A Bayesian Method for Estimating the Strength of Natural Selection Driving Parallel Evolution

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Bioinformatics and Computational Biology

in the

College of Graduate Studies

University of Idaho

by

Ailene MacPherson

Major Professor: Scott Nuismer, Ph.D.

Committee Members: Lyudmyla Barannyk, Ph.D.; Paul Hohenlohe, Ph.D.; Richard Gomulkiewicz, Ph.D.; Paul Joyce, Ph.D.

Department Administrator: Eva Top, Ph.D.

May 2015

Authorization to Submit Thesis

This thesis of Ailene MacPherson, submitted for the degree of Masters of Science with a Major in Bioinformatics and Computational Biology and titled "A Bayesian Method for Estimating the Strength of Natural Selection Driving Parallel Evolution", 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: _____ Date: _____
Scott Nuismer, Ph.D.

Committee Members: _____ Date: _____
Lyudmyla Barannyk Ph.D.

_____ Date: _____
Paul Hohenlohe, Ph.D.

_____ Date: _____
Richard Gomulkiewicz, Ph.D.

_____ Date: _____
Paul Joyce, Ph.D.

Department Administrator: _____ Date: _____
Eva Top, Ph.D.

Abstract

Parallel phenotypic or genetic evolution is often assumed to result from strong repeated natural selection associated with adaptation to particular environments. Here we develop and analyze a mathematical model that predicts the probability of parallel genetic evolution as a function of the strength of phenotypic selection and constraints imposed by genetic architecture. We then develop a Bayesian approach that uses our model, along with estimates of genetic parameters derived from QTL scans, to estimate the strength of parallel phenotypic natural selection. Using extensive individual based simulations we then evaluate the performance of our Bayesian estimator across a wide range of genetic and evolutionary scenarios. These simulations demonstrate that our Bayesian approach provides a useful tool for estimating the strength of parallel phenotypic selection from genomic data. In addition, our analyses of simulated data allows us to compare the utility of two commonly used experimental methodologies and generate guidelines for future empirical studies of parallel genetic evolution.

Acknowledgements

Funding for this thesis was provided by NSF grants nos. DEB 1118947 and DMS 0540392. In addition this work was supported by the Bioinformatics and Computational Biology Program at the University of Idaho in partnership with IBEST (the Institute for Bioinformatics and Evolutionary Studies).

Table of Contents

Lists of Figures

Chapter 1: Introduction

Parallel evolution has long captivated evolutionary biologists because of the potential insights it provides into the importance of genetic constraints and the consistency and repeatability of natural selection (Schluter et al. 2004; Schluter 2009). Capitalizing on advances in sequencing technology, studies documenting the extent to which the same genetic changes underlie repeated patterns of phenotypic evolution have begun to accumulate (Robison et al. 2001; Colosimo et al. 2004; Shapiro et al. 2004; Sundin et al. 2005). These studies have revealed that the degree of genetic parallelism varies widely among systems, with identical genetic changes in some cases (Cresko et al. 2004; Protas et al. 2006; Coyle et al. 2007; Gross et al. 2009) but largely independent changes in others (Robison et al. 2001; Sundin et al. 2005; Nichols et al. 2007; Steiner et al. 2009). Although some of the variation in the degree of genetic parallelism may be explained by evolutionary relatedness (Conte et al. 2012), much remains unexplained.

One possible reason we see identical genetic changes in some systems, but a variety of independent changes in others, is the extent of genetic constraint. This possibility was explored by Orr (2005) using a model that studied novel mutations arising at a single locus in two allopatric populations. Orr's model calculated the probability that the same strongly beneficial mutation arises and fixes in both populations and showed that the number of possible mutations is inversely related to the probability of parallel evolution. This prediction is supported by experimental adaptation of the bacteriophage $\phi X174$ to high temperatures (Wichman et al. 1999) as well as in the adaptation of antifungal drug resistance in *Saccharomyces cerevisiae* (Anderson et al. 2003).

Another potential source of the variation in genetic parallelism is differing strengths of parallel phenotypic selection across systems. Strong phenotypic selection may disproportionately favor the fixation of large effect alleles, which in turn increases the extent of genetic parallelism. Although plausible, the degree to which genetic parallelism depends on the strength of parallel

phenotypic selection remains unknown. To fully understand the role of natural selection in driving parallel evolution we need to model the process of parallel evolution across a broad range of selection strengths, population sizes, and genetic architectures. Such a model will allow us to understand how the probability of observing parallel evolution depends on the nature of standing genetic variation. The ability to predict the probability of parallel evolution from standing genetic variation is critical in light of evidence suggesting standing genetic variation has played an important role in several well-studied examples of parallel evolution (Colosimo et al. 2004; Miller et al. 2012).

Our basic approach will be to develop and analyze a multi-locus mathematical model that captures the essential features of adaptation to similar novel environments from standing genetic variation. Because our model integrates multiple loci, it will enable us to incorporate the effects of linkage disequilibrium, epistasis, and an explicit distribution of allelic effect sizes. We will use this model to determine how the strength of parallel phenotypic selection impacts the probability of fixation of any particular allele, as well as the degree of parallel genetic evolution across multiple loci. We will then develop a Bayesian approach that allows us to estimate the strength of parallel phenotypic selection using data commonly collected as part of genome wide QTL scans aimed at dissecting the genetic basis of parallel evolution. Finally, we use a combination of our Bayesian approach and individual based simulations to evaluate the performance of two alternative experimental methods for studying the genetics of parallel evolution.

Chapter 2: The Model

Biological Scenario:

We envision a scenario where individuals from an ancestral population colonize two or more novel environments and establish new populations (*see* Figure 1A). After this initial colonization we assume gene flow between the ancestral and descendent populations is negligible and that individuals within populations mate at random. The descendent populations then experience identical patterns of phenotypic selection that cause population mean phenotypes to diverge in parallel from the ancestral population; for example the repeated reduction in body armor in freshwater sticklebacks from their common marine ancestors (Colosimo et al. 2004). Next we envision that the genetic basis of the parallel phenotypic evolution is studied using one of two possible experimental designs (Conte et al. 2012).

In the first design (shown in Figure 1B), parallel genetic evolution is assessed at a set of candidate genes. To identify possible candidate genes individuals from one, or more, of the descendent populations (descendent population 1 in Figure 1B) are crossed with ancestral individuals and the resulting offspring are scanned for divergent QTL's. The remaining populations (descendent population 2 in Figure 1B) are then tested for these candidates. We label this first method with the abbreviation GC. This method has been used in human populations which have independently evolved the ability to digest lactose (Tishkoff et al. 2007; Enattah et al. 2008; Ingram et al. 2009) In the second design (shown in Figure 1C), descendent populations are searched independently for the genes responsible for the repeated phenotypic divergence from the ancestral population. This is done by performing genome wide scans for QTL's in each descendent population. We label this method with the abbreviation GG. The GG experimental design was used to identify separate genes responsible for a change in developmental rate in two populations of *Oncorhyncus mykiss,* (Robison et al. 2001; Sundin et al. 2005; Nichols et al. 2007).

