mean=92.9%, SD=4.1) with high average coverage per SNP (Table 6.2). Within a given treatment, selection responses were highly similar across replicates (Figure 6.3) yet there was little correlation between replicates in alternative environments (Figure 6.4). Using short read (75bp) sequencing of RAD loci we obtained highly confident allele frequencies estimates for 2,933 SNPs across the genome (about 4% of the total SNPs identified using whole genome sequencing). Of those SNPs, 420 met our criteria for divergent selection (purple regions in Figure 6.3B) and a cutoff of 15% or more extreme difference (15% cutoff value for allelic difference = 0.145). Taking into account SNPs nearby given the previous estimate of recombination blocks for F_2s we recovered an additional 738 SNPs to have a total of 1,158 SNPs for which island size could be estimated.

The magnitude and location of individual F_{ST} peaks exhibited both high within-environment and low cross-environment repeatability (Figures 6.4 & 6.5). In agreement with the phenotypic data above, SDS evolved replicates showed the greatest degree of genetic differentiation throughout the genome with multiple moderately sized (e.g., $F_{ST} > 0.2$) peaks identified and a larger average extent of autocorrelation in the first distance class compared to NaCl evolved replicates (Figure 6.4). We identified several peaks (e.g., in chromosome IV, X, & XV) that exhibited high

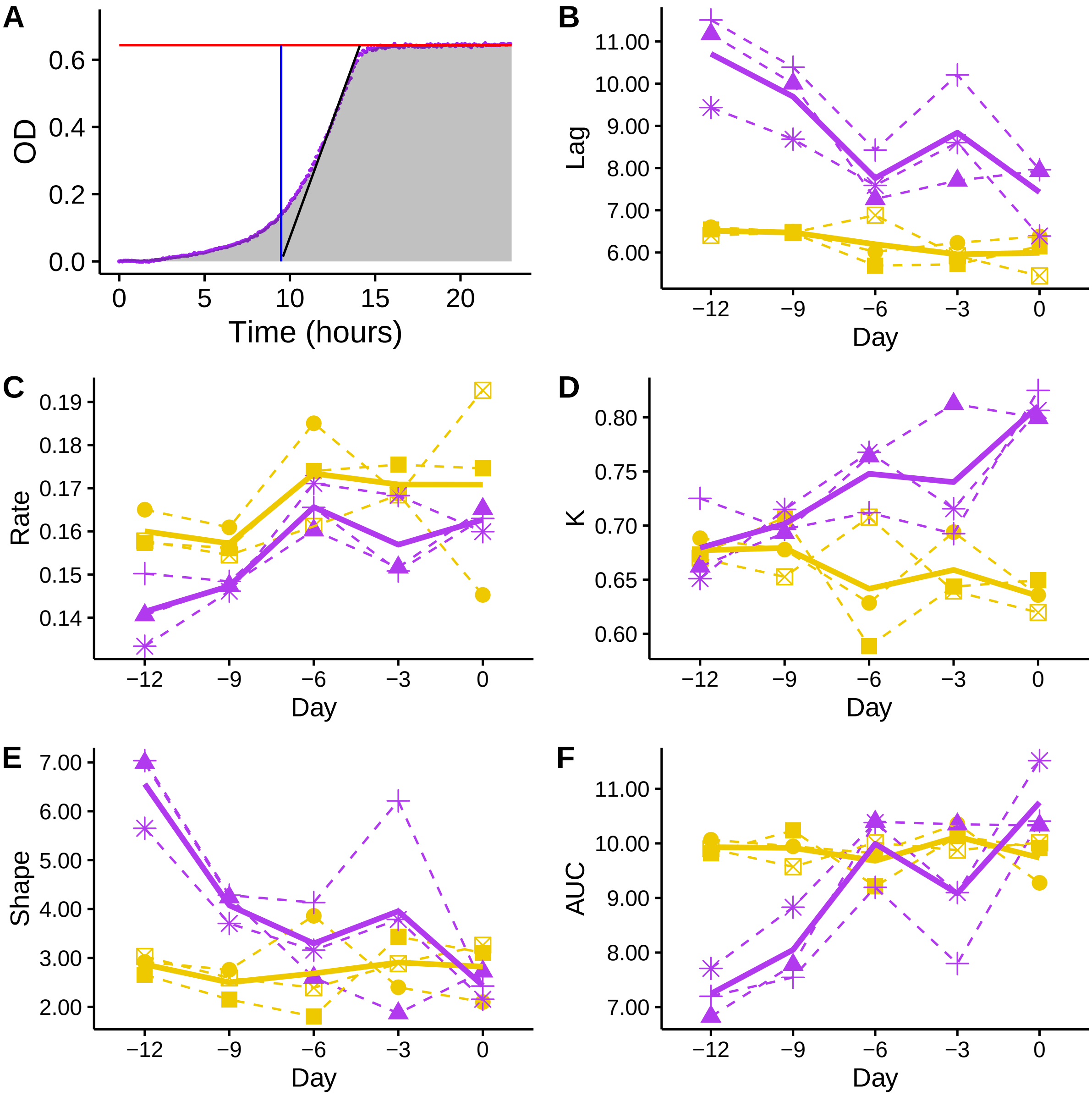


FIGURE 6.2: Raw growth curve parameter estimates through time during the initial build up of adaptation in isolation. Panel A shows a representative growth curve over time (purple dots). Here, the carrying capacity (red line, units = OD_{600}), rate (black line, units = $\triangle OD_{600} / \triangle \, time$, and lag (blue line, units = hours) are indicated. The combined statistic, AUC, is indicated by the grey shading. Panels B-F show the four parameters used in the Richards curve and their composite statistic (AUC). Within each panel populations (dashed lines) are plotted through time with the group mean (solid lines) given. Purple and gold lines denote populations evolving in NaCl and SDS, respectively.

Table 6.2: Average coverage per SNP for all 21 experimental pairwise populations. For each pairwise comparison the experimental conditions are given along with the number of SNPs (after filtering) and the allelic density (i.e., average number of alleles per SNP) for each population.

Comparison	Experiment	Migration	Replicate	SNPs used	Pop 1	Pop 1 coverage	Pop 2	Pop 2 coverage
1	I	-	1	4313	F_2	710.6	T0 (NaCl)	411.8
2	I	-	2	4167	F_2	726.9	T0 (NaCl)	147.9
3	I	-	3	4381	F_2	702.7	T0 (NaCl)	501.1
4	I	-	1	4540	F_2	683.7	T0 (SDS)	782.3
5	I	-	2	4362	F_2	704.1	T0 (SDS)	436.1
6	I	-	3	3455	F_2	801.8	T0 (SDS)	361.3
7	I	-	1	4330	T0 (SDS)	811.2	T0 (NaCl)	410.8
8	I	-	2	4132	T0 (SDS)	450.8	T0 (NaCl)	148.3
9	I	-	3	3410	T0 (SDS)	364.5	T0 (NaCl)	521.2
10	II	0 (no mating)	1	4133	T28 (SDS)	602.2	T28 (NaCl)	531.3
11	II	0 (no mating)	2	3592	T28 (SDS)	324.7	T28 (NaCl)	517.8
12	II	0 (no mating)	3	4054	T28 (SDS)	447.3	T28 (NaCl)	581.5
13	II	0 (mating)	1	4141	T28 (SDS)	528.9	T28 (NaCl)	440.5
14	II	0 (mating)	2	3362	T28 (SDS)	683.3	T28 (NaCl)	134
15	II	0 (mating)	3	3582	T28 (SDS)	439.5	T28 (NaCl)	729.1
16	II	0.2	1	838	T28 (SDS)	176.3	T28 (NaCl)	35.1
17	II	0.2	2	1531	T28 (SDS)	234.6	T28 (NaCl)	196.9
18	II	0.2	3	3951	T28 (SDS)	162.7	T28 (NaCl)	520.4
19	II	0.5	1	4175	T28 (SDS)	622.4	T28 (NaCl)	740.5
20	II	0.5	2	4296	T28 (SDS)	434.6	T28 (NaCl)	486.6
21	II	0.5	3	4279	T28 (SDS)	774.1	T28 (NaCl)	421.2

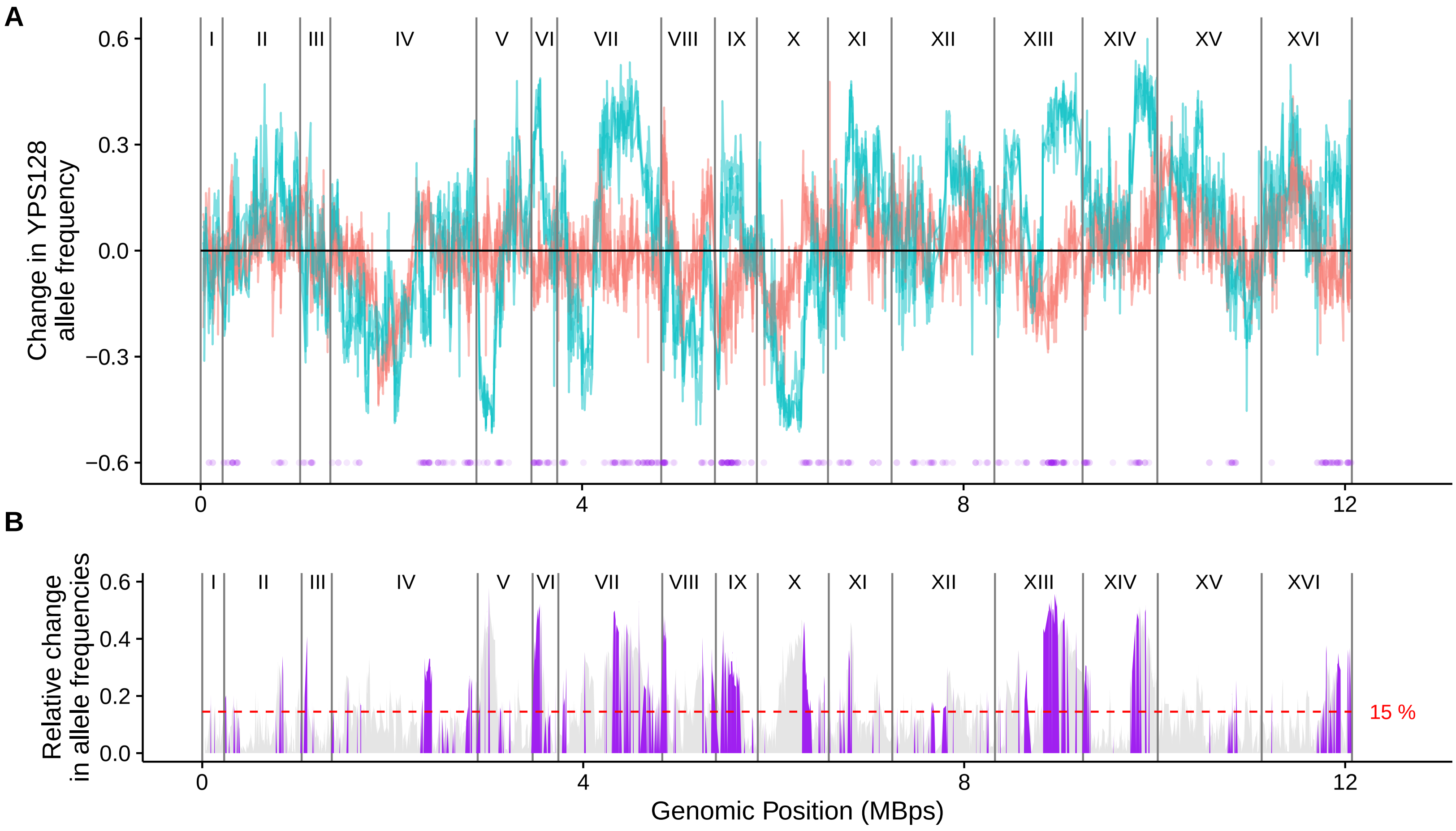


FIGURE 6.3: The genomic response to selection for each environment after 12 days of asexual growth of individuals admixed from YPS128 and DBVPG1106 ancestry. A) For each replicate population the change in allele frequency of the YPS128 allele relative to the F_2 frequency estimate is plotted. Values greater (less) than 0 indicate an increase (decrease) of the YPS128 specific allele for a given SNP-replicate combination. The three blue and three red lines are for replicates evolving in either SDS or NaCl, respectively. B) The average relative difference in allele frequencies. Regions in which there was a repeatable and opposite response between populations grown in different environments are indicated as purple dots (panel A) or purple shading (panel B).

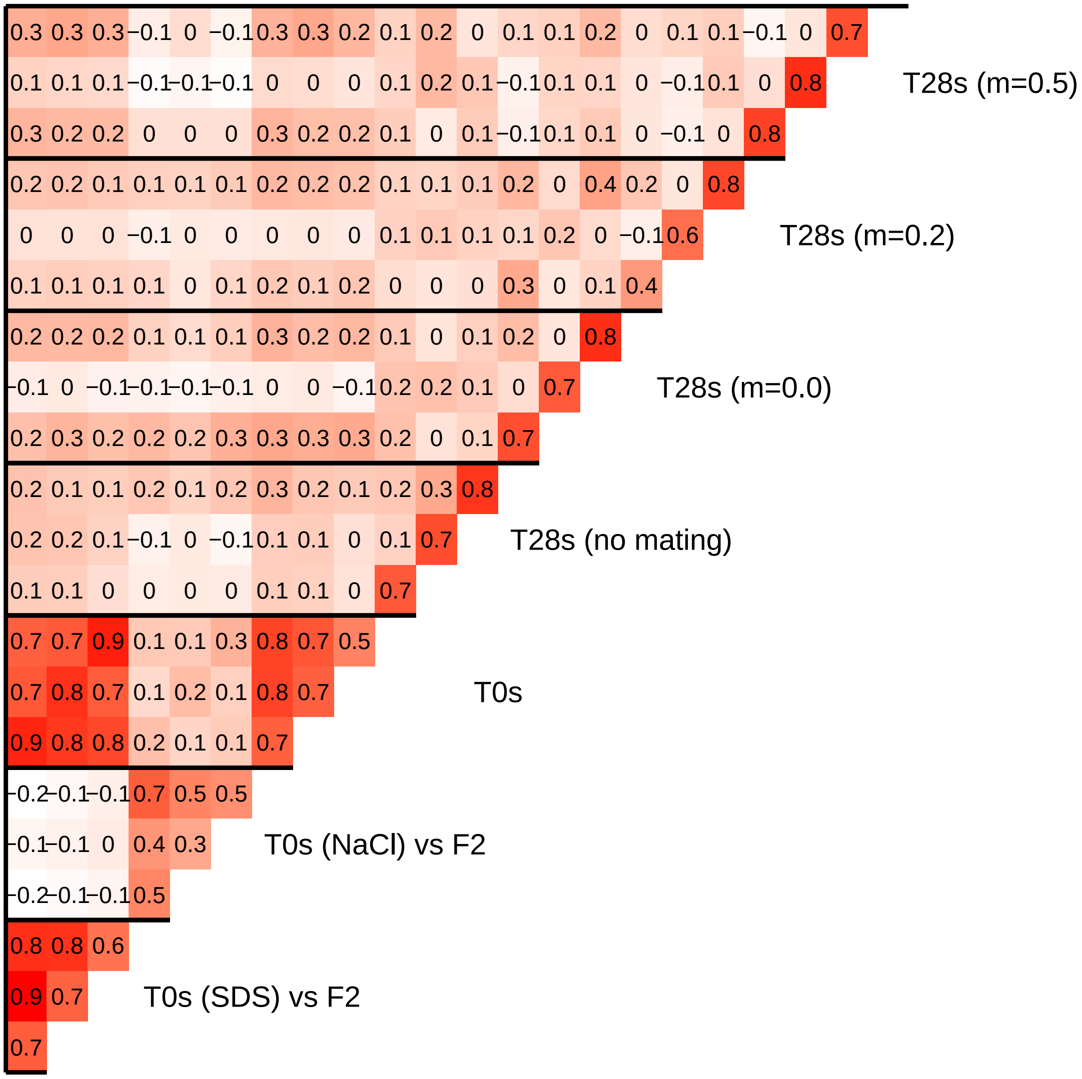


FIGURE 6.4: Pairwise correlation in F_{ST} between replicates and treatments in this study. Within a given grouping diagonal components indicate Moran's I for the first distance class (10 Kbp) within a replicate and the pairwise Pearson's correlation coefficient between pairwise observations.

F_{ST} but as a result of the differences in the degree of change in the YPS128 allele frequency (i.e., not differences in the sign of change in YPS128 allele frequency; grey shading in Figure 6.5).

6.4.2 *Local adaptation following admixture in yeast populations*

In terms of local adaptation between populations at time T0, we found that local individuals increasingly outperformed foreign individuals in lag ($P<0.027$; Figure 6.6B), carrying capacity ($P<0.0017$; Figure 6.6D), and AUC ($P<0.0009$; Figure 6.6F). Growth rate only showed a marginal significant increase across time ($P<0.066$; Figure 6.6C). While populations overall adapted to their respective environments, we did not detect a significant difference between the extent of local adaptation between environments at the end of Experiment I when taking into consideration all four local adaptation metrics together (MANOVA; Wilks' lambda=0.052, $F_{(4,1)} = 4.52$, $P<0.337$).

We found that genetic differentiation after 12 days of isolation between population pairs was dominated by the genetic differences accrued in SDS evolved populations (compare Figure 6.5A-C with Figure 6.7A-C). Indeed, we found that peaks in the T0 (SDS) vs T0 (NaCl) comparisons were highly correlated with peaks in the T0 (SDS) vs F_2 ancestral pool comparisons but not in the T0 (NaCl) vs F_2 comparisons (Figure 6.4). For much of the genome, the genomic response from evolving alternative habitats showed a similar pattern in the direction and magnitude of allele frequency change; however, we uncovered several regions that showed a differential selective response between SDS and NaCl (Figure 6.3). Notably, regions of variable size along chromosomes IV, V, VI, VII, IX, XIII, XIV, & XVI displayed an abundance of SNPs in which all three replicates of SDS evolved populations exhibited a similar response and opposite from all three replicates that evolved in NaCl.

6.4.3 *DFG from isolated, locally adapted yeast populations*

Across all migration treatments, local adaptation in both environments was generally maintained (Figure 6.8; see also Figures E.1, E.2, E.3, & E.4). At the end time point of Experiment II (T28) we found that the SDS environment contained more locally adapted individuals than the NaCl environment (MANOVA; Wilks' lambda=0.19, $F_{(4,13)} = 14.13$, $P<0.0001$). Interestingly, we did

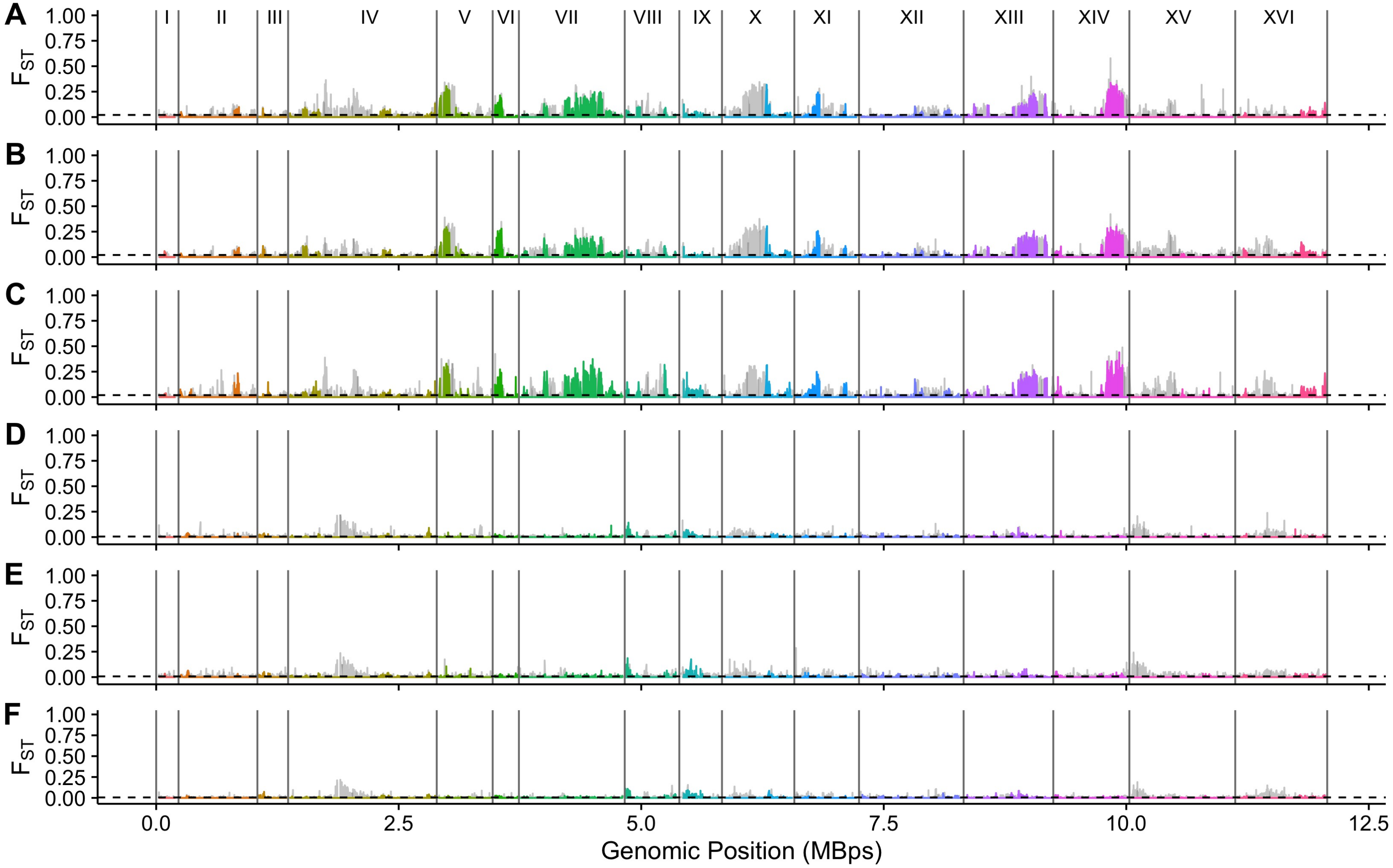


FIGURE 6.5: Genomic differentiation from initial adaptation. Shown here is F_{ST} across replicate populations evolving in SDS (A-C) or NaCl (D-F) for 12 days. Colored regions indicate putative loci of divergent selection and the different colors delineate individual chromosomes. The genome-wide average F_{ST} is given as a dashed line.

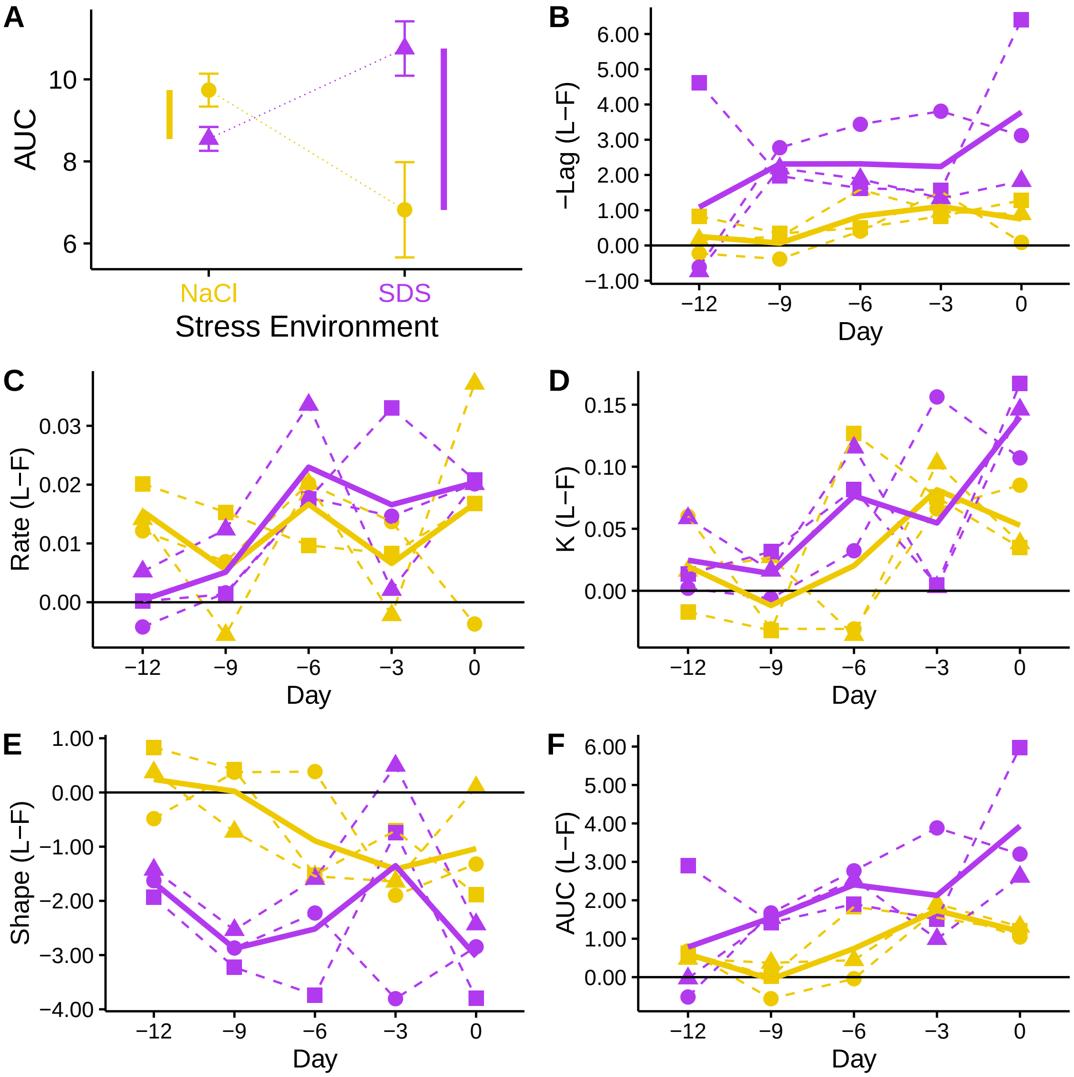


Figure 6.6: Local adaptation (LA) of local populations relative to foreign populations during divergence after secondary contact. Panel A shows an example of local adaptation using the local-foreign criterion. Here, AUC for populations are plotted as means (±SD). The color of the population averages and associated errorbars correspond to the environment in which they were grown. The mean AUC value for each of the foreign populations is subtracted from the local population's AUC value (vertical solid lines). This method produces two LA values – one for each stress environment. Panels B-F show the LA values four parameters used in the Richards curve and their composite statistic (AUC). Gold and purple lines denote NaCl and SDS habitats, respectively, and group means are denoted by solid color lines. For clarity, the LA value for lag was multiplied by -1 such that positive values indicate the temporal advantage (in hours) that the locals have compared to their foreign counterpart.

