Article:
Gompert, Z., Egan, S.P., Barrett, R. et al. (2017) Multi-locus approaches for the measurement of selection on correlated genetic loci. Molecular Ecology, 26 (1). pp. 365-382. ISSN: 0962-1083

https://doi.org/10.1111/mec.13867

Multi-locus approaches for the measurement of selection on correlated genetic loci

Zachariah Gompert[1*], Scott P. Egan[2], Rowan Barrett[3], Jeff Feder[4], and Patrik Nosil[5]

[1] Department of Biology, Utah State University, Logan, UT 84322, USA

[2] Department of BioSciences, Rice University, Houston, TX 77005, USA

[3] Redpath Museum, McGill University, Montreal, QC H3A 2K6, Canada

[4] Department of Biological Science, University of Notre Dame, South Bend, IN 46556, USA

[5] Department of Animal and Plant Sciences, University of Sheffield, Sheffield, S10 2TN, UK

Keywords: natural selection, multiple regression, polygenic models, speciation

Corresponding author: Zachariah Gompert
Department of Biology
5305 Old Main Hill
Utah State University
Logan, UT 84322-5305
Phone: (435) 797-9463
Email: zach.gompert@usu.edu

Running title: Selection on correlated genetic loci

Abstract

The study of ecological speciation is inherently linked to the study of selection. Methods for estimating phenotypic selection within a generation based on associations between trait values and fitness (e.g., survival) of individuals are established. These methods attempt to disentangle selection acting directly on a trait from indirect selection caused by correlations with other traits via multivariate statistical approaches (i.e., inference of selection gradients). The estimation of selection on genotypic or genomic variation could also benefit from disentangling direct and indirect selection on genetic loci. However, achieving this goal is difficult with genomic data because the number of potentially correlated genetic loci (p) is very large relative to the number of individuals sampled (n). In other words, the number of model parameters exceeds the number of observations ($p \gg n$). We present simulations examining the utility of whole genome regression approaches (i.e., Bayesian sparse linear mixed models) for quantifying direct selection in cases where $p \gg n$. Such models have been used for genome-wide association mapping and are common in artificial breeding. Our results show they hold promise for studies of natural selection in the wild, and thus of ecological speciation. But we also demonstrate important limitations to the approach and discuss study designs required for more robust inferences.

Introduction

Natural selection is the mechanism of adaptation and often drives speciation (Schluter, 2001; Schluter & Conte, 2009; Gompert *et al.*, 2012; Nosil, 2012). Consequently, many attempts have been made to measure phenotypic selection in the wild, with the earliest studies occurring in the late 1800s (Bumpus, 1899; Endler, 1986; Kingsolver *et al.*, 2001; Siepielski *et al.*, 2013). Phenotypic selection can be quantified from changes in the distribution of trait values in a population within a generation (due to mortality), or from the association between trait values and quantitative measures of fitness components (e.g., seed set, weight, etc.) (e.g., Lande & Arnold, 1983; Shaw *et al.*, 2008). However, correlations among characters complicate measures of selection, as direct selection on one character induces indirect selection on correlated characters (Table 1, Fig. 1). Consequently, the total selection experienced by a trait can include direct selection on that character and the indirect effects of selection on any correlated characters (Kingsolver *et al.*, 2001). Lande & Arnold (1983) showed that direct and indirect selection can be disentangled using multiple regression. Specifically, partial regression coefficients obtained from regressing fitness on a set of characters are estimates of the direct selection on each trait (these coefficients define the average gradient of the relative fitness surface). Although many modifications and refinements of this approach have been made (e.g., Schluter, 1988; Rausher, 1992; Geyer *et al.*, 2007; Reynolds *et al.*, 2016), these changes have not altered the conceptual basis of the approach.

More recently, attempts have been made to measure selection on genetic loci or genomes based on short-term (e.g., within-generation) changes in allele frequencies (e.g., Barrett *et al.*, 2008; Anderson *et al.*, 2013; Pespeni *et al.*, 2013; Anderson *et al.*, 2014; Gompert *et al.*, 2014; Egan *et al.*, 2015; Thurman & Barrett, 2016). The premise of these studies is that phenotypic selection within a generation alters the distribution of trait values and that this results in a within generation shift in allele frequencies at the causal loci affecting these traits (direct selection) and other genetic variants in linkage disequilibrium (LD) with

them (indirect selection) (Fig. 1). The extent to which phenotypic selection is transmitted down to the genetic-level depends on the heritability of the selected traits and patterns of LD. In stark contrast to our understanding of phenotypic selection, relatively little is known about individual episodes of selection on genetic loci, particularly under natural or semi-natural conditions (Barrett & Hoekstra, 2011; Thurman & Barrett, 2016). This is relevant, as measuring selection at the genetic-level could help resolve key questions about the maintenance of molecular variation in populations (e.g., Gillespie, 1991; Hahn, 2008; Huang *et al.*, 2014) and the causes of ecological specialization (e.g., Agrawal *et al.*, 2010; Anderson *et al.*, 2013; Gompert *et al.*, 2015; Gompert & Messina, 2016). Quantifying selection in the wild is also important for understanding speciation, as reproductive isolation often evolves as a direct consequence of divergent selection and local adaptation (e.g., Jiggins *et al.*, 2001; Nosil *et al.*, 2002; Lowry & Willis, 2010; Ording *et al.*, 2010). Indeed, divergent selection is a form of reproductive isolation when it causes immigrant or hybrid inviability (Wu, 2001; Nosil *et al.*, 2005). Moreover, direct or indirect selection on genetic loci and genomes can cause DNA sequence divergence that pleoitropically results in reproductive incompatibilities (e.g., Swanson & Vacquier, 2002; Tang & Presgraves, 2009). Finally, the likelihood of speciation with gene flow and the persistence of distinct species upon secondary contact depends critically on the genome-wide consequences of selection (Barton & Bengtsson, 1986; Barton & De Cara, 2009; Feder *et al.*, 2012; Flaxman *et al.*, 2013; Feder *et al.*, 2014; Flaxman *et al.*, 2014; Yeaman, 2015).

Distinguishing between the direct and indirect effects of episodes of selection on allele frequency change is a notable challenge for genomic studies. Under most conditions, the number of correlated genetic loci will greatly outnumber the number of individuals studied (genome scans typically consider tens of thousands to millions of nucleotide variants and many fewer individuals). Thus, traditional statistical methods, such as the multiple regression approach proposed by Lande & Arnold (1983) for phenotypic selection, cannot be used to obtain estimates of direct selection on each locus (such methods require the number of

observations, n, to exceed the number of model parameters, p). In other words, parsing direct and indirect selection on phenotypic and genomic variation present the same conceptual issue, but different analytical tools are needed for the latter because $p \gg n$.

We show that this problem can be approached using sparse linear mixed models that were developed for genome-wide association (GWA) mapping of polygenic traits and genomic prediction (Meuwissen *et al.*, 2001; Ober *et al.*, 2012; Habier *et al.*, 2013; Zhou *et al.*, 2013). The potential utility of GWA methods is unsurprising, as measuring episodes of selection on genetic loci is a special case of trait mapping. However, the conditions and study designs under which these methods will be most useful for inferring selection require further quantification, which we provide here. We focus on a specific model, the Bayesian sparse linear mixed model (BSLMM) introduced by Zhou *et al.* (2013), but related models and methods exist and will likely yield similar broad conclusions (e.g., Erbe *et al.*, 2012). The method we focus on uses Bayesian variable selection, model-averaging and shrinkage inducing priors to extend the Lande & Arnold (1983) multiple regression approach to cases where the number of characters (i.e., loci) exceeds the number of observations.

Herein, we demonstrate the utility and limitations of BSLMMs for studying selection by applying this method to a series of simulated data sets. We show that BSLMMs can be used to detect direct selection when fitness has a simple genetic basis. Additionally, we show that BSLMMs can generate quantitative summaries of selection across the genome, such as estimates of the additive genetic variation for fitness, under a wider variety of conditions. Whereas the quantitative summaries could also be obtained using traditional quantitative genetic breeding designs, such methods are not practical for many non-model organisms. Thus, approaches such as those considered here could help extend the direct study of selection to a broader range of organisms, an important goal if we are to achieve general understanding of ecological speciation.

Methods

Theoretical background and statistical models

We first present a general framework and issues for inferring selection, and then describe how BSLMMs can be used to infer direct selection. Multiple approaches exist to infer total selection, that is, the combined effects of direct and indirect selection on a genetic locus (e.g., Anderson *et al.*, 2014; Gompert *et al.*, 2014). Key differences include whether one estimates a selection differential (as has been done in some phenotypic studies) or a selection coefficient (as used in population genetic theory, e.g., Ewens, 2004), and how one assesses statistical significance. Selection differentials for bi-allelic genetic loci can be calculated as $\delta = p_1 - p_0$, where p_0 and p_1 are the population allele frequencies before and after selection, respectively (here we assume viability selection). While selection differentials are intuitive in phenotypic studies, selection coefficients are more useful for quantifying total selection on genotypes and are more directly related to population genetic models. Assume genotypes A_1A_1, A_1A_2, and A_2A_2 have relative expected fitnesses of w_{11}, w_{12}, and w_{22}, respectively (here marginal fitnesses are defined based on the fitness effects of the genotypes and patterns of LD with other causal variants). The selection coefficient s is then defined based on the difference in the marginal fitnesses of alternative homozygotes, such that, $w_{11} = 1 + s$, $w_{12} = 1 + hs$, and $w_{22} = 1$ (here h denotes the heterozygote effect, that is the fitness of the heterozygote relative to the difference between the two homozygotes; Gillespie, 2004). Under this formulation,

$$\hat{s} = \frac{p_1 - p_0}{p_0(1 - p_0)(p_0 + h(1 - 2p_0))} \tag{1}$$

Thus, selection coefficients represent a particular standardization of the selection differential based on genetic variation, and one that differs from the standardization used in phenotypic studies (in phenotypic studies selection differentials are standardized by the phenotypic variance; Lynch & Walsh, 1998).

In an infinite population Eqn. 1 could be used to calculate s exactly. However, stochastic processes (e.g., random mortality) in finite populations compound allele frequency changes due to drift and selection, making statistical inference of s necessary and adding uncertainty to estimates of selection. Thus, it is necessary to account for the possible contribution of drift to observed changes in allele frequencies. We present simple simulations in the on-line supplemental material (OSM) to illustrate this point, namely that genetic drift can cause substantial changes in allele frequency that can be misinterpreted as evidence of selection (distinguishing drift from selection is also an issue for phenotypic studies, although this is often not discussed).