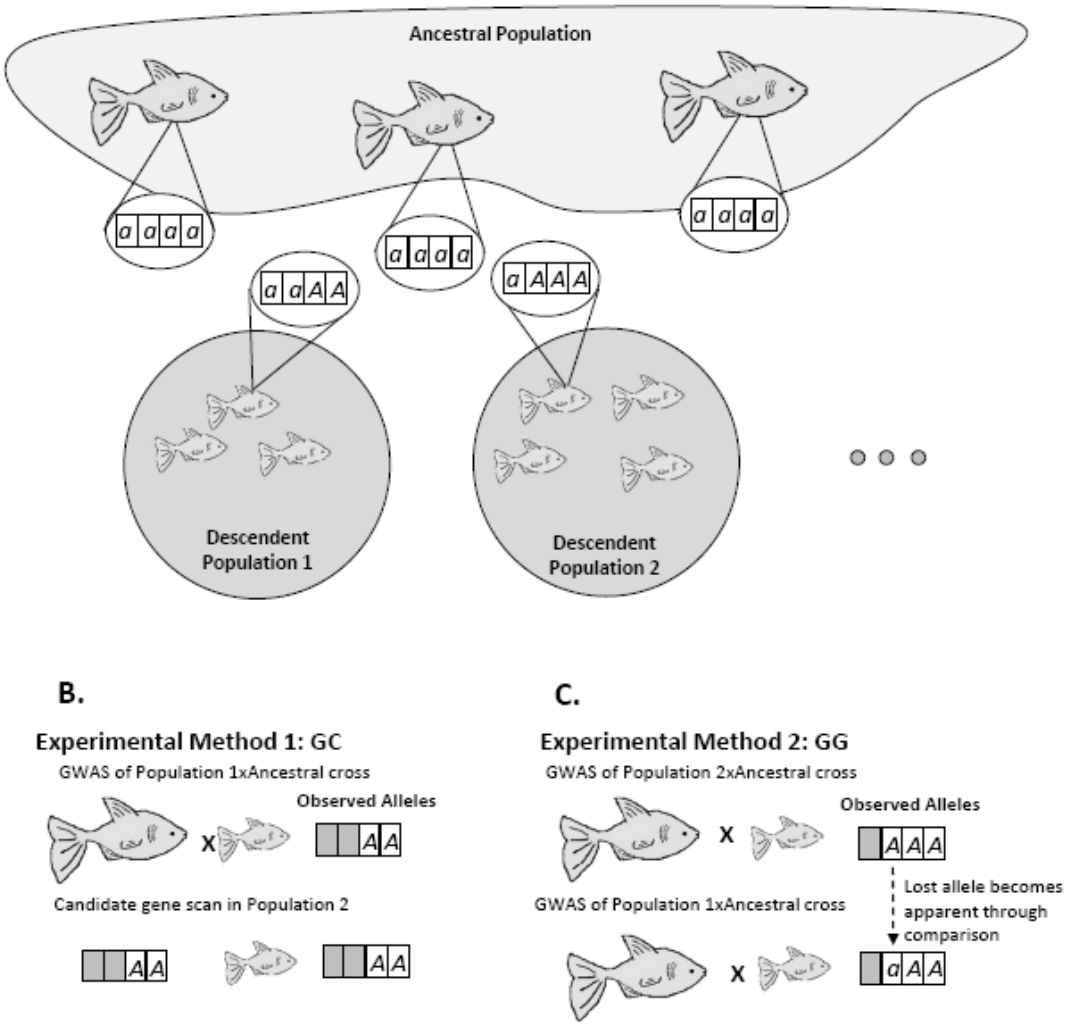

Figure 1: Schematic of biological scenario. Panel A depicts two descendent populations diverging in parallel from a common ancestral population. The a allele predominates at all four loci in the ancestral population whereas the A fixes at various loci in the two descendent populations. Panel B and C depict the two methods for deducing the underlying genetics of the repeated evolution of reduced body size in the two descendent populations depicted in panel A. Panel B showing the GC method which involves a genome wide scan of progeny from a cross between the first descendent population and the ancestral population and subsequent candidate gene search in the second descendent population. Panel C shows the GG method which involves two genome wide scans, one in each population. Compared to Method 1, Method 2 uncovers an additional locus driving divergence in population 2.

Analytical Model

Our model assumes the trait experiencing parallel selection is controlled by n additive loci. Each locus,i, has two possible alleles A_i and a_i and a phenotypic effect equal to b_i, such that the phenotype of an individual is described by:

$$z = \bar{z} + \sum_{i=1}^{n} b_i(X_i - p_i) \quad (1)$$

where X_i is an indicator variable taking the value 1 if the individual carries the A allele at locus i and the value 0 if the individual carries the a allele at locus i, p_i is the frequency of the A allele at locus i, and $\bar{z}$ is the average phenotype of the population. We assume that the phenotype of the ancestral population is small, meaning that the frequency of the A allele is low at all loci, and equal to p_{0_i}. Within the new environments, individuals experience selection for large phenotypes, favoring an increase in frequency of the A alleles. Specifically, we assume the relationship between an individual's phenotype and its fitness, $W(z)$, is linear and described by the expression:

$$W(z) = \beta z + \alpha \quad (2)$$

where the parameters β and α represent the slope an intercept of the fitness surface respectively.

To make our multi-locus model analytically tractable we further assume that the strength of linear directional selection is weak $\mathcal{O}(\epsilon)$ and the rate of recombination between loci is relatively high. Under these assumptions, linkage disequilibrium changes very quickly relative to the allele frequencies and a quasi-linkage-equilibrium (QLE) is reached where linkage disequilibrium is small, also $\mathcal{O}(\epsilon)$ (Nagylaki 1993; Nagylaki et al. 1999). Using the expression for the phenotypic trait z, given in equation (1), as well as the expression for fitness, given by equation (2), we can use the multi-locus methods developed by Barton and Turelli (1991) and expanded by Kirkpatrick et al. (2002) to derive the change in the frequency of the A_i allele at QLE over a single generation:

$$\Delta p_i = \frac{\beta}{\alpha} b_i p_i (1 - p_i) \quad (3)$$

Equation (3) reveals that at QLE, linear directional selection allows loci to evolve independently, an assumption we will later relax with individual based simulations.

The independent evolution of loci under the conditions of weak linear selection and frequent recombination, enables us to utilize the classic results of the Wright-Fisher model that describes the probability of fixation for an allele with initial frequency p_o in a population of constant size N. This probability is given by:

$$P_{fix} = \frac{e^{2Ns(1-p_0)}\left(e^{2Nsp_o}-1\right)}{e^{2Ns}-1} \quad (4)$$

(Kimura 1957; Karlin and Taylor 1981). In equation (4) s is the strength of selection acting on the allele, and p_0 is its initial frequency. Under our assumption of linear directional selection, $s = \frac{\beta}{\alpha} b_i$, and (4) can be re-written as:

$$P_{fix}(i) = \frac{e^{2N\frac{\beta}{\alpha}b_i(1-p_{o_i})}\left(e^{2N\frac{\beta}{\alpha}b_i p_{o_i}}-1\right)}{e^{2N\frac{\beta}{\alpha}b_i}-1} \quad (5)$$

Equation (5) reveals that the probability of fixation depends on initial allele frequency, local population size, the strength of phenotypic selection, and the phenotypic effect of the locus. In the next section we will use this result to explore how these important parameters influence the extent of parallel evolution.

