Unifying approaches from statistical genetics and phylogenetics for mapping phenotypes in structured populations

Joshua G. Schraiber[1], Michael D. Edge[1,3,*] & Matt Pennell[1,2,3,*]

[1]*Department of Quantitative and Computational Biology, University of Southern California, USA*
[2]*Department of Biological Sciences, University of Southern California, USA*
[3]These authors contributed equally
[*]Corresponding authors: edgem@usc.edu, mpennell@usc.edu

Abstract

In both statistical genetics and phylogenetics, a major goal is to identify correlations between genetic loci or other aspects of the phenotype or environment and a focal trait. In these two fields, there are sophisticated but disparate statistical traditions aimed at these tasks. The disconnect between their respective approaches is becoming untenable as questions in medicine, conservation biology, and evolutionary biology increasingly rely on integrating data from within and among species, and once-clear conceptual divisions are becoming increasingly blurred. To help bridge this divide, we derive a general model describing the covariance between the genetic contributions to the quantitative phenotypes of different individuals. Taking this approach shows that standard models in both statistical genetics (e.g., Genome-Wide Association Studies; GWAS) and phylogenetic comparative biology (e.g., phylogenetic regression) can be interpreted as special cases of this more general quantitative-genetic model. The fact that these models share the same core architecture means that we can build a unified understanding of the strengths and limitations of different methods for controlling for genetic structure when testing for associations. We develop intuition for why and when spurious correlations may occur using analytical theory and conduct population-genetic and phylogenetic simulations of quantitative traits. The structural similarity of problems in statistical genetics and phylogenetics enables us to take methodological advances from one field and apply them in the other. We demonstrate this by showing how a standard GWAS technique—including both the genetic relatedness matrix (GRM) as well as its leading eigenvectors, corresponding the principal components (PCs) of the genotype matrix, in a regression model—can mitigate spurious correlations in phylogenetic analyses. As a case study of this, we re-examine an analysis testing for co-evolution of expression levels between genes across a fungal phylogeny, and show that including principal components as covariates decreases the false positive rate while simultaneously increasing the true positive rate. More generally, this work provides a foundation for more integrative approaches for understanding the genetic architecture of phenotypes and how evolutionary processes shape it.

Introduction

Statistical genetics and phylogenetic comparative biology share the goal of identifying correlations between features of individuals (or populations) that are structured by historical patterns of ancestry. In the case of statistical genetics, researchers search for causal genetic variants underlying a phenotype of interest, whereas in phylogenetic comparative biology, researchers are typically interested in testing for associations among phenotypes or between a phenotype and an environmental variable. In both cases, these tests are designed to isolate the influence of a focal variable from that of many potential confounding variables. But despite the shared high-level goal, the statistical traditions in these two fields have developed largely separately, and—at least superficially—do not resemble each other. Moreover, researchers in these two statistical traditions may have different understandings of the nature of the problems they are trying to solve.

In statistical genetics, phenotypes and genotypes can be spuriously associated because of confounding due to population structure [1–4] or assortative mating [5, 6]. For example, in their famous "chopsticks" thought experiment, Lander & Schork [1] imagined that genetic variants that have drifted to higher frequency in subpopulations in which chopsticks are frequently used will appear, in a broad sample, to be associated with individual ability to use chopsticks, even though the association is due to cultural confounding and not to genetic causation. Confounding can also be genetic [7]—if a genetic variant that changes a phenotype is more common in one population than others, leading to differences in average phenotype among populations, then other, non-causal variants that have drifted to relatively high frequency in this population may appear to be associated with the phenotype in a broad sample. In addition to affecting genome-wide association study (GWAS) results, such confounding can affect heritability estimation [8, 9], genetic correlation estimates [10, 11], and prediction of phenotypes from polygenic scores [12–16]. Although many candidate solutions have been offered [17–21], the two most common approaches involve adjusting for shared ancestry using the genetic relatedness matrix (GRM, [22]), either by incorporating individual values on the first several eigenvectors of this matrix (i.e. the principal components of the genotype matrix) as fixed effects [23], or by modeling covariance among individuals attributable to genome-wide relatedness in a linear mixed model (LMM, [24–27]).

In phylogenetic comparative biology, researchers typically aim to control for the similarity of related species by incorporating the species tree into the analysis. There has been a great deal of controversy over the years as to what the underlying goals and implicit assumptions of phylogenetic comparative methods (PCMs) are [see for examples refs. 28–34]. But broadly speaking, it seems that many researchers understand the goal of PCMs to be avoiding "phylogenetic pseudoreplication" [35]—mistaking similarity due to shared phylogenetic history for similarity due to independent evolutionary events [33]. This is most commonly done by conducting a standard regression, using either generalized least squares (GLS) or a generalized linear mixed model (GLMM), but including the expected covariance structure owing to phylogeny [36–40]. (Throughout this paper, we do not make a distinction between phylogenetic GLS and phylogenetic GLMM models. We refer to them generically by the shorthand GLS for the general case and PGLS for cases where the phylogenetic variance-covariance is used; see below for definitions.) This covariance structure reflects both

the relatedness of species and the expected distribution of phenotypes under a model of phenotypic evolution [41, 42], such as a Brownian motion [43] or an Ornstein-Uhlenbeck [44, 45] process. (The "phylogenetically independent contrasts" method [46], which ushered in modern PCMs is statistically equivalent to a PGLS model assuming a Brownian model [47].)

In recent years, however, signs have emerged that these two subfields may benefit from closer conversation, as emerging approaches in both statistical genetics and phylogenetics encounter questions that call for the other subfield's expertise. For example, in humans, evolutionary conserved sequences are enriched for trait and disease heritability [48, 49], and conservation across related species can be used to prioritize medically relevant variants in fine mapping [50, 51] and rare variant association studies [52, 53]. Similarly, conservation geneticists have used multi-species alignments to estimate the fitness effects of mutations in a focal population [54]. And there is growing interest in using estimated ancestral recombination graphs (ARGs) to perform explicitly tree-based versions of QTL mapping and complex trait analysis [55, 56]. From the phylogenetics side, researchers are increasingly employing GWAS-like approaches ("PhyloG2P" methods; [57]) for mapping phenotypes of interest for which the variation primarily segregates among rather than within species. Futhermore, phylogenetic biologists have been developing phylogenetic models that consider within-species polymorphisms in genes [58, 59] and how ancient polymorphisms that are still segregating among species may confound association tests [60] .

Such emerging connections suggest that it would be beneficial to better understand the ways in which statistical genetics and phylogenetic comparative biology relate to each other. To this end, we start from first principles and develop a general statistical model for investigating associations between focal variables while controlling for shared ancestry and environment. We show that both standard GWAS and standard phylogenetic regression emerge as special cases of this more general model. We illustrate how this deep connection provides insights into why these two classes of methods can be misled by certain types of unmodeled structure, why solutions in both fields work (and when do they not), and how statistical advances in either domain can be applied straightforwardly to the other.

Results and Discussion

A general polygenic model of quantitative trait variation

We assume a standard model in which many genetic factors of small effect influence a phenotype in an additive way—that is, there is no dominance or interaction among genetic loci (epistasis). Denoting by β_l the additive effect size of the variant at the lth locus and G_{il} the genotype of the ith individual at the lth locus, we write

$$G_i = \sum_l \beta_l G_{il}, \tag{1}$$

where G_i is the genetic component of the phenotype of individual i, sometimes called a genetic value or breeding value. We then express the phenotype of individual i, Y_i, as the sum of the genetic component and

an environmental component, E_i.

$$Y_i = G_i + E_i. \tag{2}$$

Due to shared ancestry, the genotypes of individuals in the sample will be correlated; thus, the genetic components of the individuals in the sample will be correlated. Moreover, the environments experienced by individuals may be correlated, and these environmental effects may be correlated with the genetic components. Thus,

$$\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(G_i, G_j) + \mathrm{Cov}(G_i, E_j) + \mathrm{Cov}(E_i, G_j) + \mathrm{Cov}(E_i, E_j). \tag{3}$$

For the rest of the paper, our focus will be on the first term $\mathrm{Cov}(G_i, G_j)$, the covariance in phenotypes between individuals due to genetic covariance. As we will show below, many models used by both statistical geneticists and phylogenetic biologists can be understood without reference to the components that include environmental effects. Interestingly, some statistical geneticists interpret the use of the GRM in the standard GWAS model as being primarily a proxy for shared (but unmeasured) environment among individuals [7], since in many cases, individuals whose genotypes are correlated will also experience similar environments. Phylogenetic biologists, on the other hand, tend to interpret the use of the phylogenetic covariance matrix in the model exclusively as controlling for genetic effects and thereby implicitly assume that the effects of the environment on phenotypes are unstructured [29, 32]. On macroevolutionary time scales, it is not so obvious that genetic and environmental similarities will mirror each other. For example, dispersal is less limited on long timescales [61, 62], so relatively closely related species might live in very different conditions. However, for traits such as gene expression (which is increasingly studied in a phylogenetic context; [59, 63–66]) that are strongly environmentally dependent and for which measurements are taken from species in different environments, we may need to develop new ways to model the environmental terms. This is a challenge beyond the scope of the present paper. We also note that there are some circumstances in which genetic covariance in equation 3 is undefined, such as when effect sizes have an undefined variance [67], or under certain phenomenological models of evolution on phylogenies [68, 69]; we reserve these situations for future work and focus on situations in which the genetic covariance is finite in the subsequent sections.

Conceptualizations of the genetic covariance

Individuals who are more closely related will have more similar genotypes. For example, individuals in the same local population may share the same alleles identical by descent due to recent common ancestry. On the other hand, individuals in different species may not share alleles due to the species being fixed for alternative alleles at a given locus. Using equation 1,

$$\begin{aligned}\mathrm{Cov}(G_i, G_j) &= \mathrm{Cov}\left(\sum_l \beta_l G_{il}, \sum_l \beta_l G_{jl}\right) \\ &= \sum_l \mathrm{Cov}(\beta_l G_{il}, \beta_l G_{jl}) + \sum_{l \neq k} \mathrm{Cov}(\beta_l G_{il}, \beta_k G_{jk})\end{aligned} \tag{4}$$

The first term arises from the correlations at a single locus, while the second term arises from correlations among loci across individuals. We focus on the first term, though the second can be important in biologically realistic situations. A particularly important case is selection on polygenic traits, which can cause correlations among the genetic contributions to a trait across loci. For example, if a population experiences directional selection on a highly polygenic phenotype, much of the phenotypic change, compared with a related population that has not experienced such selection, is due to to small, coordinated changes in allele frequency [70, 71]. In addition, linkage can affect the evolution of polygenic traits [72] and the results of heritability estimates [73].

In full generality, it is hard to say much more, since both the effect sizes and genotypes might be viewed as random and as dependent in arbitrary ways. One way to gain further insight is to imagine a variable Z that renders the genotypes G_{il} and the effect sizes β_l conditionally independent. As an example, we will imagine the locus-level allele frequency as a candidate for such a variable below. Conditioning on Z and using the definition of covariance and the law of total expectation, the first term becomes

$$\begin{aligned}\mathrm{Cov}(\beta_l G_{il}, \beta_l G_{jl}) &= \mathbb{E}(\beta_l^2 G_{il} G_{jl}) - \mathbb{E}(\beta_l G_{il})\mathbb{E}(\beta_l G_{jl}) \\ &= \mathbb{E}_Z(\mathbb{E}(\beta_l^2 \mid Z)\mathbb{E}(G_{il} G_{jl} \mid Z)) - \mathbb{E}_Z(\mathbb{E}(\beta_l \mid Z)\mathbb{E}(G_{il} \mid Z))\mathbb{E}_Z(\mathbb{E}(\beta_l \mid Z)\mathbb{E}(G_{jl} \mid Z)).\end{aligned} \tag{5}$$

This formula is fully general, as long as the genetic covariance exists, and can be applied in any evolutionary model in which there is a variable Z—potentially a latent variable—that accounts for the relationship between effect sizes and genotypes [74]. Moreover, it applies when the variable $Z = \beta$ or $Z = G$.