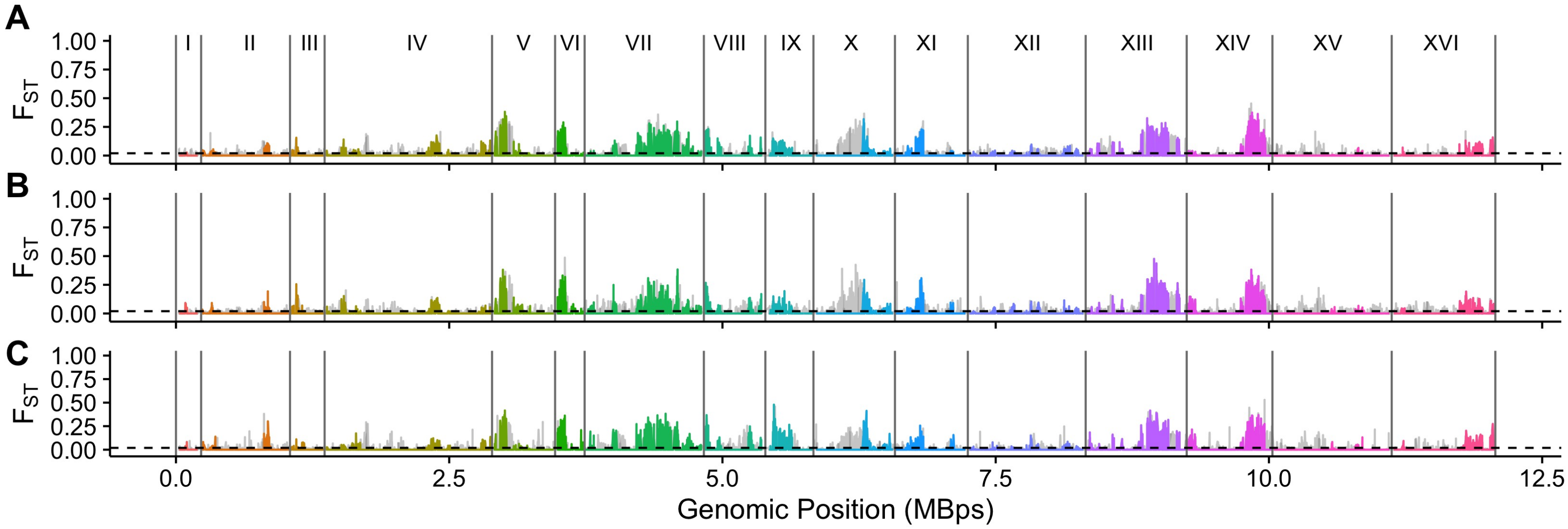


FIGURE 6.7: Genomic differentiation between replicate population pairs at the end time point of Experiment I (T0). Labels (A-C) indicate a given replicate pair of populations.

not detect any differences in local adaptation among the migration treatments (P<0.29) or in environment-by-migration interactions (P<0.47) at time point T28. However, we found that local adaptation tended to increase within a "cycle". For example, at time point T22 the only significant difference in local adaptation occurred among the migration treatments (MANOVA; Wilks' lambda=0.25, $F_{(12,34.7)} = 14.13$, P<0.04) – not between environments (P<0.17). Thus, it appears that the effect of migration (and recombination) diminishes with asexual growth in these stress environments.

Whereas there was a high degree of repeatability found at the end of Experiment I, we found a greater degree of stochasiticity and variance in genetic differentiation at the end of Experiment II (Figures 6.9 & 6.10). Within a replicate population pair, average Moran's *I* at the first distance class remained similar as Experiment I but within-treatment and across-treatment Pearson correlation coefficients were weak (Figure 6.4). We detected a significant difference between average neighborhood size and migration treatment (ANOVA; $F_{(3,105)} = 4.66$, p<0.004) but this pattern was primarily driven by the intermediate migration level, which exhibited smaller size islands than either no migration (without mating) or high migration treatments (Figure E.5). Considering only those migration treatments in which mating occurred, we found that increasing migration decreased overall mean F_{ST} (y = -0.21x + 0.19, R^2 = 0.15, p<1e-04).

6.5 DISCUSSION

It is increasingly evident that the genomic response to adaptation in nature is the result of multiple interacting evolutionary, demographic, and genic processes. Here we explore such a response in pairs of replicate populations evolving from standing genetic variation in SDS and NaCl. Our results show that many regions of the genome are responding to different types of selection and at various strengths, that genomic islands can form over short time scales, and that stochastic patterns of differentiation are likely due to genetic drift and/or latent selection pressures. We discuss each of these main findings in turn.

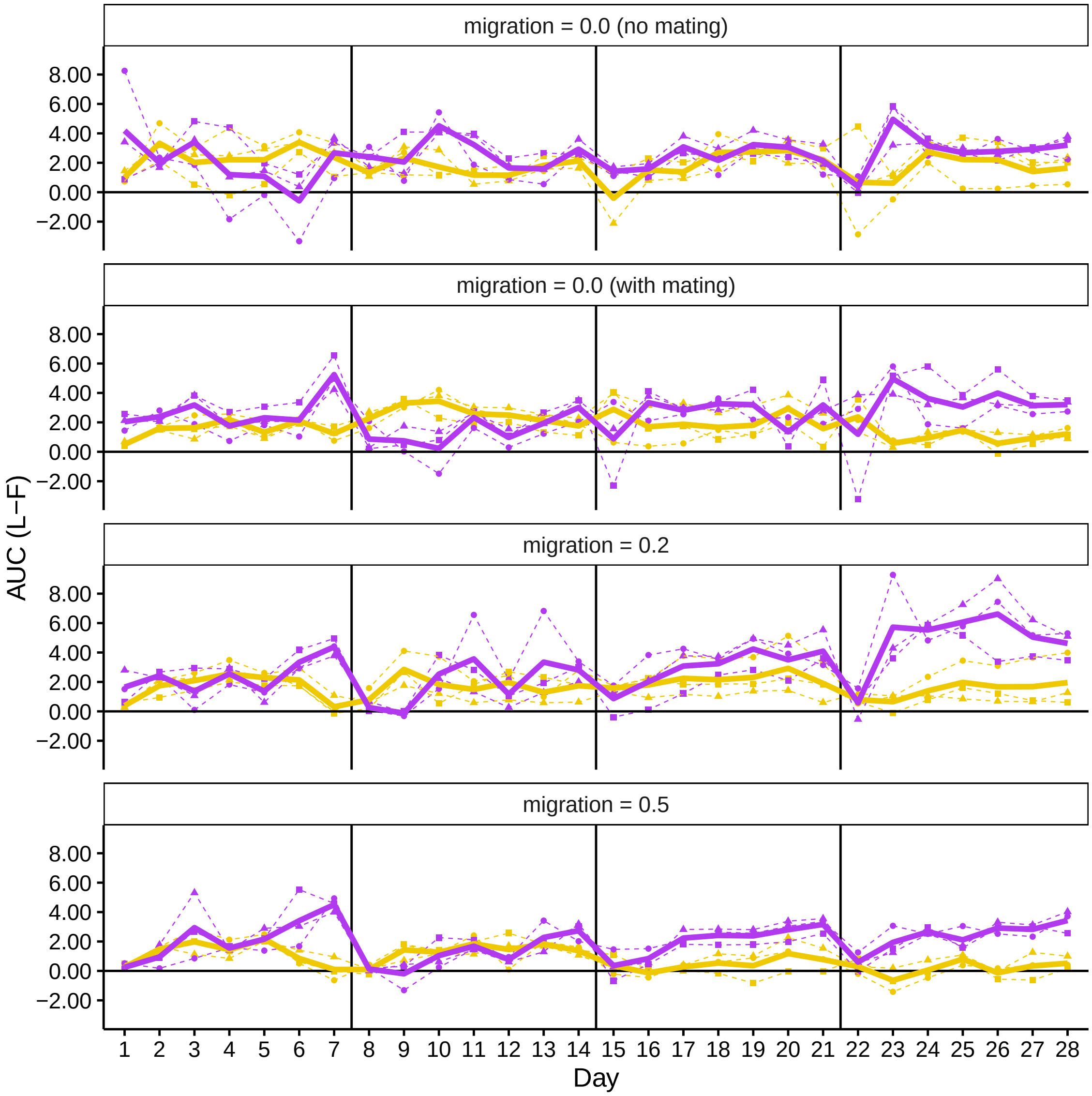


FIGURE 6.8: The extent of local adaptation (LA) during divergence with a given amount of migration. For each migration rate the extent of LA is given for each stress environment. Gold and purple denote NaCl and SDS habitats, respectively, and group means are denoted by solid color lines. Black vertical lines show where sporulation and migration (if applicable) occur between periods of asexual growth in batch.

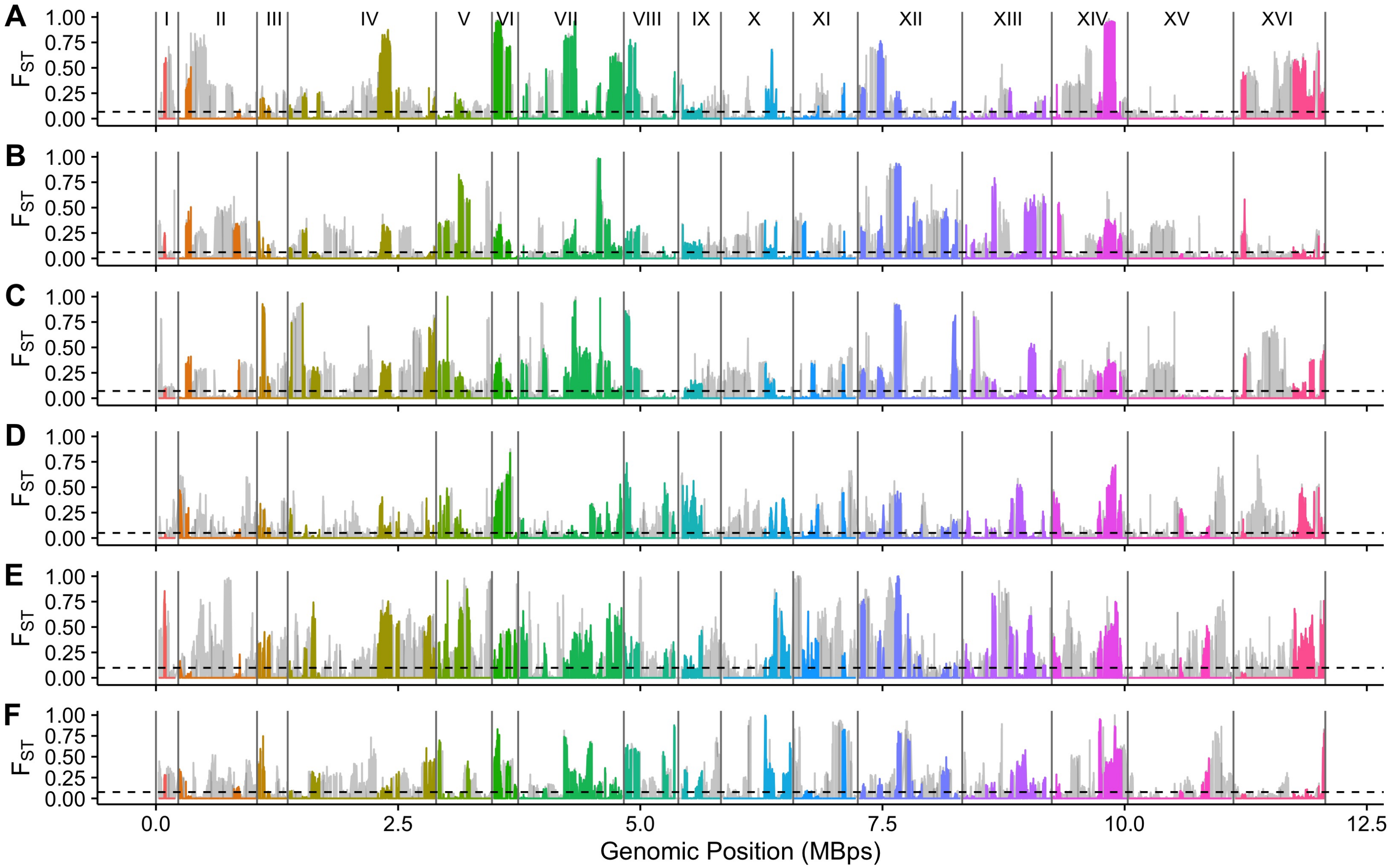


FIGURE 6.9: Genomic differentiation between replicate population pairs at the end time point of Experiment II. Shown here is F_{ST} across replicate population pairs with migration rate $m = 0$. Panels A-C show no mating replicates. Panels D-F show replicates in which mating occurred after seven days of batch transfer.

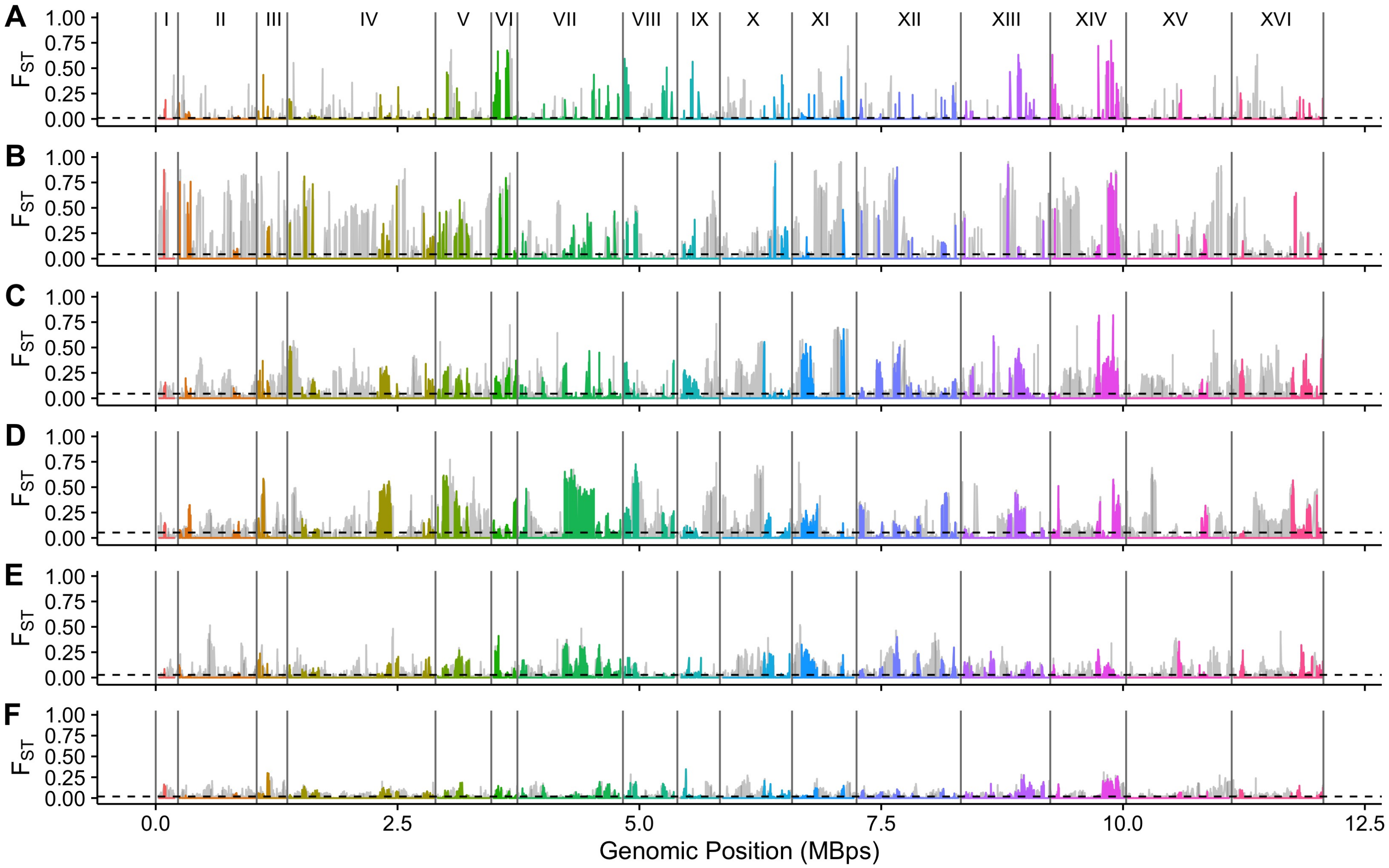


Figure 6.10: Genomic differentiation between replicate population pairs at the end time point of Experiment II. Shown here is F_{ST} across replicate population pairs with migration rate m = 0.2 (A-C) and m = 0.5 (D-F).

6.5.1 *Divergent QTL*

Natural selection can act multifariously on different genetically independent traits (Nosil et al., 2009b). When a population enters a novel environment, different types of selection pressures can act on several traits (e.g., morphological, behavioral, physiological) simultaneously. The result is a mosaic of genetic differentiation between populations evolving in alternative environments (e.g., Figure 6.5). Molecular biology approaches (e.g., bulk segregant analysis, X-QTL) have been applied to identify genomic regions associated with trait variation (Wenger et al., 2010; Bloom et al., 2013) but to our knowledge these approaches have not yet been applied in the context of divergently evolving populations. Here we used these approaches to identify and differentiate the effects of various selection pressures occurring within the genome of populations evolving in alternative environments. As a result we found hundreds of loci under selection (Figure 6.3). This approach can be further modified to more precisely delineate different types of selection. Indeed, experimentally evolving more advanced filial generations (e.g., Parts et al., 2011) would result in fine mapping of loci under directional selection in one or both environments from those evolving from divergent selection.

6.5.2 *Genomic islands from standing genetic variation*

One possible explanation for the growth of genomic islands of divergence is from divergent hitchhiking, where there is a reduced effective gene flow in genomic regions linked to divergently selected loci (Via and West, 2008). Divergence hitchhiking theory, however, may not fully explain patterns seen in nature (Yeaman, 2013). We previously modeled (Chapter 5) how genomic islands can arise from standing genetic variation via the breakdown of linkage disequilbrium within populations undergoing divergence in isolation following secondary contact (i.e., without gene flow between populations). Moreover, in a recent reanlaysis of published datasets on genomic islands, Cruickshank and Hahn (2014) found little support for the divergent hitchhiking hypothesis. Finally, in the current study we provide experimental evidence that clusters of localized genetic differentiation can occur in isolation following secondary contact. We are not arguing that the growth of genomic islands from *de novo* mutations cannot occur but rather highlight that alternative approaches can readily evolve and might better reflect natural systems.

It is worth noting that these two hypotheses need not be mutually exclusive as new mutations are continually supplying new genetic variation and that divergent hitchhiking can be acting from islands previously established by divergence originated by standing genetic variation.

We confirmed that genetic differentiation accompanying local adaptation can manifest as "islands" in less than 100 generations of populations evolving in isolation after a bout of secondary contact. We also found that the pattern of genomic islands was characterized by a mixture of divergently and non-divergently selected loci. Interestingly, some of these regions are adjacent to or overlap one another (Figures 6.5, 6.7, 6.9, & 6.10). By crossing two diverged strains of yeast we seeded populations with several thousand single nucleotide variants and even more combinations of genes for selection to act upon. In our experiment, divergence from secondary contact likely involved multifarious selection acting as well as responses to common stress-related gene networks. Further research is needed to better distinguish between genomic islands of *divergence* and the broader classification of genomic islands of *differentiation*.

To confidently estimate allele frequency differences between populations we used high coverage PoolSeq of reduced representation (RADseq) libraries and estimated F_{ST} at each SNP. The main drawback of this approach is that F_{ST}, a *relative* measure of divergence, depends on the average within population heterozygosity and that regions of limited diversity (e.g., recombination cold spots, centromeres) are expected to show inflated differentiation (Cruickshank and Hahn, 2014). An alternative method is based on *absolute* differences between populations without consideration of within population diversity. Our short read PoolSeq method, however, prohibited us from reliability estimating absolute measures of divergence (e.g., d_{XY}) because we could not reliably reconstruct large haplotype information and the variance in SNP estimates would be high (Cruickshank and Hahn, 2014). However, we circumvented this problem in our experiment by i) starting off with admixed individuals with roughly equal allele frequencies at each of the predetermined SNPs (Figure 6.1A) and ii) only including loci that showed evidence of opposing allelic responses between alternative environments when compared to the admixed ancestral pool (Figure 6.3).

6.5.3 *Stochastic islands*

We observed a highly stochastic pattern of differentiation under divergence with gene flow (Experiment II), finding little evidence that migration erodes islands of divergence on a genome-wide scale. One hypothesis for the high variance in F_{ST} seen within and between replicates here is that adaptation to one or both of the stress environments has resulted in large-scale genomic rearrangements. Indeed, Yeaman (2013) showed how genomic shuffling via transposition is one way that adaptive mutations can build up in genomic clusters with tight linkage over relatively short timescales and stress has been shown to create genomic instability – but not large scale rearrangements – in fermenting lager yeast (James et al., 2008). Since our allele calling was based on mapping reads to a reference genome, large-scale genomic rearrangements may not adequately capture the realized relationship of clusters of genes under divergent selection. However, this hypothesis is unlikely to be accepted in our case for the following reason. In our RADseq approach we sequenced paired-end reads with an insert size of approximately 400 bp. When counting alleles at each locus we discarded the paired-end reads due to variability in sequence coverage introduced from the random shearing process (see Methods). These paired-end data, however, can still be used to test if a high proportion of pairs of reads map discordantly to the reference. Mapping these paired-end data to the reference genome identified an average of 6.7% (SD=1.8) of the reads for all samples aligned discordantly when specifying a conservative 100-1000 bp insert size range. Thus, genomic rearrangements are not likely to contribute meaningfully to the high variance in F_{ST} clustering that we found within and between replicates of Experiment II.

It may be possible that additional and unforeseen selection occurred during the experimental procedure. For each of the four treatments in Experiment II we froze samples. For the three mating treatments we revived and sporulated them. After sporulation, we grew all treatments (included the no mating control) in YPD before continuing with the divergent selection batch transfer. During this procedure there are many potential selection events that could occur which would obstruct or disrupt our expected results. For example, in a previous study we created advanced intercross lines by serially mating progeny generated from crossing haploid YPS128 and DBVPG1106 strains (Chapter 4). This mating procedure was similar as performed in the current study (see Methods) but without a period of growth in a stress environment. Genotyping

96 F_6 diploid individuals revealed a strong departure from HWE expectations. The repeatability of this result is unknown since only one replicate population was serially mated. If, however, different replicates of intercrossing resulted in vastly different ancestry painting (e.g., Figure 4.6) then genetic drift in our experiment would also be likely.

6.6 CONCLUDING REMARKS

Our results underscore how multiple concurrent evolutionary and genic processes can affect the genomic response to adaptation from standing genetic variation. Here we show that such adaptation can produce a pattern of heterogeneous genetic differentiation relatively rapidly as a result of multifarious selection to novel environments. We also found that the genomic response to selection can be strongly influenced by genetic drift – either directly via population bottlenecks or indirectly as a result of strong multifarious selection.

Bibliography

Abbott, R., D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, N. Bierne, et al. 2013. Hybridization and speciation. Journal of Evolutionary Biology 26:229–246.

Acar, M., B. F. Pando, F. H. Arnold, M. B. Elowitz, and A. van Oudenaarden. 2010. A general mechanism for network-dosage compensation in gene circuits. Science 329:1656–1660.

Agrawal, A. F. and J. R. Stinchcombe. 2009. How much do genetic covariances alter the rate of adaptation? Proceedings of the Royal Society B: Biological Sciences 276:1183–1191.

Ali, O. A., S. M. O'Rourke, S. J. Amish, M. H. Meek, G. Luikart, C. Jeffres, et al. 2015. RAD Capture (Rapture): Flexible and efficient sequence-based genotyping. bioRxiv p. 1101/02.

Allen, H. L., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832–8.

Allison, A. 1954. Protection afforded by sickle-cell trait against subtertian malareal infection. British Medical Journal 1:290–4.

Alvarez-Castro, J. M. and O. Carlborg. 2007. A unified model for functional and statistical epistasis and its application in Quantitative Trait Loci analysis. Genetics 176:1151–1167.

Anderson, C. M., S. Y. Chen, M. T. Dimon, A. Oke, J. L. DeRisi, and J. C. Fung. 2011. Recombine: A suite of programs for detection and analysis of meiotic recombination in whole-genome datasets. PLoS ONE 6:e25509.