The probability of parallel genetic evolution

We begin by analyzing the simplest possible scenario, where parallel genetic evolution is assessed at only a single genetic locus. For this simple case, parallel adaptation entails the repeated fixation of the same allele in multiple descendent populations. We can use equation (5) to express the probability that the A_i allele fixes independently in each of m populations:

$$P_{\parallel} = \left(\frac{e^{2N\frac{\beta}{\alpha}b_i\left(1-p_{o_i}\right)}\left(e^{2N\frac{\beta}{\alpha}b_i p_{o_i}}-1\right)}{e^{2N\frac{\beta}{\alpha}b_i}-1} \right)^m \quad (6)$$

Equation (6), highlights three important factors that will influence the probability that parallel genetic evolution is observed. First, (6) shows that the probability of repeated fixation of an allele increases with its initial frequency. Second, (6) reveals that large effect alleles are more likely to fix in parallel. This relationship between effect size and parallel evolution is shown in Figure 2. The third effect on $P_{\parallel}$ is the stochasticity of evolution. This effect is captured by the term $N\frac{\beta}{\alpha}$, the product of the population size N and the selection gradient $\frac{\beta}{\alpha}$, which we will denote collectively as η. This combined parameter captures the balance between drift and selection; small values of η connote systems driven primarily by drift and are hence highly stochastic. Whereas if η is large, selection predominates and evolution becomes more deterministic. The probability of parallel evolution, $P_{\parallel}$, increases with η as shown by the three lines in figure 2. Since η is proportional to the strength of selection, this last result indicates that parallel genetic evolution may enable insights into the strength of natural selection driving parallel adaptation.

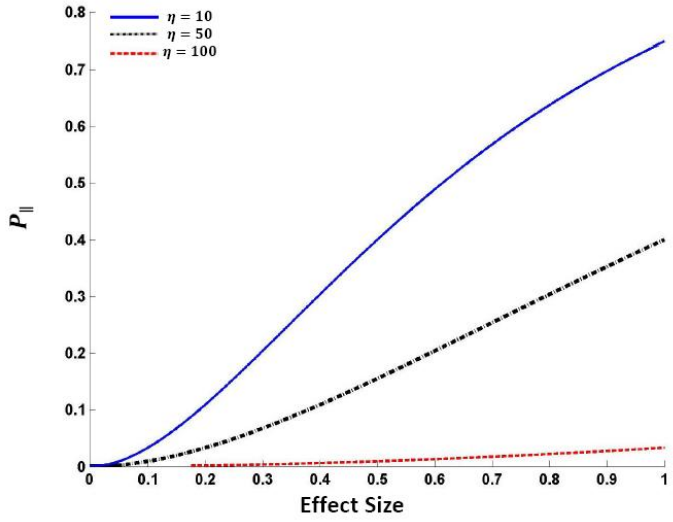

Figure 2: The probability of parallel evolution as a function of effect size, $\boldsymbol{b}$. For a given strength of selection the probability of fixation, and hence parallel evolution, increases with allelic effect size. The rate of increase however is non-linear and depends on the strength of selection s and the population size N which are given by the parameter $\eta = Ns$.

Bayesian inferences on the strength of Natural Selection

Although the single locus results of the previous section are insightful, they fall short of capturing the genetic richness of real populations, where the extent of parallel evolution must be assessed across multiple loci. In these more realistic scenarios, the genetic data can be described by a matrix, $\mathcal{D}$, whose rows represent descendent populations and columns loci. Each element of $\mathcal{D}$ takes a value of 0 or 1 depending on which allele has fixed at a particular locus in a given population (Figure 1A). Using equation (7) we can develop a likelihood function specifying the probability of observing the data, $\mathcal{D}$, given a particular strength of natural selection and population size, η. This likelihood expression consists of a product of terms, one for each locus in each population. If the A allele has fixed at a locus it contributes a term P_{fix}, as defined by equation (7). Alternatively, if the A allele is lost, it contributes a term $\left(1 - P_{fix}\right)$. Thus, for m populations and n loci, the likelihood of observing the data, $\mathcal{D}$, is given by the following product:

$$\mathcal{L}(\mathcal{D}) = \prod_{j=1}^{m} \prod_{i=1}^{n} P_{fix}(\eta, i)^{\mathcal{D}_{ij}} \left(1 - P_{fix}(\eta, i)\right)^{1-\mathcal{D}_{i,j}} \quad (7)$$

where i is an index over loci and j an index over populations.

For most genetic data $\mathcal{D}$ we can use equation (7) to formulate a maximum likelihood estimate for η. The variable η, which is the product between the population size and the strength of selection, can then be used to estimate the strength of selection if an independent estimate for population size is available. There are however, some cases where (7) will not supply a biologically meaningful estimate of η. For example, if all divergent loci fix in all descendent populations such that matrix $\mathcal{D}$ is all 1's. For such an outcome (7) yields a maximum likelihood estimate $\eta = \infty$. Although not informative on its own, such estimates can be relevant when framed in a Bayesian context where we can factor in our prior belief that extremely strong selection is unlikely.

We incorporate our knowledge about the likely strength of selection in the form of the prior distribution, $\pi(\eta)$. Bayes theorem enables us formulate estimates for η in the form of the posterior distribution $p(\eta|\mathcal{D})$ that are biologically meaningful for all possible genetic outcomes, $\mathcal{D}$.

$$p(\eta|\mathcal{D}) = \frac{\pi(\eta)\mathcal{L}(\mathcal{D})}{\int_{\eta} \mathcal{L}(\mathcal{D})} \quad (8)$$

The denominator of this expression is the integral over the likelihood surface, and cannot be easily evaluated. For this reason, we employ Markov Chain Monte Carlo simulation methods to sample from the posterior distribution and generate an estimate of the most probable value of η for the given genetic data $\mathcal{D}$. We label this estimate $\hat{\eta}$. We take two approaches to evaluating the performance of this Bayesian estimator. First we analyze its performance under the assumptions of the analytical model by generating the genetic data $\mathcal{D}$ using a Wright-Fisher model. Next, we test the robustness of our Bayesian estimator to violations of the assumptions of our analytical model by generating the genetic data $\mathcal{D}$ using multi-locus individual based simulations.