Methods across subfields of evolutionary and statistical genetics rely heavily on different versions of a matrix that encodes relationships among individuals or species by specializing equation 5. When we are referring to this matrix in general, we call it Σ. In a sample of n individuals (or n species), Σ is $n \times n$, and Σ_{ij} is proportional to some version of equation 5.

Among other names, in different settings, Σ might take the form of a "genetic relatedness matrix," "kinship matrix," "expected genetic relatedness matrix," or "phylogenetic variance-covariance matrix." Below, we consider the off-diagonal entries of each of these matrices in turn.

The genetic relatedness matrix

We can simplify equation 5 if the second term is equal to 0, which can be accomplished in a number of ways. One way is to mean-center the genotypes. In that case, we write

$$\tilde{G}_{il} = G_{il} - \frac{1}{n}\sum_j G_{jl}.$$

Then, the linear model of the phenotype (equation 2) can be written

$$\begin{aligned}
Y_i &= G_i + E_i \\
&= \sum_l \beta_l G_{il} + E_i \\
&= \sum_l \beta_l \left(G_{il} - \frac{1}{n} \sum_j G_{jl} + \frac{1}{n} \sum_j G_{jl} \right) + E_i \\
&= \sum_l \beta_l \left(G_{il} - \frac{1}{n} \sum_j G_{jl} \right) + \frac{1}{n} \sum_j \sum_l \beta_l G_{jl} + E_i \\
&= \mu + \tilde{G}_i + E_i,
\end{aligned}$$

where $\mu = \frac{1}{n} \sum_j \sum_l \beta_l G_{jl}$ is the mean genetic value in the population, and $\tilde{G}_i = \sum_l \beta_l \tilde{G}_{il}$. Then, equation 5 becomes

$$\begin{aligned}
\mathrm{Cov}(\beta_l \tilde{G}_{il}, \beta_l \tilde{G}_{jl}) &= \mathbb{E}_Z(\mathbb{E}(\beta_l^2 \mid Z)\mathbb{E}(\tilde{G}_{il}\tilde{G}_{jl} \mid Z)) - \mathbb{E}_Z(\mathbb{E}(\beta_l \mid Z)\mathbb{E}(\tilde{G}_{il} \mid Z))\mathbb{E}_Z(\mathbb{E}(\beta_l \mid Z)\mathbb{E}(\tilde{G}_{jl} \mid Z)) \\
&= \mathbb{E}_Z(\mathbb{E}(\beta_l^2 \mid Z)\mathbb{E}(\tilde{G}_{il}\tilde{G}_{jl} \mid Z)). \qquad (6)
\end{aligned}$$

The second line follows because mean-centering implies that $\mathbb{E}(\tilde{G}_{il} \mid Z) = 0$ for all i.

One reasonable candidate for Z—i.e. a variable conditional on which the genotype and allele frequency are independent—is the allele frequency p_l. Conditioning on p_l and using mean-centered genotypes,

$$\mathrm{Cov}(\beta_l \tilde{G}_{il}, \beta_l, \tilde{G}_{jl}) = \mathbb{E}_p(\mathbb{E}(\beta_l^2 \mid p_l)\mathbb{E}(\tilde{G}_{il}\tilde{G}_{jl} \mid p_l)).$$

In the presence of dense genotype data, we can approximate the expectation over genotypes by the empirical mean,

$$\mathbb{E}(\tilde{G}_{il}\tilde{G}_{jl} \mid p_l = p) \approx \frac{1}{n_p} \sum_{l:p_l=p} \tilde{G}_{il}\tilde{G}_{jl},$$

where n_p is the number of sites with frequency p. Then, provided an estimate of $\hat{\beta}_p^2 = \mathbb{E}(\beta_l^2 \mid p_l = p)$, we can compute

$$\begin{aligned}
\mathrm{Cov}(\tilde{G}_i, \tilde{G}_j) &\approx \sum_p n_p \hat{\beta}_p^2 \frac{1}{n_p} \sum_{l:p_l=p} \tilde{G}_{il}\tilde{G}_{jl} \qquad (7) \\
&= \sum_l \hat{\beta}_{p_l}^2 \tilde{G}_{il}\tilde{G}_{jl}. \qquad (8)
\end{aligned}$$

If we suppose that $\hat{\beta}_p^2 \propto (2p(1-p))^{-1}$, then this formula recapitulates the Genetic Relatedness Matrix (GRM), a realization of Σ that is commonly used to estimate heritability from SNP data [75] or to accommodate covariance due to relatedness in GWAS [24–27]. And considering that $\hat{\beta}_p^2 \propto (2p(1-p))^{-\alpha}$ recovers

the more general "alpha model" or "LDAK model" [22, 73]. In plant and animal breeding, where this framework first appeared [76], sometimes the same normalization is used as in human genetics, and sometimes genotypes are mean centered but not standardized [77, 78].

The (pedigree-based) kinship matrix

Historically, plant and animal breeders, along with human and behavior geneticists interested in resemblance of relatives, have frequently faced a situation in which they have had 1) (at least partial) pedigree data describing the parentage of sets of individual plants or animals, 2) phenotypic data on those individuals, but 3) no genome-wide genetic information. In such a situation, one can model the entries of Σ as a function of expected genetic similarity based on the pedigree information, as opposed to realized genetic sharing observed from genotypes [77, 79–81]. One can specialize equation 4 by fixing the effect sizes, leading to

$$\mathrm{Cov}\left(\sum \beta_l G_{il}, \sum \beta_l G_{jl}\right) = 2\theta_{ij} V_A, \tag{9}$$

where θ_{ij} is the kinship coefficient (obtained from the pedigree) relating individuals i and j, and V_A is the additive genetic variance. Although many derivations exist in standard texts [e.g. 81, 82], we include one in the appendix for completeness.

Methods based on this formulation include the "animal model" [79, 81, 83, 84], a widely used approach for prediction of breeding values in quantitative genetics. The connection between the animal model and genome-wide marker-based approaches was plain to the quantitative geneticists who first developed marker-based approaches to prediction [76], and it is also noted in papers aimed at human geneticists [22, 75, 85], whose initial interest in the framework focused on heritability estimation. Similarly, the animal model is known to be intimately connected to the phylogenetic methods we discuss later [38–40]. One implication is that close connections between methods used in statistical genetics and phylogenetics, which are our focus here, must exist.

The expected genetic relatedness matrix (eGRM)

We may make other assumptions to model the genetic covariances among individuals. For instance, we might let $Z = \beta$, the effect sizes themselves. Unlike in the previous subsection, the effect sizes are not fixed; they are random and independent of genotype. In this case, the covariance is computed with respect to alternative realizations of the mutational process (as in the branch-based approach in [86]) and, in this subsection, also alternative realizations of the underlying gene tree(s).

Then, if genotypes are independent of effect sizes and the expected effect size is 0, we show in the appendix that

$$\mathrm{Cov}(\beta_l G_{il}, \beta_l G_{jl}) = \mathbb{E}(\beta^2)\mathbb{E}(G_{il}G_{jl}), \tag{10}$$

so that, if all loci are equivalent

$$\mathrm{Cov}(G_i, G_j) = L\mathbb{E}(\beta^2)\mathbb{E}(G_{il}G_{jl}). \tag{11}$$

In theory, the expectation of the product of genotypes can be computed based on coalescent theory [87], focusing on the gene trees that describe the pattern of shared inheritance of alleles rather than the distribution of the alleles themselves. In practice, the gene trees underlying genetic variation in the sample cannot be observed directly. However, developing a coalescent approach provides theoretical understanding in its own right, and it also forms a basis for doing complex trait analyses using estimated genome-wide gene trees [55, 56].

Here we develop the gene-tree-based view using a different derivation from that presented by McVean [87]. For simplicity, we assume haploid genetic data, the extension to diploid data is straight-forward but tedious [87]. First, let $\mathcal{T}$ be a tree (including branch lengths) and $\mathbb{Q}$ be the measure on tree space induced by the population history. In addition, let μ be the mutation rate and T the total branch length in $\mathcal{T}$. Then, the density of a tree conditional on it having a mutation in the infinite sites model is

$$\begin{aligned} d\mathbb{Q}(\mathcal{T} \mid \text{mutation}) &= \lim_{\mu \downarrow 0} \frac{\mathbb{P}(\text{mutation} \mid \mathcal{T}) d\mathbb{Q}(\mathcal{T})}{\mathbb{P}(\text{mutation})} \\ &= \lim_{\mu \downarrow 0} \frac{\mu T e^{-\mu T} d\mathbb{Q}(\mathcal{T})}{\int_{\mathcal{T}} \mu T e^{-\mu T} d\mathbb{Q}} \\ &= \frac{T d\mathbb{Q}(\mathcal{T})}{\int_{\mathcal{T}} T d\mathbb{Q}(\mathcal{T})} \\ &= \frac{T d\mathbb{Q}(\mathcal{T})}{\mathbb{E}(T)}. \end{aligned}$$

This formula indicates that conditioning on a gene tree having a mutation results in trees with more total branch length. Hence, the distribution of gene trees that underlie SNPs is different from the unconditional distribution of gene trees; in particular, gene trees that have a SNP will tend to have more total branch length than those that do not have a SNP.

We also need that

$$\begin{aligned} \mathbb{E}(G_{il}G_{jl} \mid \mathcal{T}, \text{mutation}) &= \lim_{\mu \downarrow 0} \frac{\mathbb{P}(G_{il} = 1, G_{jl} = 1, \text{mutation} \mid \mathcal{T})}{\mathbb{P}(\text{mutation} \mid \mathcal{T})} \\ &= \lim_{\mu \downarrow 0} \frac{\mu T_{ij} e^{-\mu T_{ij}}}{\mu T e^{-\mu T}} \\ &= \frac{T_{ij}}{T}, \end{aligned}$$

where T_{ij} is the length of time in the tree shared by samples i and j. This fits the intuition that, conditional on a site being variable, individuals will share that mutation only if it occurs in a common ancestor of those

two individuals. Then,

$$\begin{aligned}
\mathbb{E}(G_{il}G_{jl} \mid \text{mutation}) &= \mathbb{E}(\mathbb{E}(G_{il}G_{jl} \mid \mathcal{T}, \text{mutation}) \mid \text{mutation}) \\
&= \mathbb{E}\left(\frac{T_{ij}}{T} \mid \text{mutation}\right) \\
&= \int_{\mathcal{T}} \frac{T_{ij}}{T} d\mathbb{Q}(\mathcal{T} \mid \text{mutation}) \\
&= \int_{\mathcal{T}} \frac{T_{ij}}{T} \frac{T d\mathbb{Q}(\mathcal{T})}{\mathbb{E}(T)} \\
&= \int_{\mathcal{T}} \frac{T_{ij}}{\mathbb{E}(T)} d\mathbb{Q}(\mathcal{T}) \\
&= \frac{\mathbb{E}(T_{ij})}{\mathbb{E}(T)}.
\end{aligned}$$

$$\mathbb{E}(G_{il}G_{jl}) = \frac{\mathbb{E}(T_{ij})}{\mathbb{E}(T)}, \tag{12}$$

We note that this formula explicitly requires conditioning on the site being variable and that we are ignoring the effects of linkage disequilibrium among sites.