Given this consideration, maximum likelihood or Bayesian methods can be used to obtain interval estimates of s from genetic data under an appropriate stochastic model that allows drift and selection to contribute to allele frequency change (e.g., Wright-Fisher or Moran models with selection; Ewens, 2004). Additionally, randomization or simulation-based methods can be used to test the null hypothesis that $s = 0$ for a particular locus, as was done by Gompert *et al.* (2014) in their null model 1, or to test the global null hypothesis that $s = 0$ for all genetic loci (i.e., that selection did not affect any of the genetic loci). This can be done by comparing the number of loci with significant evidence of selection to the number expected by chance under the global null (Gompert *et al.*, 2014). Note however, that the failure to reject null models of locus-specific or genome-wide drift is not evidence for the absence of selection, and thus this does not mean that $s = 0$ (most genetic loci will exhibit at least very low levels of LD with some causal variants in any finite population, and thus, the vast majority of cases where these null models cannot be rejected will represent type II errors; Gompert, 2016). We discuss these issues in more detail in the OSM (see 'Total Selection').

These concerns related to parsing the contributions of drift and selection apply to inference of direct selection as well, but methods for estimating direct selection must additionally account for correlations among genotypes at different loci. Lande & Arnold (1983)

proposed using multiple regression to solve the problem of trait correlations in phenotypic studies. Their approach works well as long as correlations among variables are not too strong and the number of observations (individuals) exceeds the number of traits (i.e., for $p < n$). Their approach still generally assumes that all relevant traits have been measured, which would be equivalent to assuming all causal variants have been assayed in genomic studies (the latter will rarely be true; we discuss the implications of this below). Using their approach, partial regression coefficients provide measures of direct selection (Lande & Arnold, 1983). More specifically, for bi-allelic loci with genotypes coded as 0, 1, or 2 copies of an allele, a partial regression coefficient, β, equals $\frac{1}{2}s^D$, where s^D is defined similarly to s but only includes direct selection on the genotype (here we assume perfect additivity, that is $h = 0.5$). When a relatively small number of genes or genomic regions are of interest, studies can be designed so that the number of individuals exceeds the number of genetic loci, and thus standard multiple regression approaches could be used to estimate s^D (e.g., the major effect gene *Eda* in sticklebacks; Rennison *et al.*, 2015). However, this will rarely be true for larger population genomic data sets (in such cases $p \gg n$).

BSLMMs can be applied even when $p > n$ by adopting shrinkage or sparsity-inducing priors, which pull parameter estimates back towards zero (e.g., Bernardo *et al.*, 2003; Pérez *et al.*, 2010; Guan & Stephens, 2011). This class of methods includes polygenic models and whole genome regression approaches that have been successfully applied in genome-wide association studies (GWASs) and for genomic prediction and genomic selection in plant and animal breeding (e.g., Meuwissen *et al.*, 2001; Goddard & Hayes, 2007; Heffner *et al.*, 2008; Hayes *et al.*, 2009; Resende *et al.*, 2012; Zhou *et al.*, 2013; Thomasen *et al.*, 2014). Inference of direct selection can be approached in the same manner as mapping a phenotypic trait but with fitness or some component of fitness as the phenotype. Thus, all of the lessons we have learned from decades of GWASs, such as the need for large sample sizes, apply here (e.g., Visscher *et al.*, 2012). We advance this existing knowledge by focusing on conditions most relevant for detecting selection, that is, cases where the phenotype (fitness) has a low

to moderate heritability and diffuse genetic architecture, and by considering genome-level summaries and locus-specific measures of selection.

Here we focus on and describe one such model, the BSLMM proposed by Zhou *et al.* (2013), which is part of the `gemma` software package. We show how BSLMMs can be used to estimate direct selection when numerous (tens or hundreds of thousands) genetic loci have been sequenced, while also providing higher-level summaries of the genetic architecture of fitness, such as the number of loci with measurable effects on fitness. The latter information is extracted from a few key parameters in the model (caveats and limitations of these parameters are discussed below).

BSLMMs consider the joint influence of all genetic loci on phenotype (Zhou *et al.*, 2013). These models assume phenotype, or in this case fitness, is related to multi-locus genotype, such that,

$$\mathbf{y} = 1_n\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon} \tag{2}$$

where $\mathbf{y}$ is the vector of observed fitness values (either 0 and 1 for binary outcomes such as dead vs. alive and mated vs. unmated, or a continuous metric such as survival time or seed set), μ is an intercept and $\boldsymbol{\epsilon}$ is a n vector of error terms (this captures randomness and the effect of the environment on fitness). $\mathbf{X}$ is a matrix of p genotypes for n individuals, which are generally coded as 0, 1, or 2 copies of an allele, and $\boldsymbol{\beta}$ is a vector of (partial) regression coefficients. Thus, $\boldsymbol{\beta}$ is analogous to Lande & Arnold's (1983) selection gradient, and represents the measurable effects of genotypes on fitness (i.e., direct selection). Here we use the term measurable to mean effects that are decidedly non-infinitesimal. To make the model identifiable, the regression coefficients are modeled as coming from a mixture of a normal distribution with unknown variance and a point mass at 0 (this is a shrinkage or sparsity-inducing prior). Analysis using Bayesian variable selection generates posterior inclusion probabilities (PIPs) for each genetic locus, which provide the probability of mea-

surable, direct selection on the locus. Bayesian model averaging can then be used obtain estimates of s^D (direct selection) that account for uncertainty in whether $s^D = 0$ (we refer to these estimates as $\bar{\beta}$, whereas estimates that assume $s^D \neq 0$ are denoted $\hat{\beta}$). Depending on the nature and sparsity of the genetic data, some, most or all of the causal variants may not be sequenced, particularly with reduced representation sequencing methods (e.g., GBS, RADseq, exome sequencing, etc.; Tiffin & Ross-Ibarra, 2014). However, direct selection on the causal variants can still potentially be accounted for through LD with other variants (Fig. 2). Here, we are really using indirect selection on a locus linked to the (un-sequenced) causal variant as a proxy for direct selection on the missing causal variant. Nonetheless, this can be conceptualized as an estimate of direct selection in the sense that the effects of other (i.e., correlated and sequenced) genetic loci have been accounted for (i.e., the only indirect effects are those coming from missing loci). This issue is conceptually similar to the issue of inference of direct selection on phenotypes when not all phenotypes have been measured (Lande & Arnold, 1983).

When fitness is determined by a large number of loci with very small or near infinitesimal effects, the contribution of this genetic variation to fitness might not be captured by the vector or partial regression coefficients, $\boldsymbol{\beta}$. However, even in this case, genetic variation for fitness (and thus the full contribution of direct selection to variation in realized fitness) can be inferred using information from the overall genetic similarity among individuals. In Eqn. 2 this is accounted for by the vector $\mathbf{u}$, which denotes each individual's deviation from the mean expected fitness based on their complete multi-locus genotype. More specifically, a multivariate normal prior is placed on $\mathbf{u}$ with a variance-covariance matrix that is proportional to the genetic similarity or kinship matrix, which is calculated from the data and treated as a constant in the model; $\mathbf{u}$ is then inferred from the data given this prior.

Thus, similar to classic quantitative genetic approaches, the model includes overall relatedness as a potential predictor of similarity in fitness (Lynch & Walsh, 1998). In contrast to quantitative genetic approaches, controlled crosses with specific breeding designs are not

required, and thus BSLMMs can be used in systems were controlled crosses are not practical or ethical. Nonetheless, breeding designs will affect the structure of the kinship matrix and amount of LD in the population, and patterns of relatedness can affect the efficacy of the method (see our results below). Thus, different experimental designs might be preferable for specific research questions (we discuss this point in detail below). The kinship matrix also serves to control for population structure, and can often do so more effectively than including population structure covariates (Zhao *et al.*, 2007; Kang *et al.*, 2008).

The hierarchical nature of the model provides a means to estimate parameters that summarize direct selection across the genome (Guan & Stephens, 2011; Zhou *et al.*, 2013). These include the proportion of variation in fitness explained by all of the genetic data (PVE) through $\bar{\beta}$ and $\mathbf{u}$ (PVE should approach narrow-sense heritability with sufficient genetic sampling), the proportion of the PVE explained by genetic loci with measurable effects (via the $\bar{\beta}$), which is denoted PGE, and the number of genetic variants with measurable effects on fitness (denoted n-γ). These metrics incorporate uncertainty in the specific genetic variants under selection, meaning that accurate estimates of these parameters should be possible even if the specific targets of direct selection cannot be localized. This is important, as these parameters alone can provide important information about genetic variation for fitness. Moreover, in some systems, such as hybrid zones, variation in fitness reflects components of reproductive isolation (e.g., hybrid inviability) making these measures relevant for studies of speciation.

However, inference of these parameters is affected by the extent to which causal variants are effectively tagged by LD with sequenced variants, such that PVE and n-γ will only approach the true heritability and number of causal variants if all or most causal variants are in LD with sequenced variants. This will of course depend on the sparsity of the genetic data, general patterns of LD, and the extent to which causal variants and sequenced variants have similar allele frequencies (Visscher *et al.*, 2012). More generally, the performance of BSLMMs for detecting selection will depend on numerous factors that can

usefully be explored with simulated data (as in this study).

Simulations of fitness data

We generated and analyzed data sets to assess the potential and limits of BSLMMs to quantify direct selection under different sampling designs and with different genetic architectures. The performance of this method has been evaluated in the context of genomic prediction and inference of PVE (Zhou *et al.*, 2013). Our goal here was to also evaluate performance in terms of partial regression coefficients (that is, measures of direct selection on individual genotypes in our current formulation) and to examine performance under conditions that are more relevant for studies of genome wide selection in the wild, namely low to moderate heritability and diffuse genetic architectures for fitness (Mousseau & Roff, 1987; Kruuk *et al.*, 2000; Hoffmann *et al.*, 2016). We also considered sample sizes that, while reasonably large, are more realistic for studies of natural populations (compared to sample sizes that might be obtainable for studies of human disease).

Fitness data sets were simulated under a variety of conditions and analyzed using the BSLMM implemented in `gemma`. We considered accuracy of inference with respect to individual estimates of s^D and summaries of the genetic basis of variation in fitness (e.g., PVE). We used previously generated genotyping-by-sequencing (GBS) genotype data as the starting point for simulations of fitness values. That is, we assigned selection coefficients to GBS genotypes and used these to compute the expected fitness for each individual based on the GBS data. This approach was used because it captures realistic patterns of genetic variation and linkage disequilibrium. We did not make inferences about selection in these specific species or populations (i.e., the fitness values were assigned by us in the aforementioned simulation context). Although we used GBS data, BSLMM could be used with whole genome sequences, or even data sets that include a mixture of SNPs and structural variants. Our primary genetic data set included 592 *Timema cristinae* stick insects collected from a single population with genotypes for 246,258 SNPs (mean minor allele frequency = 0.09). A

full description of these data can be found in Comeault *et al.* (2015). We first considered a quantitative metric of fitness (e.g., adult weight, longevity, seed set, flower number, etc.).