Wright-Fisher Simulation

We simulated the data $\mathcal{D}$ for two populations under the Wright-Fisher model by drawing a random number for each locus and population and setting $\mathcal{D}_{i,j}$ to 1 if the random number was less than p_{fix}, given by equation (5), and to 0 otherwise. The value of p_{fix} depends on the initial allele frequency at each locus,p_{0_i}, the allelic effect sizes of each locus, b_i, as well as the variable η. For each simulation we drew the values of these parameters independently and at random. Initial allele frequencies ranged between 0 and 0.5 and were drawn independently for each locus. Because our model envisions divergence of descendent populations from a common ancestor we assumed that the initial frequency at any one locus was the same in both populations. Allelic effect sizes were drawn from a uniform distribution between 0 and 1, whereas the value of η ranged from 0 to 50. The genetic outcome $\mathcal{D}$ simulated in this manner may not however resemble what would be measured using experimental methods. For example, using current genomic techniques it is not possible to identify loci that have not diverged from the ancestral state. To address how experimental methodologies effect our Bayesian estimates we considered two modified forms of $\mathcal{D}$ that resemble sampling under the two experimental methods described previously. The first of these methods, GC, (Figure 1B) assesses parallel genetic evolution at candidate genes which are known to have generated the phenotypic divergence in the first descendent population. Hence under this method we only consider the columns of $\mathcal{D}$ (ie loci) where the A allele has fixed in the first population. The effective number of loci under this method is denoted by n_{GC}. The second experimental method, GG, (Figure 1C) independently assesses divergent loci in all populations. $\mathcal{D}$ under this method therefore contains all columns (loci) which have fixed in at least one population. We denote the effective number of loci under this method by n_{GG}.

For each simulated $\mathcal{D}$, as well as for $\mathcal{D}$ modified by the two experimental methods, we estimated η using a metropolis algorithm as described in the supplementary material. For the prior

distribution $\pi(\eta)$ we used a uniform distribution on the interval $\eta = \pm 60$. To analyze the performance of the Bayesian estimator we ran a regression of the estimated values of $\hat{\eta}$ on the true values η using 200 data points. Overall, this analysis revealed that the Bayesian estimator was quite accurate, explaining between 30% and 60% of the variation (see Table S2). In addition, our analysis showed that the accuracy of the estimates increases with the number of loci. This trend holds true regardless of the experimental method used. However, the effective number of loci under the GG experimental method is always greater than that of the GC method. The results of these simulations suggest our estimator performs quite well when data meet the assumptions of our analytical model; however, this may not be the case for real data. In the next section, we explore the performance of our estimator using individual based simulations that allow us to violate key assumptions of our analytical model such as weak selection and frequent recombination.

Individual Based Simulation

Our individual based simulations (IBS) consider two allopatric populations, each of which has a constant size of $N = 1000$ individuals. Initial allele frequencies and effect sizes at each diallelic loci as well as the value of η were drawn randomly as described above under the Wright-Fisher model. Individuals within each population undergo a two stage life cycle. During the first stage, "selection", the probability that an individual survives is given by its fitness, fitness is computed using either equation (1) which describes linear selection or an expression for stabilizing selection described below. Surviving individuals enter the second life cycle stage, "reproduction", which consists of generating an offspring population from the remaining parental population. This is done by drawing a pair of parents at random from the pool of surviving individuals and producing an offspring from these parents by recombining the parental genomes at a specified rate r and allowing mutation between the two allelic states at a per locus mutation rate of $\mu = 10^{-6}$. This process is continued with replacement of parents until the offspring population reaches the pre-selection size of N. This

life cycle is repeated until all loci approach fixation or loss (allele frequencies > 0.99 or < 0.01) at which point the simulations were terminated and the matrix of genetic data $\mathcal{D}$ filled by rounding. As in the previous section, we formulate modified versions of $\mathcal{D}$ that resemble sampling under the two experimental methods. Then, using the metropolis algorithm, we compute estimates for the value of η using the original outcome $\mathcal{D}$ as well as the two modified forms of $\mathcal{D}$ (*see Supplementary Material*).

We used the Individual based simulations to test the robustness of the Bayesian estimator when selection is strong and/or non-linear. To test the effect of non-linear selection, individual based simulations were run where an individual's fitness was determined by one of two alternative forms of selection: linear directional selection described by (2), or stabilizing selection toward a phenotypic optimum:

$$W(z) = e^{-\gamma(z-\theta)^2} \quad (9)$$

where θ is the phenotypic optima and γ is the strength of stabilizing selection. Including simulations where selection is stabilizing is important because it relaxes our previous assumption that loci evolve independently. Stabilizing selection is particularly useful in testing this assumption because the extent of interdependence between loci can be manipulated by changing the value of the phenotypic optima, θ. Specifically, when θ is greater than the largest possible phenotype, z_{max}, loci remain relatively independent as directional selection predominates over epistatic selection. However when $\theta < z_{max}$ this is no longer true as epistatic selection now dominates. Therefore, when $\theta > z_{max}$ evolution is much more likely to resemble linear selection as our analytical model assumed. As expected, analysis of simulated data shows that the accuracy of our Bayesian estimates depends on the form of selection. Specifically, estimates for η are most accurate under linear selection, somewhat less accurate under stabilizing selection toward a distance optimum($\theta > z_{max}$), and least accurate under stabilizing selection toward a close optimum ($\theta < z_{max}$). In addition to assuming

that selection is linear we also assumed that selection is weak. Indeed, the accuracy of the Bayesian estimates decrease with increasing η, shown by the increasing variance about the regression line. In addition, a systematic underestimation of large η values results in regression lines with consistently positive intercepts (see Figure 3).