Andolfatto, P., D. Davison, D. Erezyilmaz, T. T. Hu, J. Mast, T. Sunayama-Morita, et al. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Research 21:610–617.

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics (in press).

Arnold, M. L. and N. H. Martin. 2009. Adaptation by introgression. Journal of Biology 8:82.

Arnold, S. J., M. E. Pfrender, and A. G. Jones. 2001. The adaptive landscape as a conceptual bridge between micro- and macroevolution. Genetica 112-113:9–32.

Arnold, S. J., B. Reinhard, P. A. Hohenlohe, B. C. Ajie, and A. G. Jones. 2008. Understanding the evolution and stability of the G-matrix. Evolution 62:2451–2461.

Aylor, D. L. and Z.-B. Zeng. 2008. From classical genetics to quantitative genetics to systems biology: modeling epistasis. PLoS Genetics 4:e1000029.

Babu, M. M., N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. 2004. Structure and evolution of transcriptional regulatory networks. Current Opinion in Structural Biology 14:283–291.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376.

Bank, C., R. Bürger, and J. Hermisson. 2012. The limits to parapatric speciation: Dobzhansky-Muller incompatibilities in a continent-Island model. Genetics 191:845–863.

Barrett, R. D. H. and D. Schluter. 2008. Adaptation from standing genetic variation. Trends in Ecology & Evolution 23:38–44.

Beaumont, M. A. and R. A. Nichols. 1996. Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society B: Biological Sciences 263:1619–1626.

Benfey, P. N. and T. Mitchell-Olds. 2008. From genotype to phenotype: systems biology meets natural variation. Science 320:495–497.

Berg, J. and M. Lässig. 2004. Local graph alignment and motif search in biological networks. Proceedings of the National Academy of Sciences of the United States of America 101:14689–14694.

Birch, C. P. 1999. A new generalized logistic sigmoid growth equation compared with the richards growth equation. Annals of Botany 83:713–723.

Björklund, M., a. Husby, and L. Gustafsson. 2013. Rapid and unpredictable changes of the G-matrix in a natural bird population over 25 years. Journal of Evolutionary Biology 26:1–13.

Bloom, J. S., I. M. Ehrenreich, W. T. Loo, T.-L. V. Lite, and L. Kruglyak. 2013. Finding the sources of missing heritability in a yeast cross. Nature 494:234–7.

Blount, Z. D., C. Z. Borland, and R. E. Lenski. 2008. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America 105:7899–906.

Buerkle, C. A. and Z. Gompert. 2013. Population genomics based on low coverage sequencing: how low should we go? Molecular Ecology 22:3028–35.

Carlborg, O., L. Jacobsson, P. Ahgren, P. Siegel, and L. Andersson. 2006. Epistasis and the release of genetic variation during long-term selection. Nature Genetics 38:418–420.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124–40.

Chen, J. M., D. N. Cooper, N. Chuzhanova, C. Ferec, and G. P. Patrinos. 2007. Gene conversion: mechanisms, evolution and human disease. Nature Reviews Genetics 8:762–775.

Cherry, J. M., C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, et al. 1997. Genetic and physical maps of Saccharomyces cerevisiae. Nature 387:67–73.

Cherry, J. M., E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Research 40:D700–5.

Cirulli, E. T., R. M. Kliman, and M. A. F. Noor. 2007. Fine-scale crossover rate heterogeneity in Drosophila pseudoobscura. Journal of molecular evolution 64:129–35.

Codón, A., J. Gasent-Ramírez, and T. Benítez. 1995. Factors which affect the frequency of sporulation and tetrad formation in Saccharomyces cerevisiae baker's yeasts. Appl. Envir. Microbiol. 61:630–638.

Cole, F., S. Keeney, and M. Jasin. 2012. Preaching about the converted: How meiotic gene conversion influences genomic diversity. Annals of the New York Academy of Sciences 1267:95–102.

Colosimo, P. F. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin Alleles. Science 307:1928–1933.

Costanzo, M., A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, et al. 2010. The genetic landscape of a cell. Science 327:425–431.

Cruickshank, T. E. and M. W. Hahn. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular Ecology 23:3133–57.

Cubillos, F. a., E. J. Louis, and G. Liti. 2009. Generation of a large set of genetically tractable haploid and diploid Saccharomyces strains. FEMS Yeast Research 9:1217–25.

Davey, J. W., T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi, and M. L. Blaxter. 2013. Special features of RAD Sequencing data: implications for genotyping. Molecular Ecology 22:3151–64.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12:499–510.

Denby, C. M., J. H. Im, R. C. Yu, C. G. Pesce, and R. B. Brem. 2012. Negative feedback confers mutational robustness in yeast transcription factor regulation. Proceedings of the National Academy of Sciences 109:3874–3878.

Dieckmann, U. and M. Doebeli. 1999. On the origin of species by sympatric speciation distribution K(x) is unimodal and varies according to a gaussian. Nature 400:354–357.

Dixon, S., M. Costanzo, A. Baryshnikova, B. Andrews, and C. Boone. 2009. Systematic Mapping of Genetic Interaction Networks. Annual Review of Genetics 43:601–25.

Dobzhansky, T. and T. G. Dobzhansky. 1937. Genetics and the Origin of Species. Columbia University Press.

Draghi, J. A. and M. C. Whitlock. 2012. Phenotypic plasticity facilitates mutational variance, genetic variance, and evolvability along the major axis of environmental variation. Evolution 66:2891–2902.

Durand, E., F. Jay, O. E. Gaggiotti, and O. Franc. 2008. Spatial inference of admixture proportions and secondary contact zones. Molecular Biology and Evolution 26:1963–1973.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. 1st ed. Cambridge University Press, New York.

Eddelbuettel, D. and R. François. 2011. Rcpp : Seamless R and C++ Integration. Journal of Statistical Software 40:1–18.

Ehrenreich, I. M., N. Torabi, Y. Jia, J. Kent, S. Martis, J. a. Shapiro, et al. 2010. Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature 464:1039–42.

Erwin, D. H. and E. H. Davidson. 2009. The evolution of hierarchical gene regulatory networks. Nature Reviews Genetics 10:141–148.

Félix, M.-A. 2012. Evolution in developmental phenotype space. Current Opinion in Genetics & Development 22:593–9.

Fierst, J. L. and T. F. Hansen. 2010. Genetic architecture and postzygotic reproductive isolation: Evolution of Bateson-Dobzhansky-Muller incompatibilities in a polygenic model. Evolution 64:675–693.

Fierst, J. L. and P. C. Phillips. 2012. Variance in epistasis links gene regulation and evolutionary rate in the yeast genetic interaction network. Genome Biology and Evolution 4:1080–1087.

Fisher, R. A. 1930. The Genetical Theory of Natural Selection: A Complete Variorum Edition. Oxford University Press, Oxford.

Fishman, L. and J. H. Willis. 2001. Evidence for Dobzhansky-Muller incompatibilites contributing to the sterility of hybrids between Mimulus guttatus and M. nasutus. Evolution 55:1932–1942.

Freking, B. A., S. K. Murphy, A. A. Wylie, S. J. Rhodes, J. W. Keele, K. A. Leymaster, et al. 2002. Identification of the single base change causing the callipyge muscle hypertrophy phenotype, the only known example of polar overdominance in mammals. Genome Research 12:1496–506.

Gavrilets, S. 2000. Rapid evolution of reproductive barriers driven by sexual conflict. Nature 403:886–889.

Gavrilets, S. and A. Vose. 2005. Dynamic patterns of adaptive radiation. Proceedings of the National Academy of Sciences 102:18040–18045.

Gay, J., S. Myers, and G. McVean. 2007. Estimating meiotic gene conversion rates from population genetic data. Genetics 177:881–94.

Gemmell, N. J. and J. Slate. 2006. Heterozygote advantage for fecundity. PLoS ONE 1:e125.

Gibson, G. and I. Dworkin. 2004. Uncovering cryptic genetic variation. Nature Reviews Genetics 5:681–90.

Gietz, R. D. and R. H. Schiestl. 2008. High-efficiency yeast transformation using the LiAc / SS carrier DNA / PEG method. Nature Protocols 2:31–35.

Gjuvsland, A. B., B. J. Hayes, T. H. E. Meuwissen, E. Plahte, and S. W. Omholt. 2007a. Nonlinear regulation enhances the phenotypic expression of trans-acting genetic polymorphisms. BMC Systems Biology 1:32.

Gjuvsland, A. B., B. J. Hayes, S. W. Omholt, and O. Carlborg. 2007b. Statistical epistasis is a generic feature of gene regulatory networks. Genetics 175:411–420.

Goddard, M. R., J. C. J. Godfray, and A. Burt. 2005. Sex Increases the efficacy of natrual selection in experimental yeast populations. Nature 434:636–640.

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

Guillaume, F. and M. C. Whitlock. 2007. Effects of migration on the genetic covariance matrix. Evolution 61:2398–2409.

Hadfield, J. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software 33:1–22.

Haldane, J. B. S. 1919. The probable errors of calculated linkage values, and the most accurate method of determining gametic from certain zygotic series. Journal of Genetics 8:291–297.

Hansen, T. F. and D. Houle. 2008. Measuring and comparing evolvability and constraint in multivariate characters. Journal of Evolutionary Biology 21:1201–1219.

Hartl, D. L. and A. G. Clark. 2007. Principles of Population Genetics. 4th ed. Sinauer Associates Inc., Sunderland, MA.

Hecker, M., S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke. 2009. Gene regulatory network inference: data integration in dynamic models-a review. Bio Systems 96:86–103.

Hedrick, P. W. 2011. Genetics of Populations. 4th ed. Jones and Bartlett Publishers, Sudbury, Massachusetts.

Hermisson, J. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335–2352.

Herskowitz, I. 1988. Life cycle of the budding yeast Saccharomyces cerevisiae. Microbiological Reviews 52:536–53.

Hether, T. D. and P. A. Hohenlohe. 2014. Genetic regulatory network motifs constrain adaptation through curvature in the landscape of mutational (co)variance. Evolution 68:950–64.

Hohenlohe, P. A., S. Bassham, M. Currey, and W. A. Cresko. 2012. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. Proceedings of the Royal Society B: Biological Sciences 367:395–408.

Hollick, J. B. and V. L. Chandler. 1998. Epigenetic allelic states of a maize transcriptional regulatory locus exhibit overdominant gene action. Genetics 150:891–7.

Houle, D. 1998. How should we explain variation in the genetic variance of traits? Genetica 102-103:241–253.

Houle, D. and J. Fierst. 2013. Properties of spontaneous mutational variance and covariance for wing size and shape in Drosophila melanogaster. Evolution 67:1116–30.

Houle, D., D. R. Govindaraju, and S. Omholt. 2010. Phenomics: the next challenge. Nature Reviews Genetics 11:855–866.

Houle, D., B. Morikawa, and M. Lynch. 1996. Comparing mutational variabilities. Genetics 143:1467–1483.

Hu, Y., C. Willer, X. Zhan, H. M. Kang, and G. Abecasis. 2013. Accurate local-ancestry inference in Exome-sequenced admixed individuals via off-target sequence reads. The Amierican Journal of Human Genetics 93:891–899.

Huang, Y., H. Li, H. Hu, X. Yan, M. S. Waterman, H. Huang, et al. 2007. Systematic discovery of functional modules and context-specific functional annotation of human genome. Bioinformatics 23:i222–9.

Illingworth, C. J. R., L. Parts, A. Bergström, G. Liti, and V. Mustonen. 2013. Inferring genome-wide recombination landscapes from advanced intercross lines: application to yeast crosses. PLoS ONE 8:e62266.

James, T. C., J. Usher, S. Campbell, and U. Bond. 2008. Lager yeasts possess dynamic genomes that undergo rearrangements and gene amplification in response to stress. Current Genetics 53:139–152.

Jeong, S., A. Rokas, and S. B. Carroll. 2006. Regulation of body pigmentation by the abdominal-B xox protein and its gain and loss in Drosophila evolution. Cell 125:1387–1399.

Johnson, N. a. 2009. One hundred years after Bateson: a pair of incompatible genes underlying hybrid sterility between yeast species. Heredity 103:360–361.

Johnson, N. A. and A. H. Porter. 2000. Rapid Speciation via Parallel, Directional Selection on Regulatry Genetic Pathways. Journal of Theoretical Biology 205:527–542.

Jones, A. G., S. J. Arnold, and R. Bürger. 2003. Stability of the G-Matrix in a population experiencing pleiotropic mutation, stabilizing selection, and genetic drift. Evolution 57:1747–1760.

Jones, A. G., S. J. Arnold, and R. Burger. 2004. Evolution and stability of the G-matrix on a landscape with a moving optimum. Evolution 58:1639–1654.

Jones, A. G., S. J. Arnold, and R. Bürger. 2007. The mutation matrix and the evolution of evolvability. Evolution 61:727–45.

———. 2012. The mutation matrix and the evolution of evolability. Evolution 61:727–745.

Jost, J. 2008. Riemannian Geometry and Geometric Analysis. 5th ed. Springer Science & Business Media, Springer-Verlag, Berlin.

Jost, M. C. and K. L. Shaw. 2006. Phylogeny of Ensifera (Hexapoda : Orthopter) using three ribosomal loci, with implications for the evolution of acoustic communication. Molecular Phylogenetics and Evolution 38:510–530.

Kahm, M., G. Hasenbrink, H. Lichtenberg-Fraté, J. Ludwig, and M. Kschischo. 2010. grofit: fitting biological growth curves with R. Journal of Statistical Software 33:1–21.

Kauppi, L., A. J. Jeffreys, and S. Keeney. 2004. Where the crossovers are: recombination distributions in mammals. Nature Reviews Genetics 5:413–424.

Kawecki, T. J. and D. Ebert. 2004. Conceptual issues in local adaptation. Ecology Letters 7:1225–1241.

Khare, A. K., B. Singh, and J. Singh. 2011. A fast and inexpensive method for random spore analysis in Schizosaccharomyces pombe. Yeast 28:527–533.

Kimura, M. 1965. A stochastic model concerning the maintenance of genetic variability in quantitative characters. Proceedings of the National Academy of Sciences of the United States of America 54:731–736.

Kingsolver, J. G., R. Gomulkiewicz, and P. a. Carter. 2001. Variation, selection and evolution of function-valued traits. Genetica 112-113:87–104.

Kirkpatrick, M. 2009. Patterns of quantitative genetic variation in multiple dimensions. Genetica 136:271–284.

Kofler, R., R. V. Pandey, and C. Schlötterer. 2011. PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27:3435–3436.

Kosambi, D. D. 1943. the Estimation of Map Distances From Recombination Values. Annals of Eugenics 12:172–175.

Kruuk, L. E. B. 2004. Estimating genetic parameters in natural populations using the "animal model". Philosophical transactions of the Royal Society of London. Series B, Biological sciences 359:873–90.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, et al. 2004. Versatile and open software for comparing large genomes. Genome Biology 5:R12.

Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. Evolution 33:402–416.

Lande, R. and S. J. Arnold. 1983. The Measurement of Selection on Correlated Characters. Evolution 37:1210–1226.

Langmead, B. and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–9.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9.

Liang, M. and R. Nielsen. 2014. The lengths of admixture tracts. Genetics 197:953–967.

Lichten, M. and B. De Massy. 2011. The impressionistic landscape of meiotic recombination. Cell 147:267–270.

Liti, G., D. M. Carter, A. M. Moses, J. Warringer, L. Parts, S. A. James, et al. 2009. Population genomics of domestic and wild yeasts. Nature 458:337–341.

Liu, C., D. Rubin, and Y. Wu. 1998. Parameter expansion to accelerate EM: the PX-EM algorithm. Biometrika 85:755–770.

Liu, Y., T. Nyunoya, S. Leng, S. a. Belinsky, Y. Tesfaigzi, and S. Bruse. 2013. Softwares and methods for estimating genetic ancestry in human populations. Human genomics 7:1.

Lowry, D. B., J. L. Modliszewski, K. M. Wright, C. a. Wu, and J. H. Willis. 2008. The strength and genetic basis of reproductive isolating barriers in flowering plants. Proceedings of the Royal Society B: Biological Sciences 363:3009–3021.

Lynch, M. 2007. The evolution of genetic networks by non-adaptive processes. Nature Reviews Genetics 8:803–813.

Lynch, M. and B. Walsh. 1998. Genetics and Analysis of Quantitative Traits. 1st ed. Sinauer, Sunderland.

Lynch, M., S. Xu, T. Maruki, X. Jiang, P. Pfaffelhuber, and B. Haubold. 2014. Genome-wide linkage-disequilibrium profiles from single individuals. Genetics 198:269–81.

Magoč, T. and S. L. Salzberg. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–63.

Magwene, P. M., Ö. Kayikçi, J. a. Granek, J. M. Reininga, Z. Scholl, and D. Murray. 2011. Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America 108:1987–1992.

Mancera, E., R. Bourgon, A. Brozzi, W. Huber, and M. Steinmetz. 2008. High-resolution mapping of meiotic crossovers and noncrossovers in yeast. Nature 454:479–485.

Mckinney, B. A., N. M. Pajewski, M. D. Ritchie, and T. Pennsylvania. 2012. Six degrees of epistasis: statistical network models for GWAS. Frontiers in Genetics 2:1–6.

Mersha, T. B. 2015. Mapping asthma-associated variants in admixed populations. Frontiers in Genetics 6:1–21.

Mersmann, l. 2011. microbenchmark: Sub microsecond accurate timing functions. The R Project for Statistical Computing. 1:1-3.

Michel, A. P., S. Sim, T. H. Q. Powell, M. S. Taylor, P. Nosil, and J. L. Feder. 2010. Widespread genomic divergence during sympatric speciation. Proceedings of the National Academy of Sciences 107:9724–9729.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network motifs: simple building blocks of complex networks. Science 298:824–827.

Mitteroecker, P. 2009. The developmental basis of variational modularity: Insights from quantitative genetics, morphometrics, and developmental biology. Evolutionary Biology 36:377–385.

Muller, H. J. 1942. Isolation mechanisms, evolution and temperature. Biol. Symp. 6:71–125.

Nadeau, N. J., A. Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra, S. W. Baxter, et al. 2012. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Proceedings of the Royal Society B: Biological Sciences 367:343–353.

Nishant, K. T., W. Wei, E. Mancera, J. L. Argueso, A. Schlattl, N. Delhomme, et al. 2010. The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. PLoS Geneticsenetics 6:e1001109.

Nosil, P., J. L. Feder, and P. T. R. S. B. 2012. Genomic divergence during speciation: causes and consequences. Philosophical Transactions of the Royal Society B: Biological Sciences 367:332–342.

Nosil, P., D. J. Funk, and D. Ortiz-Barrientos. 2009a. Divergent selection and heterogeneous genomic divergence. Molecular Ecology 18:375–402.

Nosil, P., L. J. Harmon, and O. Seehausen. 2009b. Ecological explanations for (incomplete) speciation. Trends in Ecology and Evolution 24:145–156.

O'Malley, M. A. 2012. Evolutionary systems biology: historical and philosophical perspectives on an emerging synthesis. *In* O. Soyer (ed.) Evolutionary systems biology: advances in experimental medicine and biology, p. 751. Springer Science & Business Media.

Omholt, S. W., E. Plahte, L. Oyehaug, and K. Xiang. 2000. Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. Genetics 155:969–980.

Orr, H. A. 1998. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution 52:935–949.

———. 2005. The genetic theory of adaptation: a brief history. Nature Reviews Genetics 6:119–127.

———. 2006. The distribution of fitness effects among beneficial mutations in Fisher's geometric model of adaptation. Journal of Theoretical Biology 238:279–85.

Orr, H. A. and A. J. Betancourt. 2001. Haldane's sieve and adaptation from the standing genetic variation. Genetics 157:875884.

Ott, J., J. Wang, and S. M. Leal. 2015. Genetic linkage analysis in the age of whole-genome sequencing. Nature Reviews Genetics 16:275–284.

Otto, S. P. and A. C. Gerstein. 2008. Primer: The evolution of haploidy and diploidy. Current Biology 18:R1121–R1124.

Padhukasahasram, B. and B. Rannala. 2011. Bayesian population genomic inference of crossing over and gene conversion. Genetics 189:607–19.

———. 2013. Meiotic gene-conversion rate and tract length variation in the human genome. European Journal of Human Genetics pp. 1–8.

Palmer, M. E. and M. W. Feldman. 2009. Dynamics of hybrid incompatibility in gene networks in a constant environment. Evolution 63:418–431.

Pâques, F. and J. E. Haber. 1999. Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. Microbiology and Molecular Biology Reviews 63:349–404.

Parts, L., F. a. Cubillos, J. Warringer, K. Jain, F. Salinas, S. J. Bumpstead, et al. 2011. Revealing the genetic structure of a trait by sequencing a population under selection. Genome Research 21:1131–1138.

Patterson, N., N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler, J. R. Oksenberg, et al. 2004. Methods for high-density admixture mapping of disease genes. American Journal of Human Genetics 74:979–1000.

Paulsen, M., S. Legewie, R. Eils, E. Karaulanov, and C. Niehrs. 2011. Negative feedback in the bone morphogenetic protein 4 (BMP4) synexpression group governs its dynamic signaling range and canalizes development. Proceedings of the National Academy of Sciences of the United States of America 108:10202–7.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 7:e37135.

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

Pigliucci, M. 2010. Genotype-phenotype mapping and the end of the 'genes as blueprint' metaphor. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 365:557–566.

Presgraves, D. C. 2010. The Molecular Evolutionary Basis of Species Formation. Nature Reviews Genetics 11:175–180.

Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, T. H. Beaty, et al. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genetics 5:e1000519.

Qi, J., A. J. Wijeratne, L. P. Tomsho, Y. Hu, S. C. Schuster, and H. Ma. 2009. Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in Saccharomyces cerevisiae. BMC Genomics 10:1–12.

Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77:257–286.

Rice, S. H. 2004. Evolutionary Theory: Mathematical and Conceptual Foundations. 1st ed. Sinauer, Sunderland, MA.

Rosenblum, E. B., B. A. J. Sarver, J. W. Brown, S. Des Roches, K. M. Hardwick, T. D. Hether, et al. 2012. Goldilocks meets Santa Rosalia: an ephemeral speciation model explains patterns of diversification across time scales. Evolutionary Biology 39:255–261.

Scannell, D. R., O. A. Zill, A. Rokas, C. Payen, M. J. Dunham, M. B. Eisen, et al. 2011. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 1:11–25.

Schluter, D. 1996. Adaptive radiation along genetic lines of least resistance. Evolution 50:1766–1774.

———. 2000. Ecological character displacement in adaptive radiation. The American Naturalist 156:S4–S16.

Sherman, F. 2002. Getting started with yeast. *In* C. Guthrie and G. R. Fink (eds.) Guide to Yeast Genetics and Molecular Biology, pp. 3–21. Academic Press, San Diego.

Slatkin, M. 2008. Linkage disequilibrium âĂŤ understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9:477–485.

Smith, J. M. and J. Haigh. 2009. The hitch-hiking effect of a favourable gene. Genetical Research 23:23.

Stadler, P. 2000. Population Dependent Fourier Decomposition of Fitness Landscapes over Recombination Spaces: Evolvability of Complex Characters. Bulletin of Mathematical Biology 62:399–428.

Steppan, S. J., P. C. Phillips, and D. Houle. 2002. Comparative quantitative genetics: evolution of the G matrix. Trends in Ecology & Evolution 17:320–327.

Struhl, K., D. T. Stinchcomb, S. Scherer, and R. W. Davis. 1979. High-frequency transformation of yeast: autonomous replication of hybrid DNA molecules. Proceedings of the National Academy of Sciences of the United States of America 76:1035–1039.

Stuart, J. M. 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science 302:249–255.

Sullivan, J., J. R. Demboski, K. C. Bell, S. Hird, B. Sarver, N. Reid, et al. 2014. Divergence with gene flow within the recent chipmunk radiation (Tamias). Heredity 113:185–94.

Szostak, J. W., T. L. Orr-Weaver, R. J. Rothstein, and F. W. Stahl. 1983. The double-strand-break repair model for recombination. Cell 33:25–35.

Taylor, E. B. and J. D. McPhail. 2000. Historical contingency and ecological determinism interact to prime speciation in sticklebacks, Gasterosteus. Proceedings of the Royal Society B: Biological Sciences 267:2375–84.

Tøndel, K., U. G. Indahl, A. B. Gjuvsland, J. O. Vik, P. Hunter, S. W. Omholt, et al. 2011. Hierarchical cluster-based partial least squares regression (HC-PLSR) is an efficient tool for metamodelling of nonlinear dynamic models. BMC Systems Biology 5:1–17.

Travisano, M. and R. G. Shaw. 2013. Lost in the Map. Evolution 67:305–314.

Tsai, I. J., A. Burt, and V. Koufopanou. 2010. Conservation of recombination hotspots in yeast. Proceedings of the National Academy of Sciences of the United States of America 107:7847–7852.

Turelli, M. 1984. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. Theoretical population biology 25:138–193.

Turner, T. L., M. W. Hahn, and S. V. Nuzhdin. 2005. Genomic islands of speciation in Anopheles gambiae. PLoS Biology 3:1572–1578.

Tyler, A. L., F. W. Asselbergs, S. M. Williams, and J. H. Moore. 2009. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. BioEssays 31:220–7.

Unckless, R. L. and H. A. Orr. 2009. Dobzhansky-Muller incompatibilities and adaptation to a shared environment. Heredity 102:214–217.

Via, S. 2012. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. Philosophical Transactions of the Royal Society B: Biological Sciences 367:451–460.

Via, S. and J. West. 2008. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. Molecular Ecology 17:4334–4345.