We initially simulated 50 replicate data sets with a narrow sense heritability of fitness (h^2) of 0.3 or 0.05 and with 10, 100, or 1000 causal variants (we use L to denote the number of causal variants). We sampled the fitness effect of each causal variant from a standard normal distribution and assumed that the causal variants affected fitness additively with incomplete dominance ($h = 0.5$). Causal variants were chosen randomly from the set of genotyped SNPs and used to calculate expected fitness values. We then analyzed each data set with and without the causal variants included as potential covariates in the model. We did this because many causal variants will not be sequenced with partial genome sequencing approaches (Tiffin & Ross-Ibarra, 2014), such as GBS, but can still potentially be accounted for through LD with other variants. As mentioned previously, when causal variants are missing from the data set, we are really measuring indirect selection on a linked locus as a proxy for direct selection on the missing causal variant.

Additional simulations were conducted to further test how different conditions influence the efficacy of this method. First, the simulations described above were repeated using a binary metric of fitness, such as survival. We converted each individual's quantitative score into a binary score by assuming that 50% of individuals with the highest quantitative score had a viability of 1, whereas the rest of the individuals had a viability of 0. Another set of simulations assessed the performance improvement through increased sample size (i.e., larger n). We sampled 2500 individuals from the set of genotyped individuals with replacement, and then simulated phenotypic data as described above for the initial set of simulations, but without the 1000 causal variants treatment. Genotypes (i.e., individuals) were replicated to obtain this sample size; this alters the structure of the kinship matrix and could affect performance independent of sample size. To test the effect of replicating genotypes (versus increasing sample sizes), we generated another series of data sets where we randomly chose 148 of the 592 individuals and replicated them each four times (with N kept constant at

592). This also allowed us to evaluate the benefits and costs of more structured experimental designs (e.g., experiments involving full or half-sib families or even clones).

We simulated a final series of fitness data sets using GBS data from *Rhagoletis pomonella* (Dryad DOI:10.5061/dryad.mb2tj). These data were described by Egan *et al.* (2015). Whereas this was a smaller data set (149 individuals and 33,723 SNPs), it is of interest because inversion polymorphisms result in large blocks of elevated LD, and more generally, LD is higher in *R. pomonella* (e.g., significant LD often extends beyond 10 cM) than in *T. cristinae* (e.g, average LD between SNPs ranges from 0.007 [SNPs $<$ 100 bp apart] to 0.004 [SNPs $>$ 100 bp apart]) (Feder *et al.*, 2003; Gompert *et al.*, 2014; Egan *et al.*, 2015). Thus, it allowed us to ask whether increased LD offset the negative effect of a smaller sample size (for simplicity, we focus on the effect on PVE and n-γ). To this end, we replicated genotypes in a subset of simulations to obtain the same sample size as we had for the *T. cristinae* data (N = 592 individuals). Note that higher levels of LD generally make it easier to tag causal variants, but more difficult to localize them (see, e.g., Rieseberg & Buerkle, 2002), but that LD should in general improve estimates of PVE as this only requires tagging causal variants. As with the initial set of simulations, we generated replicate data sets with h^2 equal to 0.3 or 0.05 and 10, 100, or 1000 causal variants (we only considered a quantitative metric of fitness, and only only 10 or 100 causal variants for the simulations with 592 individuals).

Analyses of the simulated data

We fit a BSLMM for each data set using `gemma` with two replicate MCMC runs, each with a 1 million iteration burnin, 2 million sampling iterations and a thinning interval of 100. Kinship matrixes were calculated as $K = \frac{1}{p}\mathbf{X}\mathbf{X}^T$, where $\mathbf{X}$ is the matrix of genotypic data and p is the number of loci.

We quantified the evidence of direct selection on individual SNPs based on posterior

inclusion probabilities, model-averaged estimates of selection ($\bar{\beta} = \frac{1}{2}s^D$), and point estimates of β assuming $\beta \neq 0$ (denoted $\hat{\beta}$). Both estimates of selection coefficients account for correlations among genotypes at different loci. We then assessed performance based on the correlation between true and inferred selection coefficients, and the normalized root-mean square error (RMSE) (normalized by the range of β). SNP effects were only considered for data sets that included the causal variants to make comparisons with true results readily interpretable.

We summarized posterior distributions for genetic architecture parameters (we focused mostly on PVE and n-γ, but also present estimates of PGE) based on the posterior mode and the 90% highest posterior density interval (HPDI), as calculated with the `R` package `coda`. The accuracy and precision of these parameter estimates were then quantified based on the RMSE and 90% HPDI coverage, where the latter is the proportion of the time that the true parameter value was included in the 90% HPDIs. Thus, lower RMSE and higher 90% HPDI coverage equate to greater accuracy and precision of the BSLMM approach for inferring our parameters of interest.

Results

Estimating direct selection

Under most conditions, partial regression coefficients (i.e., measures of direct selection or $\frac{1}{2}s^D$) were only weakly correlated with their true values (Fig. 3), such that distribution of true versus estimated effect sizes differed (Fig. 4). A notable exception occurred when fitness had a high heritability ($h^2 = 0.3$) and was determined by a modest number of variants ($L = 10$). Under these conditions estimates of selection ($\bar{\beta}$) were highly correlated with their true values (mean $r = 0.73$, s.d. 0.16) and the inferred and true effect size distributions were similar (Fig. 4c). Correlations between true and estimated effects were also higher when

only causal variants were considered (Fig. 3), or when the sample size was increased to 2500 (Fig. S1). In contrast, replicating genotypes (without increasing N) caused a decrease in correlations between true and inferred measures of selection (Fig. S2).

The mean posterior inclusion probability (PIP) for causal variants was relatively high for $h^2 = 0.3$ and $L = 10$ (0.26, s.d. 0.10), but was near-zero for more diffuse genetic architectures or when h^2 was low (Fig. 5a). Average PIPs for causal variants nearly doubled when the sample size was increased from 592 to 2500 individuals (0.48 for $h^2 = 0.3$ and $L = 10$, and 0.13 for $h^2 = 0.05$ and $L = 10$; Fig. 5b), but decreased notably when genotypes were replicated without increasing N (Fig. 5c). The accuracy of estimates of direct selection was also affected by the genetic architecture of fitness and the estimator used. For example, estimates of partial regression coefficients were the least accurate (i.e., had the greatest RMSE) when data sets were simulated with diffuse genetic architectures or when point estimates of selection ($\hat{\beta}$) were used rather than model-averaged estimates ($\bar{\beta}$) (Fig. S3). As with other metrics, increasing sample size to 2500 resulted in a decline in normalized RMSE (Fig. S4), but using replicated genotypes while keeping the sample size at 592 increased normalized RMSE (Fig. S5).

Quantitative estimation of genetic variation for fitness

Even with moderately large sample sizes (e.g., 100s of individuals), considerable uncertainty was observed for estimates of the proportion of variation in fitness explained by the genetic data (PVE) and the number of causal variants with measurable effects (n-γ) (e.g., Figs. S6, S7, S8). Despite this overall lack of precision, posterior point estimates of PVE were reasonably accurate (e.g., for the *T. cristinae* data with $N = 592$, RMSE varied from 0.06 to 0.23; Table 2, Fig. 6). The accuracy of point estimates increased with sample size and replication of individual genotypes, with much lower RMSE (and higher 90% HPDI coverage) for $N = 2500$ or $N = 592$ with replicates than $N = 592$ with unique genotypes (0.01 to 0.02

for $N = 2500$ compared to 0.09 to 0.19 for similar conditions with $N = 592$; Table 2, Fig. S9).

PVE was often lower for binary fitness metrics than for quantitative fitness metrics, though this did not have a consistent effect on accuracy (i.e., in some cases this gave better estimates as results for the quantitative metric were upwardly biased; Table 2; Fig. S10a). Simulations based on the *R. pomonella* data gave more variable and less accurate estimates of PVE than did those from *T. cristinae*, particularly with $h^2 = 0.3$ and $L = 100$ or 1000 (Table S1; Fig. S10b). However, results based on the *R. pomonella* data were similar to *T. cristinae* when we replicated genotypes to obtain the same sample sizes, suggesting that the poorer performance with the *R. pomonella* data was due to low sample sizes rather than high LD (Table S1; Fig. S10). 90% HPDIs for PVE generally included the true parameter value (the worst performance was observed for binary metrics; Table 2).

Estimation of the number of casual variants

Performance was notably poorer in terms of estimating the number of causal variants (that is, for inference of n-γ compared to PVE), but these results were also more difficult to interpret (Table 2, S1). Specifically, we seldom found evidence for greater than 10 variants with measurable effects on fitness, regardless of conditions (the greatest exception was for the case of 100 causal variants with $h^2 = 0.3$ and $N = 2500$; Table 2). Thus, estimates of n-γ were mostly (but not entirely) independent of simulation conditions (that is, of the true parameter values). However, because the magnitude of fitness effects varied among causal variation (which were normally distributed) and many had very small effects (this is particularly true for the case where 1000 variants explained only 5% of the variation in fitness), not all of these variants necessarily had "measurable" effects on fitness and many were likely subsumed in the polygenic term (i.e., via their contribution to overall genetic similarity captured by the kinship matrix).

This interpretation is consistent with the fact that our estimates of PVE were fairly accurate, and that the proportion of the PVE that was attributable to loci with measurable, rather than infinitesimal effects (PGE in `gemma`) decreased with the number of causal variants. For example, mean estimates of PGE based on the *Timema* data with $h^2 = 0.3$ were 0.79, 0.41, and 0.03 for simulations with $L = 10$, 100 and 1000, respectively. Also in support of this, SNP posterior inclusion probabilities (PIPs), which measure the probability a locus has a measurable effect on fitness and are the basis for estimates of the number of causal variants (n-γ), were positively correlated with effect sizes. Average correlations (Pearson's r values) between PIPs and effect sizes for these same data sets were 0.61 ($L = 10$), 0.27 ($L = 100$) and 0.05 ($L = 1000$).