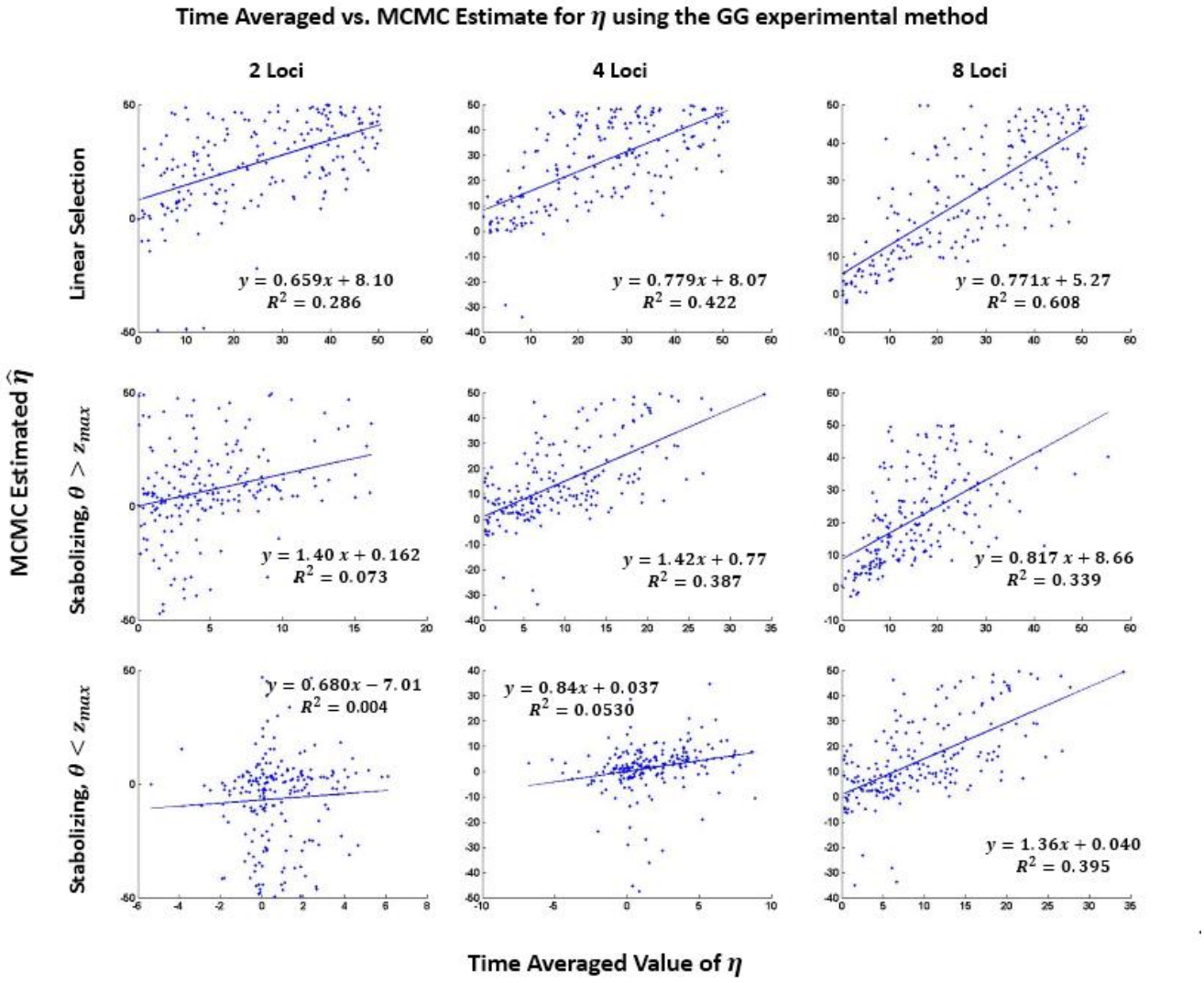

Figure 3: Regression Fit of IBS Data under the three forms of selection. Data and linear regression fit between the time averaged values of η and the Bayesian estimate $\hat{\eta}$ for 200 replicates of the individual based simulation. The Bayesian estimator uses genetic data filtered to resemble sampling using the GG experimental method.

We also used our simulations to explore the sensitivity of our estimator to infrequent or restricted recombination among candidate loci. Not surprisingly, these simulations revealed that our Bayesian estimator performs best under free recombination ($r = 0.5$) and poorest when recombination is absent altogether ($r = 0$). The effect of constrained recombination is more drastic for stabilizing selection than linear selection, and is particularly pronounced when $\theta < z_{max}$. Stabilizing selection is more susceptible to error from constrained recombination because in such cases epistasis causes the buildup of linkage disequilibrium between loci, exacerbating the violation of our assumption that loci evolve independently.

Chapter 3: Discussion

The results of our analytical model demonstrate that parallel genetic evolution becomes increasingly likely as the strength of parallel selection increases. The extent of this increase, however, is determined by the genetic architecture of the trait under selection with parallel evolution at large effect loci being more sensitive to natural selection then loci of smaller effect. Our Bayesian estimator utilizes this connection between genetic effect size and sensitivity to selection to formulate estimates for the strength of natural selection driving parallel genetic evolution. Our individual based simulations illustrate that such estimates are accurate under a broad range of parameters and robust to violations in assumption of the underlying model.

Evaluating the performance of our Bayesian estimator using simulated data reveals that the number of loci used for inference is a critical. Specifically, increasing the number of loci from 2 to 8 yields a fourfold increase in accuracy. The increase in accuracy with the number of loci is particularly pronounced when selection is stabilizing, and holds regardless of the experimental method used to identify the loci involved in phenotypic adaptation. This result suggests that if wish to make inferences about the processes underlying repeated patterns of phenotypic evolution, it is important to identify as many loci as possible. The two experimental methods, GG and GC, discussed throughout differ consistently in the number of loci they detect with GG methods always exceeding GC methods. Hence it follows that inferences drawn from GG methods should typically outperform inferences drawn from GC methods.

The number of candidate loci identified in known examples of parallel evolution ranges anywhere from 2 loci (Wilkens and Strecker 2003; Protas et al. 2006; Gross et al. 2009) to as many as 8 loci (Robison et al. 2001; Sundin et al. 2005), with the majority of studies at the low end of this range. Given the importance of having many loci to accurately determine the strength of selection,

current data may be unprepared to address the role of natural selection in driving parallel evolution in any one system. However we may be able to make general inferences about the strength of selection by looking at parallel genetic evolution across systems. Our model predicts that strong natural selection should produce an overabundance of parallel genetic evolution at large effect loci. Indeed, many large effect genes, such as EDA and Pitx1 in stickleback (Colosimo et al. 2004; Shapiro et al. 2004), as well as Mc1r in lizards and fish (Wilkens and Strecker 2003; Protas et al. 2006; Vidal et al. 2007; Gross et al. 2009; Rosenblum et al. 2010) underlie parallel phenotypic evolution. However in this regard, our current assessment of the genetics of parallel evolution may be substantially skewed as populations are much more likely to be tested for large effect candidate genes rather than genes of small effect. It is therefore important that future studies of parallel evolution focus on understanding the role of genes from across the effect size spectrum.

Although the results of our individual based simulation suggest that the Bayesian estimator is quite robust to violations in many of our assumptions, there are still several assumptions remain untested. We still assume that after initial colonization there is no ongoing gene flow between the ancestral and descendent populations. This assumption enables accurate estimation of initial allele frequencies in descendent population from analysis of the standing genetic variation of the ancestral population. Another assumption is constant population size across time, this does not allow for extreme bottlenecks or extensive founder effects. Historically small population sizes may be critically important in cases like the repeated evolution of reduced skin pigmentation in European and Asian human populations for which there is evidence for an extensive bottlenecks (Schmegner et al. 2005; Amos and Hoffman 2010). Finally we assume that selection is identical in each population. We can extend our results to compute independent Bayesian estimates of selection in each descendent population by considering the loci from each population separately. However this reduces the

number of loci used for each inference by a factor of the number of populations, and hence greatly reduces accuracy.

We have developed and tested a Bayesian approach to estimating the strength of parallel phenotypic selection using genetic data. This approach offers a novel method for capitalizing on rapidly increasing genetic data derived from studies of parallel evolution. By analyzing parallel genetic evolution within and across biological systems in the light of our theoretical predictions it may be possible to formulate predictions for the role of natural selection in driving parallel evolution at large. Extensions of our model to repeated phenotypic evolution within more distantly related populations may finally be able to assess the relative roles of natural selection and genetic constraint in driving parallel evolution.

References

Amos, W. and J. I. Hoffman. 2010. Evidence that two main bottleneck events shaped modern human genetic diversity. P Roy Soc B-Biol Sci 277:131-137.

Anderson, J. B., C. Sirjusingh, A. B. Parsons, C. Boone, C. Wickens, L. E. Cowen, and L. M. Kohn. 2003. Mode of selection and experimental evolution of antifungal drug resistance in Saccharomyces cerevisiae. Genetics 163:1287-1298.