Wagner, A., G. P. Wagner, and P. Similiont. 1994. Epistasis can facilitate the evolution of reproductive isolation by peak shifts. Genetics 138:533–545.

Wagner, G. P. and L. Altenbery. 1996. Perspective: Complex adaptations and the evolution of evolvability. Evolution 50:967–976.

Walsh, B. and M. W. Blows. 2009. Abundant genetic variation + strong selection = multivariate genetic constraints: A geometric view of adaptation. Annual Review of Ecology, Evolution, and Systematics 40:41–59.

Wang, Y., A. B. Gjuvsland, J. O. Vik, N. P. Smith, P. J. Hunter, and S. W. Omholt. 2012. Parameters in dynamic models of complex traits are containers of missing heritability. PLoS Computational Biology 8:e1002459.

Weir, B. S., L. R. Cardon, A. D. Anderson, D. M. Nielsen, and W. G. Hill. 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Research 15:1468–1476.

Wenger, J. W., K. Schwartz, and G. Sherlock. 2010. Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from Saccharomyces cerevisiae. PLoS genetics 6:e1000942.

Wilson, A. J., D. Réale, M. N. Clements, M. M. Morrissey, E. Postma, C. A. Walling, et al. 2010. An ecologist's guide to the animal model. The Journal of Animal Ecology 79:13–26.

Wittkopp, P. J., K. Vaccaro, and S. B. Carroll. 2002. Evolution of yellow gene regulation and pigmentation in Drosophila. Current Biology 12:1547–1556.

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

———. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *In* Proceedings of The Sixth Congress on Genetics, pp. 356–366.

Wu, C. 2001. The genic view of the process of speciation. Journal of Evolutionary Biology 14:851–865.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42:565–9.

Yanowitz, J. 2010. Meiosis: making a break for it. Current Opinion in Cell Biology 22:744–751.

Yeaman, S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. Proceedings of the National Academy of Sciences of the United States of America 110:E1743–51.

Yeaman, S. and S. P. Otto. 2011. Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. Evolution 65:2123–2129.

Yeaman, S. and M. C. Whitlock. 2011. The genetic architecture of adaptation under migration-selection balance. Evolution 65:1897–1911.

Zeyl, C. 2006. Experimental evolution with yeast. FEMS Yeast Research 6:685–91.

Zhan, H. and S. Xu. 2011. Generalized linear mixed model for segregation distortion analysis. BMC Genetics 12:1–14.

Zhang, F., H. Q. Zhai, A. H. Paterson, J. L. Xu, Y. M. Gao, T. Q. Zheng, et al. 2011. Dissecting genetic networks underlying complex phenotypes: The theoretical framework. PLoS ONE 6:e14541.

Zhao, H. and T. P. Speed. 1996. On genetic map functions. Genetics 142:1369–1377.

Zhbannikov, I. Y. 2015. Preprocessing algorithms and software for genomics studies with high-throughput sequencing data. Phd dissertation, University of Idaho.

Zhu, M., M. Yu, and S. Zhao. 2009. Understanding quantitative genetics in the systems biology era. International Journal of Biological Sciences 5:161–170.

APPENDIX A

Supplementary Information to Chapter 2

A.1 GENOTYPE-TO-PHENOTYPE EQUATIONS

The equations in this section describe the "Genotype-to-Phenotype" map for the system of ODEs. They are in terms of genotypic values α_1 and α_2, which represent the summed additive contribution of both parental alleles for gene 1 and 2, respectively. Parameters are the amount of regulator needed to yield a 50% response (θ) and decay rate of expressed product (γ).

A.1.1 *Motif "A" - single dependency, positive*

$$x_1 = \frac{\alpha_1}{\gamma} \tag{A.1}$$

$$x_2 = \frac{\alpha_1\alpha_2}{\gamma(\theta + \alpha_1)} \tag{A.2}$$

A.1.2 *Motif "B" - single dependency, negative*

$$x_1 = \frac{\alpha_1}{\gamma} \tag{A.3}$$

$$x_2 = \frac{\theta\alpha_2}{\theta\gamma + \alpha_1} \tag{A.4}$$

A.1.3 *Motif "C" - double dependency, negative feedback loop*

$$x_1 = \frac{2\theta\alpha_1}{(\theta\gamma - \alpha_1) + \sqrt{4\alpha_1\alpha_2 + (\theta\gamma + \alpha_1)^2}} \tag{A.5}$$

$$x_2 = \frac{(\theta\gamma - \alpha_1) + \sqrt{4\alpha_1\alpha_2 + (\theta\gamma + \alpha_1)^2}}{2\gamma} \tag{A.6}$$

A.1.4 *Motif "D" - double dependency, both positive*

$$x_1 = \frac{-\theta\gamma + \alpha_1 - \alpha_2 + \sqrt{4\theta\gamma\alpha_1 + (-\theta\gamma + \alpha_1 - \alpha_2)^2}}{2\gamma} \tag{A.7}$$

$$x_2 = \frac{-\theta\gamma - \alpha_1 + \alpha_2 + \sqrt{4\theta\gamma\alpha_1 + (-\theta\gamma + \alpha_1 - \alpha_2)^2}}{2\gamma} \tag{A.8}$$

A.1.5 *Motif "E" - double dependency, both negative*

$$x_1 = \frac{\alpha_1\alpha_2 - \theta^2}{\gamma(\theta + \alpha_2)} \tag{A.9}$$

$$x_2 = \frac{\alpha_1\alpha_2 - \theta^2}{\gamma(\theta + \alpha_1)} \tag{A.10}$$

A.1.6 *Motif "F" - no dependency*

$$x_1 = \frac{\alpha_1}{\gamma} \tag{A.11}$$

$$x_2 = \frac{\alpha_2}{\gamma} \tag{A.12}$$

A.2 STABILITY OF EQUILIBRIA

Stability of equilibrium values was determined by analyzing the eigenvalues of the Jacobian matrix for each system of ODEs. Below we provide the Jacobian and the eigenvalues.

A.2.1 *Motif "A" - single dependency, positive*

$$J(x_1, x_2) = \begin{pmatrix} -\gamma & 0 \\ \frac{\alpha_2}{(\theta+x_2)^2} & -\gamma \end{pmatrix} \tag{A.13}$$

$$\lambda_1, \lambda_2 = -\gamma \tag{A.14}$$

A.2.2 *Motif "B" - single dependency, negative*

$$J(x_1, x_2) = \begin{pmatrix} -\gamma & 0 \\ \frac{-\alpha_2\theta}{(\theta+x_1)^2} & -\gamma \end{pmatrix} \tag{A.15}$$

$$\lambda_1, \lambda_2 = -\gamma \tag{A.16}$$

A.2.3 *Motif "C" - double dependency, negative feedback loop*

$$J(x_1, x_2) = \begin{pmatrix} -\gamma & \frac{-\alpha_1\theta}{(\theta+x_2)^2} \\ \frac{\alpha_2\theta}{(\theta+x_1)^2} & -\gamma \end{pmatrix} \tag{A.17}$$

$$\begin{aligned}\lambda_1, \lambda_2 \quad &= \{\pm\sqrt{-\alpha_1\alpha_2\theta^2(\theta+x_1)^2(\theta+x_2)^2} \\ &-\gamma\theta^4 - 2\gamma\theta^3 x_1 - 2\gamma\theta^3 x_2 - \gamma\theta^2 {x_1}^2 \\ &-4\gamma\theta^2 x_1 x_2 - \gamma\theta^2 {x_2}^2 - 2\gamma\theta {x_1}^2 x_2 \\ &-2\gamma\theta x_1 x_2^2 - \gamma x_1^2 x_2^2\} \cdot ((\theta+x_1)(\theta+x_2))^{-1}\end{aligned} \tag{A.18}$$

A.2.4 *Motif "D" - double dependency, both positive*

$$J(x_1, x_2) = \begin{pmatrix} -\gamma & \frac{-\alpha_1\theta}{(\theta+x_2)^2} \\ \frac{-\alpha_2\theta}{(\theta+x_1)^2} & -\gamma \end{pmatrix} \tag{A.19}$$

$$\begin{aligned}\lambda_1, \lambda_2 \quad &= \{\pm\sqrt{-\alpha_1\alpha_2\theta^2(\theta+x_1)^2(\theta+x_2)^2} \\ &-\gamma\theta^4 - 2\gamma\theta^3 x_1 - 2\gamma\theta^3 x_2 - \gamma\theta^2 {x_1}^2 \\ &-4\gamma\theta^2 x_1 x_2 - \gamma\theta^2 {x_2}^2 - 2\gamma\theta {x_1}^2 x_2 \\ &-2\gamma\theta x_1 x_2^2 - \gamma x_1^2 x_2^2\} \cdot ((\theta+x_1)(\theta+x_2))^{-1}\end{aligned} \tag{A.20}$$

A.2.5 *Motif "E" - double dependency, both negative*

$$J(x_1, x_2) = \begin{pmatrix} -\gamma & \frac{\alpha_1\theta}{(\theta+x_2)^2} \\ \frac{\alpha_2\theta}{(\theta+x_1)^2} & -\gamma \end{pmatrix} \tag{A.21}$$

$$\begin{aligned}\lambda_1, \lambda_2 \quad &= \{\pm\sqrt{-\alpha_1\alpha_2\theta^2(\theta+x_1)^2(\theta+x_2)^2} \\ &-\gamma\theta^4 - 2\gamma\theta^3 x_1 - 2\gamma\theta^3 x_2 - \gamma\theta^2 {x_1}^2 \\ &-4\gamma\theta^2 x_1 x_2 - \gamma\theta^2 {x_2}^2 - 2\gamma\theta {x_1}^2 x_2 \\ &-2\gamma\theta x_1 x_2^2 - \gamma x_1^2 x_2^2\} \cdot ((\theta+x_1)(\theta+x_2))^{-1}\end{aligned} \tag{A.22}$$

A.2.6 *Motif "F" - no dependency*

$$J(x_1, x_2) = \begin{pmatrix} -\gamma & 0 \\ 0 & -\gamma \end{pmatrix} \tag{A.23}$$

$$\lambda_1, \lambda_2 = -\gamma \tag{A.24}$$

A.3 PHENOTYPE-TO-GENOTYPE EQUATIONS

Below are equations that describe the "Phenotype-to-Genotype" map – equations that provide the genotypic values, calculated as the sum of parental allelic values, required to give a particular two-trait phenotype, given the parameters θ and γ.

A.3.1 *Motif "A" - single dependency, positive*

$$\alpha_1 = \gamma x_1 \tag{A.25}$$

$$\alpha_2 = \frac{\gamma x_2(\theta + x_1)}{x_1} \tag{A.26}$$

A.3.2 *Motif "B" - single dependency, negative*

$$\alpha_1 = \gamma x_1 \tag{A.27}$$

$$\alpha_2 = \frac{\gamma x_2(\theta + x_1)}{\theta} \tag{A.28}$$

A.3.3 *Motif "C" - double dependency, negative feedback loop*

$$\alpha_1 = \frac{\gamma x_1(\theta + x_2)}{\theta} \tag{A.29}$$

$$\alpha_2 = \frac{\gamma x_2(\theta + x_1)}{x_1} \tag{A.30}$$

A.3.4 *Motif "D" - double dependency, both positive*

$$\alpha_1 = \frac{\gamma x_1(\theta + x_2)}{\theta} \tag{A.31}$$

$$\alpha_2 = \frac{\gamma x_2(\theta + x_1)}{\theta} \tag{A.32}$$

A.3.5 *Motif "E" - double dependency, both negative*

$$\alpha_1 = \frac{\gamma x_1(\theta + x_2)}{x_2} \tag{A.33}$$

$$\alpha_2 = \frac{\gamma x_2(\theta + x_1)}{x_1} \tag{A.34}$$

A.3.6 *Motif "F" - no dependency*

$$\alpha_1 = \gamma x_1 \tag{A.35}$$

$$\alpha_2 = \gamma x_2 \tag{A.36}$$

A.4 LICENSE AGREEMENT

The body of CHAPTER 2 was previously published in *Evolution* (see Hether and Hohenlohe, 2014). Figures A.10 through A.14 are the license terms and conditions for reprinting in this dissertation.

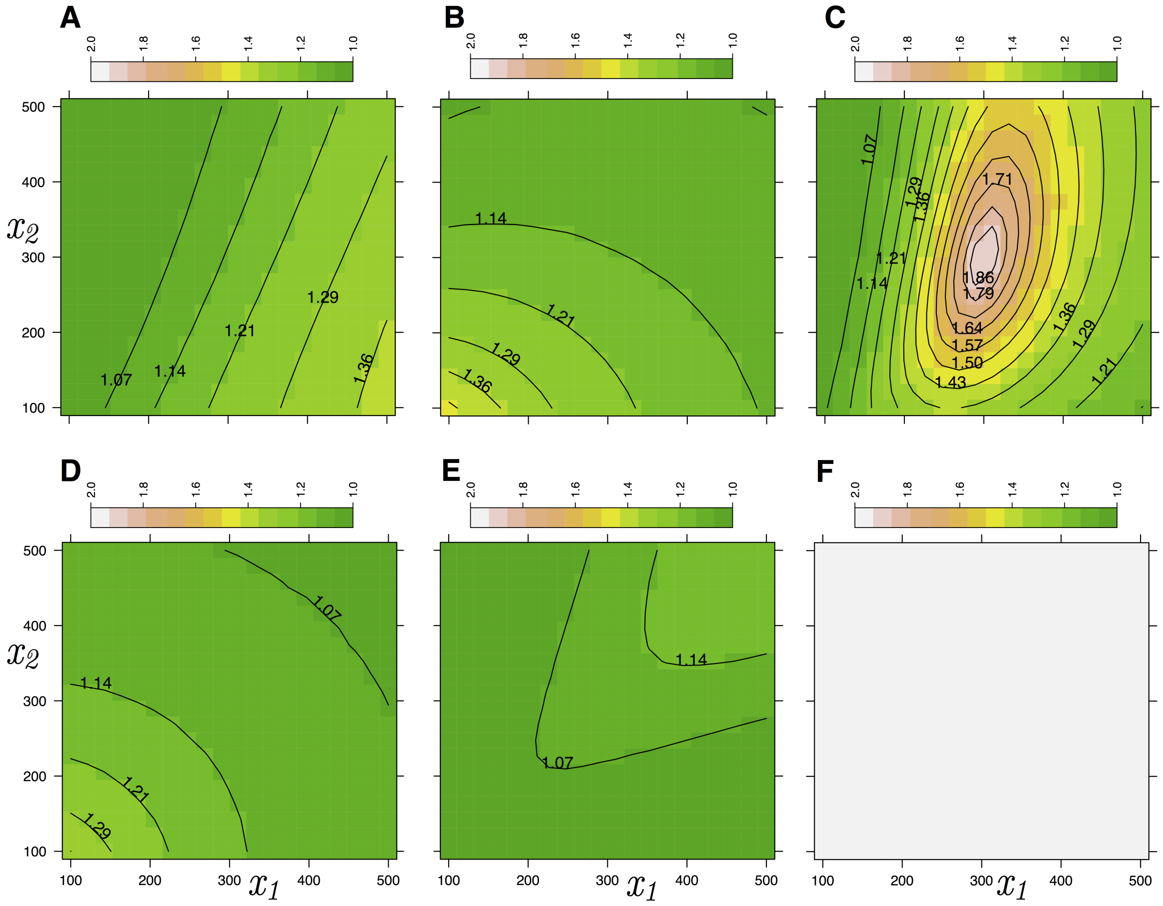


FIGURE A.1: Dimensionality of **M** across trait space. For each network motif (A-F), dimensionality was calculated as the sum of eigenvalues divided by the leading eigenvalue for each **M**-matrix along a 20-by-20 grid.

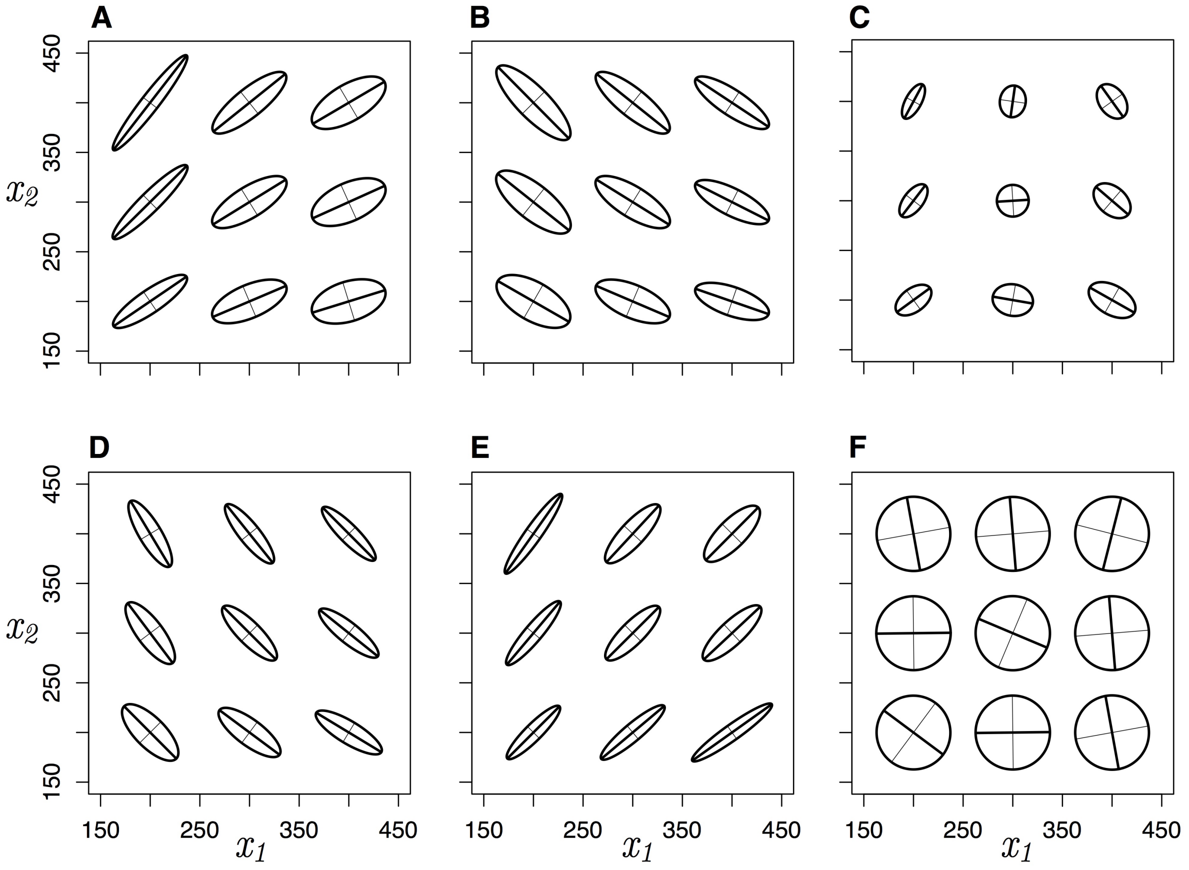


Figure A.2: The additive genetic (co)variance matrix **G** across phenotypic space. For each network motif (A-F), **G** matrices for nine populations are plotted as 95% confidence ellipses of breeding values (i.e., posterior mode of individual's random effects for each trait).

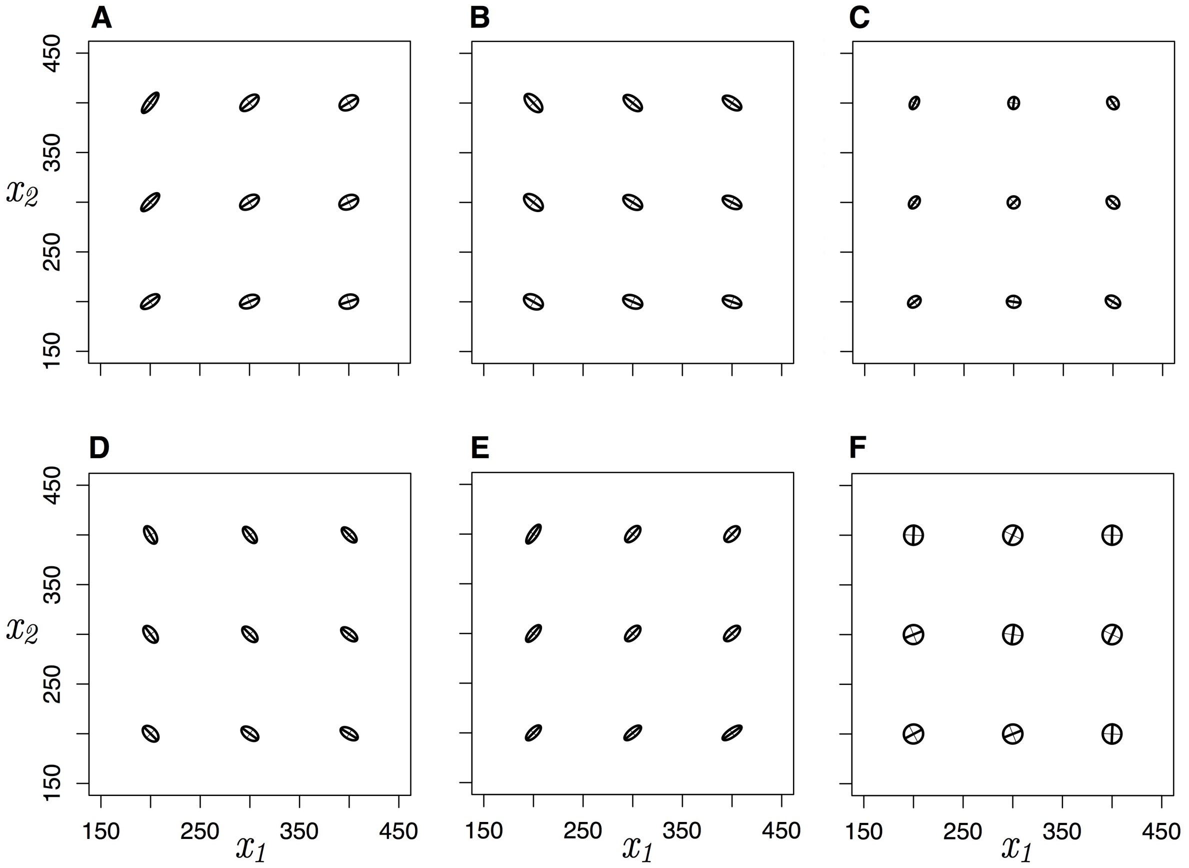


Figure A.3: The epistatic (co)variance matrix **E** across phenotypic space. For each network motif (A-F), **E** matrices for nine populations are plotted as 95% confidence ellipses of the posterior mode of individual's residual effects for each trait.

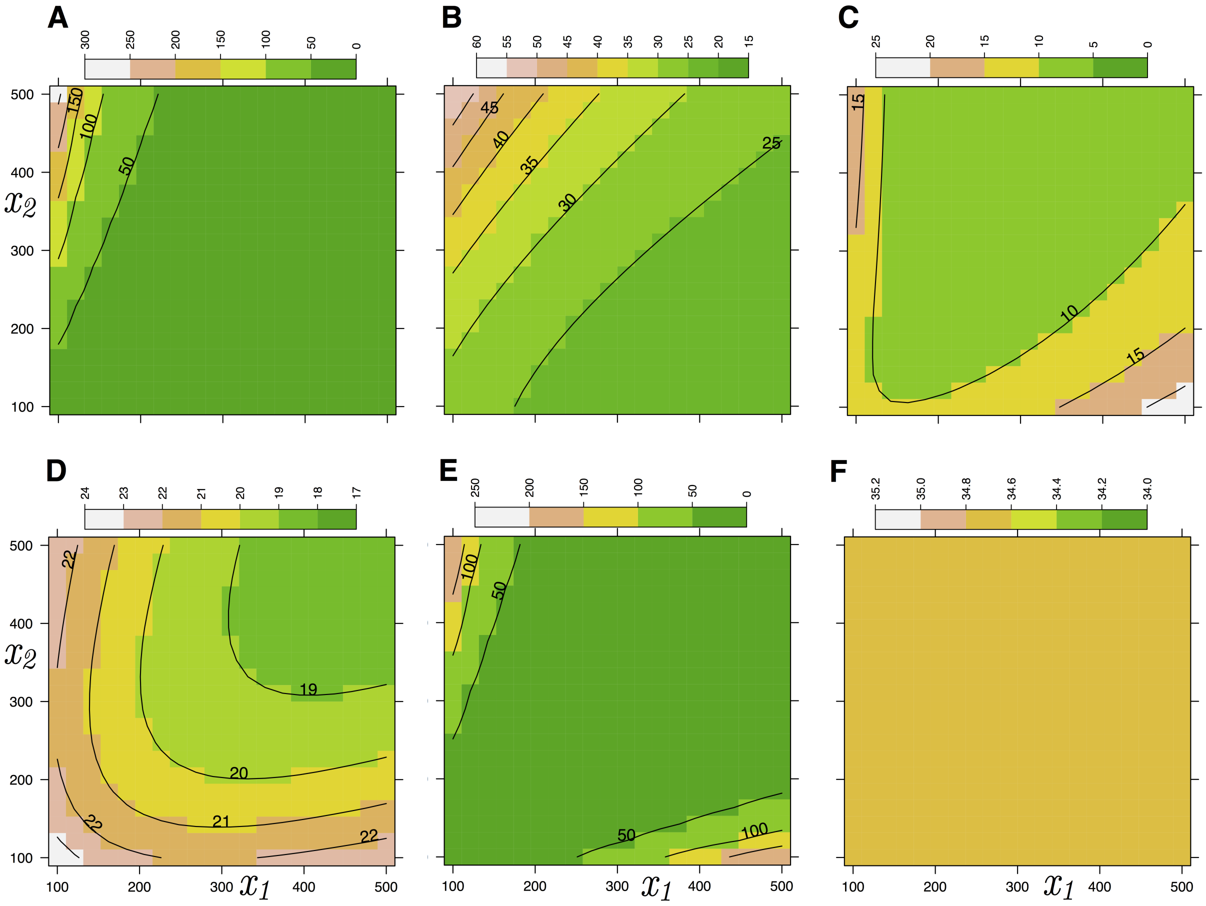


Figure A.4: Evolvability across trait space when mutation is limiting. For each network motif (A-F), evolvability was calculated as the average of the eigenvalues of **M**. Note that the scales are different across panels.