Discussion

Estimating direct selection

We found that BSLMMs could provide useful information about individual bouts of direct selection on genetic loci under at least some conditions, but that important and sometimes strong limitations exist. For example, we showed that reasonably accurate estimates of selection coefficients could be obtained when sample sizes were large (N = 2500), the genetic architecture of fitness was relatively concentrated ($L = 10$) and fitness was more heritable ($h^2 = 0.3$). With that said, even very large sample sizes gave poor estimates of direct selection when fitness had a diffuse genetic architecture (e.g., $h^2 = 0.05$ and $L = 1000$). Thus, when heritability is low or fitness is highly polygenic, it might not be practical or even possible to obtain large enough samples for accurate estimates of direct selection on individual loci. These results are consistent with the general finding from GWASs over the past few decades that large sample sizes are often required but not always sufficient to map phenotypes for complex or quantitative traits onto genotypes (Manolio *et al.*, 2009; Visscher *et al.*, 2012).

Replicating genotypes (while holding N constant) actually degraded performance with respect to estimating direct selection. We suspect this occurred because fewer independent data points were available to isolate the effects of individual loci on fitness. With this in mind, our results suggest that experiments designed to detect direct selection on individual genes should maximize sample sizes without necessarily attempting to include multiple individuals from the same family or replicate clones (when this is an option). In some systems it might be possible to obtain larger total sample sizes by studying multiple experimental populations in a block design (as in Gompert *et al.*, 2014), perhaps at the expense of sample sizes within populations or blocks. Moreover, such replicated block designs could provide additional information about the consistency of selection across space or genomic backgrounds. In the end, the large experiments required to accurately measure direct selection on genes might benefit from (or even require) multi-investigator collaborative efforts on the same scale as those currently used to map human diseases (e.g., $N > 100{,}000$ as in IL6R Genetics Consortium Emerging Risk Factors Collaboration, 2012).

In addition to study design, we found that the estimator used to infer selection coefficients mattered. In particular, we obtained more accurate estimates of direct selection (lower RMSE and a higher correlation with the true values) with model-averaged coefficients (i.e., $\bar{\beta}$) than with those that assumed a non-zero effect (i.e., $\hat{\beta}$). A notable exception occurred for concentrated genetic architectures when only considering causal variants. Here, $\hat{\beta}$ consistently outperformed $\bar{\beta}$ with respect to RMSE and the correlation with the true parameter value. But, because causal variants will rarely be known *a priori*, we still recommend using model-averaged regression coefficients to estimate direct selection on genetic loci.

Quantifying genetic variation for fitness

Some key questions about selection can be addressed directly from statistical summaries of direct selection at the genome-level (e.g., via the model parameters PVE, PGE and n-γ). When the heritability of fitness is low or fitness is highly polygenic, focusing on these

453 questions and parameters might be the most productive way forward (Rockman, 2012). For
454 example, estimates of PVE can be converted into measures of additive genetic variation
455 for fitness and these could be productively compared across environments, populations or
456 fitness components. In turn, these measures are of interest for studies of speciation as
457 genetic variation for fitness determines the evolutionary response to selection and thereby
458 affects the possibility for colonization of new habitats. Whereas such information could also
459 be obtained using traditional quantitative genetic breeding designs (Falconer & Mackay,
460 1996), these methods are not practical for many non-model organisms.

461 We found that fairly accurate estimates of PVE could be obtained under a wider
462 variety of conditions than estimates of direct selection on genes. The accuracy of PVE point
463 estimates was determined mostly by sample size (bigger was of course better) and whether
464 or not genotypes were replicated. Specifically and in contrast to the results for estimating
465 selection coefficients (see above), replication of genotypes increased the accuracy of PVE
466 estimates, likely by both increasing LD and increasing the explanatory power of overall
467 genetic similarity. Thus, when possible, studies designed to estimate PVE should include
468 replicate clones or inbred lines. Note however, that this will come at the cost of decreasing
469 one's ability to parse individual genotypic effects (compared to an analysis of the same
470 number of unrelated individuals). When clones are not available other structured designs,
471 such as studies of siblings or hybrids, should have a similar albeit less pronounced effect.
472 Because structured designs increase LD and thereby make it easier to tag a greater proportion
473 of causal variants with fewer sequenced loci, they could be particularly appropriate when
474 generating GBS data.

475 Unfortunately, n-γ was routinely underestimated, particularly when L was large,
476 although performance did improve with $N = 2500$. This however does not necessarily reflect
477 a failure of the method, as the effects of many causal variants were simply subsumed in
478 the polygenic term when the number of causal variants was large. As such, these smaller
479 effect causal variants did not contribute to estimates n-γ. Nonetheless, based on our results,

estimates of n-γ should be interpreted with extreme caution.

Additional considerations and future directions

Further refinements and extensions of BSLMMs have the potential to increase the utility of these models for studying direct selection. For example, current BSLMMs do not account for dominance or epistasis, which are central to many theories of speciation (e.g., Orr, 1995; Turelli & Orr, 2000; Gavrilets, 2004; Orr, 2005). Dominance can readily be incorporated into whole genome regression models, such as BSLMMs, and the same is true in principle for epistasis but the number of genotype combinations present a daunting, but not insurmountable, computational challenge (Zhang & Liu, 2007; Jiang *et al.*, 2009; Wang *et al.*, 2010; Ritchie, 2011, 2015). Our understanding of speciation would benefit from measures of selection that explicitly incorporate genotype-environment interactions or that tie selection to trait genetics. Genotype-environment interactions for fitness are central to ecological speciation and have been tested for in many studies, but often by *post hoc* comparisons rather than formal inference within a model (e.g., Gompert *et al.*, 2014). With that said, adding additional model parameters for genotype-environment interactions or epistasis will further increase the sample size required for accurate inferences. Thus, trade-offs exist between extending the realism of models and obtaining reliable estimates of parameters with limited sample sizes. Notably, methods now exist that take trait architectures into account when testing for selection based on spatial patterns of genetic variation (Berg & Coop, 2014). Similar approaches could be used to powerfully connect fitness to phenotype and genotype in short-term studies of selection, and doing so should not entail a cost (unlike adding epistasis) as this would decrease the number of free parameters in the model. Such an integrative framework has the potential to truly advance our understanding of the causes and dynamics of speciation in nature.

Beyond methodological refinements, progress in understanding selection's role in speciation can be made by combining information from studies of direct selection with genome

scans of natural populations or even long-term evolve and re-sequence experiments. Population genomic methods (e.g., F_{ST} outlier analyses and tests for allele frequency–environment correlations; Beaumont & Balding, 2004; Foll & Gaggiotti, 2008; Coop *et al.*, 2010; Günther & Coop, 2013) gain power to detect selection by compounding the evolutionary consequences of selection over many generations (Lewontin & Krakauer, 1973). However, such approaches rarely provide actual estimates of selection (Thurman & Barrett, 2016), do not parse direct vs. indirect selection and can be confounded by demographic processes (Excoffier *et al.*, 2009). In contrast, short-term studies of direct selection can employ experimental designs where demography is known precisely and where processes other than selection and drift (e.g., gene flow, mutation, and recombination) are eliminated (e.g., Gompert *et al.*, 2014). Consistency of patterns between these types of studies would implicate direct selection as a key driver of divergence and suggest selection has acted in a consistent manner through time. Conversely, a lack of consistency could suggest methodological shortcomings, indicate a greater role for other evolutionary processes (such as drift and linked selection), or show that selection or LD varies through time. Such temporal variation in selection has been detected in phenotypic and genetic studies (Barrett *et al.*, 2008; Siepielski *et al.*, 2009; Anderson *et al.*, 2014; Bergland *et al.*, 2014; Thurman & Barrett, 2016), but has rarely been incorporated into models of speciation.

Evolve and re-sequence experiments provide a powerful means to measure selection by compounding information over many generations (e.g., Cooper *et al.*, 2003; Blount *et al.*, 2008; Burke *et al.*, 2010, 2014; Long *et al.*, 2015; Gompert & Messina, 2016), and could be used to distinguish between direct and indirect selection (using, e.g., "driver" "passenger" models as in Illingworth & Mustonen, 2011). However, such studies have been mostly restricted to organisms with short generation times that can be maintained in the lab (e.g., viruses, bacteria, yeast, and *Drosophila*), and lab conditions may fail to capture the complexity of nature. In contrast, experiments that measure one or several bouts of selection within a generation can be conducted with a greater diversity of organisms under natural

or semi-natural conditions. Indeed, hundreds or even thousands of such within-generation estimates of phenotypic selection have increased our awareness of how variable selection can be across traits, time periods, and populations, and refinement of this awareness continues (Kingsolver *et al.*, 2001; Siepielski *et al.*, 2009). It will thus be important to recognize when multi-generation experiments are needed (e.g., to measure the effect size distribution of mutations fixed during a bout of adaptation), versus when replicated within-generation experiments might be more productive (e.g., to contrast directions of selection on genotypes across a suite of environments or to distinguish between mechanisms by eliminating mutation, recombination, etc.). When possible, short-term measures of selection should be compared to results from longer-term evolve and re-sequence experiments on the same species to determine whether the former can be extrapolated to predict evolutionary trajectories over greater time-scales (which are clearly relevant for speciation).

Alternative approaches

Some questions in speciation can only be addressed by disentangling direct and indirect selection. For example, measures of direct selection are most relevant for identifying the specific genes or alleles that cause reproductive isolation. Nonetheless and despite our focus on direct selection in this manuscript, there are cases where the combined effects of direct and indirect selection (that is, total selection) are of interest, and thus where the "problem" of correlated genetic loci disappears.

First, the expected genomic response to an episode of selection (i.e., genome wide changes in genotype and gamete frequencies) is dictated by total selection, not direct selection alone. This means that evolutionary change from one generation to the next is best predicted from total selection. With that said, longer-term predictions will only be valid if LD is maintained through time, for example by tight physical linkage or by selection and gene flow as can occur in hybrid zones (Barton & Hewitt, 1985). Otherwise, patterns of LD will change via recombination and changes in allele or haplotype frequencies.

Second, several important evolutionary phenomena depend on the total selection experienced by genetic loci each generation (that is, direct selection and LD with causal variants), including genetic hitchhiking (Maynard-Smith & Haigh, 1974), genome-wide congealing during speciation with gene flow (Flaxman *et al.*, 2013, 2014), and the reduction in effective gene flow across a hybrid zone (i.e., the barrier to gene flow; Barton, 1983; Barton & Bengtsson, 1986; Gavrilets, 2004; Barton & De Cara, 2009). Thus, under a range of conditions, whether populations can speciate with gene flow or remain distinct upon secondary contact depends on the total selection (specifically total selection in the context of divergent selection or selection against hybrids) rather than only direct selection on causal variants (Barton, 1983; Flaxman *et al.*, 2014). In conclusion, total selection matters because it is not always just individual genes that respond to selection, but potentially sets of genes or genomes (Lewontin, 1974), and thus measures of total selection provide key information about evolutionary processes in general, and speciation in particular.