Barton, N. H. and M. Turelli. 1991. Natural and Sexual Selection on Many Loci. Genetics 127:229-255.

Colosimo, P. F., C. L. Peichel, K. Nereng, B. K. Blackman, M. D. Shapiro, D. Schluter, and D. M. Kingsley. 2004. The genetic architecture of parallel armor plate reduction in threespine sticklebacks. Plos Biol 2:635-641.

Conte, G. L., M. E. Arnegard, C. L. Peichel, and D. Schluter. 2012. The probability of genetic parallelism and convergence in natural populations. P Roy Soc B-Biol Sci 279:5039-5047.

Coyle, S. M., F. A. Huntingford, and C. L. Peichel. 2007. Parallel evolution of Pitx1 underlies pelvic reduction in Scottish threespine stickleback (Gasterosteus aculeatus). J Hered 98:581-586.

Cresko, W. A., A. Amores, C. Wilson, J. Murphy, M. Currey, P. Phillips, M. A. Bell, C. B. Kimmel, and J. H. Postlethwait. 2004. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. P Natl Acad Sci USA 101:6050-6055.

Enattah, N. S., T. G. K. Jensen, M. Nielsen, R. Lewinski, M. Kuokkanen, H. Rasinpera, H. El-Shanti, J. K. Seo, M. Alifrangis, I. F. Khalil, A. Natah, A. Ali, S. Natah, D. Comas, S. Q. Mehdi, L. Groop, E. M. Vestergaard, F. Imtiaz, M. S. Rashed, B. Meyer, J. Troelsen, and L. Peltonen. 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. Am J Hum Genet 82:57-72.

Gross, J. B., R. Borowsky, and C. J. Tabin. 2009. A Novel Role for Mc1r in the Parallel Evolution of Depigmentation in Independent Populations of the Cavefish Astyanax mexicanus. Plos Genet 5.

Ingram, C. J. E., C. A. Mulcare, Y. Itan, M. G. Thomas, and D. M. Swallow. 2009. Lactose digestion and the evolutionary genetics of lactase persistence. Hum Genet 124:579-591.

Karlin, S. and H. M. Taylor. 1981. A second course in stochastic processes. Academic Press, New York.

Kimura, M. 1957. Some Problems of Stochastic-Processes in Genetics. Ann Math Stat 28:882-901.

Kirkpatrick, M., T. Johnson, and N. Barton. 2002. General models of multilocus evolution. Genetics 161:1727-1750.

Miller, M. R., J. P. Brunelli, P. A. Wheeler, S. X. Liu, C. E. Rexroad, Y. Palti, C. Q. Doe, and G. H. Thorgaard. 2012. A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. Mol Ecol 21:237-249.

Nagylaki, T. 1993. The Evolution of Multilocus Systems under Weak Selection. Genetics 134:627-647.

Nagylaki, T., J. Hofbauer, and P. Brunovsky. 1999. Convergence of multilocus systems under weak epistasis or weak selection. J Math Biol 38:103-133.

Nichols, K. M., K. W. Broman, K. Sundin, J. M. Young, P. A. Wheeler, and G. H. Thorgaard. 2007. Quantitative trait loci x maternal cytoplasmic environment interaction for development rate in Oncorhynchus myhiss. Genetics 175:335-347.

Orr, H. A. 2005. The probability of parallel evolution. Evolution 59:216-220.

Protas, M. E., C. Hersey, D. Kochanek, Y. Zhou, H. Wilkens, W. R. Jeffery, L. I. Zon, R. Borowsky, and C. J. Tabin. 2006. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nat Genet 38:107-111.

Robison, B. D., P. A. Wheeler, K. Sundin, P. Sikka, and G. H. Thorgaard. 2001. Composite interval mapping reveals a major locus influencing embryonic development rate in rainbow trout (Oncorhynchus mykiss). J Hered 92:16-22.

Rosenblum, E. B., H. Rompler, T. Schoneberg, and H. E. Hoekstra. 2010. Molecular and functional basis of phenotypic convergence in white lizards at White Sands. P Natl Acad Sci USA 107:2113-2117.

Schluter, D. 2009. Evidence for Ecological Speciation and Its Alternative. Science 323:737-741.

Schluter, D., E. A. Clifford, M. Nemethy, and J. S. McKinnon. 2004. Parallel evolution and inheritance of quantitative traits. Am Nat 163:809-822.

Schmegner, C., J. Hoegel, W. Vogel, and G. Assum. 2005. Genetic variability in a genomic region with long-range linkage disequilibrium reveals traces of a bottleneck in the history of the European population. Hum Genet 118:276-286.

Shapiro, M. D., M. E. Marks, C. L. Peichel, B. K. Blackman, K. S. Nereng, B. Jonsson, D. Schluter, and D. M. Kingsley. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428:717-723.

Steiner, C. C., H. Rompler, L. M. Boettger, T. Schoneberg, and H. E. Hoekstra. 2009. The Genetic Basis of Phenotypic Convergence in Beach Mice: Similar Pigment Patterns but Different Genes. Mol Biol Evol 26:35-45.

Sundin, K., K. H. Brown, R. E. Drew, K. M. Nichols, P. A. Wheeler, and G. H. Thorgaard. 2005. Genetic analysis of a development rate QTL in backcrosses of clonal rainbow trout, Oncorhynchus mykiss. Aquaculture 247:75-83.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S.

Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39:31-40.

Vidal, N., A. S. Delmas, P. David, C. Cruaud, A. Coujoux, and S. B. Hedges. 2007. The phylogeny and classification of caenophidian snakes inferred from seven nuclear protein-coding genes. Cr Biol 330:182-187.

Wichman, H. A., M. R. Badgett, L. A. Scott, C. M. Boulianne, and J. J. Bull. 1999. Different trajectories of parallel evolution during viral adaptation. Science 285:422-424.

Wilkens, H. and U. Strecker. 2003. Convergent evolution of the cavefish Astyanax (Characidae, Teleostei): genetic evidence from reduced eye-size and pigmentation. Biol J Linn Soc 80:545-554.

Appendix A: ***Evolution of Allele Frequencies and Linkage Disequilibrium***

As discussed in the main text, our model studies parallel phenotypic evolution in two or more descendent populations which were colonized by a single common ancestral population (Figure 1). Within the ancestral population selection is assumed to favor a small values of the phenotype, z. Whereas larger values of the phenotype z are favored in the descendent populations. We further assume the phenotype z is determined by the additive action of n diallelic loci with alleles A and a such that:

$$z = \bar{z} + \sum_{i=1}^{n} b_i \zeta_i + e_z \quad \text{(S1)}$$

where $\bar{z}$ is the average phenotype of the population,b_i is the effect of the A allele relative to the a allele, $\zeta_i = (X_i - p_i)$ where X_i is an indicator variable which takes on the value 1 if the individual carries the A_i allele and a value of 0 if it carries the a_i allele, and p_i is the allele frequency of the A allele. The variable e_z describes the random environmental component of the phenotype. For simplicity we assume there is no environment effect and hence $e_z = 0$. Given this simplification equation (S1) reduces to equation (1) of the main text. Because we assume selection favors a small value of the phenotype, z, in the ancestral population and ignore mutation, allele frequencies, p, within the ancestral population will be near zero. We assume that the approximate values of these allele frequencies are known.