The connection between the genotype-based view of the GRM and the gene-tree-based view, in which Σ is viewed as the conditional (on the gene tree(s)) expectation of the GRM over random mutations, developed by McVean [87], has been the basis for several recent methods in statistical genetics [55, 56, 88] and in phylogenetics [89]. (The expectation can also be computed for versions of the GRM that standardize genotypes by a function of the allele frequency [56, 88].) For example, Link and colleagues [55] use estimated gene trees in a region of the genome to compute the expectation of a local GRM formed from neutral variants falling on the estimated trees as a Poisson process. These matrices are then used as input to a variance-components model, which brings some advantages in mapping QTLs. Specifically, the resulting (conditional) expected genetic relatedness matrices naturally incorporate LD, providing better estimates of local genetic relatedness than could be formed from a handful of SNPs in a local region [55, 88].

The phylogenetic variance-covariance matrix

In an extreme case, we might consider only variation among long-separated species. In this case, there may be only a single tree that describes the relationships among species, and the expectation over gene trees used in the previous subsection can be dropped. Then the entries of the relatedness matrix Σ, which in the case of phylogenetic methods is referred to as the phylogenetic variance-covariance (or vcv), are given by

$$\mathrm{Cov}(G_i, G_j) = \frac{L\mathbb{E}(\beta^2)}{T} T_{ij}. \tag{13}$$

This can be recognized as the covariance under the Brownian motion model [43] commonly used to model continuous traits in phylogenetics, given a phylogenetic tree, when setting the diffusion rate σ^2 of the Brow-

nian motion process to

$$\sigma^2 = \frac{L\mathbb{E}(\beta^2)}{T}. \tag{14}$$

The formulation of equation 14 may look unfamiliar in phylogenetics, where the Brownian motion rate is typically taken to be V_G/N, where V_G is the additive genetic variance and N is the effective population size, following Lande [90] or simply $U\mathbb{E}(\beta^2)$, where U represents the total mutation rate toward causative alleles, following Lynch and Hill [91]. To reconcile our result with the existing literature, note that $\mathbb{E}(L) = UT$, so that

$$\begin{aligned}\mathbb{E}(\sigma^2) &= \frac{\mathbb{E}(L)\mathbb{E}(\beta^2)}{T} \\ &= U\mathbb{E}(\beta^2),\end{aligned}$$

as shown by Lynch and Hill [91]. Further, under a neutral model, the equilibrium additive genetic variance V_G is proportional to $NU\mathbb{E}(\beta^2)$ [92]. Thus, under neutrality,

$$\begin{aligned}\frac{V_G}{N} &\propto \frac{NU\mathbb{E}(\beta^2)}{N} \\ &= U\mathbb{E}(\beta^2),\end{aligned}$$

showing that under a neutral model, the Lande formulation is equivalent to the Lynch and Hill formulation, up to constants that depend on ploidy. Thus, we see that our equation 14 matches familiar formulations in the literature [93].

Consistent with previous arguments (e.g., [34]), this result also implies that one straightforward interpretation of the standard PGLS model is that it stratifies the regression between focal variables by an unobserved variable (or variables) that evolved primarily by drift. Hansen and colleagues have pointed out that this may not be an appropriate model for testing for adaptation [31, 32, 94], which was the primary motivation for developing many comparative methods in the first place [95]. Moreover, recently, standard PGLS has fallen into question in scenarios in which there is discordance between the gene tree and the species [96, 97]. Our formulation makes it clear that the standard PGLS formulation only applies when there is a single tree underlying all loci; if there is instead a distribution of gene trees, equation 12 suggests that the appropriate thing to do is to average over gene trees, as suggested by Hibbins *et al.* [97]. Nonetheless, as we illustrate below, the fact that the standard phylogenetic regression falls out as a special case of the same general model as standard statistical genetics models is useful in practice, even if not always in theory.

How the same type of unmodeled structure misleads both GWAS and phylogenetic regressions

The result of the previous section that standard models in statistical genetics and phylogenetics are closely related immediately suggests that these models might suffer the same pathologies under model misspecification, and that solutions to these pathologies could be shared across domains. Here we illustrate this by studying the problem of how unmodeled (phylo)genetic structure biases estimates of regression covariates.

This problem has received much attention in both the statistical genetics [98, 99] and phylogenetics literature [33, 34, 100], but the approaches taken in the two fields differ.

We assume that we have a sample of size n with a vector of a predictor, $x = (x_1, x_2, \ldots, x_n)^T$, and a trait, $y = (y_1, y_2, \ldots, y_n)^T$. In the context of genome-wide association studies, x may be the (centered) genotypes at a locus to be tested for associaiton, while in the context of phylogenetics, x is often an environmental variable or another trait that is hypothesized to influence y. Then, the regression model is

$$y_i = x_i\beta + G_i + E_i, \tag{15}$$

where G_i and E_i are the genetic and environmental components, as in equation 2, and β is the effect of x on y. In genome-wide association studies, β is the effect size of the locus being examined, while in phylogenetics it may quantify the effect of an environmental variable or other continuous trait, rather than the effect of an allele.

Theoretical analysis

To understand the purpose and limitations of corrections for (phylo)genetic structure, we examined the properties of the estimators of regression coefficients with and without correction for (phylo)genetic structure. The simplest estimator, $\hat{\beta}$, is the ordinary least squares estimator. In that case, the estimated effect size is

$$\hat{\beta}_x^{(OLS)} = \frac{x^T y}{x^T x}.$$

Because equation 15 shows that y is (phylo)genetically structured with covariance matrix Σ, we can expand $\hat{\beta}^{(OLS)}$ in terms of the eigenvectors and eigenvalues of Σ, which form an orthonormal basis of $\mathbb{R}^n$, by diagonalizing,

$$\Sigma = V\Lambda V^T$$

where $V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$ is a matrix whose columns are the eigenvectors of Σ, and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a diagonal matrix whose diagonal is the eigenvalues of Σ. This decomposition is always possible because Σ is symmetric and positive semidefinite by virtue of being a covariance matrix. Then,

$$\hat{\beta}^{(OLS)} = \frac{\sum_{j=1}^n (v_j^T x)(v_j^T y)}{\sum_{j=1}^n (v_j^T x)(v_j^T x)}. \tag{16}$$

This shows that we can conceptualize the ordinary least squares estimator as adding up the correlations between x and y projected onto each eigenvector of Σ. Loosely, large-magnitude slope estimates arise when x and y both project with large magnitude onto one or more eigenvectors of Σ. If an eigenvector of Σ is correlated with a confounding variable, such as the underlying (phylo)genetic structure, then x and y may both have substantial projections onto it, even if x and y are only spuriously associated due to the confound.

Two seemingly distinct approaches have been proposed to address this issue. First, researchers have

proposed including the eigenvectors of Σ as covariates. In the phylogenetic setting, this is known as phylogenetic eigenvector regression [101]. (In practice, researchers often use the eigenvectors of a distance matrix derived from the phylogenetic tree rather than Σ itself, but these two matrices have a straightforward mathematical connection [102]). In the statistical genetics setting, the analogous approach is to include the principal component projections of the data that are used to generate the genetic relatedness matrix—i.e., the principal components of the genotype matrix [23]—in the regression. To see that the eigenvectors of Σ are equivalent to the projections of the data onto the principal components, suppose we have $n \times L$ genotype matrix G whose rows represent individuals and whose columns represent genetic loci. Entry i, j contains the number of copies of a non-reference allele carried by individual i at locus j. Then, the projections of the data onto the kth principal component is given by

$$v_k = Gw_k$$

where w_k is the kth eigenvector of the covariance matrix of $L \times L$ covariance matrix of the genotypes, i.e. $G^T G w_k = \lambda_k w_k$. Then,

$$\begin{aligned} G^T G w_k &= \lambda_k w_k \\ \implies \underbrace{GG^T}_{\Sigma} \underbrace{Gw_k}_{v_k} &= \lambda_k \underbrace{Gw_k}_{v_k} \\ \implies \Sigma v_k &= \lambda_k v_k, \end{aligned}$$

where the second line follows from left multiplying the first line by G. For clarity, we emphasize that this is distinct from computing PCs on the phenotypic matrix (if multiple phenotypes are measured), which is intended to solve a different problem [103, 104].

Now, we construct a design matrix $X = \begin{bmatrix} x & v_1 & v_2 & \cdots & v_J \end{bmatrix}$, whose columns are the predictor x, followed by the first J eigenvectors of Σ. (In theory, any set of eigenvectors of Σ might be included in the design matrix, but in statistical-genetic practice, it is typically the leading eigenvectors.) Then, using standard theory, the vector of coefficient estimates is

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

After some linear algebra (see Appendix), the estimate of the coefficient of the predictor x is

$$\hat{\beta}_x^{(OLS,PC)} = \frac{\sum_{j=J+1}^{n} (v_j^T x)(v_j^T y)}{\sum_{j=J+1}^{n} (v_j^T x)(v_j^T x)}, \tag{17}$$

i.e., it is the OLS estimator, but the first J eigenvectors of Σ are removed. This shows why inclusion of the PCs as covariates can correct for (phylo)genetic structure: it simply eliminates some of the dimensions on which x and y may covary spuriously. However, it also shows the limitations of including principal components as covariates. First, because it is simply cutting out entire dimensions, it can result in a loss of power. Second, confounding that aligns with principal components that are not included in the design matrix is not corrected.

The second approach to including the eigenvectors of Σ as covariates is to use Σ itself to model the residual correlation structure. In phylogenetic biology, this is accomplished using phylogenetic generalized least squares (PGLS) [37, 38], whereas in statistical genetics this is accomplished using linear mixed models (LMM) [27, 105]. In both settings, it is common to add a "white noise" or "environmental noise" term, such that the residual covariance structure is $\sigma_G^2\Sigma + \sigma_E^2 I$, where σ_G^2 scales the contribution of genetics, σ_E^2 scales the contribution of environment, and I is the identity matrix. In the context of phylogenetics, the relative sizes of σ_G^2 and σ_E^2 are of interest when estimating the phylogenetic signal measurement Pagel's lambda [106, 107], whereas in statistical genetics, they are the subject of heritability estimation [108]. Then, both PGLS and LMM approaches model the data as

$$y \sim \text{Multivariate Normal}(\beta x, \sigma_G^2\Sigma + \sigma_E^2 I),$$

where σ_G^2 and σ_E^2 are typically estimated, for example by maximum likelihood [109], residual maximum likelihood [75], Haseman-Elston regression [110], or other methods [27, 105, 111]. For the theoretical analysis that follows, we assume $\sigma_G^2 = 1$ and $\sigma_E^2 = 0$. This does not restrict the applicability of our analysis, because $\sigma_G^2\Sigma + \sigma_E^2 I$ has the same eigenvectors as Σ, with corresponding eigenvalues $\sigma_G^2\lambda_i + \sigma_E^2$, where λ_i are the eigenvalues of Σ.