Acknowledgments

We would like to thank and S. Xu, P. Schluter, and S. Rogers for organizing this special issue, and four anonymous reviewers for comments that improved this manuscript. ZG was funded by the US National Science Foundation (NSF DEB #1638768) and PN was supported by the Royal Society of London via a University Research Fellowship. Computing, storage and other resources from the Division of Research Computing in the Office of Research and Graduate Studies at Utah State University and the Center for High-Performance Computing at the University of Utah are gratefully acknowledged.

References

Agrawal AA, Conner JK, Rasmann S (2010) *Evolution Since Darwin: the First 150 Years*, chap. Tradeoffs and negative correlations in evolutionary ecology, pp. 243–268. Sinauer Associates, Inc.

Anderson JT, Lee CR, Mitchell-Olds T (2014) Strong selection genome-wide enhances fitness trade-offs across environments and episodes of selection. *Evolution*, **68**, 16–31.

Anderson JT, Lee CR, Rushworth CA, Colautti RI, Mitchell-Olds T (2013) Genetic trade-offs and conditional neutrality contribute to local adaptation. *Molecular Ecology*, **22**, 699–708.

Barrett RD, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767–780.

Barrett RDH, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. *Science*, **322**, 255–257.

Barton N, Bengtsson B (1986) The barrier to genetic exchange between hybridizing populations. *Heredity*, **57**, 357–376.

Barton NH (1983) Multilocus clines. *Evolution*, **37**, 454–471.

Barton NH, De Cara MAR (2009) The evolution of strong reproductive isolation. *Evolution*, **63**, 1171–1190.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Berg JJ, Coop G (2014) A population genetic signal of polygenic adaptation. *PLoS Genetics*, **10**, e1004412.

603 Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA (2014) Genomic evidence
604 of rapid and stable adaptive oscillations over seasonal time scales in *Drosophila*. *PLoS*
605 *Genetics*, **10**, e1004775.

606 Bernardo J, Bayarri M, Berger J, *et al.* (2003) Bayesian factor regression models in the "large
607 p, small n" paradigm. *Bayesian Statistics*, **7**, 733–742.

608 Blount ZD, Borland CZ, Lenski RE (2008) Historical contingency and the evolution of a key
609 innovation in an experimental population of *Escherichia coli*. *Proceedings of the National*
610 *Academy of Sciences*, **105**, 7899–7906.

611 Bumpus HC (1899) The elimination of the unfit as illustrated by the introduced sparrow,
612 *Passer domesticus*. *Biol. Lect., Woods Hole Mar. Biol. Station*, **6**, 209–226.

613 Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, Long AD (2010) Genome-
614 wide analysis of a long-term evolution experiment with *Drosophila*. *Nature*, **467**, 587–590.

615 Burke MK, Liti G, Long AD (2014) Standing genetic variation drives repeatable experimental
616 evolution in outcrossing populations of *Saccharomyces cerevisiae*. *Molecular Biology and*
617 *Evolution*.

618 Comeault AA, Flaxman SM, Riesch R, *et al.* (2015) Selection on a genetic polymorphism
619 counteracts ecological speciation in a stick insect. *Current Biology*, **25**, 1975–1981.

620 Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations
621 to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

622 Cooper TF, Rozen DE, Lenski RE (2003) Parallel changes in gene expression after 20,000
623 generations of evolution in *Escherichia coli*. *Proceedings of the National Academy of Sci-*
624 *ences*, **100**, 1072–1077.

625 Egan SP, Ragland GJ, Assour L, *et al.* (2015) Experimental evidence of genome-wide impact

of ecological selection during early stages of speciation-with-gene-flow. *Ecology Letters*, **18**, 817–825.

Endler JA (1986) *Natural Selection in the Wild*. Princeton University Press, Princeton, NJ.

Erbe M, Hayes B, Matukumalli L, *et al.* (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, **95**, 4114–4129.

Ewens WJ (2004) *Mathematical Population Genetics: I. Theoretical Introduction*, vol. 27. Springer Science & Business Media.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Falconer DS, Mackay TFC (1996) *Introduction to Quantitative Genetics*. Pretince Hall Publishers.

Feder JL, Gejji R, Yeaman S, Nosil P (2012) Establishment of new mutations under divergence and genome hitchhiking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 461–474.

Feder JL, Nosil P, Flaxman SM (2014) Assessing when chromosomal rearrangements affect the dynamics of speciation: implications from computer simulations. *Frontiers in Genetics*, **5**.

Feder JL, Roethele JB, Filchak K, Niedbalski J, Romero-Severson J (2003) Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, rhagoletis pomonella. *Genetics*, **163**, 939–953.

Flaxman SM, Feder JL, Nosil P (2013) Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. *Evolution*, **67**, 2577–2591.

649 Flaxman SM, Wacholder AC, Feder JL, Nosil P (2014) Theoretical models of the influence
650 of genomic architecture on the dynamics of speciation. *Molecular Ecology*, **23**, 4074–4088.

651 Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for
652 both dominant and codominant markers: A Bayesian perspective. *Genetics*, **180**, 977–993.

653 Gavrilets S (2004) *Fitness Landscapes and the Origin of Species*. Princeton University Press.

654 Geyer CJ, Wagenius S, Shaw RG (2007) Aster models for life history analysis. *Biometrika*,
655 **94**, 415–426.

656 Gillespie J (2004) *Populations Genetics : a Concise Guide*. 2nd edn., Johns Hopkins Uni-
657 versity Press.

658 Gillespie JH (1991) *The Causes of Molecular Evolution*. Oxford University Press, USA.

659 Goddard ME, Hayes B (2007) Genomic selection. *Journal of Animal Breeding and Genetics*,
660 **124**, 323–330.

661 Gompert Z (2016) Bayesian inference of selection in a heterogeneous environment from ge-
662 netic time-series data. *Molecular Ecology*, **25**, 121–134.

663 Gompert Z, Comeault AA, Farkas TE, *et al.* (2014) Experimental evidence for ecological
664 selection on genome variation in the wild. *Ecology Letters*, **17**, 369–379.

665 Gompert Z, Jahner JP, Scholl CF, *et al.* (2015) The evolution of novel host use is unlikely to
666 be constrained by trade-offs or a lack of genetic variation. *Molecular Ecology*, **24**, 2777–
667 2793.

668 Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA (2012) Genomic
669 regions with a history of divergent selection affect fitness of hybrids between two butterfly
670 species. *Evolution*, **66**, 2167–2181.

671 Gompert Z, Messina FJ (2016) Genomic evidence that resource-based trade-offs limit host-
672 range expansion in a seed beetle. *Evolution*, **70**, 1249–1264.

673 Guan Y, Stephens M (2011) Bayesian variable selection regression for genome-wide associa-
674 tion studies and other large-scale problems. *Annals of Applied Statistics*, **5**, 1780–1815.

675 Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies.
676 *Genetics*, **195**, 205–220.

677 Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black
678 box of genomic prediction. *Genetics*, **194**, 597–607.

679 Hahn MW (2008) Toward a selection theory of molecular evolution. *Evolution*, **62**, 255–265.

680 Hayes B, Bowman P, Chamberlain A, Goddard M (2009) Genomic selection in dairy cattle:
681 Progress and challenges. *Journal of Dairy Science*, **92**, 433–443.

682 Heffner EL, Sorrells ME, Jannink JL (2008) Genomic selection for crop improvement. *Crop*
683 *Science*, **49**, 1–12.

684 Hoffmann AA, Merilä J, Kristensen TN (2016) Heritability and evolvability of fitness and
685 nonfitness traits: Lessons from livestock. *Evolution*, **70**, 1770–1779.

686 Huang Y, Wright SI, Agrawal AF (2014) Genome-wide patterns of genetic variation within
687 and among alternative selective regimes. *PLoS Genetics*, **10**, e1004527.

688 IL6R Genetics Consortium Emerging Risk Factors Collaboration (2012) Interleukin-6 recep-
689 tor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. *The*
690 *Lancet*, **379**, 1205–1213.

691 Illingworth CJR, Mustonen V (2011) Distinguishing driver and passenger mutations in an
692 evolutionary history categorized by interference. *Genetics*, **189**, 989–1000.

693 Jiang R, Tang W, Wu X, Fu W (2009) A random forest approach to the detection of epistatic
694 interactions in case-control studies. *BMC Bioinformatics*, **10**, S65.

695 Jiggins C, Naisbit R, Coe R, Mallet J (2001) Reproductive isolation caused by colour pattern
696 mimicry. *Nature*, **411**, 302–305.

697 Kang HM, Zaitlen NA, Wade CM, *et al.* (2008) Efficient control of population structure in
698 model organism association mapping. *Genetics*, **178**, 1709–1723.

699 Kingsolver JG, Hoekstra HE, Hoekstra JM, *et al.* (2001) The strength of phenotypic selection
700 in natural populations. *The American Naturalist*, **157**, 245–261.

701 Kruuk LE, Clutton-Brock TH, Slate J, Pemberton JM, Brotherstone S, Guinness FE (2000)
702 Heritability of fitness in a wild mammal population. *Proceedings of the National Academy*
703 *of Sciences*, **97**, 698–703.

704 Lande R, Arnold S (1983) The measurement of selection on correlated characters. *Evolution*,
705 **37**, 1210–1226.

706 Lewontin R (1974) *The genetic basis of evolutionary change*. Columbia University Press,
707 New York, NY, USA.

708 Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective
709 neutrality of polymorphisms. *Genetics*, **74**, 175–195.

710 Long A, Liti G, Luptak A, Tenaillon O (2015) Elucidating the molecular architecture of
711 adaptation via evolve and resequence experiments. *Nature Reviews Genetics*, **16**, 567–
712 582.

713 Lowry DB, Willis JH (2010) A widespread chromosomal inversion polymorphism contributes
714 to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Bi-*
715 *ology*, **8**, e1000500.

Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland, MA, USA.

Manolio TA, Collins FS, Cox NJ, *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Maynard-Smith J, Haigh J (1974) Hitch-hiking effect of a favorable gene. *Genetical Research*, **23**, 23–35.

Meuwissen THE, Hayes B, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Mousseau TA, Roff DA (1987) Natural selection and the heritability of fitness components. *Heredity*, **59**, 181–197.

Nosil P (2012) *Ecological speciation*. Oxford University Press.

Nosil P, Crespi BJ, Sandoval CP (2002) Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature*, **417**, 440–443.

Nosil P, Vines TH, Funk DJ (2005) Reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution*, **59**, 705–719.

Ober U, Ayroles JF, Stone EA, *et al.* (2012) Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1002685.

Ording GJ, Mercader RJ, Aardema ML, Scriber J (2010) Allochronic isolation and incipient hybrid speciation in tiger swallowtail butterflies. *Oecologia*, **162**, 523–531.