We begin our analysis by focusing on the simplest possible selective scenario capable of generating parallel phenotypic evolution: directional selection of identical strength within each of the descendent populations. Specifically, we assume that selection is linear such that absolute fitness as a function of the phenotype is given by the line:

$$W(z) = \beta z + \alpha \quad \text{(S2)}$$

Here β and α describe the slope and intercept of the selection surface respectively. Equation (S2) is the same as equation (2) of the main text. Averaging (S2) over individuals gives us the following expression for the average fitness of a population.

$$\bar{W} = \beta\bar{z} + \alpha$$

Where $\bar{z}$ is the average phenotype of the population as defined in equation (S1). The relative fitness of an individual with phenotype z is given by the ratio $w(z) = \frac{W(z)}{\bar{W}}$. The resulting expression for relative fitness is simplified by assuming that selection is weak and Taylor expanding about $\beta = 0$:

$$w(z) = \frac{\beta z+\alpha}{\beta\bar{z}+\alpha} \approx 1 + \frac{\beta}{\alpha}(z - \bar{z}) + \mathcal{O}(\beta^2) \quad \text{(S3)}$$

Substituting in the definition of the trait z from equation (S1) we get an expression for relative fitness as a function of the individual effect of each locus:

$$w(z) = 1 + \sum_{i=1}^{n} \frac{\beta}{\alpha} b_i \zeta_i \quad \text{(S4)}$$

Expression (S4) is very useful as it allows us to determine the selection on each locus. To do so we begin with equation (7) from of Kirkpatrick et al. (2002) which gives a general expression for relative fitness:

$$w(z) = 1 + \sum_{i}^{n} a_i(\zeta_i) + \sum_{i}^{n} \sum_{j<i}^{n} a_{i,j}\left(\zeta_{ij} - D_{ij}\right) + \cdots \quad \text{(S5)}$$

where D_{ij} is the linkage disequilibrium between the alleles at the i^{th} and j^{th} loci, a_i is the selection coefficient on the A_i allele, and $a_{i,j}$ is a selection coefficient describing the correlational selection acting on the combination of the A_i and A_j alleles. We can solve for these selection coefficients (the $a's$) by comparing like terms between (S4) and (S5). This reveals that:

$$a_i = \frac{\beta}{\alpha} b_i \quad and \quad a_{i,j} = 0 \quad \text{(S6)}$$

With these selection coefficients in hand we can describe the change in the allele frequency of the A_i allele over a single generation using equation (10) from Kirkpatrick et al. (2002).

$$\Delta p_i = a_i p_i(1 - p_i) + \sum_{j \neq i}^{n} a_j D_{ij} \quad \text{(S7)}$$

Where D_{ij} is the linkage disequilibrium between the A_i and A_j alleles. To further simplify this expression we assume that recombination is frequent relative to the strength of selection, an assumption which allows the populations to reach a state known as quasi-linkage-equilibrium (QLE). At QLE, the D_{ij} terms are small and on the same order as the selection, $\mathcal{O}(\beta)$ (Nagylaki 1993; Nagylaki et al. 1999) This allows us to simplify the change in allele frequency in (S7) to:

$$\Delta p_i \approx a_i p_i(1 - p_i) + \mathcal{O}(\beta^2) \quad \text{(S8)}$$

which simplifies to equation (3) of the main text upon substituting the value of a_i from equation (S6).

Because (S8) shows that at QLE the evolution of allele frequencies across loci is independent, we can make use of several classical results derived from the Wright-Fisher model to study the balance between selection and drift within the colonizing populations. Specifically, this can be easily seen by comparing equation (S8) to the following equation which is the classical expression for the change in allele frequency under the Wright-Fisher model.

$$\Delta p = s\, p(1 - p) \quad \text{(S9)}$$

(Hartl and Clark 2007) Comparison of (S8) to (S9) reveals that under linear selection at QLE $s = \frac{\beta}{\alpha} b_i$. We can use this expression for s to utilize another classical result of the Wright-Fisher model, the probability of fixation of an allele under linear directional selection (Karlin and Taylor 1981).

$$P_{fix} = \frac{e^{2Ns(1-p_0)}\left(e^{2Nsp_o} - 1\right)}{e^{2Ns} - 1} = \frac{e^{2N\frac{\beta}{\alpha}b_i(1-p_0)}\left(e^{2N\frac{\beta}{\alpha}b_i p_o} - 1\right)}{e^{2N\frac{\beta}{\alpha}b_i} - 1} \quad \text{(S10)}$$

Which is equation (4) and (5) of the main text.

Extending equation (S10) to capture parallel genetic evolution is not straight-forward. If there are m populations undergoing parallel evolution parallel genetic evolution at locus i represents the independent fixation of the A_i allele in each of these populations and is hence given by the m^{th} power of (S10). See equation (6) of the main text. QLE enables us to assume a similar independence between the fixation of alleles at loci as exists between the fixation of alleles across populations. This important simplification enables us to easily extend equation (S10) to capture the probability of any particular genetic outcome $\mathcal{D}$ as described in equation (7) of the main text.

Appendix B: ***Markov Chain Monte Carlo Simulation of Posterior Distributions.***

At the conclusion of the individual based simulation we have the genetic data $\mathcal{D}$, which consists of a $2 \text{ X } n$ matrix of 0's and 1's indicating the fixation or loss of alleles at the n loci in the two descendent populations. As inputs to our simulation we also know the values for the allelic effect sizes and initial allele frequencies. We use the following Metropolis algorithm to sample the posterior distribution, $p(\eta|\mathcal{D})$ where $\eta = N\frac{\beta}{\alpha}$. The basic algorithm can be described in 7 steps, the first of these steps initializes the algorithm, the 2nd through 4th steps are recursive and generate samples from the posterior (Marjoram et al. 2003), whereas the 5th and 6th steps address the convergence and termination of the algorithm respectively. Convergence to the true posterior distribution can be computed by simulating multiple independent sequences of points and then comparing the variance between versus within these sequences, we will simulate $m = 5$ sequences to assess convergence. One important feature of MCMC sampling of the posterior distribution is called "burn in", a phenomena describing how the beginning portion of a sequence of sampled points depends largely on the initial starting point rather than on the posterior distribution, therefore the first simulated points are often uninformative in describing the posterior. To eliminate "burn in"

effects, when testing convergence and computing the estimated posterior distribution it is a common practice to only use the last half of the simulated points in a sequence (Gelman 2004).

Initialization

1. Draw m initial estimates of $\eta = N\frac{\beta}{\alpha}$ prior distribution. These estimates will serve as the starting points for each of the m sequences.