With these assumptions, the regression coefficient can be estimated via generalized least squares,

$$\hat{\beta}_x^{(GLS)} = (x^T\Sigma^{-1}x)^{-1}x^T\Sigma^{-1}y.$$

By diagonalizing Σ (see Appendix), the estimated regression coefficient is

$$\hat{\beta}_x^{(GLS)} = \frac{\sum_{j=1}^n \lambda_j^{-1}(v_j^T x)(v_j^T y)}{\sum_{j=1}^n \lambda_j^{-1}(v_j^T x)(v_j^T x)}. \tag{18}$$

Like the ordinary least squares estimator in equation 16, this expression includes all the eigenvectors of Σ. However, it downweights each eigenvector according to its eigenvalue. Thus, GLS downweights dimensions according to their importance in Σ, which aims to describe the structure according to which x and y may be spuriously correlated. However, unlike equation 17, it retains all dimensions. Compared with adjusting for the leading eigenvectors of Σ using OLS, the GLS approach retains some ability to detect contributions to associations that align with the leading eigenvectors. It also adjusts for Σ in its entirety, rather than just its leading eigenvectors. This means that it adjusts for even very recent (phylo)genetic structure, which will likely not be encoded by the leading eigenvectors. That said, one disadvantage of GLS is that it assumes that all eigenvectors of Σ contribute to confounding in proportion to their eigenvalues, potentially resulting in an inability to completely control for confounding if the effect of an eigenvector of Σ is not proportional to its eigenvalue, as may be the case with, for example, environmental confounding.

Where sample sizes and computational resources allow it, the state-of-the-art practice in statistical genetics is to use a linear mixed model framework while *also* including principal components as covariates [27, 105, 111]. This at first may seem surprising, because it seems to be controlling for Σ twice. However,

the analysis above suggests that including the principal components as covariates and using generalized least squares have different, and perhaps complementary, outcomes. To see how they interact, again costruct the design matrix $X = (x\, v_1\, v_2\, \ldots\, v_J)$, whose columns are the predictor x, followed by the first J eigenvectors of Σ, and use the generalized least square estimator,

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y$$

After some linear algebra (see Appendix), we again obtain the estimate of the regression coefficient of x,

$$\hat{\beta}_x^{(GLS,PC)} = \frac{\sum_{j=J+1}^{n} \lambda_j^{-1} (v_j^T x)(v_j^T y)}{\sum_{j=J+1}^{n} \lambda_j^{-1} (v_j^T x)(v_j^T x)}. \tag{19}$$

Thus, combining the principal components with Σ in generalized least squares may provide the benefits of both approaches: if there is confounding in a eigenvector of Σ that is "too large"—that is, it is out of proportion with its associated eigenvalue—then if that eigenvector is included in the design matrix, it will simply be excised from the estimator, as in equation 17. However, we still maintain the ability to control for spurious association between x and y due to the structure of Σ but not along included eigenvectors, as in equation 19. The major difficulty is in identifying the eigenvectors of Σ that might be associated with confounding effects larger than their corresponding eigenvalues would suggest.

Simulation analysis

To put the intution developed from the previous subsection into practice, we performed simulations in both phylogenetic and statistical-genetic settings. First, to explore how the different approaches outlined above correct for both (phylo)genetic structure and environmental confounding, we performed simulations inspired by Felsenstein's "worst case" scenario [34, 46]. Felsenstein's worst case supposes that there are two diverged groups of samples that are measured for two variables x, and y, which are then tested for association; the only (phylo)genetic structure is between the two groups. In the phylogenetic setting, we represent the two clades as star trees with 100 tips each, connected by internal branches, and we simulate x and y as arising from independent instances of Brownian motion along the tree (see Methods). In the statistical genetics setting, we use `msprime` [112] to simulate 100 diploid samples from each of two populations, and then simulated quantitative traits using the alpha model [22] (see Methods). In this setting, McVean [87] showed that the first principal component captures population membership; hence, we only include the first principal component to capture any residual confounding. To perform inference in the phylogenetic case, we used the package `phylolm` [109], and for the statistical-genetic case, we used a custom implementation of `GREML` [75].

We first explored the impact of deepening the divergence between the two clades, starting from no divergence and increasing to high divergence (Figure 1A,C). As expected, we see ordinary least squares fails to control for the population stratification as the divergence time becomes large, resulting in high false positive rates. However, all of the other approaches appropriately controlled for the population stratification. This is as expected: in the case of two populations, all of the (phylo)genetic stratification is due to the accumulation

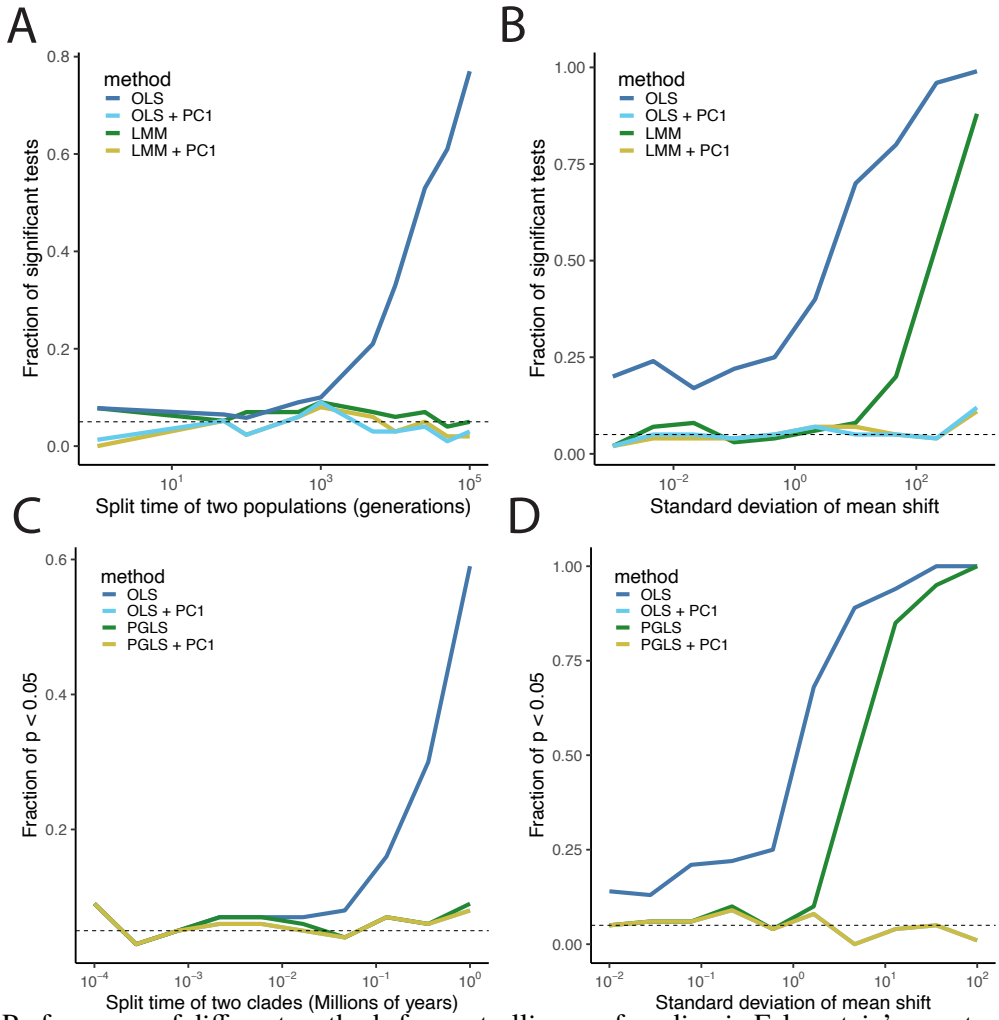

Figure 1: Performance of different methods for controlling confounding in Felsenstein's worst case. A) The false-positive rate of each method in a simulated GWAS as a function of divergence time between the two groups. The horizontal axis shows the divergence time, while the vertical axis shows the fraction of tests that would be significant at the 0.05 level. Each line represents a different method. B) The false-positive rate of each method in a simulated GWAS as a function of the size of an environmental shift. The horizontal axis shows the standard deviation of the normal distribution from which the shift was drawn, and the vertical axis shows the fraction of tests that would be significant at the 0.05 level. C) The false-positive rate of each method in a simulated phylogenetic regression as a function of divergence time between the two groups. The horizontal axis shows the divergence time, while the vertical axis shows the fraction of tests that would be significant at the 0.05 level. Each line represents a different method. The lines for OLS + PC1 and PGLS + PC1 are completely overlapping. D) The false positive rate of each method in a simulated phylogenetic regression as a function of the size of non-Brownian shifts in both predictor and response variables. The horizontal axis shows the standard deviation of the normal distribution from which the shift was drawn, and the vertical axis shows the fraction of tests that would be significant at the 0.05 level. The lines for OLS + PC1 and PGLS + PC1 are completely overlapping.

of genetic variants in each group. Hence, either discarding the correlation between x and y on the dimension corresponding to group membership as in equation 17 or downweighting it as in equation 19 is sufficient to remove the confounding effect of the population stratification.

Despite the success of both OLS with principal components and generalized least squares in controlling for population stratification, it has recently been recognized that phylogenetic generalized least squares does not control for all types of confounding in Felsenstein's worst case: for example, if there is a large shift in x and y on the branch leading to one of the groups, phylogenetic generalized least squares produces high false positive rates [34]. Because including the first principal component will completely eliminate the contribution to the estimated coefficient that projects along group membership, whereas generalized least squares will only downweight it, we reasoned that including the first principal component in either ordinary or generalized least squares should restore control even in the presence of large shifts.

We tested our hypothesis using simulations with divergence time in which ordinary least squares was not sufficient to correct for population stratification. In the phylogenetic case, we simulated an additional shift in one of the clades for both x and y by sampling from independent normal distributions, while in the statistical-genetic case, we simulated an environmental shift sampled from a normal distribution in one of the clades (Figure 1B,D). As expected, ordinary least squares is insufficient to address the confoudning, and becomes increasingly prone to false positives as the size of the shift increases. In line with our hypothesis, phylogenetic generalized least squares and linear mixed modeling also fail to control for the shift as it becomes large, while including just a single PC in each case is sufficient to regain control over false positives.

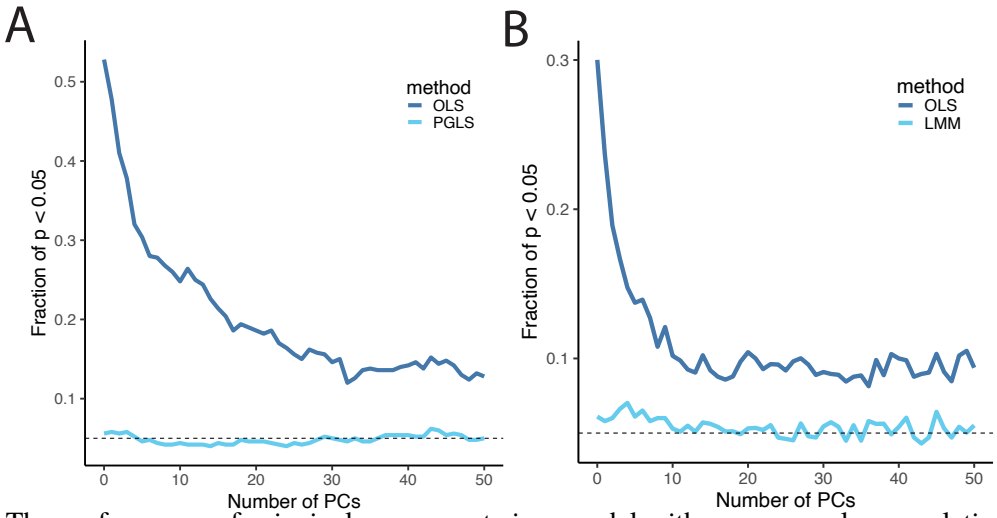

Figure 2: The performance of principal components in a model with more complex population structure. A) Performance of ordinary least squares and phylogenetic least squares in a model with 200 tips related by a pure birth tree. The horizontal axis shows the number of principal components included as covariates, and the vertical axis shows the fraction of tests that would be significant at the 0.05 level. B) Performance of ordinary least squares and a linear mixed model in a model with twenty populations related by a pure birth tree, and ten diploid individuals per population. The horizontal axis shows the number of principal components included as covariates, and the vertical axis shows the fraction of tests that would be significant at the 0.05 level.