Orr H (1995) The population-genetics of speciation - the evolution of hybrid incompatibilities. *Genetics*, **139**, 1805–1813.

Orr HA (2005) The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, **6**, 119–127.

739 Pérez P, de los Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based
740 on molecular markers and pedigree using the bayesian linear regression package in R. *The*
741 *Plant Genome*, **3**, 106–116.

742 Pespeni MH, Sanford E, Gaylord B, *et al.* (2013) Evolutionary change during experimental
743 ocean acidification. *Proceedings of the National Academy of Sciences*, **110**, 6937–6942.

744 Rausher MD (1992) The measurement of selection on quantitative traits: biases due to
745 environmental covariances between traits and fitness. *Evolution*, pp. 616–626.

746 Rennison DJ, Heilbron K, Barrett RD, Schluter D (2015) Discriminating selection on lateral
747 plate phenotype and its underlying gene, Ectodysplasin, in threespine stickleback. *The*
748 *American Naturalist*, **185**, 150–156.

749 Resende M, Munoz P, Acosta J, *et al.* (2012) Accelerating the domestication of trees us-
750 ing genomic selection: accuracy of prediction models across ages and environments. *New*
751 *Phytologist*, **193**, 617–624.

752 Reynolds RJ, de los Campos G, Egan SP, Ott JR (2016) Modelling heterogeneity among
753 fitness functions using random regression. *Methods in Ecology and Evolution*, **7**, 70–79.

754 Rieseberg LH, Buerkle CA (2002) Genetic mapping in hybrid zones. *American Naturalist*,
755 **159**, S36–S50.

756 Ritchie MD (2011) Using biological knowledge to uncover the mystery in the search for
757 epistasis in genome-wide association studies. *Annals of Human Genetics*, **75**, 172–182.

758 Ritchie MD (2015) Finding the epistasis needles in the genome-wide haystack. In: *Epistasis*,
759 pp. 19–33, Springer.

760 Rockman MV (2012) The QTN program and the alleles that matter for evolution: All that's
761 gold does not glitter. *Evolution*, **66**, 1–17.

762 Schluter D (1988) Estimating the form of natural selection on a quantitative trait. *Evolution*,
763 pp. 849–861.

764 Schluter D (2001) Ecology and the origin of species. *Trends in Ecology and Evolution*, **16**,
765 372–380.

766 Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proceedings of National*
767 *Academy of Sciences*, **106**, 9955–9962.

768 Shaw RG, Geyer CJ, Wagenius S, Hangelbroek HH, Etterson JR (2008) Unifying life-history
769 analyses for inference of fitness and population growth. *The American Naturalist*, **172**,
770 E35–E47.

771 Siepielski AM, DiBattista JD, Carlson SM (2009) It's about time: the temporal dynamics
772 of phenotypic selection in the wild. *Ecology Letters*, **12**, 1261–1276.

773 Siepielski AM, Gotanda KM, Morrissey MB, Diamond SE, DiBattista JD, Carlson SM (2013)
774 The spatial patterns of directional phenotypic selection. *Ecology Letters*, **16**, 1382–1392.

775 Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. *Nature*
776 *Reviews Genetics*, **3**, 137–144.

777 Tang S, Presgraves DC (2009) Evolution of the *Drosophila* nuclear pore complex results in
778 multiple hybrid incompatibilities. *Science*, **323**, 779–782.

779 Thomasen JR, Egger-Danner C, Willam A, Guldbrandtsen B, Lund MS, Sørensen AC (2014)
780 Genomic selection strategies in a small dairy cattle population evaluated for genetic gain
781 and profit. *Journal of Dairy Science*, **97**, 458–470.

782 Thurman TJ, Barrett RDH (2016) The genetic consequences of selection in natural popula-
783 tions. *Molecular Ecology*, **25**, 1429–1448.

784 Tiffin P, Ross-Ibarra J (2014) Advances and limits of using population genetics to understand
785 local adaptation. *Trends in Ecology & Evolution*, **29**, 673–680.

786 Turelli M, Orr HA (2000) Dominance, epistasis and the genetics of postzygotic isolation.
787 *Genetics*, **154**, 1663–1679.

788 Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. *The*
789 *American Journal of Human Genetics*, **90**, 7–24.

790 Wang Y, Liu X, Robbins K, Rekaya R (2010) AntEpiSeeker: detecting epistatic interac-
791 tions for case-control studies using a two-stage ant colony optimization algorithm. *BMC*
792 *Research Notes*, **3**, 117.

793 Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*,
794 **14**, 851–865.

795 Yeaman S (2015) Local adaptation by alleles of small effect. *The American Naturalist*, **186**,
796 S74–S89.

797 Zhang Y, Liu JS (2007) Bayesian inference of epistatic interactions in case-control studies.
798 *Nature Genetics*, **39**, 1167–1173.

799 Zhao K, Aranzana MJ, Kim S, *et al.* (2007) An *Arabidopsis* example of association mapping
800 in structured samples. *PLoS Genetics*, **3**, e4.

801 Zhou X, Carbonetto P, Stephens M (2013) Polygenic modeling with Bayesian sparse linear
802 mixed models. *PLoS Genetics*, **9**, e1003264.

Data Accessibility

Simulated data sets and scripts used for analysis will be archived with DRYAD (DOI pending).

Author Contributions

ZG generated and analyzed the simulated data sets. All authors wrote and revised the manuscript.

Tables and Figures

Table 1: Glossary of key terms.

term	**definition**
direct selection	selection on a genetic locus resulting from its effect on fitness
indirect selection	selection on a genetic locus caused by LD with directly selected genotypes at other loci
total selection	combined effects direct and indirect selection on a genetic locus
linkage disequilibrium (LD)	statistical correlations between genotypes at different loci (physical linkage can facilitate LD but is not required for it)
selection coefficient (s)	measure of the strength of selection (direct or total), often expressed as the difference in expected fitness between alternative homozygotes
polygenic modeling	methods for connecting phenotypes to genotypes that consider many loci at once and do not rely on binary classifications of loci as associated or un-associated with phenotype
PVE	proportion of the phenotypic variation explained by the genetic data, which should approach the narrow-sense heritability of the trait (fitness) as the genome becomes saturated with genetic markers
PGE	the proportion of the PVE explained by loci with measurable effects on a trait (fitness); the remainder of the PVE comprises loci with near infinitesimal effects
n-γ	number of genetic markers with measurable effects on the phenotype (fitness)
PIP	posterior inclusion probability, that is the posterior probability that a genetic marker is under direct selection (or is in high LD with an un-sequenced locus under direct selection)
HPDI	highest posterior density interval, that is the interval that contains the most probable parameter values such that every value in the interval is more probable than any value not in the interval

Table 2: Accuracy of genome-level parameter estimates under different conditions. Results are shown for data sets generated from the *T. cristinae* genetic data; see (Table S1) for results from the *R. pomonella* data. Average metrics across replicates are reported with and without causal variants included in the analysis. 'estimate' denotes the point estimate of the parameter (posterior mode), 'RMSE' is the root mean square error, and '90% cov.' gives the proportion of times the true parameter value was included in the 90% HDPIs. 'no. loci' gives the actual number of causal variants (L), whereas 'no. SNPs' refers to the number of causal variants inferred from the model. 'N' is the sample size (N) and ^{a} denotes cases where genotypes were replicated (see the main text for details).

h^2	no. loci	metric	causal	N	PVE			no. SNPs		
					estimate	RMSE	90% cov.	estimate	RMSE	90% cov.
0.3	1000	quantitative	true	592	0.26	0.20	0.92	8.7	991.7	0.00
0.3	100	quantitative	true	592	0.34	0.19	0.86	18.3	85.6	0.84
0.3	10	quantitative	true	592	0.39	0.14	0.80	7.3	5.6	0.88
0.05	1000	quantitative	true	592	0.09	0.14	0.96	3.5	996.5	0.00
0.05	100	quantitative	true	592	0.08	0.09	0.98	3.6	96.4	0.82
0.05	10	quantitative	true	592	0.07	0.09	0.94	3.5	6.6	1.00
0.3	1000	binary	true	592	0.12	0.23	0.72	8.8	991.8	0.00
0.3	100	binary	true	592	0.16	0.18	0.84	4.6	95.4	0.74
0.3	10	binary	true	592	0.26	0.15	0.90	6.0	7.0	0.94
0.05	1000	binary	true	592	0.05	0.06	1.00	3.8	996.2	0.00
0.05	100	binary	true	592	0.05	0.07	0.96	3.6	96.4	0.83
0.05	10	binary	true	592	0.07	0.10	0.96	4.1	6.1	1.00
0.3	100	quantitative	true	2500	0.30	0.02	0.90	63.2	45.3	0.62
0.3	10	quantitative	true	2500	0.31	0.02	0.90	7.2	3.7	0.78
0.05	100	quantitative	true	2500	0.05	0.02	0.80	9.1	99.1	0.68
0.05	10	quantitative	true	2500	0.05	0.01	0.94	3.9	6.8	0.84
0.3	100	quantitative	true	592^a	0.31	0.03	0.96	4.8	99.5	0.74
0.3	10	quantitative	true	592^a	0.30	0.05	0.84	4.3	6.1	0.74
0.05	100	quantitative	true	592^a	0.05	0.03	0.92	3.3	96.7	0.66
0.05	10	quantitative	true	592^a	0.04	0.03	0.88	3.0	7.1	1.00
0.3	1000	quantitative	false	592	0.24	0.19	0.88	4.2	995.8	0.00
0.3	100	quantitative	false	592	0.25	0.19	0.94	5.2	94.9	0.92
0.3	10	quantitative	false	592	0.26	0.19	0.92	3.8	6.4	0.98
0.05	1000	quantitative	false	592	0.08	0.14	0.96	3.6	996.5	0.00
0.05	100	quantitative	false	592	0.08	0.10	0.98	3.8	96.4	0.82
0.05	10	quantitative	false	592	0.07	0.09	0.96	3.5	6.4	1.00