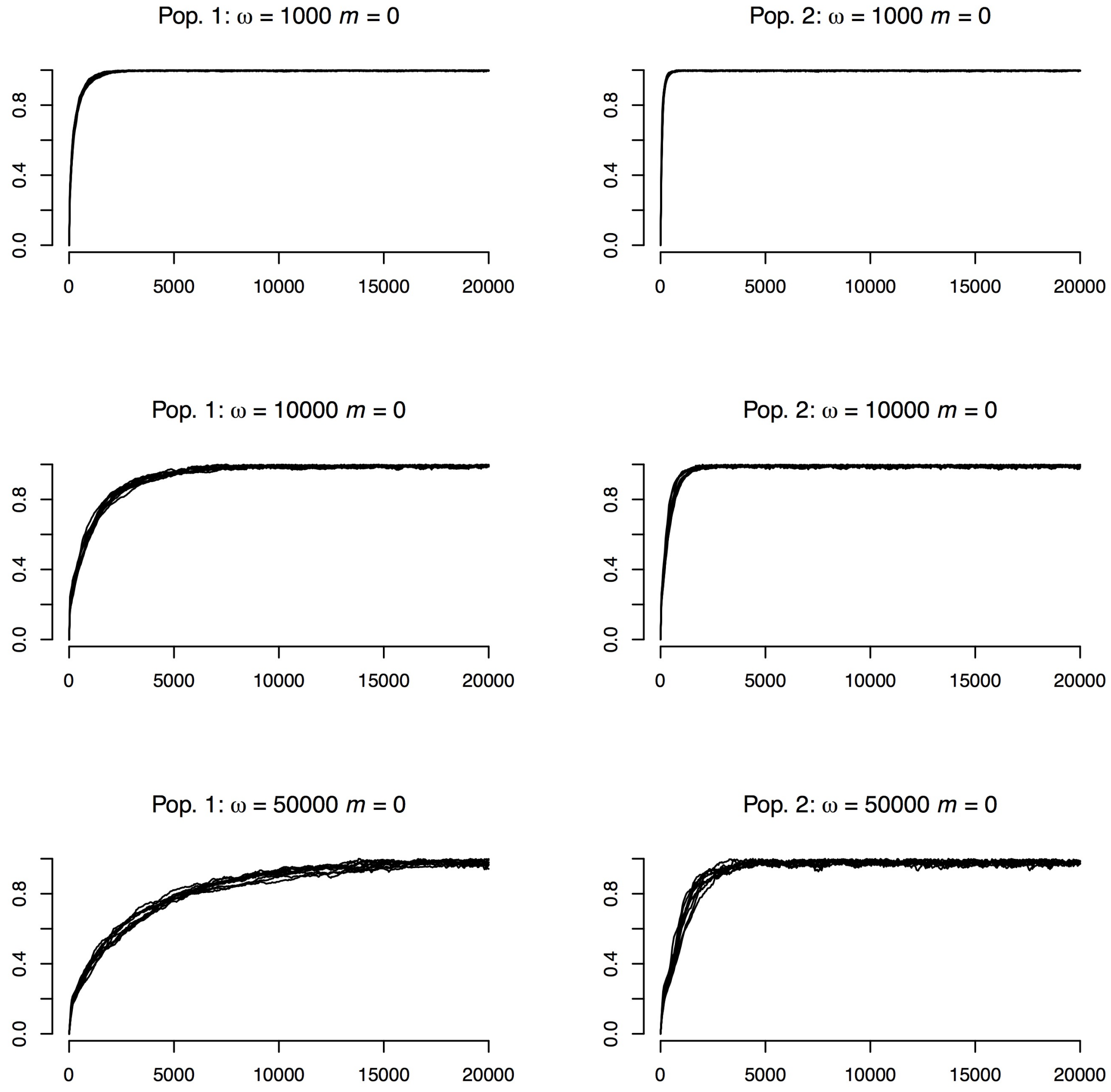


Figure A.5: The extent of local adaptation through time across 15 simulated replicates for Population "1" (left column) and "2" (right column). Rows denote seperate variance of stabilizing selection, ω. Shown here are data from migration = 0.

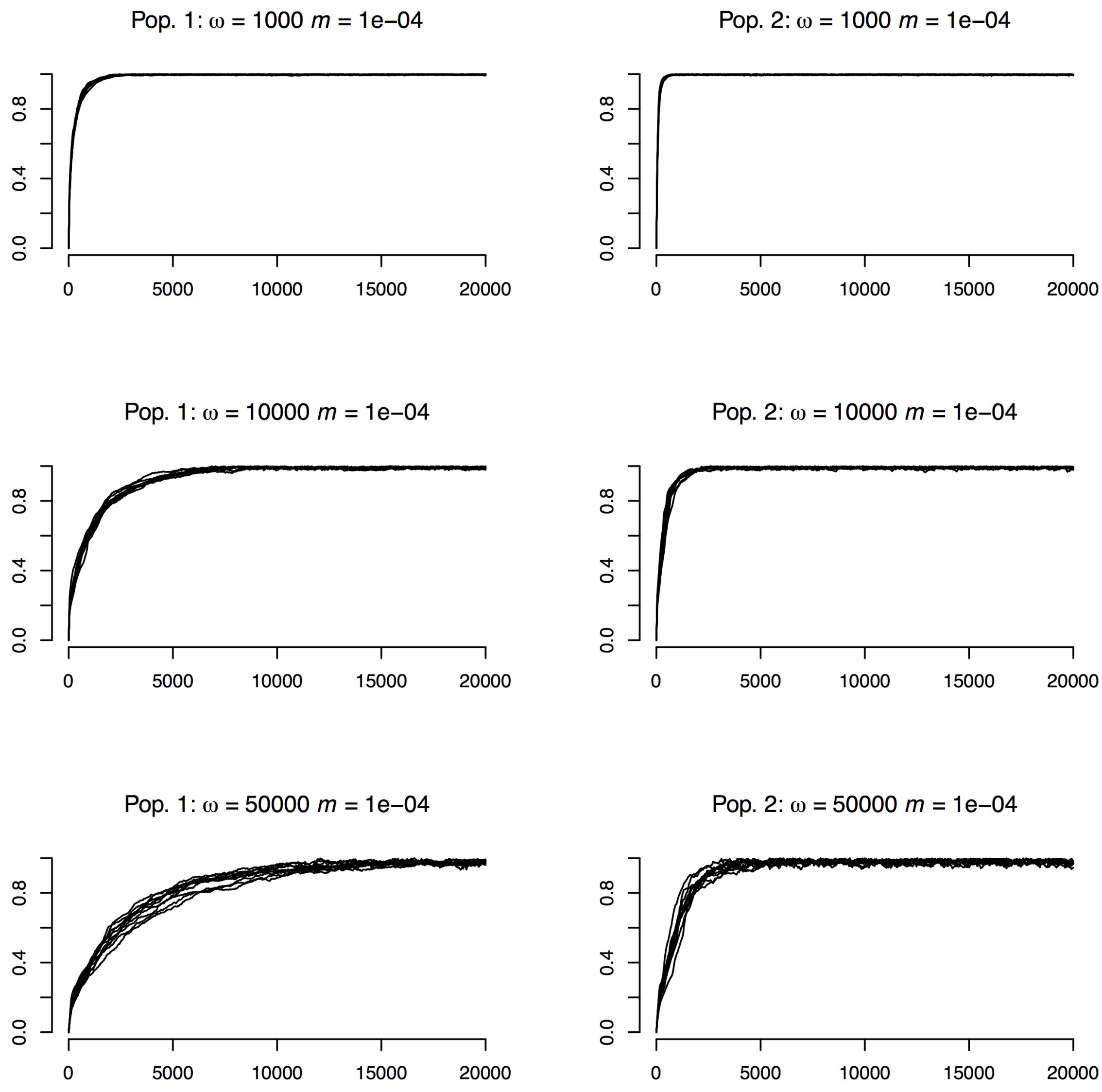


Figure A.6: The extent of local adaptation through time across 15 simulated replicates for Population "1" (left column) and "2" (right column). Rows denote seperate variance of stabilizing selection, ω. Shown here are data from migration = 0.0001.

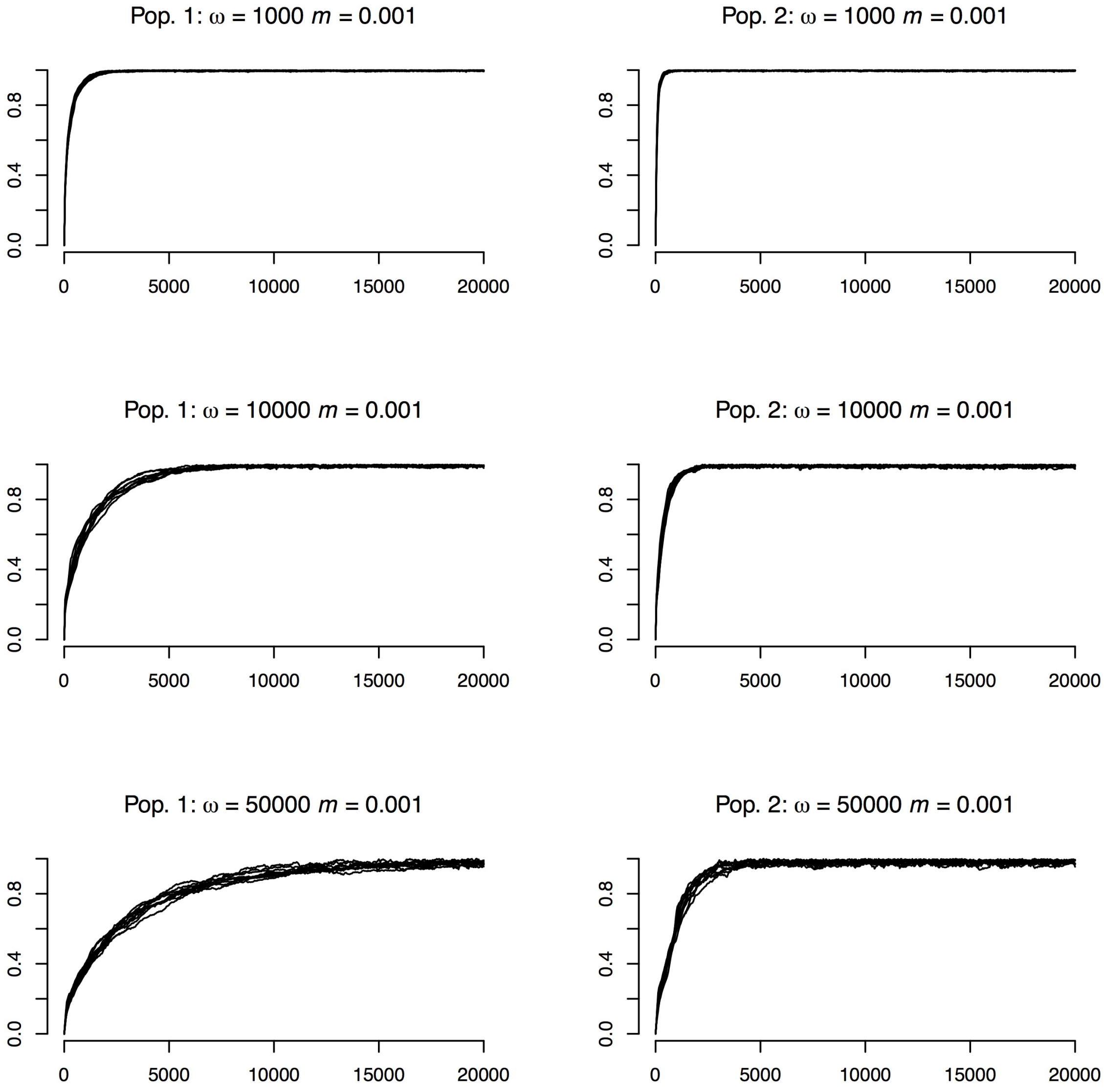


Figure A.7: The extent of local adaptation through time across 15 simulated replicates for Population "1" (left column) and "2" (right column). Rows denote seperate variance of stabilizing selection, ω. Shown here are data from migration = 0.001.

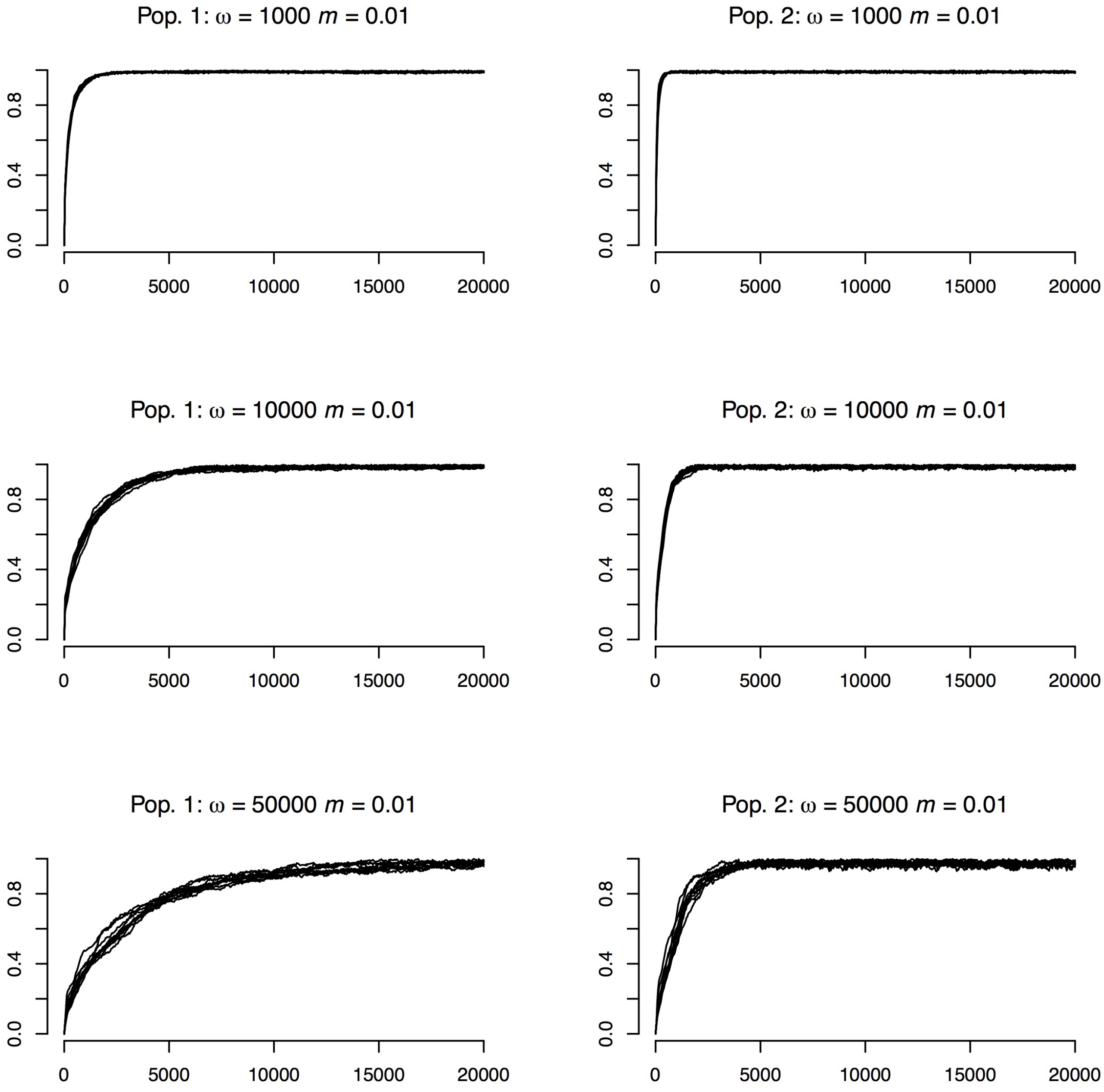


Figure A.8: The extent of local adaptation through time across 15 simulated replicates for Population "1" (left column) and "2" (right column). Rows denote seperate variance of stabilizing selection, ω. Shown here are data from migration = 0.01.

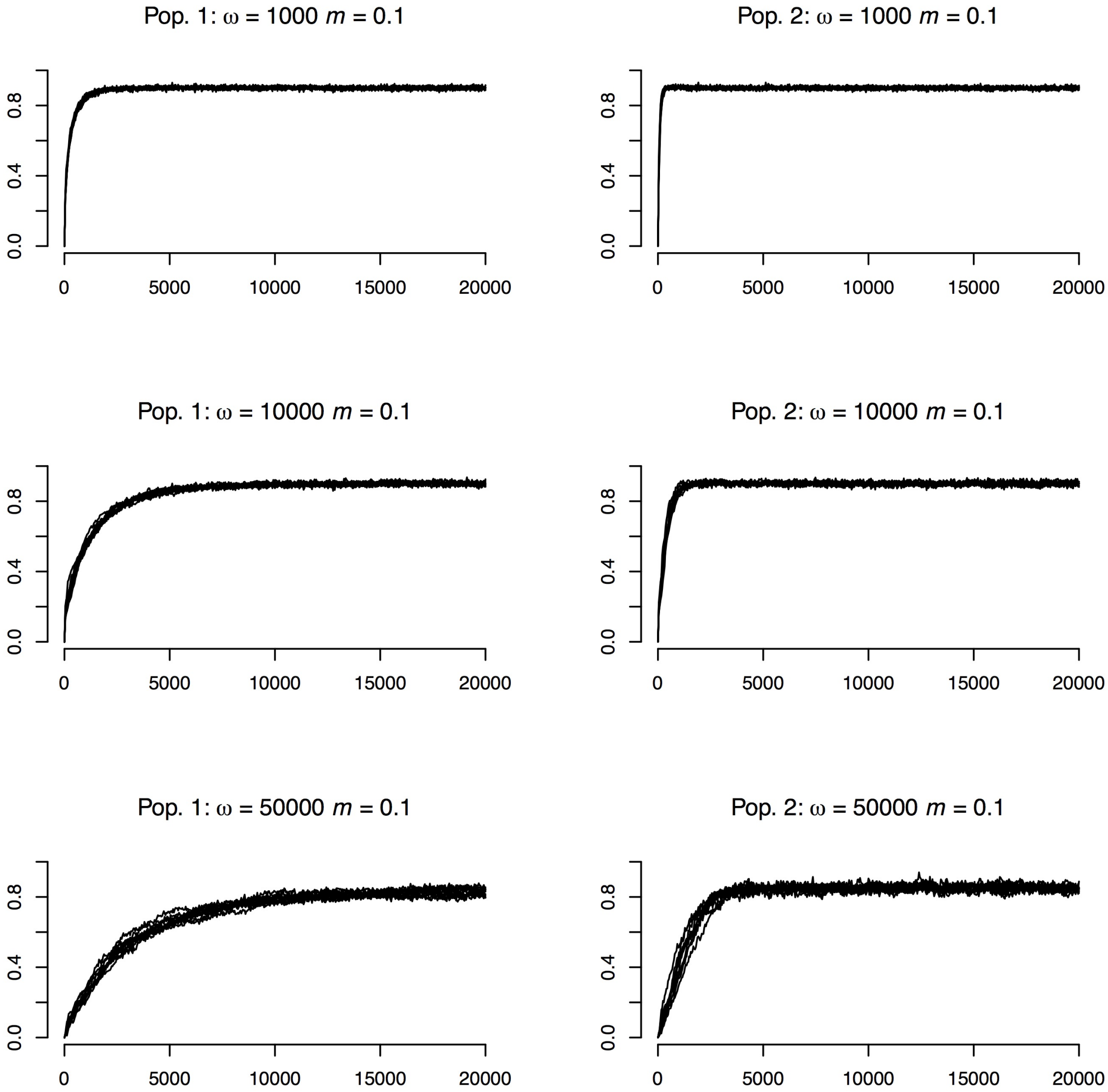
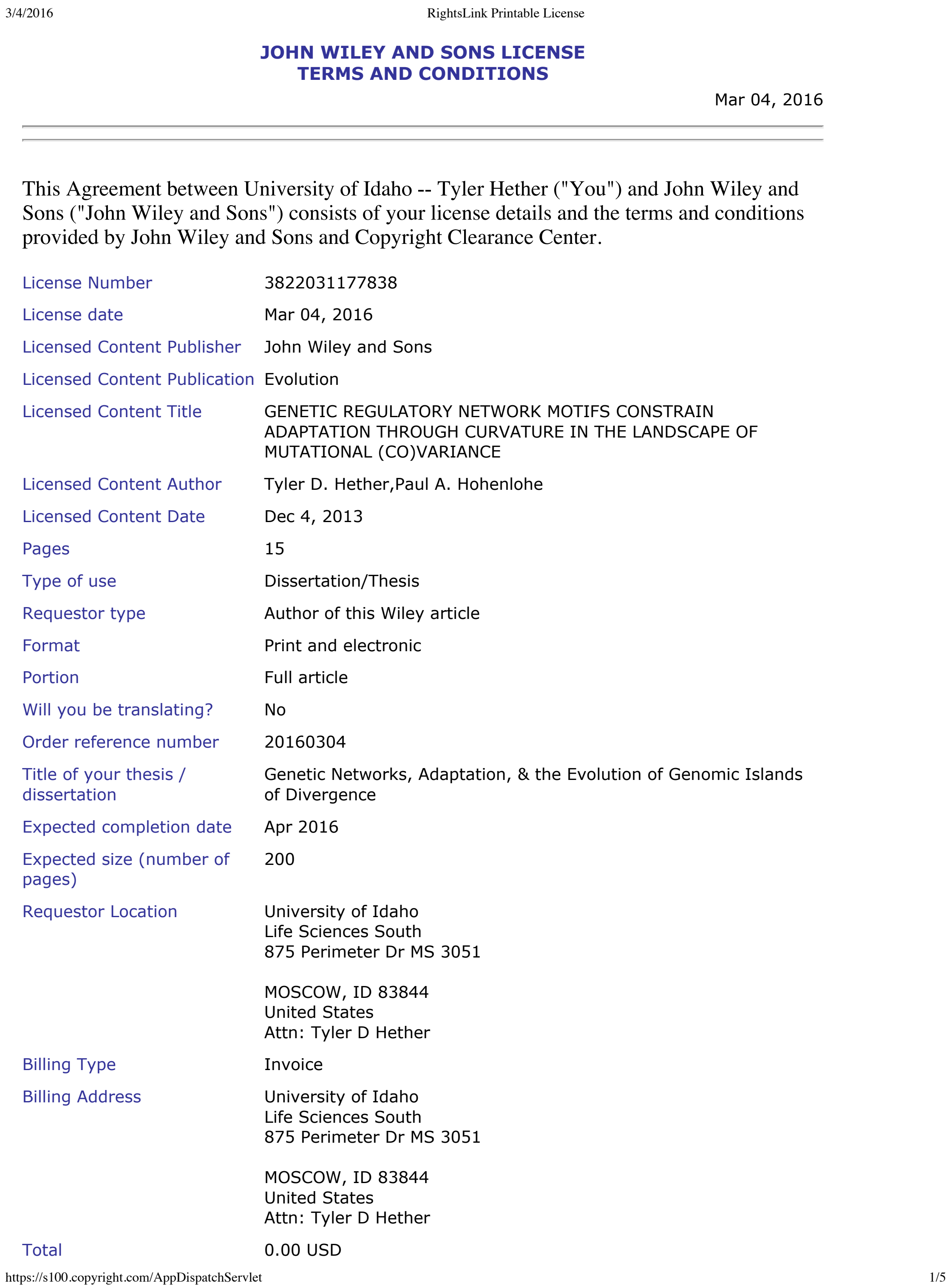


Figure A.9: The extent of local adaptation through time across 15 simulated replicates for Population "1" (left column) and "2" (right column). Rows denote seperate variance of stabilizing selection, ω. Shown here are data from migration = 0.1.