The preceding analysis might suggest that including principal components as covariates is sufficient to adjust for (phylo)genetic structure while also being superior to generalized least squares in dealing with environmental confounding. Recent work, however, suggests that principal components may not be able to

adjust for more subtle signatures of population structure [8, 15, 99, 113]. To explore this, we simulated both phylogenetic regression and a variant association test using a more complicated model of population structure. For the phylogenetic case, we simulated pure birth trees with 200 tips, while in the statistical genetics case, we simulated pure birth trees with 20 tips and sampled 10 diploids from each tip using `msprime`. Then, as before, we simulated using a Brownian motion model in the phylogenetic case, or an additive model for the statistical genetic case.

As expected, using ordinary least squares without any principal components does not control for population structure in either the phylogenetic or the statistical-genetic setting, while the methods that use generalized least squares estimates of the regression coefficients appropriately model population structure (Figure 2). Although adding additional principal components reduces the false positive rate of ordinary least squares, it does not reach the desired false positive rate of 5%. This is in line with our theoretical analysis: as seen in equation 17, including principal components in ordinary least squares eliminates dimensions that explain the most genetic differentiation, but the correlations on the remaining dimensions are not adjusted. Because there is substantial fine-scale population structure in these simulations, removal of just a few dimensions with large eigenvalues is not sufficient to control for the subtle signature of population structure. On the other hand, generalized least-squares approaches, as shown in equation 19, will continue to correct for population structure that is found deeper into the eigenvectors of the correlation matrix (echoing points previously raised in the phylogenetics literature [114–116]). We also note that while the our analysis is focused on the eigenvectors of Σ, we suspect similar lines of reasoning may apply to other situations where eigenvector regression is used, such as in spatial ecology [117].

A case study of applying PCs to PCMs

Although the eigenvectors of the phylogenetic variance-covariance matrix (or closely related quantities) have often been included in regression models by researchers using phylogenetic eigenvector regression [101], to the best of our knowledge, phylogenetic biologists have not previously used these eigenvectors as fixed effects in a PGLS model—which we have shown above to be an effective strategy in theory. To illustrate the approach in practice, we re-examine a recent study by Cope *et al.* [118] that tested for co-evolution in mRNA expression counts across 18 fungal species. More specifically, these researchers were interested in testing whether genes whose protein products physically interacted (using independent data from [119]) were more likely to have correlated expression counts than those whose protein products did not. They found support for this prediction. While we suspect the core finding is robust, and there are some theoretical reasons to expect that RNA expression counts should be Brownian-like under some selective scenarios [120], other studies have shown expression counts for many genes in this dataset (and many others) are not well-described by a Brownian process [65, 121]. As such, some of their observed correlations could be spurious due to unmodeled phylogenetic structure [34].

We re-analyzed the data of Cope *et al.* [118] with the addition of PCs of (phylogenetic) Σ as fixed effects in the PGLS model (see Methods and Materials for details). Cope *et al.* used a correlated multivariate

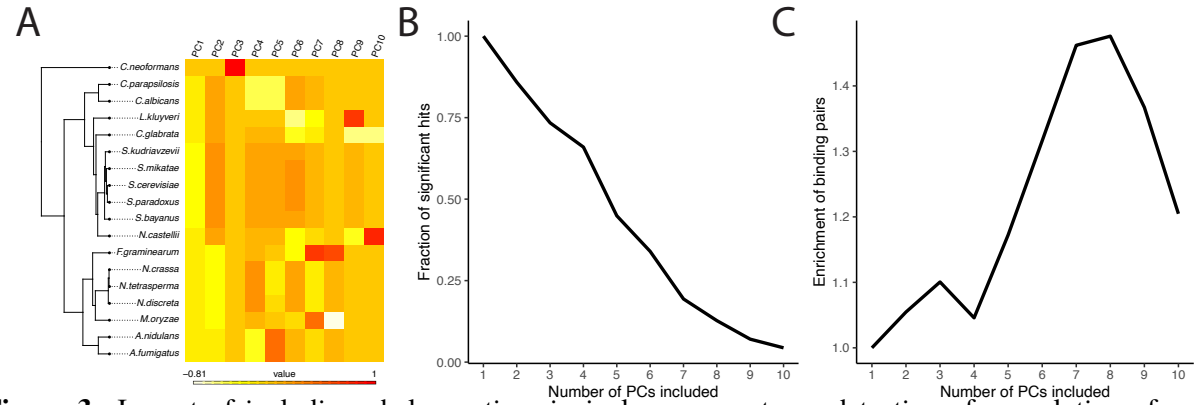

Figure 3: Impact of including phylogenetic principal components on detection of coevolution of gene-expression levels in fungi. A) The fungal tree; colors indicate each species' position in the first ten dimensions of principal component space. B) The overall number of significant pairs decreases as more principal components are included in the regression. The horizontal axis indicates the number of principal components included as fixed effects, and the vertical axis shows the proportion of significant pairs compared with a model that includes no principal components as fixed effects. C) The enrichment of known binding pairs as a function of principal components included. The horizontal axis indicates the number of principal components included as fixed effects, and the vertical axis shows the enrichment of known binding pairs relative to a model in which no principal components are included.

Brownian model to test their hypothesis, which is slightly different than the more common PGLS approach [122], but they are close enough for our purposes. We conducted several iterations of the analyses, varying the number of PCs included from 1 to 10; Figure 3A shows how the different species project onto each principal component. We found that, as anticipated, the number of significant correlations decreased as more PCs were included (Figure 3B). However, as more PCs were included, the proportion of significant correlations in gene-expression count data in which the genes are known to physically interact increased (up to about 8 PCs; Figure 3C). If we assume that the significant correlations for physically interacting genes are more likely to be true positives than those for pairs of genes not known to interact physically, then the results would suggest that including the PCs in the analysis might reduce the false positive rate while still finding many of the true positives.

Uyeda *et al.* [34] suggest that one way to mitigate the spurious correlations arising from large, unreplicated events (see above sections) would be to simply use indicator variables in the regression model that encode the part of the phylogeny from which a tip descends (i.e., if there was a concordant shift in the means of both traits along a single branch, the indicator variable would be 0 if a species was on one side of that branch and 1 if on the other). This is similar in spirit to the use of hidden Markov models for the evolution of discrete traits [100, 123]. However, as Uyeda *et al.* point out, this leaves open the hard problem of identifying the branches on which to stratify. It is not possible to include an indicator for every branch, as the model would then be overdetermined. Using the simple method borrowed from GWAS studies of including PCs of Σ as fixed effects in the typical phylogenetic regression may be a promising (partial) solution to the problem of spurious correlations.

The genetic model vs. the statistical model

We began by adding assumptions to equation 2 in order to show that common practices in disparate areas of genetics can be seen as special cases of the same model. One notable assumption is that of a purely additive model [124] for the phenotype (equation 1). There are two reasons we might be suspicious of this assumption. First, it is debatable to what extent most traits obey the additive model, given evidence of non-additive genetic contributions to traits across species [125, 126]. However, even if non-additive contributions are important for determining individual phenotypes or for understanding traits' biology, they might still contribute a relatively small fraction of trait variance, meaning they might be safely ignored for some purposes [127–129] (but see [130]). Second, we used a neutral coalescent model to find an expression for the Brownian motion diffusion parameter in terms of the effect sizes of individual loci (equation 14). Although this provides a satisfying justification for the use of a phylogenetic regression model with a Brownian covariance structure and for averaging over gene trees to accommodate ILS (*sensu* [97]), it is likely unreasonable in many situations. It has long been appreciated that, while a population-mean phenotype will be expected to evolve according to a Brownian process under simple quantitative-genetic models of genetic drift [41, 90, 93, 131] the Brownian rate estimated from phylogenetic comparative data is orders of magnitude too slow to be consistent with plausible values for the quantitative-genetic parameters used to derive the Brownian model [93, 131–133]. There are more elaborate explanations than pure genetic drift for why long-term evolution may show relatively simple dynamics [134] but understanding the coalescent patterns of loci under these scenarios is likely challenging [135] and beyond the scope of the present paper.

However, even if one finds the genetic model unreasonable, the equivalence of the *statistical* models used in statistical genetics and phylogenetics still holds: that is, the core structures of the models are the same, whether or not one is willing to interpret the parameters in the same way or not. Indeed, phylogenetic biologists have been here before, with the realization that PGLMMs are structurally equivalent to the pedigree-based analyses using the animal model from quantitative genetics [38–40] even though the recognition that they were equivalent did not rely on a specific genetic model for phenotypes (above, we prove that they can both be derived from the same genetic model). Nonetheless, the recognition of a structural equivalency between the animal model and the phylogenetic model made it possible to use techniques from quantitative genetics to solve hard problems in phylogenetic comparative methods. For example, inspired by a similar model from [136], Felsenstein developed a phylogenetic threshold model [137, 138], in which discrete phenotypes are determined by a continuous liability that itself evolves according to a Brownian process. Hadfield [139] proved this model was identical to a variant of the animal model and that existing MCMC algorithms could be used to efficiently estimate parameters and extend the threshold to the multivariate case, which had not been previously derived.

Incorporating natural selection

All of the connections highlighted in this manuscript depend in some way on the assumption that causal variants are neutral. However, in many species, including humans, there is direct evidence of selection act-

ing on genetic and heritable phenotypic variation [140–147]. In both phylogenetic and statistical genetic traditions, there are attempts to capture the qualitative features of natural selection when performing regression analyses. In phylogenetics, selection is often modeled by the use of the Ornstein-Uhlenbeck process [44, 148–151], which can be derived from a quantitative genetic model of stabilizing selection [90]. Under the Ornstein-Uhlenbeck model, phenotypes on a phylogeny will be drawn toward a common optimum value; hence, covariances in Σ are higher than expected under a Brownian motion model. Modifying PGLS to include an Ornstein-Uhlenbeck model of the genetic covariance is straightforward, because the data remains multivariate Gaussian [37, 109]. Assuming a single optimum over an entire phylogenetic tree may be unrealistic. To mitigate that issue, several approaches for detecting optimum shifts have been developed [152, 153].

One the other hand, in statistical genetics, natural selection is sometimes incorporated via the *alpha*-model, in which the effect size of an allele is proportional to $(2p(1-p))^{-\alpha}$ [75, 154, 155]. This qualitatively captures the idea that rare alleles will have larger effect sizes than common alleles; however, the realism and accuracy of the alpha model has been questioned [74, 155–157]. A promising path forward in statistical genetics is for the development of models that explicitly include the action of natural selection.

Towards a more integrative study of the genetic bases of phenotypes

Building up this general framework is a step towards inference methods that coherently integrate intra- and interspecific variation to understand the genotype-to-phenotype map and how evolutionary processes, acting at different time scales, shape it. We can envision a number of paths forward for both statistical genetics and phylogenetics that can build on the work we present here.