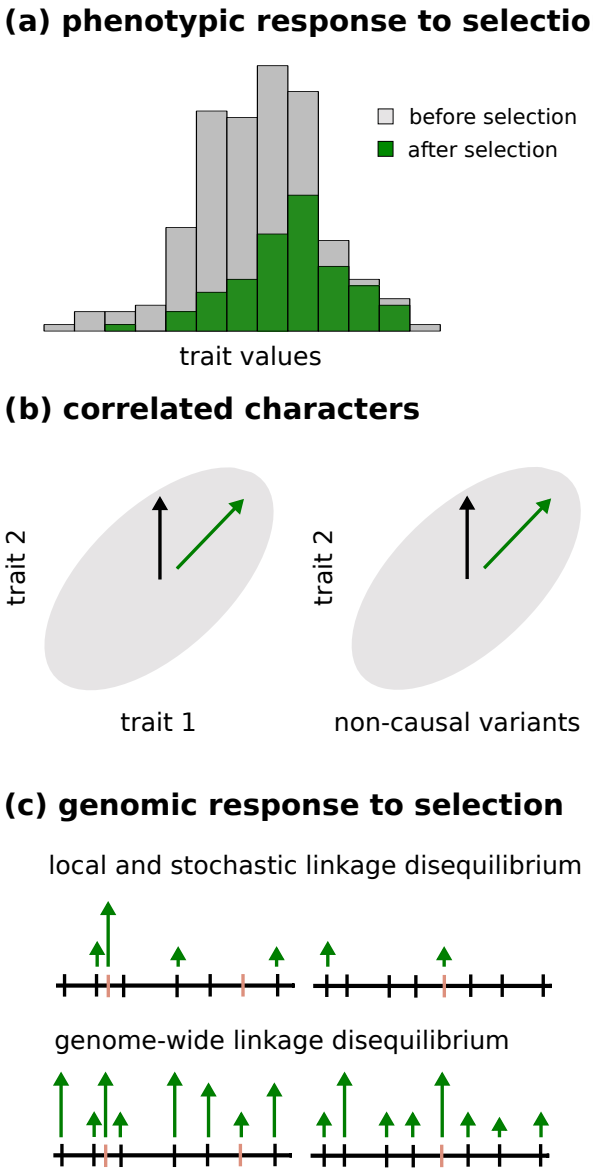

Figure 1: Schematic representation of how phenotypic selection drives allele frequency change across the genome, either directly or indirectly because of correlations among traits and non-causal loci. Panel (a) shows how direct phenotypic selection on a trait (in this case trait 2) alters the distribution of that trait. Panel (b) shows how selection on trait 2 (black arrows denote the direction of selection) can cause a response to selection at a correlated trait (trait 1) that itself has no effect on fitness, and thus at genetic variants that underlie variation in the correlated trait (green arrows give the direction of the response) when correlations exist as denoted by the gray ellipses. Panel (c) shows how the response to selection depends on patterns of LD. Here horizontal lines denote chromosomes, vertical bars correspond to genetic variants with (peach) or without (black) effects on trait 2 (that is, the trait that affect fitness), and vertical arrows indicate the magnitude of the response to selection (direct selection only occurs on the causal variants).

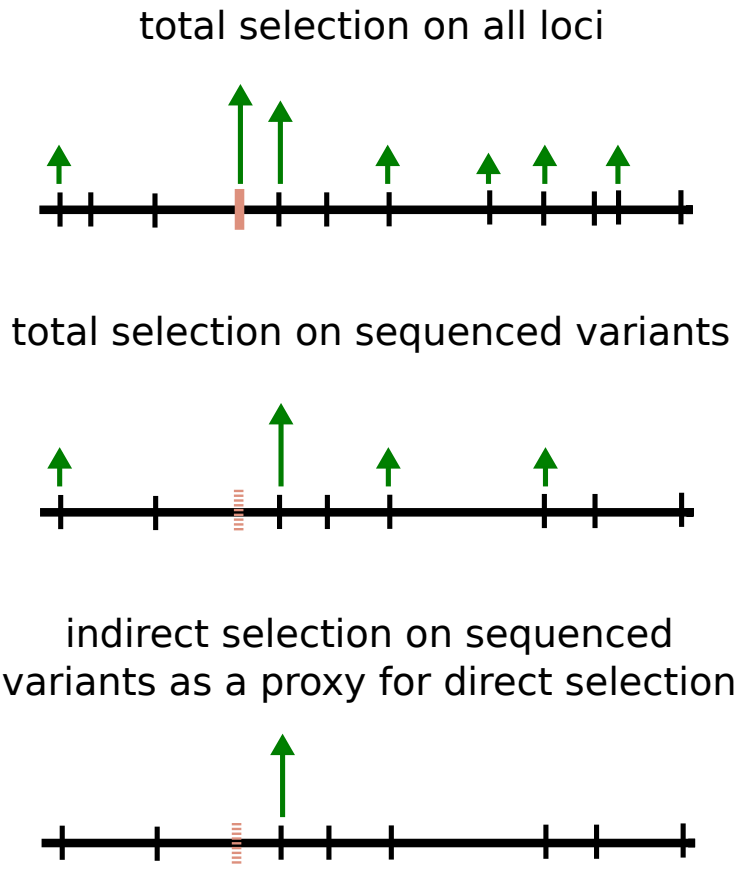

Figure 2: Graphical depiction of total and direct selection when causal variants are not sequenced in an empirical study. The top image ('selection on all loci') shows selection on a series of genetic variants. The horizontal line denotes a chromosome, vertical bars correspond to variants with (peach) or without (black) effects on fitness, and vertical arrows indicate the magnitude of selection. In the next two images, information is presented for the subset of variants that were sequenced; the causal variant was not sequenced but its position is noted with a dashed line. The middle image shows that all genetic markers in LD with the causal variant experienced indirect selection ('total selection on sequenced variants'). Whereas, the bottom image shows that, at least in this example, direct selection on the un-sequenced causal variant is fully accounted for as direct selection on the genetic variant most associated with the un-sequenced causal variant ('indirect selection on sequenced variants as a proxy for direct selection'). Because of imperfect LD, the strength of direct selection on the missing causal variant is underestimated, but the number of causal variants (one) is correctly inferred.

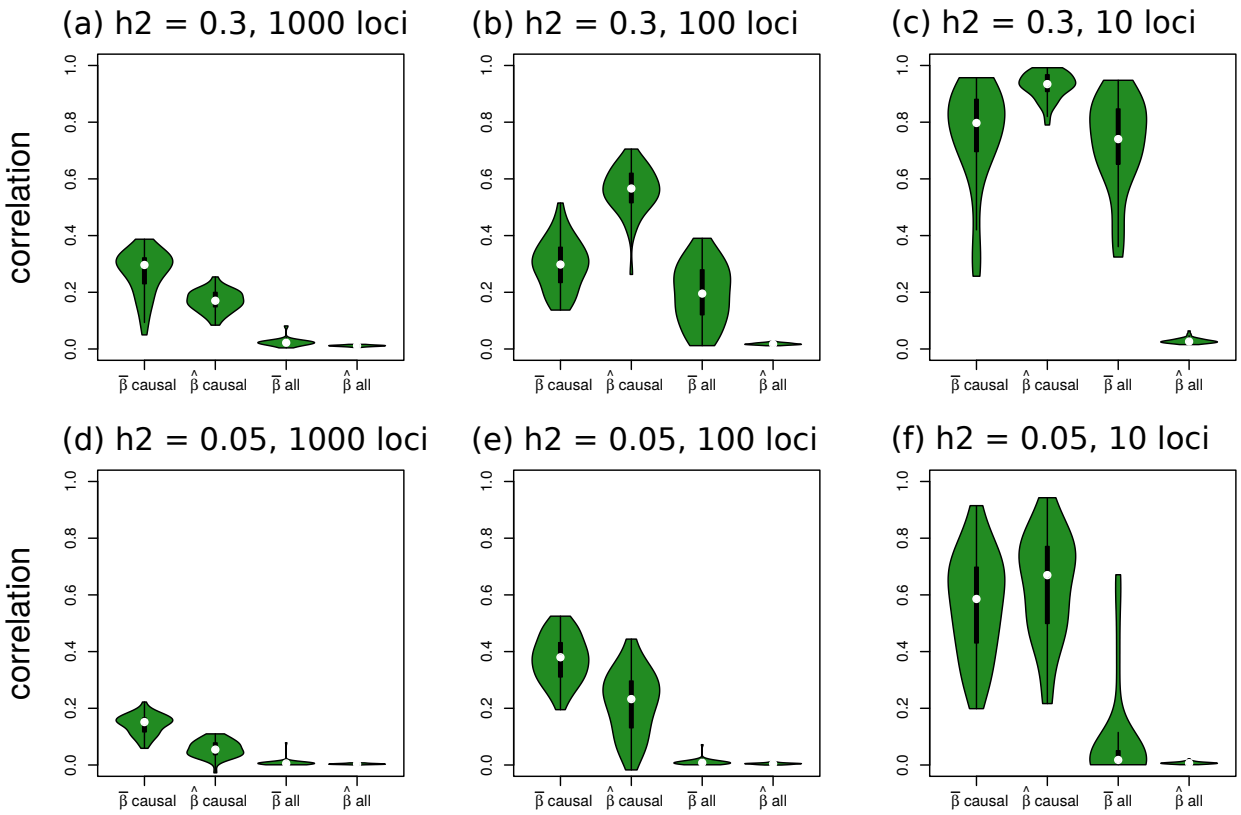

Figure 3: Violin plots summarize the distribution (across data sets) of Pearson correlations between true and estimated regression coefficients (i.e., measures of direct selection). Results shown here are from the *Timema cristinae* GBS data with $N = 592$ (without genotype replication) and a quantitative fitness metric. Results for different genetic architectures (i.e., h^2 = narrow-sense heritability and L = number of causal variants) are shown in each panel. Correlations for different combinations of h^2 and L are shown in different panels. Correlations were calculated for model-average ($\bar{\beta}$) and raw ($\hat{\beta}$) estimates of direct selection, and were calculated based on all SNPs or only the causal variants.