3/4/2016 RightsLink Printable License

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

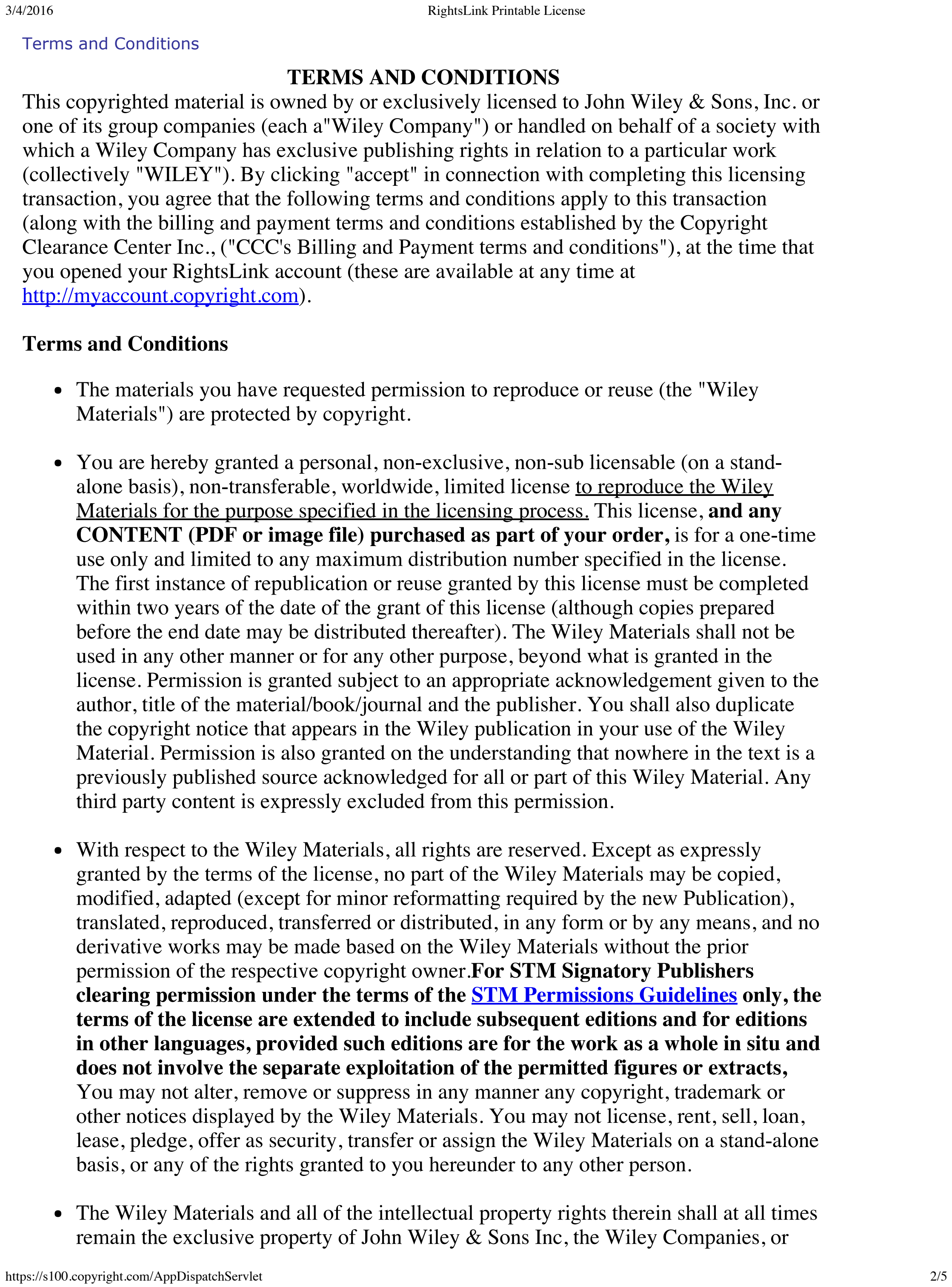
Mar 04, 2016

This Agreement between University of Idaho -- Tyler Hether ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	3822031177838
License date	Mar 04, 2016
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Evolution
Licensed Content Title	GENETIC REGULATORY NETWORK MOTIFS CONSTRAIN ADAPTATION THROUGH CURVATURE IN THE LANDSCAPE OF MUTATIONAL (CO)VARIANCE
Licensed Content Author	Tyler D. Hether,Paul A. Hohenlohe
Licensed Content Date	Dec 4, 2013
Pages	15
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Order reference number	20160304
Title of your thesis / dissertation	Genetic Networks, Adaptation, & the Evolution of Genomic Islands of Divergence
Expected completion date	Apr 2016
Expected size (number of pages)	200
Requestor Location	University of Idaho Life Sciences South 875 Perimeter Dr MS 3051 MOSCOW, ID 83844 United States Attn: Tyler D Hether
Billing Type	Invoice
Billing Address	University of Idaho Life Sciences South 875 Perimeter Dr MS 3051 MOSCOW, ID 83844 United States Attn: Tyler D Hether
Total	0.00 USD

https://s100.copyright.com/AppDispatchServlet 1/5

Figure A.10: License Agreement 1 of 5.



3/4/2016 RightsLink Printable License

Terms and Conditions

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

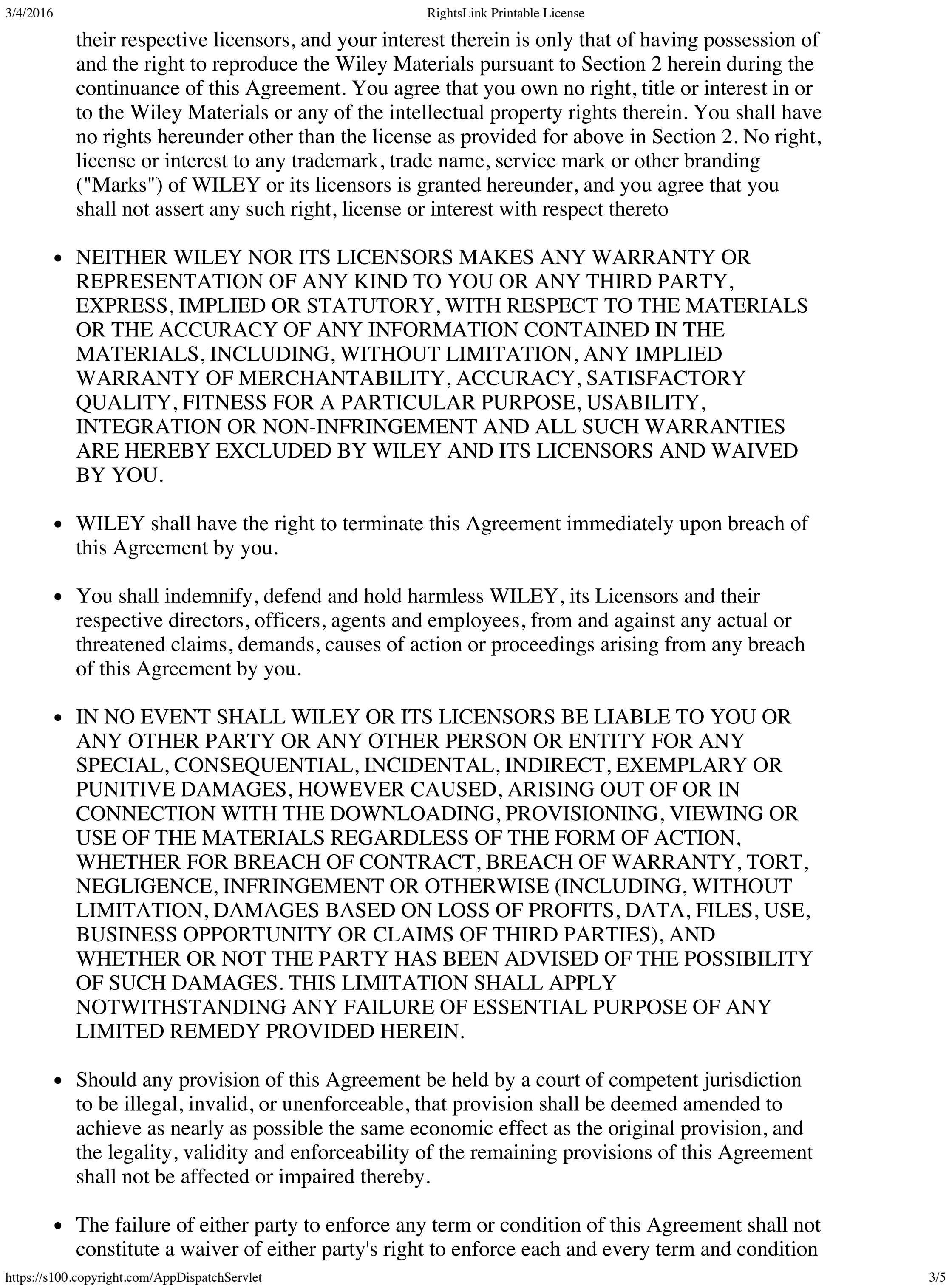
- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or

https://s100.copyright.com/AppDispatchServlet 2/5

FIGURE A.11: License Agreement 2 of 5.



their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

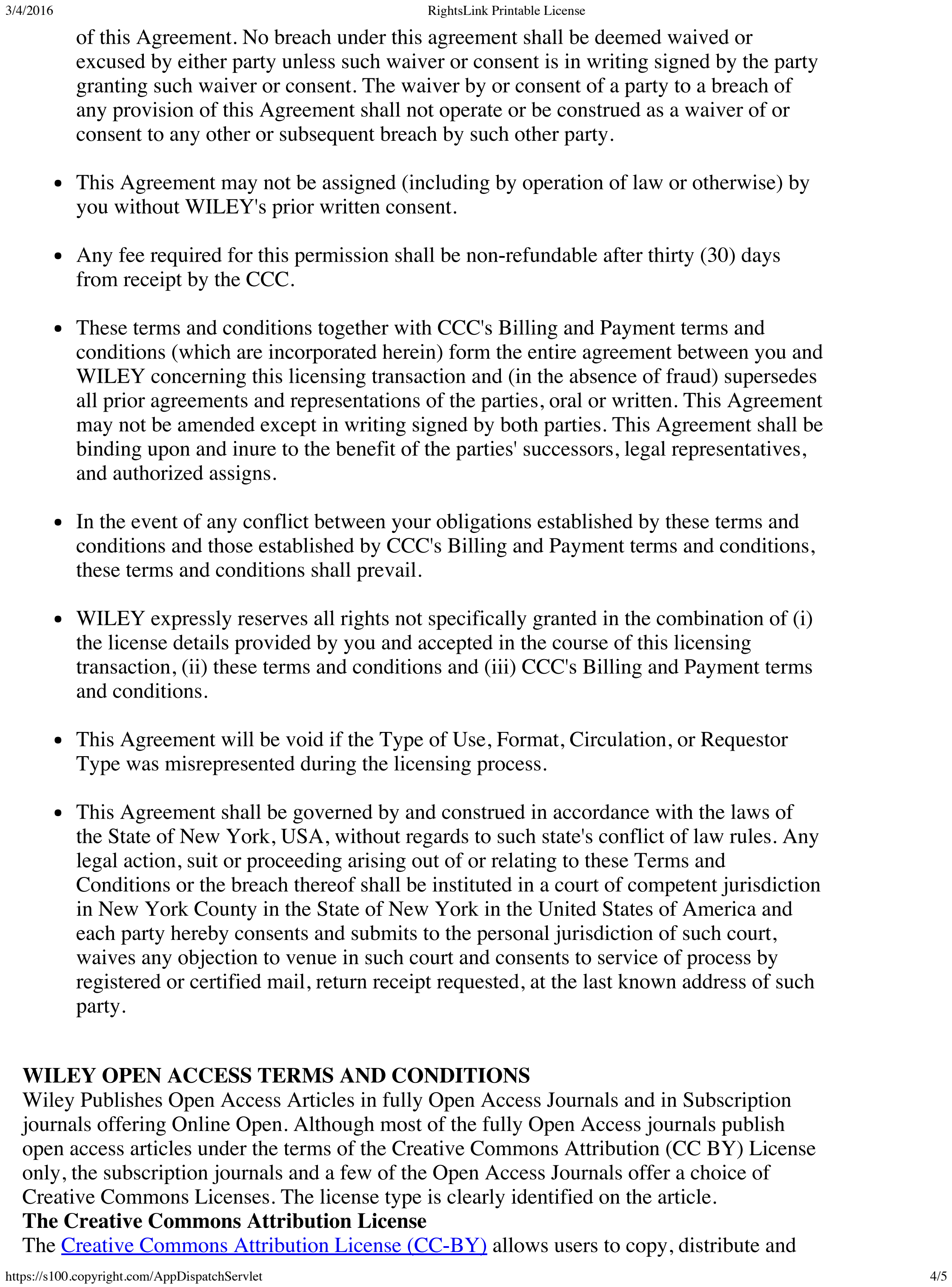
- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition

Figure A.12: License Agreement 3 of 5.



3/4/2016 RightsLink Printable License

of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

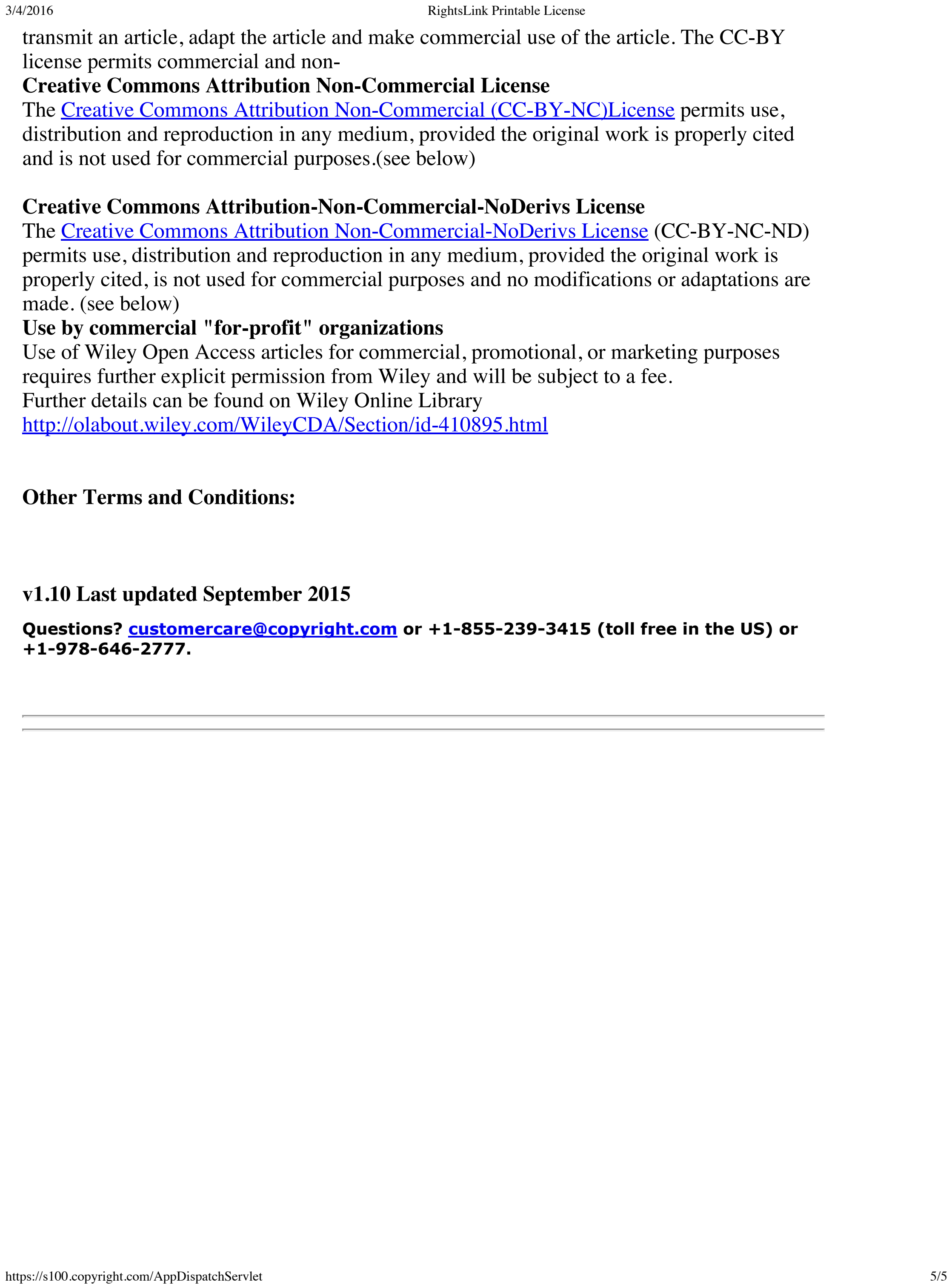
- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS
Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.
The Creative Commons Attribution License
The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and

https://s100.copyright.com/AppDispatchServlet 4/5

Figure A.13: License Agreement 4 of 5.



3/4/2016 RightsLink Printable License

transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

https://s100.copyright.com/AppDispatchServlet 5/5

Figure A.14: License Agreement 5 of 5.

APPENDIX B

Supplementary Information to Chapter 3

B.1 JOINT ALLELE FREQUENCY CHANGE IN A SINGLE LOCUS, THREE ALLELE MODEL

In this section we describe the deterministic change in allele frequencies for a single locus under selection that contains three alleles. We denote the 3 alleles – labeled 1, 2, & 3 – in vector form:

$$\mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} \tag{B.1}$$

To calculate the joint change in allele frequencies after selection we modified formula given in Rice (2004). Consider a matrix of genotype fitness values **F** for genotype w_{ij} derived from the simulated example in Chapter 3:

$$\mathbf{F} = \begin{pmatrix} w_{11} = 0.07 & w_{12} = 0.99 & w_{13} = 0.78 \\ w_{21} = 0.99 & w_{22} = 0.07 & w_{23} = 0.23 \\ w_{31} = 0.78 & w_{32} = 0.23 & w_{33} = 0.53 \end{pmatrix} \tag{B.2}$$

The change of allele frequencies after selection can be calculated as:

$$\triangle\,\mathbf{p} = \frac{1}{2\bar{w}}\mathbf{G}(2\mathbf{w}^{\star})^{T} \tag{B.3}$$

where **G** is the allelic (co)variance matrix:

$$\mathbf{G} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) \end{pmatrix} \tag{B.4}$$

In Equation B.3, $\mathbf{w}^{\star}$ is equal to **pF** and $\bar{w}$ is calculated similarly to Equation 5.5 but here we only consider a single locus. That is, $\bar{w}$ is the sum of the dot product of **p** and $\mathbf{w}^{\star}$.

A ternary plot showing the change in allele frequencies for different starting values of **p** is given in Figure B.1. There are two stable heterozygous equilibrium points that correspond to ridges B and C in Figure 3.11. The first occurs when the '-' allele and the '0' allele are equal at 0.5. The second occurs when the '-' allele and '+' allele is approximately 0.26 and 0.74, respectively.

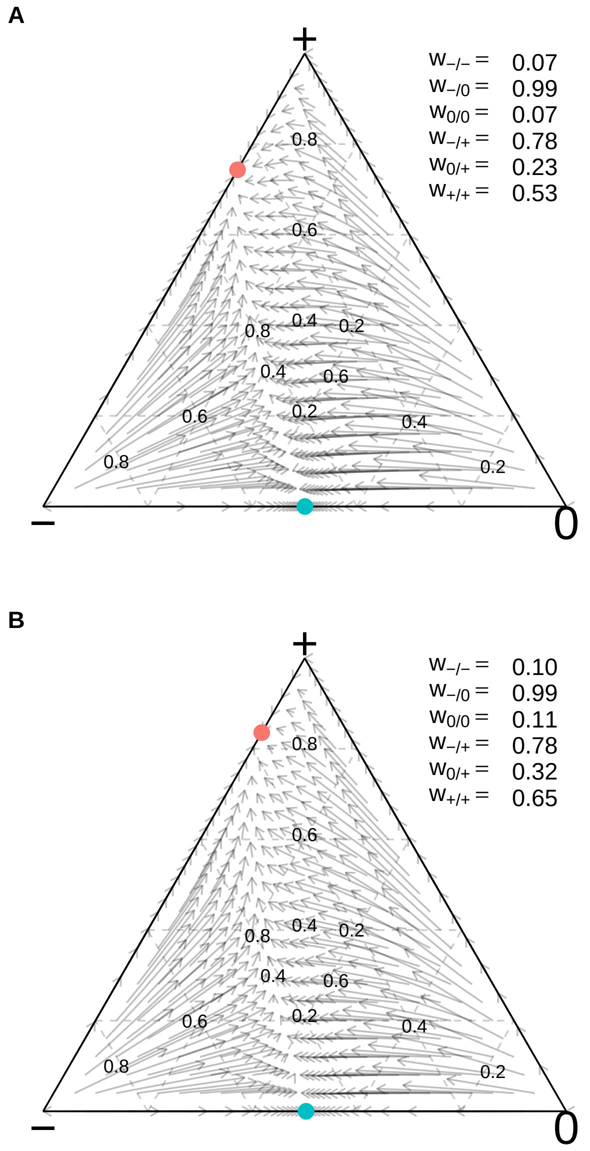


Figure B.1: Vector field of change in allele frequencies after a bout of selection given the two sets of genotype fitness values dervived from the simulated example in Chapter 3 (see also Equation B.2). Each vertix indicates the fixation of a given allele and allelic polymorphism occurs within the simplex. Vectors show the direction of allele frequency change after selection given regularly spaced starting allele frequencies. Values inside the plot show the relative allele frequencies. The two stable, heterzygote allele frequencies are denoted by dots (blue = "-/0" heterozgyote; red = "-/+" heterozgyote). A) Fitnesses during which the '+' allele failed to fix (i.e., dashed vertical lines in Figure 3.9; generation 1,942,134). B) Fitnesses while the '+' allele rose to fixation (i.e., dotted vertical lines in Figure 3.9; generation 1,951,254).

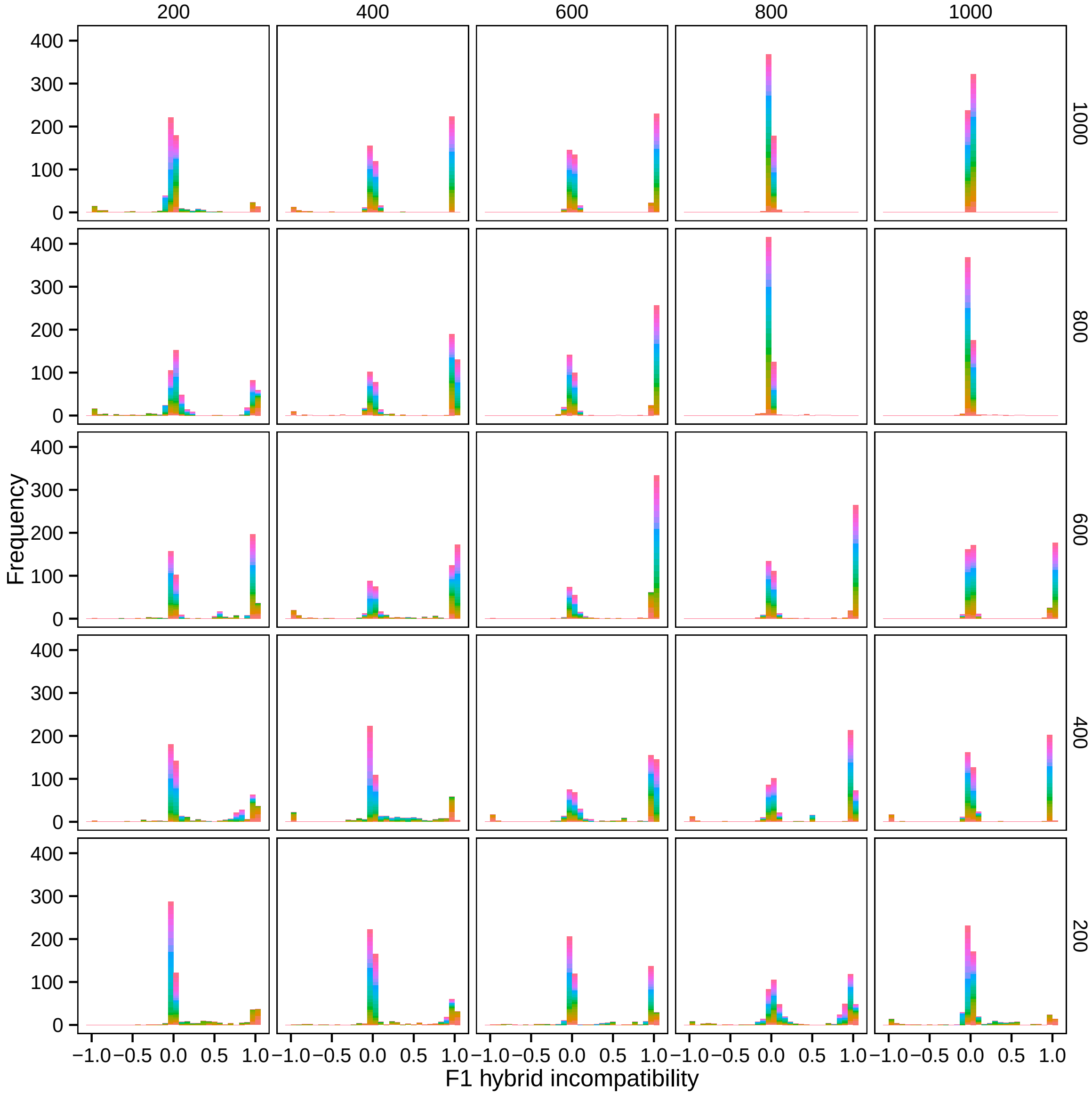


Figure B.2: Distribution of F_1 hybrid incompatibility between 20 population pairs through time. At the beginning of the simulation each population was initialized at point $x_1 = 1000$, $x_2 = 1000$ (upper right panel). Each sub-panel shows frequency counts of F_1 incompatibility through time (orange to purple shows early to later generations, respectively). Columns and rows show the location of the new x_1 and x_2 optimum, respectively.