Nearly two decades of GWAS in humans have revealed thousands of genetic variants associated with complex traits and disease [158–160]. Analyzing the identified variants and their estimated effect sizes shows two key trends: effect sizes tend to be larger for rarer alleles [154, 155, 161–164], and effect sizes tend to be larger in more evolutionarily conserved regions [50, 165–169]. A likely explanation for these patterns is that many traits in humans are evolving subject to stabilizing selection [156, 170–172], which in turn impacts their genetic architecture by determining the distribution of effect sizes as a function of allele frequency [154–156]. Though the importance of evolutionary conservation in triaging functional variants in the human genome has long been appreciated, it is becoming increasingly important as we collect larger and larger samples of human variants, the vast majority of which are extremely rare [173, 174]. In fact, recent work has shown that evolutionary conservation accounts for the vast majority of the predictive power of a state-of-the-art deep learning approach to variant annotation [169, 175]. A limitation of current approaches for utilizing evolutionary conservation is that typically there is no way to include information about the phylogenetic structure of the data (i.e., only multiple sequence alignments are used). Overcoming this is not straightforward and will require mechanistic modeling: the observed level of conservation is a nonlinear function of the strength of selection acting against variants at a locus; small changes in the strength of negative selection can greatly decrease the amount of variability seen on phylogenetic timescales, and this can cause

counter-intuitive behavior of conservation scores [54, 176]. We focus here on applications to human genetics, but the same challenges exist whatever the focal organism is. For example, a recent study identified a variant of large effect for body size in canids by aggregating data from multiple species (both extinct and extant) [177]; while an outstanding discovery, we anticipate that variants with smaller effects might also be identified using similar data with a comprehensive approach for including phylogenetic structure. As such, the paucity of interpretable evolutionary methodology that combines trait-related information within and across species is a barrier to understanding the forces that shape trait-related genetic variation. There is a rich tradition of phylogenetic models that have been developed to depict various evolutionary scenarios [42, 122] that could potentially be leveraged in order to make such inferences. One of our goals in laying out the underlying connections among these modeling traditions is to facilitate the transfer of these models across disciplinary divides.

In phylogenetics, the PhyloG2P research program [57] also encompasses a wide spectrum of processes at different levels of divergence. The so-called PhyloGWAS method [178] is used to identify causal variants for a phenotype that segregated in ancestral populations but have fixed in descendant lineages. In other cases, researchers have used phylogenetic approaches to investigate convergent mutations associated with phenotypic convergence across lineages (e.g., [179]). And then there are methods, such as RERConverge [174], and PhyloACC [180], that identify regions with a relatively high number of substitutions — but not necessarily the same ones — in phylogenetically distinct lineages that have convergently evolved the same phenotype, presumably due to positive selection or relaxed negative selection [57]. For example, Sackton *et al.* [181] used such an approach to identify regulatory regions that had high rates of evolution in lineages of flightless birds; they also demonstrated that some of these regions influence wing development using experimental perturbations. Such rate association tests seem to be very similar, both conceptually and likely statistically, to techniques used in rare-variant association studies, which look for local enrichment of rare variants in cases vs. controls, rather than associating single variants with phenotype [52, 182–184].

More generally, integration over gene trees may be necessary for a wide range of phylogenetic methods. In a series of papers, Hahn and colleagues [60, 96, 97, 185, 186] have pointed out that conducting comparative tests using the species tree can be misleading: if the genes underlying a phenotype of interest have different topologies than that of a species tree—due to incomplete lineage sorting (ILS) and other processes [187]—researchers may mistake similarity due to common ancestry for convergence. This phenomenon, known as hemiplasy [188] may be relatively common [185, 186], motivating the development of ILS-aware phylogenetic regresssion [97] and PhyloG2P methods [189].

There are clear biological rationales explaining why various types of analyses will be more or less informative at different timescales. But this is a difference of degree and not of kind. And the different methodological traditions in statistical genetics and phylogenetics are just that — traditions. There is no reason a researcher should think about the problem of trait mapping in a fundamentally distinct way just because she happened to be trained in a statistical genetics or phylogenetics lab. Ultimately, we should work to take the best ideas from both of these domains and blend them into a more cohesive paradigm that will facilitate richer insights into the molecular basis of phenotypes.

Materials and Methods

Simulation details

To perform phylogenetic simulations, we used the `fastBM` function from the `phytools` `R` package [190]. In all cases, Brownian motions were simulated independently and with rate 1. When performing phylogenetic simulations of Felsenstein's Worst Case, we used `stree` from `ape` [191] to simulate two star trees of 100 tips, where each tip in the star tree had length 0.5. We then connected the two star trees using internal branches of varying length. To add a non-Brownian confounder, in each simulation we added an independent normal random variable with varying standard deviations to the x and y values for individuals from clade 1. (Within a given simulation, all individuals in clade 1 were augmented by the *same* value for each trait, while between simulations, the confounding effect was a random draw.) When performing simulations in a more complicated phylogeny, we used `TreeSim` [192] to generate pure-birth trees with birth rate = 1 and complete taxon sampling. Each simulation replicate used a different tree. For ordinary least squares on phylogenetic data, we used the `R` function `lm`. For PGLS on phylogenetic data, we used the `R` package `phylolm` [109] with the Brownian motion model and no environmental noise.

To perform genome-wide association study simulations, we first generated neutral tree sequences and mutations using `msprime` [112]. To ensure our results were not simply due to genetic linkage, we simulated a high recombination of 10^{-5} per generation with a mutation rate an order of magnitude lower, 10^{-6} per generation. We first simulated causal variants on a sequence of length 100000, and generated phenotypes by sampling an effect size for each variant from a normal distribution with mean 0 and variance $\frac{0.01}{2p_l(1-p_l)}$ where p_l is the allele frequency of variant l. We then created each individual's phenotype using the additive model, equation equation 1. We then added environmental noise so that the trait's heritability was less than 1. In all simulations, every population had diploid population size 10000. To simulate the variant being tested for association, we simulated independent tree sequences and mutations and selected a random variant with allele frequency greater than 0.1. When simulating a GWAS analogue of Felsenstein's Worst Case, we drew 100 diploid samples from each population, and varied the divergence time of the two populations. To include an environmental shift in one population, we added a normal random variable with varying standard deviation only to individuals in population 1. To simulate under a more complicated population structure, we simulated 20-tip pure birth trees using `TreeSim` with a birth rate of 5. We then multiplied all branch lengths by 10000 to convert them into generations, and imported them into `msprime` using the `from_species_tree` function. We then generated tree sequences and mutations, sampling 10 diploid individuals from each population. Note that each replicate simulation was performed on an independent random population tree. We performed association testing using a custom `python` implementation of the linear mixed model. We first used restricted maximum likelihood to estimate σ_G^2 and σ_E^2, followed by using generalized least squares to estimate the regression coefficients and their standard errors.

Phylogenetic analysis of yeast gene expression data

We obtained the species tree, gene expression matrix, and list of physically interacting genes from `https://github.com/acope3/GeneExpression_coevolution` [118]. We then randomly subsampled 500 genes that had measurements in at least 15 of the 20 species to test for association, resulting in 124750 pairs. Because of differential missingness among genes, we computed phylogenetic PC loadings only on the subtree for which both genes had data present, meaning that each pair may have had slightly different PC loadings. We then used `phylolm` [109] with no measurement error to estimate the regression coefficient. For each number of principal components included, we corrected for multiple testing by controlling the FDR at 0.05 using the Benjamini-Hochberg procedure [193].

Data availability

Code used to generate the results in this paper can be found at `https://github.com/Schraiber/PGLS_GWAS`

Acknowledgements

We thank Matt Hahn, Nick Mancuso, and members of the Pennell, Edge, and Mooney labs for their thoughtful comments on parts of this study. Alex Cope provided additional guidance on our analysis of the yeast gene expression data.

Appendix

Covariance of the genetic component of the phenotype with pedigree-based kinship

With a known pedigree but no genotype data, we might derive the covariances among individuals in the genetic component of a trait by fixing the (unobserved) effect sizes and allele frequencies. For simplicity, we will also assume that individuals are not inbred. Given the pedigree, at any given locus, diploid individuals i and j have some probability of inheriting 0, 1, or 2 alleles identical by descent (IBD). Call these probabilities r_0, r_1, and r_2, respectively. For example, for a parent-offspirng pair, $r_1 = 1$, and for a pair of full siblings, $r_0 = 1/4$, $r_1 = 1/2$, and $r_2 = 1/4$.

Following equation 4, the covariance of the genetic component of the trait for individuals i and j is

$$\mathrm{Cov}(\sum \beta_l G_{il}, \sum \beta_l G_{jl}) = \sum_l \mathrm{Cov}(\beta_l G_{il}, \beta_l G_{jl}) + \sum_{l \neq k} \mathrm{Cov}(\beta_l G_{il}, \beta_k G_{jk})$$

As in the main text, we assume that genotypes at distinct loci are independent, causing the second term to vanish. With the effect sizes treated as fixed constants, we have

$$\mathrm{Cov}\left(\sum \beta_l G_{il}, \sum \beta_l G_{jl}\right) = \sum_l \beta_l^2 \,\mathrm{Cov}(G_{il}, G_{jl}) = \sum_l \beta_l^2 (\mathbb{E}(G_{il}G_{jl}) - \mathbb{E}(G_{il})\mathbb{E}(G_{jl})).$$

The G values are allelic counts of a non-reference allele (we assume the locus is biallelic), and if we fix the non-reference allele frequencies p_l, then $\mathbb{E}(G_{il}) = \mathbb{E}(G_{jl}) = 2p_l$, giving

$$\mathrm{Cov}\left(\sum \beta_l G_{il}, \sum \beta_l G_{jl}\right) = \sum_l \beta_l^2 (\mathbb{E}(G_{il}G_{jl}) - 4p_l^2). \tag{20}$$

To proceed, we need $\mathbb{E}(G_{il}G_{jl})$ given the IBD probabilities r_0, r_1, and r_2. We thus compute the desired expectation conditional on each IBD state. $G_{il}G_{jl} = 1$ if individuals are heterozygous, $G_{il}G_{jl} = 2$ if one individual is heterozygous and the other is homozygous for the non-reference allele, and $G_{il}Gjl = 4$ if both individuals are homozygous for the non-reference allele. If individuals i and j share no alleles IBD, using Hardy–Weinberg genotype probabilities (per the assumption of no inbreeding), the conditional expectation is therefore

$$\begin{aligned}
\mathbb{E}(G_{il}G_{jl} \mid IBD = 0) &= 1 \times (2p_l(1-p_l))^2 + 2 \times 4p_l(1-p_l)p_l^2 + 4 \times p^4 \\
&= 4p_l^2(1-p_l)^2 + 8p_l^3(1-p_l) + 4p_l^4 \\
&= 4p_l^2((1-p_l)^2 + 2p_l(1-p_l) + p_l^2) \\
&= 4p_l^2.
\end{aligned}$$

A similar calculation given one allele inherited IBD gives

$$\begin{aligned}
\mathbb{E}(G_{il}G_{jl}|IBD = 1) &= 1 \times (p_l(1-p_l)^2 + p_l^2(1-p_l)) + 2 \times 2p_l^2(1-p_l) + 4 \times p_l^3 \\
&= p_l + 3p_l^2.
\end{aligned}$$

(To explain the first line: for the first term, the individuals are both heterozygous if the IBD allele is of the non-reference type and the other two are reference alleles, or if the IBD allele is non-reference and both of the non-IBD alleles are reference alleles. For the second term, since one individual is homozygous for the non-reference allele, the IBD allele must be of the non-reference allele, and of the other two alleles, one must be of each type. For the third term, both individuals are homozygous if the IBD allele and both non-IBD alleles are non-reference alleles.)