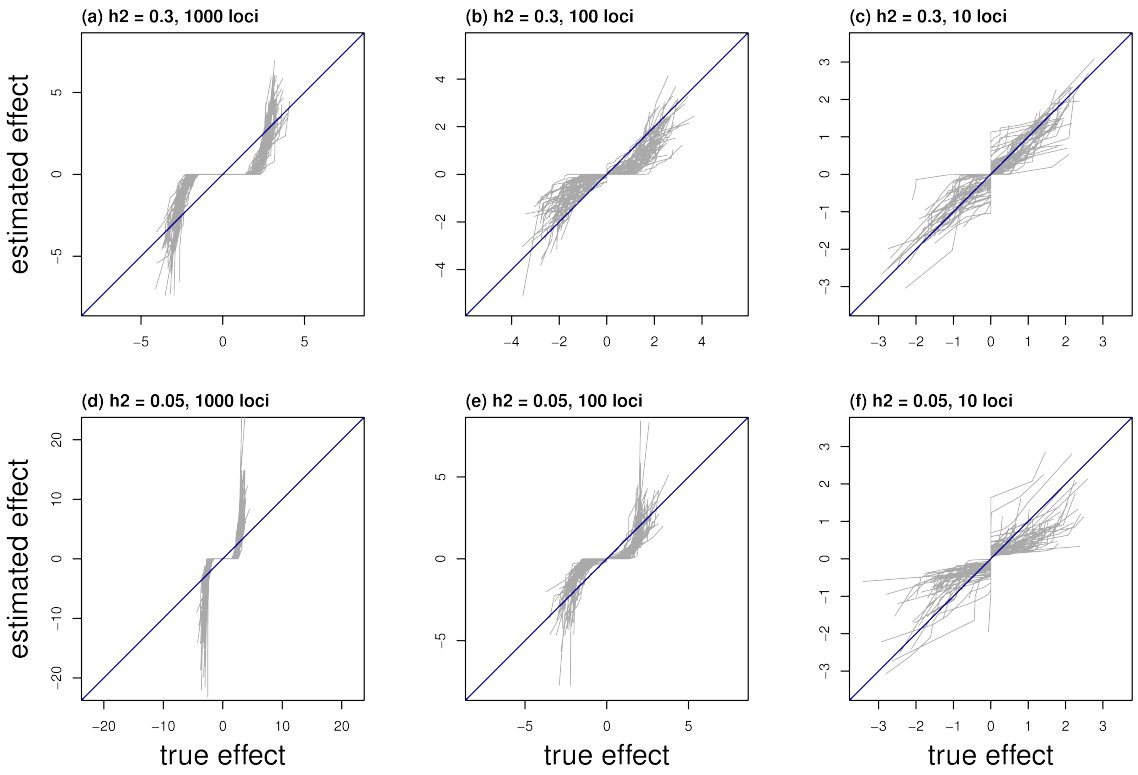

Figure 4: Quantile-quantile plots compare distributions of true (simulated) and estimated effect sizes. Each gray line corresponds to a single simulated data set. Results shown here are based on the *Timema cristinae* GBS data set with $N = 592$ (without genotype replication) and a quantitative fitness metric. Results for different genetic architectures (i.e., h^2 = narrow-sense heritability and L = number of causal variants) are shown in each panel (50 replicate data sets per conditions). One-to-one diagonal lines are included for reference. Effect size distributions for each simulated data set were obtained by averaging distributions over ten random draws from the posterior distribution of the `gemma` model parameters γ and β.

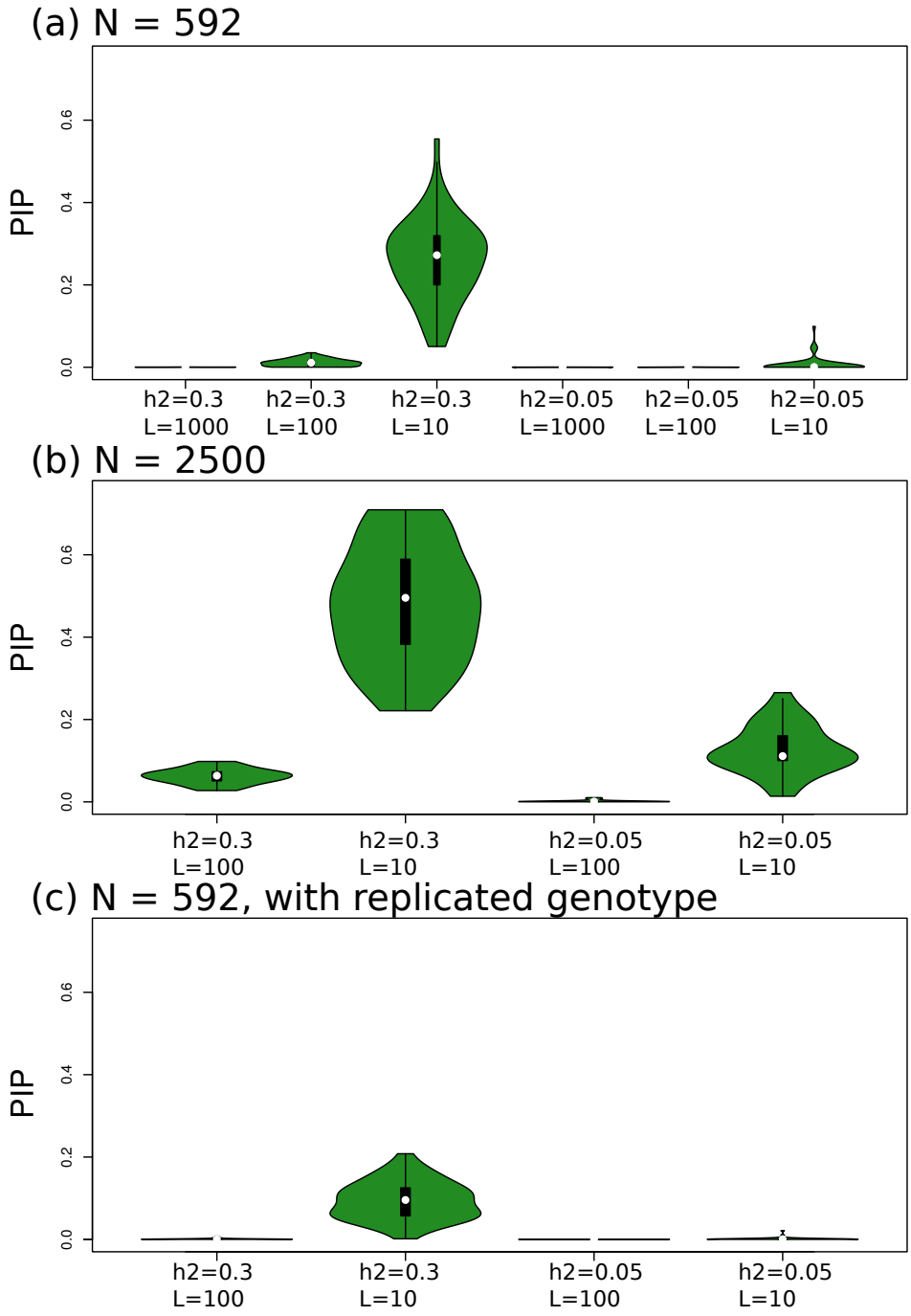

Figure 5: Violin plots summarize the distribution (across data sets) of posterior inclusion probabilities (PIPs) for causal variants, that is for variants directly affecting fitness. Results are shown for the *Timema cristinae* GBS data with a quantitative fitness metric with different sampling sizes and schemes (a-c) and genetic architectures (i.e., values of h^2 = narrow-sense heritability and L = number of causal variants).

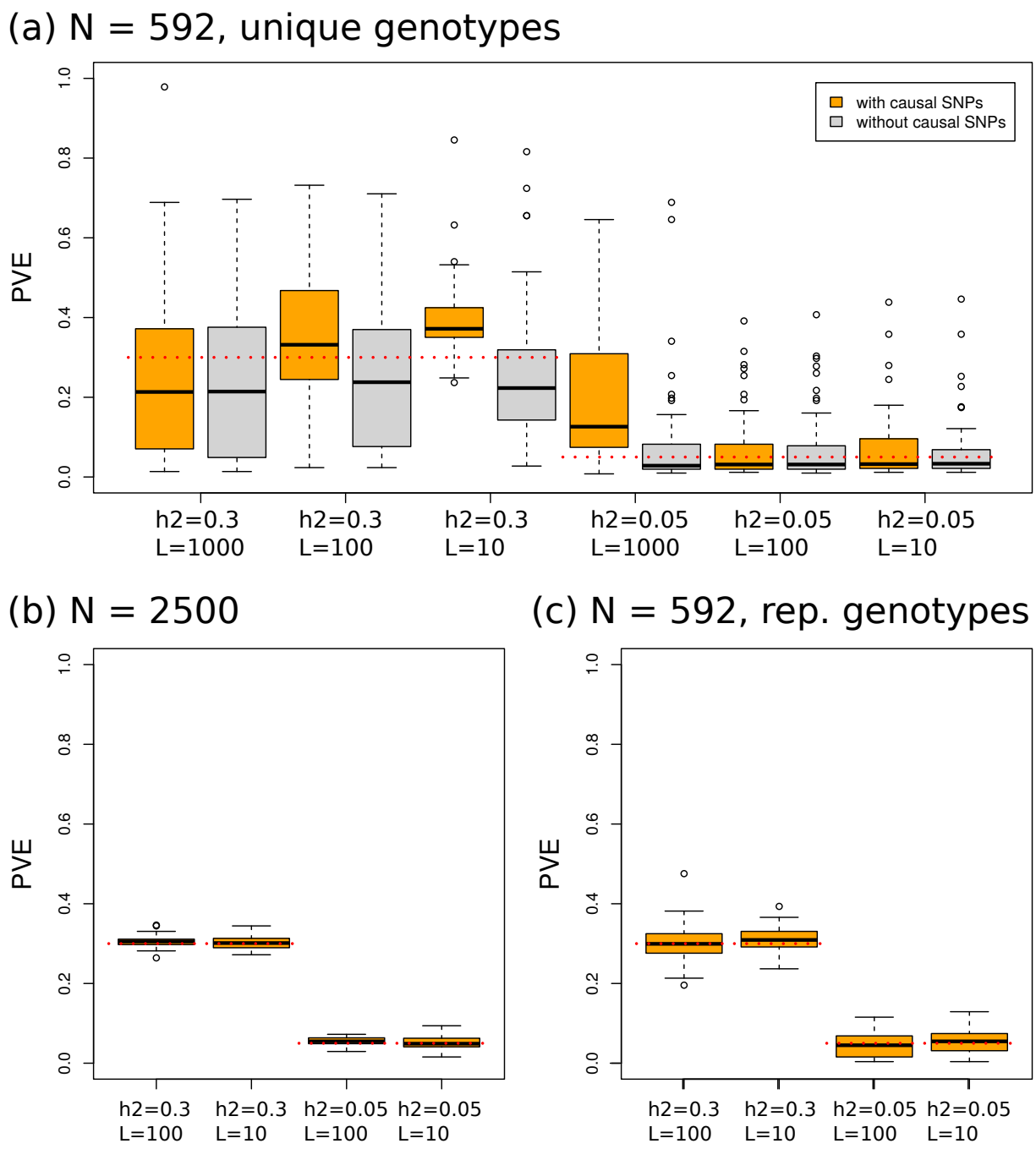

Figure 6: Box plots illustrate the distribution of point estimates for the proportion of variation in fitness explained by the genetic data (PVE). We show the distribution of point estimates (posterior mode) across replicates for different conditions. Dotted red-lines indicate the true parameter value. Panels (a), (b), and (c) give results for different sample sizes and schemes. Results shown here are based on the *Timema cristinae* GBS data with a quantitative metric of fitness and a range of genetic architectures (h^2 = narrow-sense heritability, L = number of causal variants, N = number of individuals).