Recursive algorithm: (Repeat for each of the m sequences)

2. From the current estimate η propose a move from to a new point η^*, where $\eta^*$is drawn from the jump distribution $J(\eta, \eta^*)$.
3. Calculate the probability of accepting the point η^*, h.

$$h = \min\left(1, \frac{P(D|\eta^*)\pi(\eta^*)J(\eta, \eta^*)}{P(D|\eta)\pi(\eta)J(\eta^*, \eta)}\right)$$

4. Move to $\eta^*$with probability h, otherwise stay at η.

Assessing Convergence: (After simulating 2n points in each of the m sequences)

5. Denote these m sequences by $\psi_j\ j \in \{1,2 \dots m = 5\}$. Discard the first n points in each sequence for "burn in" and denoting the final n points in the 5 sequences as $\psi_{i,j}\ i \in \{1,2 \dots n\},\ j \in \{1,2 \dots m = 5\}$, calculate the following ratio of the variance in the points between versus within each sequence:

$$R = \sqrt{\frac{\frac{n-1}{n}W + \frac{1}{n}B}{W}}$$

Where:

$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{i,j} + \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n}\psi_{i,j}\right)^2$ and

$$W = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n-1}\sum_{i=1}^{n}\left(\psi_{i,j} - \frac{1}{n}\sum_{i=1}^{n}\psi_{i,j}\right)^2$$

6. If $R < 1$ simulating additional points will not significantly improve the estimation of the peak of the posterior distribution (Gelman 2004).

After enough points have been simulated to reach a ratio of $R < 1$, we use the last n points in each of the $m = 5$ sequences to generate a single histogram which approximates the posterior distribution. We then find the bin of this histogram containing the most points. The minimum value of this bin was used as our estimate $\hat{\eta}$.

Appendix C: ***Sensitivity of Bayesian Estimator to error in parameters***

The main text discusses at length the robustness of the Bayesian estimator to violations in the assumptions of the analytical model. The tests addressing this however were performed using exact values for the initial frequencies and effect sizes of the alleles. Although these values can be estimated in natural systems such estimates will likely be associated with substantial error. It is therefore also important to address the robustness of the Bayesian estimator to error in these parameters as well. To do so we ran a series of individual based simulations under linear directional selection using the same procedure as described in the main text. However, rather than using the true values for the initial frequencies and effect sizes when estimating $\hat{\eta}$, we drew estimates for these parameters allowing for a specified amount of error. Specifically, we drew the values of these parameters from a Gaussian distribution centered about the true value and with a variance of either 1%, 5%, or 10% of the true value. **Table S3** describes the results from these simulations. In general the Bayesian estimator was highly insensitive to error in initial allele frequency and only moderately sensitive to error in effect size.

Appendix D: Supplementary Tables:

A. Single Locus

Form of Selection	r	Regression slope	Regression intercept	Regression R^2
Linear	0.5	1.01	-0.012	0.961
$\theta < z_{max}$	0.5	0.379	0.0449	0.517
$z_{max} > \theta$	0.5	0.994	0.0256	0.780
Linear	0.05	1.07	-0.0414	0.981
$\theta < z_{max}$	0.05	3.03	-0.0315	0.852
$z_{max} > \theta$	0.05	1.07	0.624	0.937

B. Multiple Loci

Form of Selection	r	Regression slope	Regression intercept	Regression R^2
Linear	0.5	1.07	-0.001	0.944
$\theta < z_{max}$	0.5	0.21	0.003	0.937
$z_{max} > \theta$	0.5	1.07	-0.001	0.941
Linear	0.05	.974	0.000	0.984
$\theta < z_{max}$	0.05	.921	0.001	0.969
$z_{max} > \theta$	0.05	0.830	0.003	0.898

Table S1: Single locus and Multiple loci predictions. Regression line coefficients and R^2 values for the fit between simulated and predicted rates of parallel evolution at a single locus (A), and a probability of a specific genetic outcome across multiple loci (B).

		GG Method		
Selection	**# of loci**	**Slope**	**Itercept**	**R^2**
Linear	2	0	1	0.233554
		1.022567	-7.97479	0.307461
Linear	4	0	1	0.276214
		0.746356	5.526915	0.314167
Linear	8	0	1	0.500288
		0.867688	3.253798	0.51223
$\theta > z_{max}$	2	0	1	0.026395
		2.205828	-14.7943	0.154519
$\theta > z_{max}$	4	0	1	0.199309
		1.490425	-9.16855	0.246104
$\theta > z_{max}$	8	0	1	0.396366
		0.960228	5.07128	0.447337
$\theta < z_{max}$	2	0	1	-0.15149
		1.403294	-11.0413	0.019578
$\theta < z_{max}$	4	0	1	0.001686
		1.220801	-6.16244	0.05796
$\theta < z_{max}$	8	0	1	0.012296
		0.796777	6.950119	0.114301

Table S2: Regression fit for three forms of natural selection. Regression line coefficients and R^2 values for the three forms of natural selection with free and constrained recombination. Estimates for the strength of selection were made using data filtered to resemble the GG experimental methods. The average number of loci used to estimate η under each experimental method is also recorded. WF indicates data simulated using the Wright-Fisher Model, a perfect fit to the assumptions of the analytical model.

				GG Method		
Selection	# of loci	$Error\ in\ p_0$	$Error\ in\ b$	Slope	Itercept	R^2
Linear	2	10%	10%	0	1	0.103721
				0.783463	-2.20934	0.207322
Linear	4	10%	10%	0	1	0.334789
				0.889694	0.120192	0.355744
Linear	8	10%	10%	0	1	0.496155
				0.851563	3.581057	0.511715
Linear	2	10%	0%	0	1	0.139702
				0.972201	-9.05249	0.250067
Linear	4	10%	0%	0	1	0.4312
				1.035959	-4.47945	0.456793
Linear	8	10%	0%	0	1	0.538903
				0.883528	3.016746	0.548438
Linear	2	0%	10%	0	1	0.050792
				0.700849	1.604204	0.130831
Linear	4	0%	10%	0	1	0.243258
				0.916524	-4.28903	0.306573
Linear	8	0%	10%	0	1	0.497547
				0.816885	3.087144	0.531526

Table S3: Accuracy of Bayesian estimate with error in parameters. The regression line coefficients and R^2 values computed when initial frequencies and effect sizes were known with error. Initial frequencies and effect sizes were drawn from Gaussian distributions centered at the true values and with variances of 10% or the true value.

Appendix References:

Gelman, A. 2004. Bayesian data analysis. Chapman & Hall/CRC, Boca Raton, Fla.

Hartl, D. L. and A. G. Clark. 2007. Principles of population genetics. Sinauer Associates, Sunderland, Mass.

Karlin, S. and H. M. Taylor. 1981. A second course in stochastic processes. Academic Press, New York.

Kirkpatrick, M., T. Johnson, and N. Barton. 2002. General models of multilocus evolution. Genetics 161:1727-1750.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavare. 2003. Markov chain Monte Carlo without likelihoods. P Natl Acad Sci USA 100:15324-15328.

Nagylaki, T. 1993. The Evolution of Multilocus Systems under Weak Selection. Genetics 134:627-647.

Nagylaki, T., J. Hofbauer, and P. Brunovsky. 1999. Convergence of multilocus systems under weak epistasis or weak selection. J Math Biol 38:103-133.