If the individuals share two alleles IBD, then

$$\begin{aligned}\mathbb{E}(G_{il}G_{jl}|IBD=2) &= 1\times 2p_l(1-p_l)+4\times p_l^2\\ &= 2p_l+2p_l^2.\end{aligned}$$

Combining these terms weighted by the IBD probabilities gives the desired expectation,

$$\begin{aligned}\mathbb{E}(G_{il}G_{jl}) &= r_0\mathbb{E}(G_{il}G_{jl}|IBD=0)+r_1\mathbb{E}(G_{il}G_{jl}|IBD=1)+r_2\mathbb{E}(G_{il}G_{jl}|IBD=2)\\ &= r_0 4p_l^2+r_1(p_l+3p_l^2)+r_2(2p_l+2p_l^2)\\ &= p_l(r_1+2r_2)+p_l^2(4r_0+3r_1+2r_2)\\ &= 4p_l\theta_{ij}+p_l^2(4-4\theta_{ij}),\end{aligned} \tag{21}$$

where the last line comes from noticing that $r_0+r_1+r_2=1$ and defining the kinship coefficient $\theta_{ij}=r_1/4+r_2/2$, equal to the probability that a pair of alleles chosen at random, one from individual i and one from individual j, is IBD.

We can now return to the main covariance of interest by plugging the expression for $\mathbb{E}(G_{il}G_{jl})$ from equation 21 into equation 20, giving

$$\begin{aligned}\mathrm{Cov}\left(\sum\beta_l G_{il},\sum\beta_l G_{jl}\right) &= \sum_l \beta_l^2(\mathbb{E}(G_{il}G_{jl})-4p_l^2)\\ &= \sum_l \beta_l^2(4p_l\theta_{ij}+p_l^2(4-4\theta_{ij})-4p_l^2)\\ &= \sum_l \beta_l^2(4p_l\theta_{ij}-4p_l^2\theta_{ij})\\ &= 2\theta_{ij}\sum_l 2\beta_l^2 p_l(1-p_l)\\ &= 2\theta_{ij}V_A.\end{aligned}$$

The final line comes from noting that the sum in the previous line is the additive genetic variance (V_A) under the assumptions here, that is, the variance of the genetic component of the phenotype among outbred individuals. This is the result required in the main text.

Covariance of genotypes when effect sizes and allele frequencies are independent

We have

$$\begin{aligned}
&\mathbb{E}_\beta(\beta_l^2\mathbb{E}(G_{il}G_{jl} \mid \beta_l)) - \mathbb{E}_\beta(\beta_l\mathbb{E}(G_{il} \mid \beta))\mathbb{E}_\beta(\beta_l\mathbb{E}(G_{jl} \mid \beta)) \\
&\quad = \mathbb{E}(G_{il}G_{jl})\mathbb{E}(\beta_l^2) - \mathbb{E}(G_{il})\mathbb{E}(G_{jl})\mathbb{E}(\beta_l)^2 \\
&\quad = \mathbb{E}(\beta_l^2)\operatorname{Cov}(G_{il}, G_{jl}) + \operatorname{Var}(\beta_l^2)\mathbb{E}(G_{il})\mathbb{E}(G_{jl}),
\end{aligned}$$

where the second line comes from the assumption that genotypes and effect sizes are independent, and the third line comes from adding and subtracting $\mathbb{E}(G_{il})\mathbb{E}(G_{jl})\mathbb{E}(\beta_l^2)$ to the second line and simplifying. Further making the assumption that $\mathbb{E}(\beta)$ = 0,

$$\begin{aligned}
\mathbb{E}(\beta_l^2)\operatorname{Cov}(G_{il}, G_{jl}) + \operatorname{Var}(\beta_l^2)\mathbb{E}(G_{il})\mathbb{E}(G_{jl}) &= \mathbb{E}(\beta_l^2)\operatorname{Cov}(G_{il}, G_{jl}) + \mathbb{E}(\beta_l^2)\mathbb{E}(G_{il})\mathbb{E}(G_{jl}) \\
&= \mathbb{E}(\beta^2)\left(\operatorname{Cov}(G_{il}, G_{jl}) + \mathbb{E}(G_{il})\mathbb{E}(G_{jl})\right) \\
&= \mathbb{E}(\beta^2)\mathbb{E}(G_i, G_j).
\end{aligned}$$

The estimated regression coefficient from ordinary least squares with principal components

Recall the design matrix $X = \begin{bmatrix} x & v_1 & v_2 & \cdots & v_J \end{bmatrix}$, where the v_i are orthonomal eigenvectors of Σ. Then,

$$X^TX = \begin{bmatrix} x^Tx & v_1^Tx & v_2^Tx & \cdots & v_J^Tx \\ v_1^Tx & 1 & 0 & \cdots & 0 \\ v_2^Tx & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_J^Tx & 0 & 0 & \cdots & 1 \end{bmatrix},$$

i.e. the first row and columns are the projection of x along each eigenvector and the rest of the matrix is an identity matrix. The identity matrix arises because the eigenvectors annihilate each other. Then,

$$(X^TX)^{-1} = \frac{1}{x^Tx - \sum_{j=1}^J (v_j^Tx)^2}\begin{bmatrix} 1 & -v_1^Tx & -v_2^Tx & \cdots & -v_J^Tx \\ -v_1^Tx & x^Tx - \sum_{j\neq 1}^J (v_j^Tx)^2 & -(v_1^Tx)(v_2^Tx) & \cdots & -(v_1^Tx)(v_J^Tx) \\ -v_2^Tx & -(v_2^Tx)(v_1^Tx) & x^Tx - \sum_{j\neq 2}^J (v_j^Tx)^2 & \cdots & -(v_2^Tx)(v_J^Tx) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -v_J^tx & -(v_J^Tx)(v_1^Tx) & -(v_J^Tx)(v_2^Tx) & \cdots & x^Tx - \sum_{j\neq J}^J (v_j^Tx)^2 \end{bmatrix}.$$

Ultimately, we only need the first row of $(X^T X)^{-1}$ to estimate the regression coefficient for x. Note that by expanding $x^T x$ in terms of the eigenvectors, we have that

$$\begin{aligned} x^T x - \sum_{j=1}^{J} (v_j^T x)^2 &= \sum_{j=1}^{n} (v_j^T x)(v_j^T x) - \sum_{j=1}^{J} (v_j^T x)^2 \\ &= \sum_{j=J+1}^{n} (v_j^T x)(v_j^T x), \end{aligned}$$

so that the first row of $(X^T X)^{-1}$ is

$$(X^T X)_1^{-1} = \frac{1}{\sum_{j=J+1}^{n} (v_j^T x)(v_j^T x)} \begin{bmatrix} 1 & -v_1^T x & -v_2^T x & \cdots & -v_J^T x \end{bmatrix}$$

We also compute

$$X^T y = \begin{bmatrix} x^T y \\ v_1^T y \\ v_2^T y \\ \vdots \\ v_J^T y \end{bmatrix}$$

so that

$$\begin{aligned} (X^T X)_1^{-1} X^T y &= \frac{x^T y - \sum_{j=1}^{J} (v_j^T x)(v_j^T y)}{\sum_{j=J+1}^{n} (v_j^T x)(v_j^T x)} \\ &= \frac{\sum_{j=1}^{n} (v_j^T x)(v_j^T y) - \sum_{j=1}^{J} (v_j^T x)(v_j^T y)}{\sum_{j=J+1}^{n} (v_j^T x)(v_j^T x)} \\ &= \frac{\sum_{j=J+1}^{J} (v_j^T x)(v_j^T y)}{\sum_{j=J+1}^{n} (v_j^T x)(v_j^T x)} \end{aligned}$$

as desired. Note that when $J = 0$, this recovers the ordinary least squares estimator (equation 16), as expected.

The estimated regression coefficient from generalized least squares

Recall that we can diagonalize the covariance matrix Σ as

$$\Sigma = V \Lambda V^T,$$

where $V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$ is a matrix whose columns are the eigenvectors of Σ and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a diagonal matrix whose entries are the eigenvalues of Σ. If we have design matrix $X = \begin{bmatrix} x & v_1 & v_2 & \cdots & v_J \end{bmatrix}$,

we can rewrite the genrealized least square estimator as

$$\begin{aligned}(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y &= (X^T(V\Lambda V^T)^{-1}X)^{-1}X^T(V\Lambda V^T)^{-1}y \\ &= (X^TV\Lambda^{-1}V^TX)^{-1}X^TV\Lambda^{-1}V^Ty\end{aligned}$$

where $\Lambda^{-1} = \text{diag}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_n}\right)$, using the fact that V is an orthonormal basis. Next, compute

$$X^TV = \begin{bmatrix} v_1^Tx & v_2^Tx & \cdots & v_n^Tx \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \end{bmatrix}$$

which is an $(J+1) \times n$ matrix, so the pattern extends $J+1$ rows and the identity elements come from the eigenvectors annihilating each other. Then,

$$X^TV\Lambda^{-1} = \begin{bmatrix} \frac{1}{\lambda_1}v_1^Tx & \frac{1}{\lambda_2}v_2^Tx & \cdots & \frac{1}{\lambda_n}v_n^Tx \\ \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \end{bmatrix}$$

Note that $V^TX = (X^TV)^T$, so that

$$X^TV\Lambda^{-1}V^TX = \begin{bmatrix} \sum_{j=1}^n \frac{1}{\lambda_i}(v_j^Tx)^2 & \frac{1}{\lambda_1}v_1^Tx & \frac{1}{\lambda_2}v_2^Tx & \cdots & \frac{1}{\lambda_J}v_J^Tx \\ \frac{1}{\lambda_1}v_1^Tx & \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ \frac{1}{\lambda_2}v_2^Tx & 0 & \frac{1}{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\lambda_J}v_J^Tx & 0 & 0 & \cdots & \frac{1}{\lambda_J} \end{bmatrix}.$$

As in the case for ordinary least squares, we will only need the first row of the inverse,

$$(X^TV\Lambda^{-1}V^TX)_1^{-1} = \frac{1}{\sum_{J+1}^n \frac{1}{\lambda_j}(v_j^Tx)^2}\begin{bmatrix} 1 & -v_1^Tx & -v_2^Tx & \cdots & -v_J^Tx \end{bmatrix}.$$

Next, compute

$$V^Ty = \begin{bmatrix} v_1^Ty \\ v_2^Ty \\ \vdots \\ v_n^Ty \end{bmatrix}$$

so that

$$X^T V \Lambda^{-1} V^T y = \begin{bmatrix} \sum_{j=1}^{n} \frac{1}{\lambda_i}(v_j^T x)(v_j^T y) \\ \frac{1}{\lambda_1} v_1^T y \\ \frac{1}{\lambda_2} v_2^T y \\ \vdots \\ \frac{1}{\lambda_J} v_J^T y \end{bmatrix}$$

and finally

$$\begin{aligned} (X^T V \Lambda^{-1} V^T X)_1^{-1} X^T V \Lambda^{-1} V^T y &= \frac{\sum_{j=1}^{n} \frac{1}{\lambda_i}(v_j^T x)(v_j^T y) - \sum_{j=1}^{J} \frac{1}{\lambda_j}(v_j^T x)(v_j^T y)}{\sum_{j=J+1}^{n} \frac{1}{\lambda_j}(v_j^T x)(v_J^T x)} \\ &= \frac{\sum_{j=J+1}^{n} \frac{1}{\lambda_i}(v_j^T x)(v_j^T y)}{\sum_{j=J+1}^{n} \frac{1}{\lambda_j}(v_j^T x)(v_J^T x)} \end{aligned}$$

as desired. Setting $J = 0$ recovers the GLS estimator without any PCs included as covariates, as in equation 18.

References

[1] Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994). URL `https://www.science.org/doi/abs/10.1126/science.8091226`. `https://www.science.org/doi/pdf/10.1126/science.8091226`.