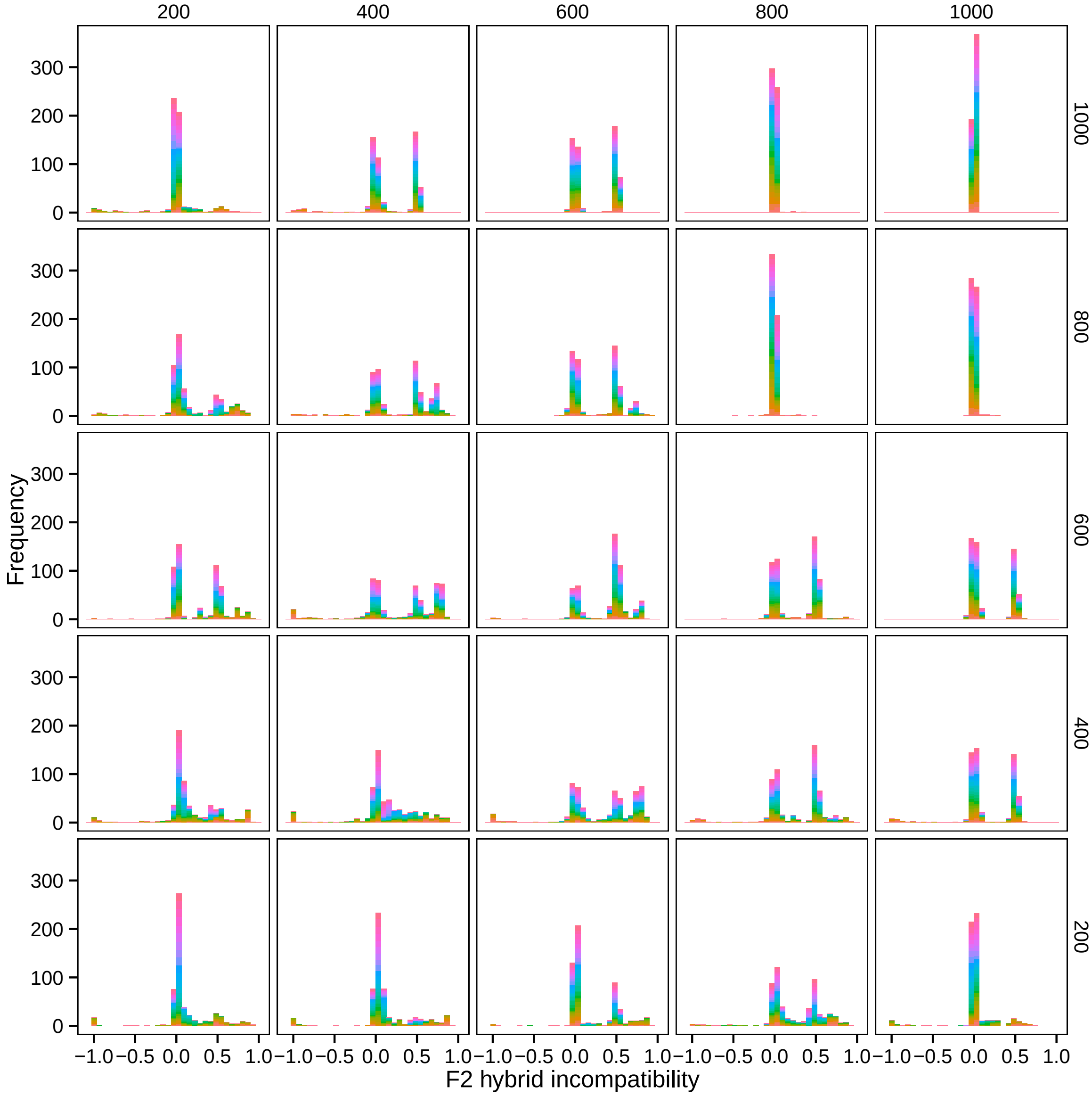


Figure B.3: Distribution of F_2 hybrid incompatibility between 20 population pairs through time. At the beginning of the simulation each population was initialized at point $x_1 = 1000$, $x_2 = 1000$ (upper right panel). Each sub-panel shows frequency counts of F_2 incompatibility through time (orange to purple shows early to later generations, respectively). Columns and rows show the location of the new x_1 and x_2 optimum, respectively.

Appendix C

Supplementary Information to Chapter 4

This section contains supplemental figures for Chapter 4.

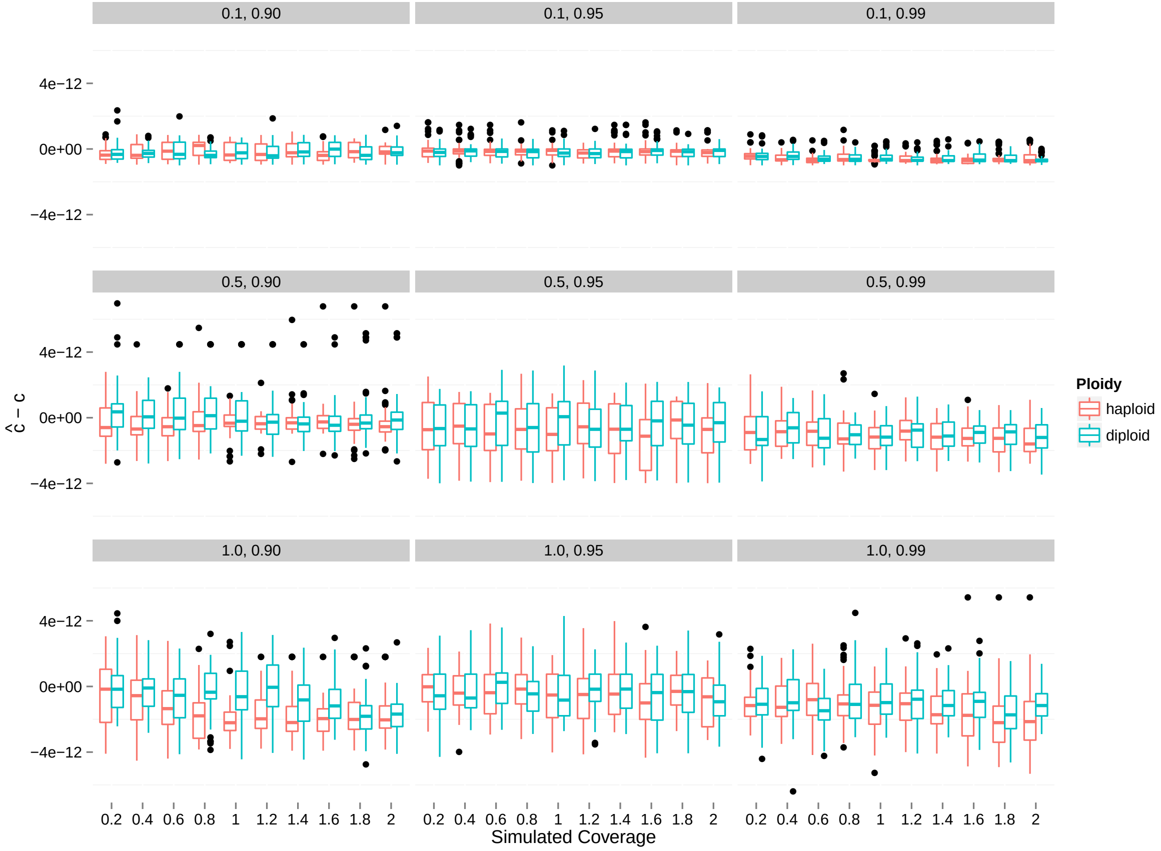


Figure C.1: Performance of *HMMancestry* in estimating genome-wide recombination rate. Shown is the difference between the ML estimate and the simulated (true) value of recombination rate c in cM/kb, for different levels of recombination rate (rows), assignment probability (columns), ploidy (colors), and sequencing coverage (x-axis). Box and whisker plots show median (plus first and second quartiles) of 50 simulated replicates for each parameter combination.

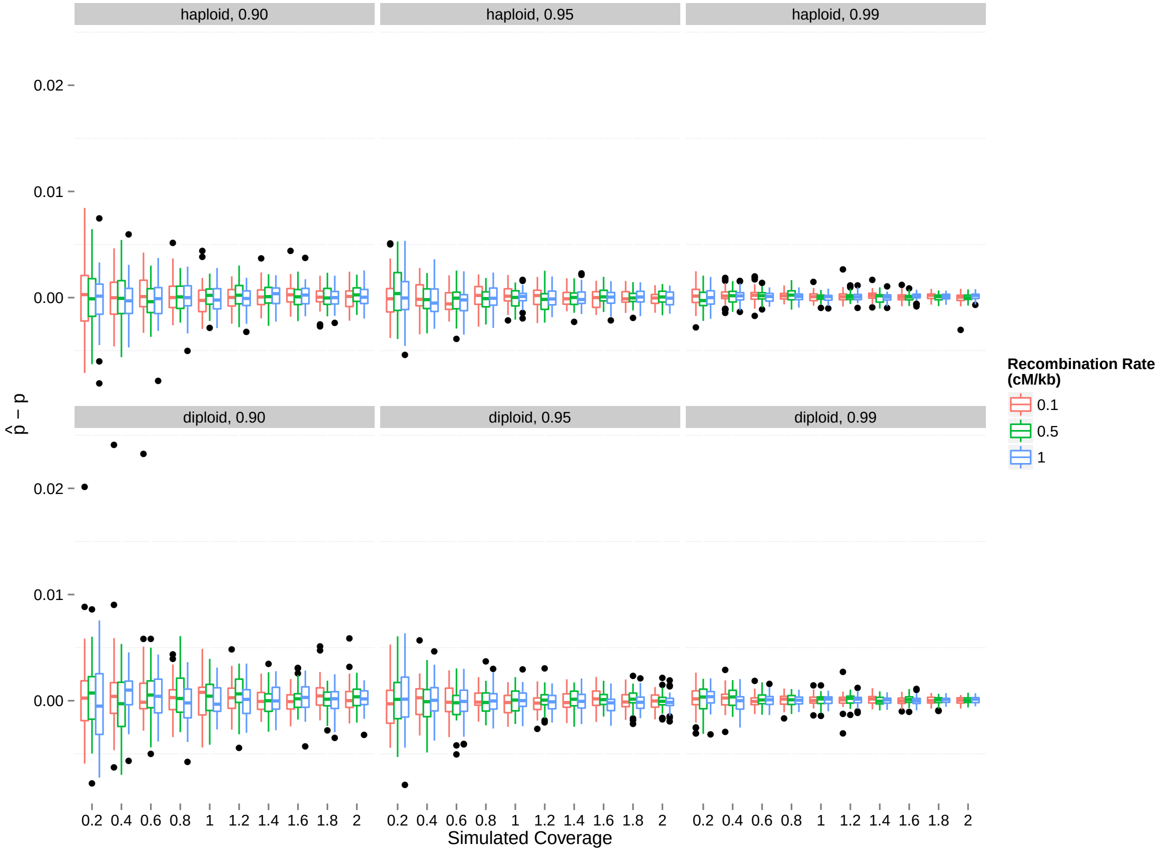


Figure C.2: Performance of *HMMancestry* in estimating assignment probability.Shown is the difference between the ML estimate and the simulated (true) value of assignment probability *p*, which is the probability of a sequence read correctly assigning to one parent, for different levels of ploidy (rows), true assignment probability (columns), recombination rate (colors), and sequencing coverage (x-axis). Box and whisker plots show median (plus first and second quartiles) of 50 simulated replicates for each parameter combination.

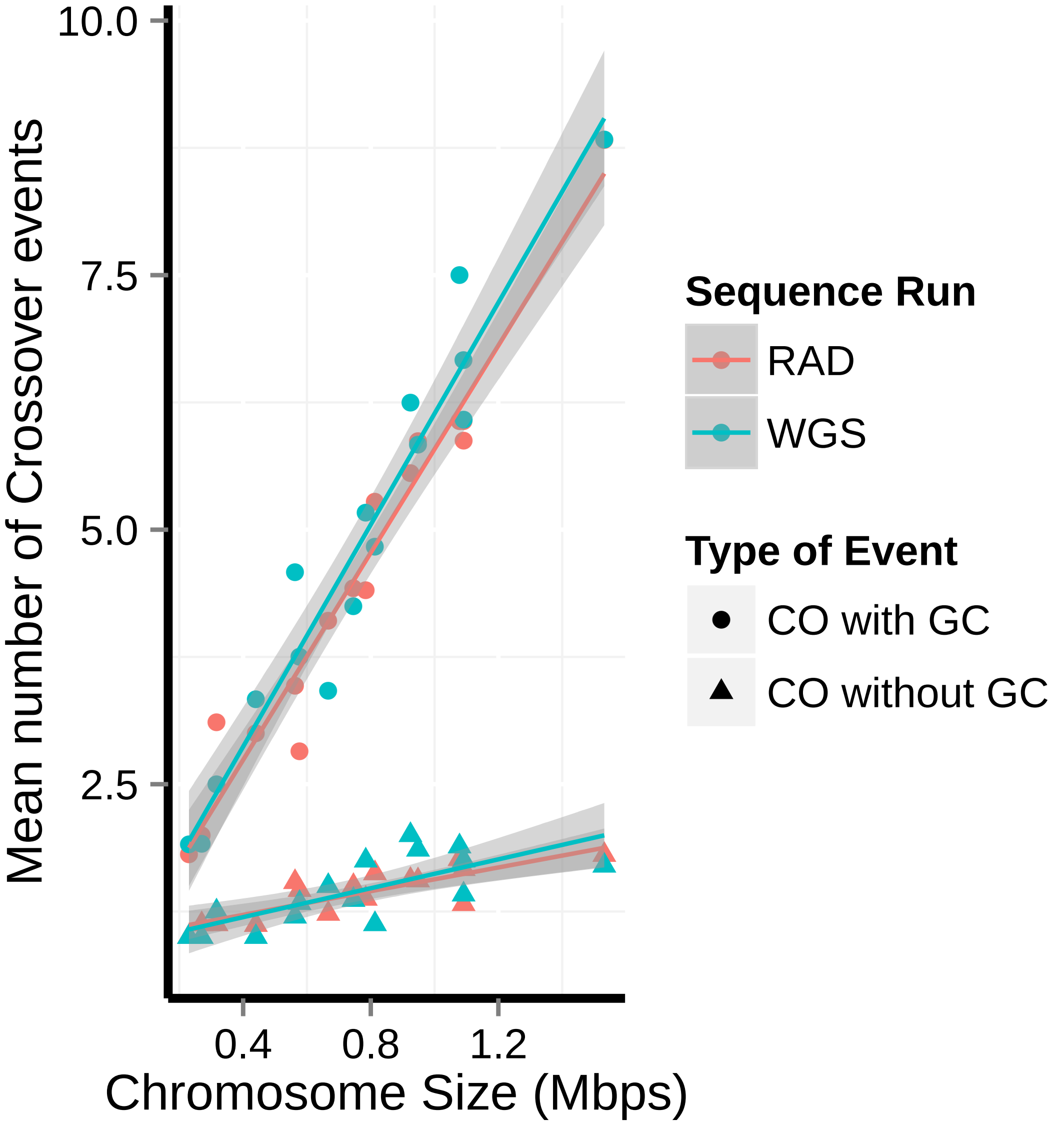


FIGURE C.3: The average number of crossover events with or without a gene conversion tract as a function of chromosome size for WGS and RAD datasets.

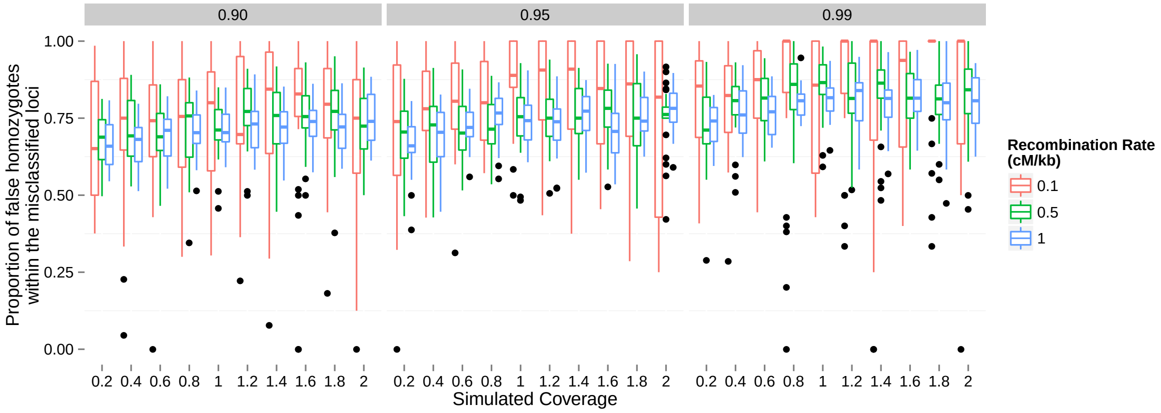


FIGURE C.4: Misclassification of homozygous sites as heterozygous from simulated data. Box plots show variation across 50 replicate runs. Y-axis shows the proportion of the misclassified loci there were erroneously assigned to one of two parental types. Columns denote the simulated assignment probability (p).

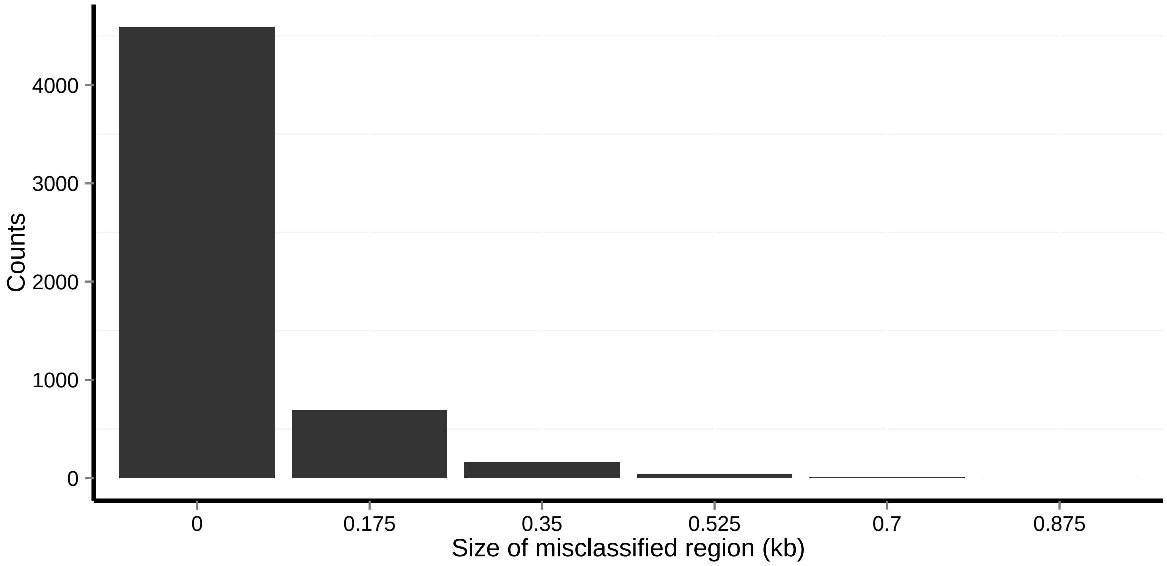


Figure c.5: Frequency spectrum of the size of misclassified regions in simulated data. Histograms show the chromosome size of continuous loci that were misclassified in the HMM. Singletons, in which a single, misclassified locus was flanked correctly assigned loci on either side were given a size of zero. Parameter values in the simulated data we derived from empirical estimates. Number of loci = 73,294, displacement between each SNP = 175 bp, c = 3.1, p = 0.993, coverage = 2.8X, frequency of COs = 0.51, frequency of conversion = 0.85, length of conversion = 1,920 bps.

appendix d

Supplementary Information to Chapter 5

This section contains supplemental figures for Chapter 5.

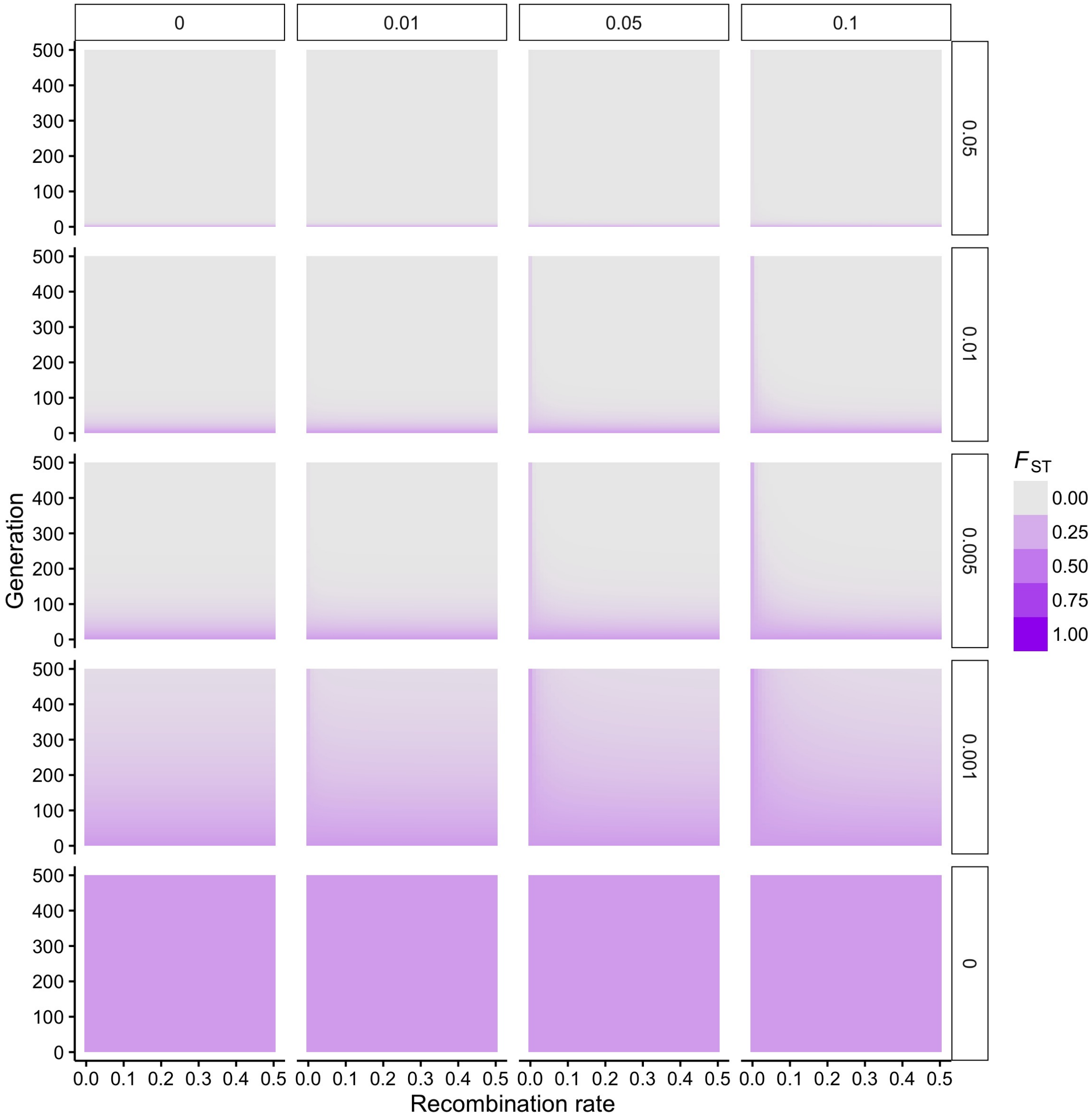


FIGURE D.1: Dynamics of divergence with gene flow under scenario #2 – obligate sexual, $m_f = 1$, LD=0, X-LD=0.125. For each panel, the extent of divergence (F_{ST}) at neutral loci are given across time. Rows indicate migration rate, m, between diverging populations and columns indicate the strength of divergent selection, s, at selected locus A.

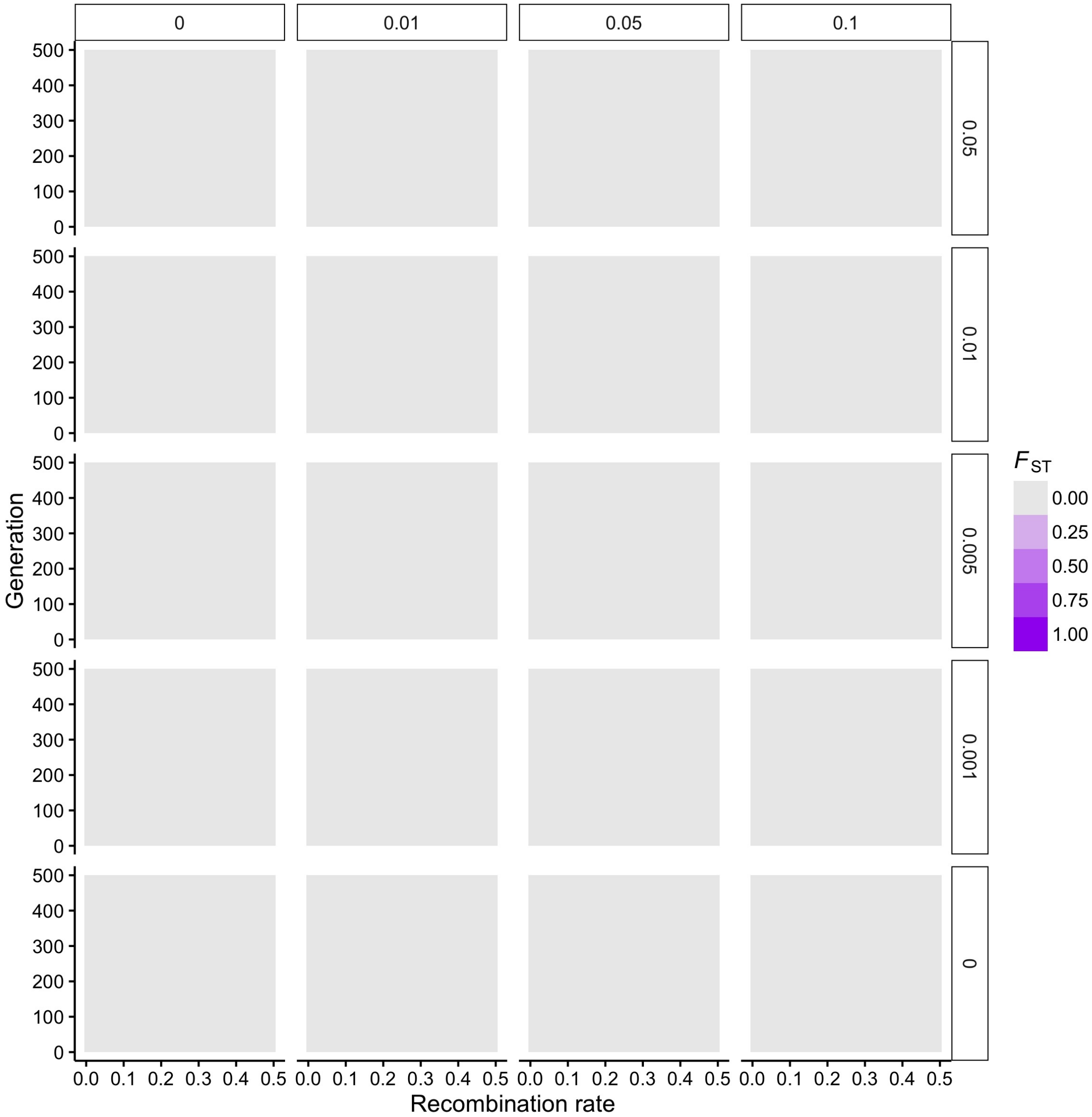


FIGURE D.2: Dynamics of divergence with gene flow under scenario #3 – obligate sexual, $m_f = 1$, LD=0, X-LD=0. For each panel, the extent of divergence (F_{ST}) at neutral loci are given across time. Rows indicate migration rate, m, between diverging populations and columns indicate the strength of divergent selection, s, at selected locus A.

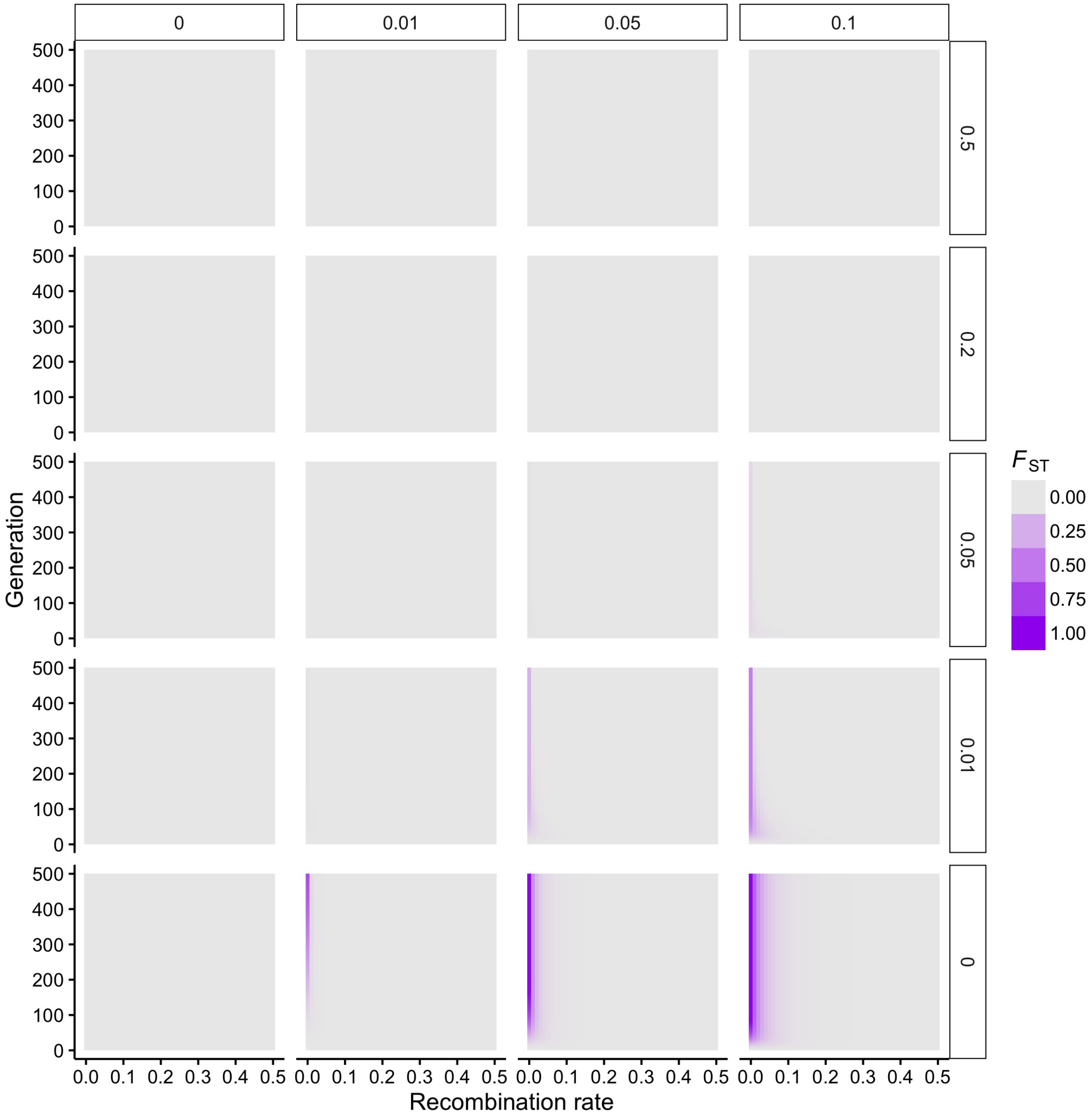


FIGURE D.3: Dynamics of divergence with gene flow under scenario #5 – obligate sexual, $m_f = 1$, LD=0.25 - 0.25r, X-LD=0.25 - 0.25r. For each panel, the extent of divergence (F_{ST}) at neutral loci are given across time. Rows indicate migration rate, m, between diverging populations and columns indicate the strength of divergent selection, s, at selected locus A. Note the change of migration rates investigated in this figure compared to Figure 5.2

.

APPENDIX E

Supplementary Information to Chapter 6

This section contains supplemental figures for Chapter 6.

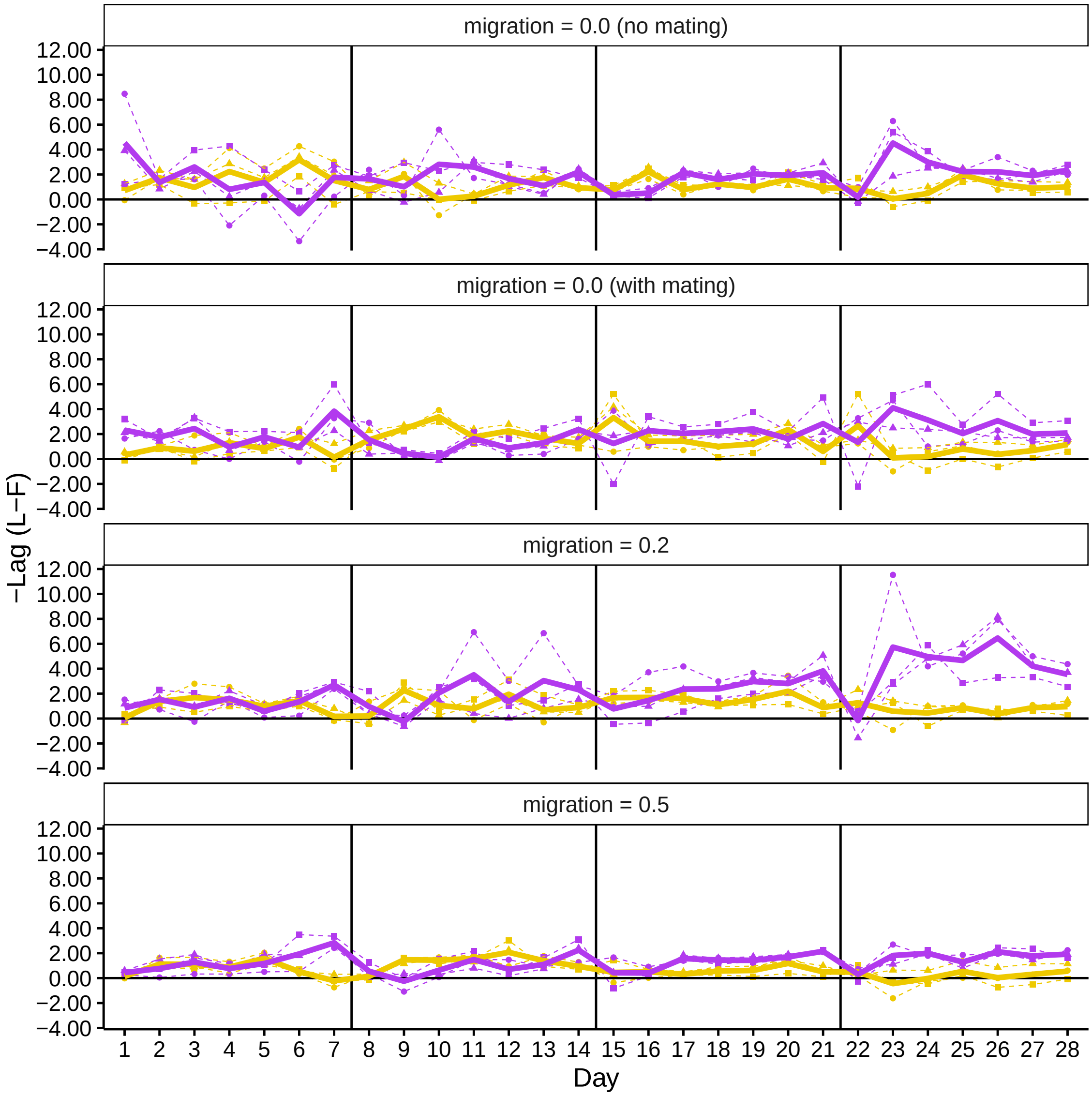


Figure E.1: The extent of local adaptation (LA) for lag during divergence with a given amount of migration. For each migration rate the extent of LA is given for each stress environment. Gold and purple denote NaCl and SDS habitats, respectively, and group means are denoted by solid color lines. Black vertical lines show where sporulation and migration (if applicable) occur between periods of asexual growth in batch.

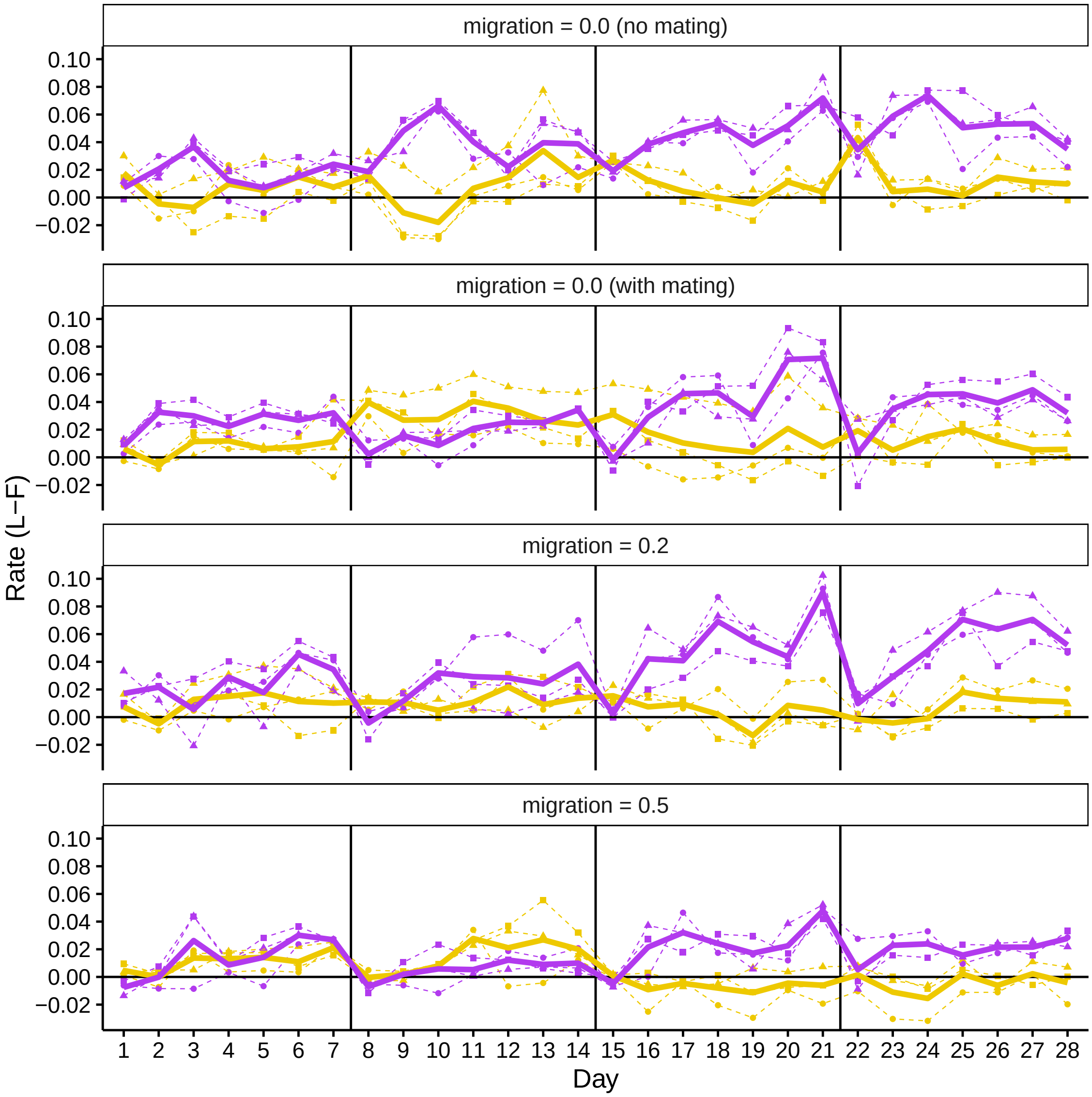


Figure E.2: The extent of local adaptation (LA) for rate during divergence with a given amount of migration. For each migration rate the extent of LA is given for each stress environment. Gold and purple denote NaCl and SDS habitats, respectively, and group means are denoted by solid color lines. Black vertical lines show where sporulation and migration (if applicable) occur between periods of asexual growth in batch.

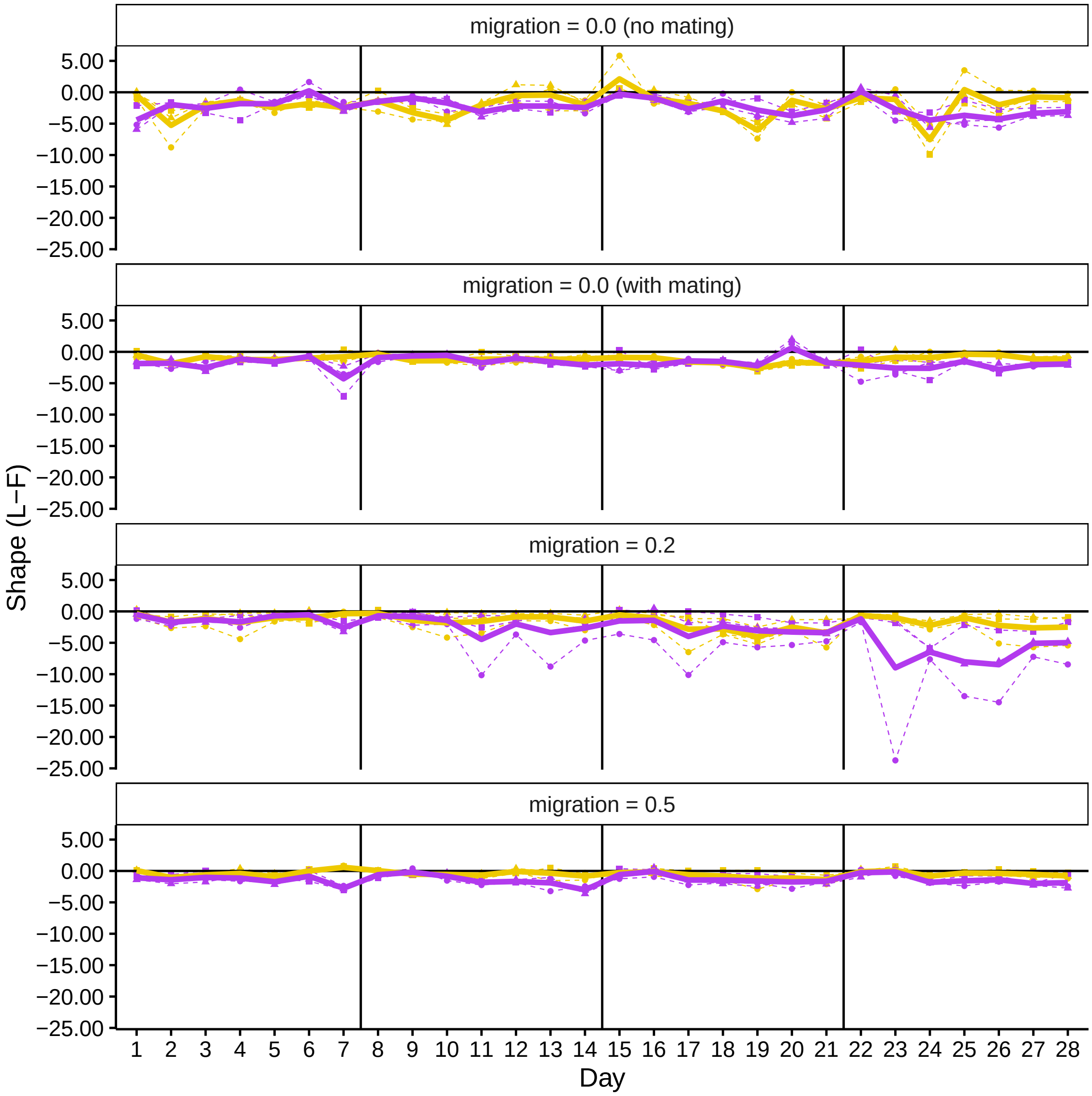


FIGURE E.3: The extent of local adaptation (LA) for rate during divergence with a given amount of migration. For each migration rate the extent of LA is given for each stress environment. Gold and purple denote NaCl and SDS habitats, respectively, and group means are denoted by solid color lines. Black vertical lines show where sporulation and migration (if applicable) occur between periods of asexual growth in batch.

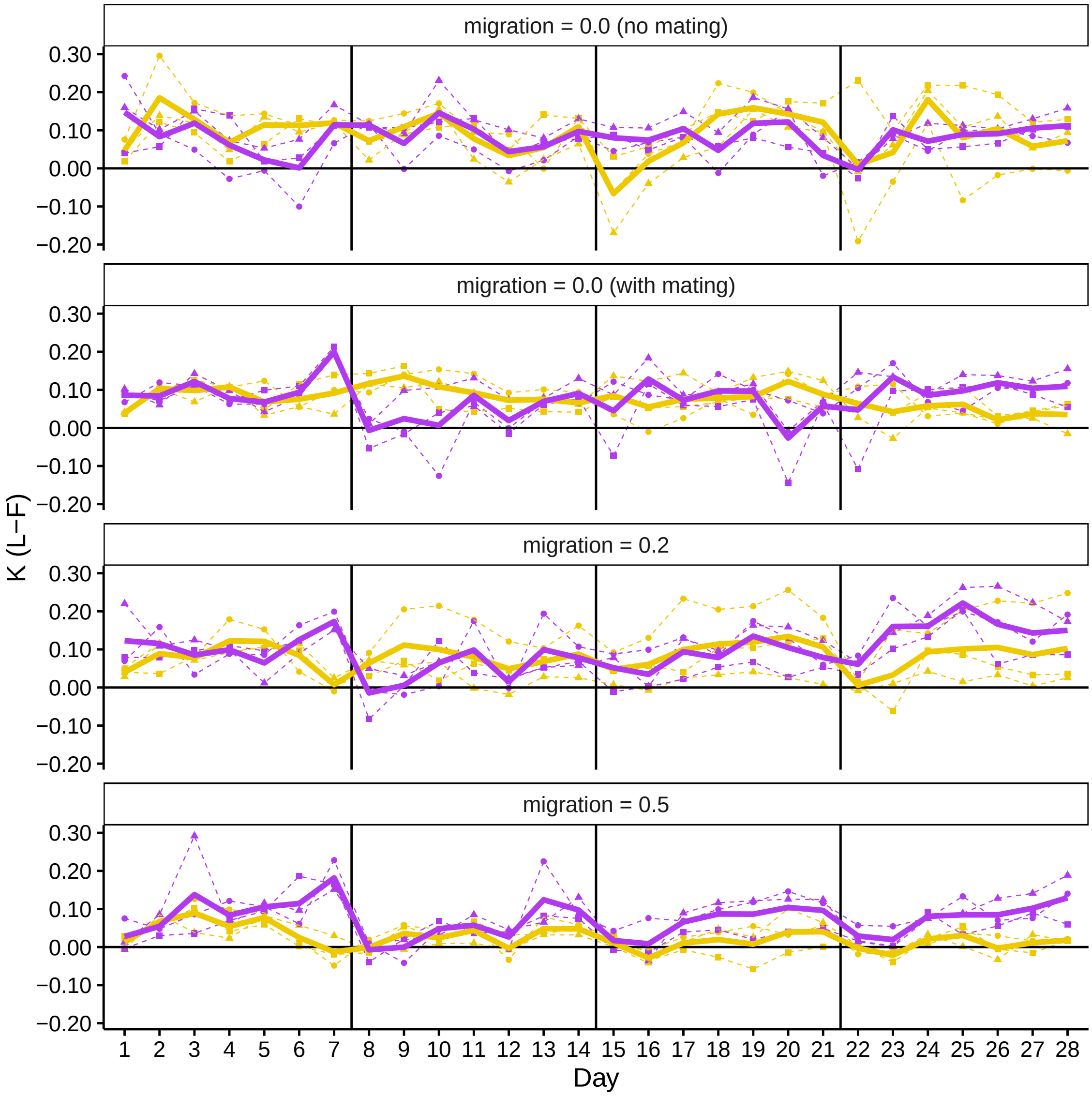


FIGURE E.4: The extent of local adaptation (LA) for rate during divergence with a given amount of migration. For each migration rate the extent of LA is given for each stress environment. Gold and purple denote NaCl and SDS habitats, respectively, and group means are denoted by solid color lines. Black vertical lines show where sporulation and migration (if applicable) occur between periods of asexual growth in batch.

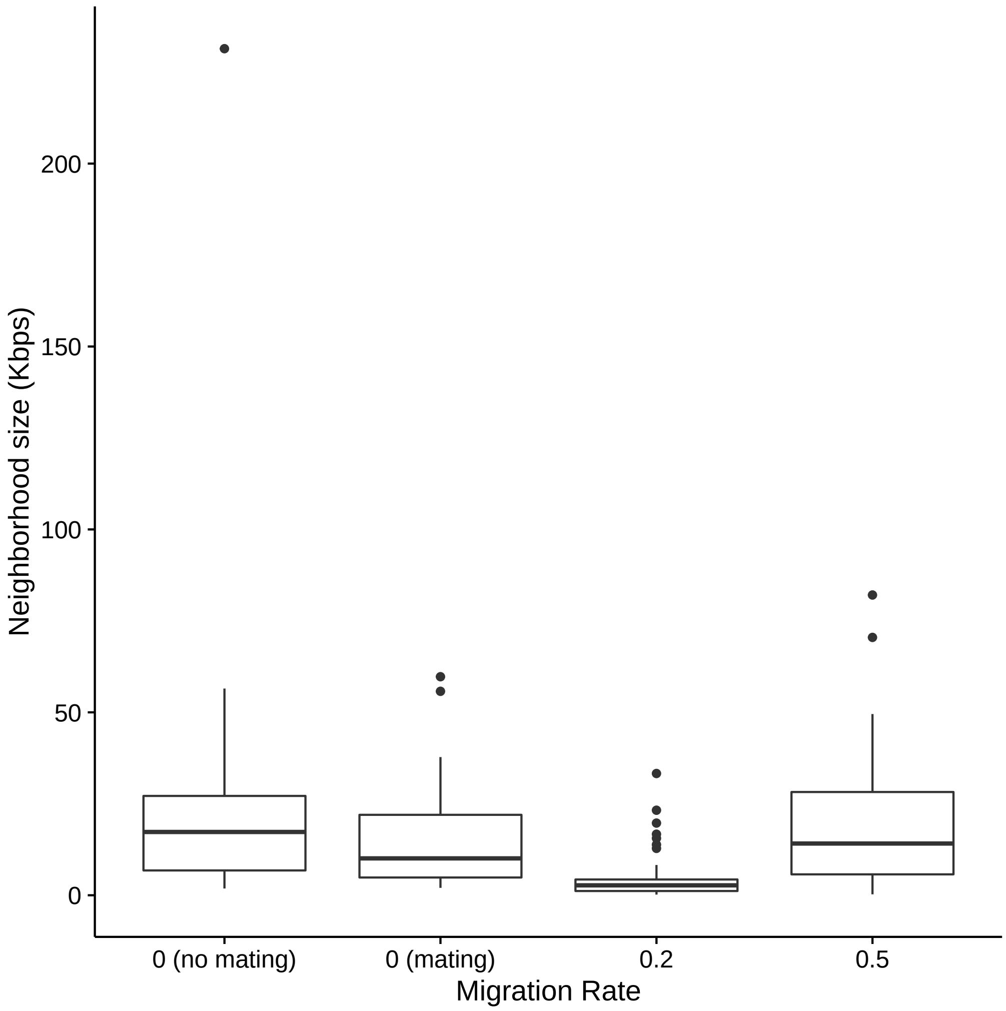


Figure E.5: Average island size vs migration rate. A significant difference (Tukey's HSD) occur between migration $m = 0.2$ and non-mating $m = 0.0$ (P<0.041).