[2] Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *The Lancet* **361**, 598–604 (2003). URL `https://doi.org/10.1016/S0140-6736(03)12520-2`. Publisher: Elsevier.

[3] Rosenberg, N. A. & Nordborg, M. A General Population-Genetic Model for the Production by Population Structure of Spurious Genotype–Phenotype Associations in Discrete, Admixed or Spatially Distributed Populations. *Genetics* **173**, 1665–1678 (2006). URL `https://doi.org/10.1534/genetics.105.055335`.

[4] Edge, M. D., Gorroochurn, P. & Rosenberg, N. A. Windfalls and pitfalls: Applications of population genetics to the search for disease genes. *Evolution, Medicine, and Public Health* **2013**, 254–272 (2013). URL `https://doi.org/10.1093/emph/eot021`.

[5] Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019). URL `https://www.science.org/doi/abs/10.1126/science.aax3710`. `https://www.science.org/doi/pdf/10.1126/science.aax3710`.

[6] Carl Veller & Graham Coop. Interpreting population and family-based genome-wide association stud-

ies in the presence of confounding. *bioRxiv* 2023.02.26.530052 (2023). URL http://biorxiv.org/content/early/2023/02/27/2023.02.26.530052.abstract.

[7] Vilhjálmsson, B. J. & Nordborg, M. The nature of confounding in genome-wide association studies. *Nature Reviews Genetics* **14**, 1–2 (2013). URL https://doi.org/10.1038/nrg3382.

[8] Gaurav Bhatia *et al.* Subtle stratification confounds estimates of heritability from rare variants. *bioRxiv* 048181 (2016). URL http://biorxiv.org/content/early/2016/04/12/048181.abstract.

[9] Young, A. I. Solving the missing heritability problem. *PLOS Genetics* **15**, e1008222 (2019). URL https://doi.org/10.1371/journal.pgen.1008222. Publisher: Public Library of Science.

[10] Gianola, D. Assortative mating and the genetic correlation. *Theoretical and Applied Genetics* **62**, 225–231 (1982). URL https://doi.org/10.1007/BF00276244.

[11] Border, R. *et al.* Cross-trait assortative mating is widespread and inflates genetic correlation estimates. *Science* **378**, 754–761 (2022). URL https://doi.org/10.1126/science.abo2059. Publisher: American Association for the Advancement of Science.

[12] Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019). URL https://doi.org/10.7554/eLife.39725. Publisher: eLife Sciences Publications, Ltd.

[13] Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**, e39702 (2019). URL https://doi.org/10.7554/eLife.39702. Publisher: eLife Sciences Publications, Ltd.

[14] Barton, N., Hermisson, J. & Nordborg, M. Why structure matters. *eLife* **8**, e45380 (2019). URL https://doi.org/10.7554/eLife.45380. Publisher: eLife Sciences Publications, Ltd.

[15] Zaidi, A. A. & Mathieson, I. Demographic history mediates the effect of stratification on polygenic scores. *eLife* **9**, e61548 (2020). URL https://doi.org/10.7554/eLife.61548.

[16] Jennifer Blanc & Jeremy J. Berg. Testing for differences in polygenic scores in the presence of confounding. *bioRxiv* 2023.03.12.532301 (2023). URL http://biorxiv.org/content/early/2023/08/22/2023.03.12.532301.abstract.

[17] Devlin, B. & Roeder, K. Genomic Control for Association Studies. *Biometrics* **55**, 997–1004 (1999). URL https://doi.org/10.1111/j.0006-341X.1999.00997.x. Publisher: John Wiley & Sons, Ltd.

[18] Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association Mapping in Structured Populations. *The American Journal of Human Genetics* **67**, 170–181 (2000). URL https://www.sciencedirect.com/science/article/pii/S0002929707624422.

[19] Reich, D. E. & Goldstein, D. B. Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology* **20**, 4–16 (2001). URL https://doi.org/10.1002/1098-2272(200101)20:1<4::AID-GEPI2>3.0.CO;2-T. Publisher: John Wiley & Sons, Ltd.

[20] Epstein, M. P., Allen, A. S. & Satten, G. A. A Simple and Improved Correction for Population Stratification in Case-Control Studies. *The American Journal of Human Genetics* **80**, 921–930 (2007). URL https://doi.org/10.1086/516842. Publisher: Elsevier.

[21] Gorroochurn, P., Hodge, S. E., Heiman, G. A. & Greenberg, D. A. A Unified Approach for Quantifying, Testing and Correcting Population Stratification in Case-Control Association Studies. *Human Heredity* **64**, 149–159 (2007). URL https://doi.org/10.1159/000102988.

[22] Speed, D. & Balding, D. J. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* **16**, 33–44 (2015). URL https://doi.org/10.1038/nrg3821.

[23] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909 (2006). URL https://doi.org/10.1038/ng1847.

[24] Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208 (2006). URL https://doi.org/10.1038/ng1702.

[25] Kang, H. M. *et al.* Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**, 1709–1723 (2008). URL https://doi.org/10.1534/genetics.107.080101.

[26] Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824 (2012). URL https://doi.org/10.1038/ng.2310.

[27] Loh, P.-R. *et al.* Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics* **47**, 284–290 (2015).

[28] Westoby, M., Leishman, M. R. & Lord, J. M. On misinterpreting thephylogenetic correction'. *Journal of Ecology* **83**, 531–534 (1995).

[29] Westoby, M., Leishman, M. & Lord, J. Further remarks on phylogenetic correction. *Journal of Ecology* **83**, 727–729 (1995).

[30] Harvey, P. H., Read, A. F. & Nee, S. Why ecologists need to be phylogenetically challenged. *Journal of Ecology* **83**, 535–536 (1995).

[31] Hansen, T. F. & Orzack, S. H. Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: the need for controlled comparisons. *Evolution* **59**, 2063–2072 (2005).

[32] Hansen, T. F. & Bartoszek, K. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Systematic biology* **61**, 413–425 (2012).

[33] Maddison, W. P. & FitzJohn, R. G. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Systematic biology* **64**, 127–136 (2015).

[34] Uyeda, J. C., Zenil-Ferguson, R. & Pennell, M. W. Rethinking phylogenetic comparative methods. *Systematic Biology* **67**, 1091–1109 (2018).

[35] Read, A. F. & Nee, S. Inference from binary comparative data. *Journal of Theoretical Biology* **173**, 99–108 (1995).

[36] Grafen, A. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **326**, 119–157 (1989).

[37] Martins, E. P. & Hansen, T. F. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist* **149**, 646–667 (1997).

[38] Lynch, M. Methods for the analysis of comparative data in evolutionary biology. *Evolution* **45**, 1065–1080 (1991).

[39] Housworth, E. A., Martins, E. P. & Lynch, M. The phylogenetic mixed model. *The American Naturalist* **163**, 84–96 (2004).

[40] Hadfield, J. & Nakagawa, S. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of evolutionary biology* **23**, 494–508 (2010).

[41] Felsenstein, J. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* **19**, 445–471 (1988).

[42] Pennell, M. W. & Harmon, L. J. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences* **1289**, 90–105 (2013).

[43] Felsenstein, J. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American journal of human genetics* **25**, 471 (1973).

[44] Hansen, T. F. Stabilizing selection and the comparative analysis of adaptation. *Evolution* **51**, 1341–1351 (1997).

[45] Butler, M. A., Schoener, T. W. & Losos, J. B. The relationship between sexual size dimorphism and habitat use in greater antillean anolis lizards. *Evolution* **54**, 259–272 (2000).

[46] Felsenstein, J. Phylogenies and the comparative method. *The American Naturalist* **125**, 1–15 (1985).

[47] Blomberg, S. P., Lefevre, J. G., Wells, J. A. & Waterhouse, M. Independent contrasts and pgls regression estimators are equivalent. *Systematic biology* **61**, 382–391 (2012).

[48] Hujoel, M. L., Gazal, S., Hormozdiari, F., van de Geijn, B. & Price, A. L. Disease heritability enrichment of regulatory elements is concentrated in elements with ancient sequence age and conserved function across species. *The American Journal of Human Genetics* **104**, 611–624 (2019). URL https://www.sciencedirect.com/science/article/pii/S0002929719300539.

[49] Sullivan, P. F. *et al.* Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023). URL https://www.science.org/doi/abs/10.1126/science.abn2937. https://www.science.org/doi/pdf/10.1126/science.abn2937.

[50] Gao, B. & Zhou, X. Mesusie enables scalable and powerful multi-ancestry fine-mapping of causal variants in genome-wide association studies. *Nature Genetics* 1–10 (2024).

[51] Pankratov, V. *et al.* Prioritizing autoimmunity risk variants for functional analyses by fine-mapping mutations under natural selection. *Nature Communications* **13**, 7069 (2022).

[52] Li, Z. *et al.* A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature methods* **19**, 1599–1611 (2022).

[53] Liu, Y. *et al.* Acat: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics* **104**, 410–421 (2019).

[54] Huber, C. D., Kim, B. Y. & Lohmueller, K. E. Population genetic models of gerp scores suggest pervasive turnover of constrained sites across mammalian evolution. *PLoS genetics* **16**, e1008827 (2020).

[55] Link, V. *et al.* Tree-based qtl mapping with expected local genetic relatedness matrices. *bioRxiv* 2023–04 (2023).

[56] Zhang, B. C., Biddanda, A., Gunnarsson, F., Cooper, F. & Palamara, P. F. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nature Genetics* **55**, 768–776 (2023). URL https://doi.org/10.1038/s41588-023-01379-x.

[57] Smith, S. D., Pennell, M. W., Dunn, C. W. & Edwards, S. V. Phylogenetics is the new genetics (for most of biodiversity). *Trends in Ecology & Evolution* **35**, 415–425 (2020).

[58] De Maio, N., Schlötterer, C. & Kosiol, C. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Molecular biology and evolution* **30**, 2249–2262 (2013).

[59] Rohlfs, R. V., Harrigan, P. & Nielsen, R. Modeling gene expression evolution with an extended ornstein–uhlenbeck process accounting for within-species variation. *Molecular biology and evolution* **31**, 201–211 (2014).

[60] Hahn, M. W. & Nakhleh, L. Irrational exuberance for resolved species trees. *Evolution* **70**, 7–17 (2016).

[61] Donoghue, M. J. A phylogenetic perspective on the distribution of plant diversity. *Proceedings of the National Academy of Sciences* **105**, 11549–11555 (2008).

[62] Crisp, M. D. *et al.* Phylogenetic biome conservatism on a global scale. *Nature* **458**, 754–756 (2009).

[63] Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).

[64] Schraiber, J. G., Mostovoy, Y., Hsu, T. Y. & Brem, R. B. Inferring evolutionary histories of pathway regulation from transcriptional profiling data. *PLoS computational biology* **9**, e1003255 (2013).

[65] Chen, J. *et al.* A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome research* **29**, 53–63 (2019).

[66] Price, P. D. *et al.* Detecting signatures of selection on gene expression. *Nature Ecology & Evolution* **6**, 1035–1045 (2022).

[67] Schraiber, J. G. & Landis, M. J. Sensitivity of quantitative traits to mutational effects and number of loci. *Theoretical population biology* **102**, 85–93 (2015).

[68] Landis, M. J., Schraiber, J. G. & Liang, M. Phylogenetic analysis using lévy processes: finding jumps in the evolution of continuous traits. *Systematic biology* **62**, 193–204 (2013).

[69] Bastide, P. & Didier, G. The Cauchy Process on Phylogenies: A Tractable Model for Pulsed Evolution. *Systematic Biology* syad053 (2023). URL `https://doi.org/10.1093/sysbio/syad053`. `https://academic.oup.com/sysbio/advance-article-pdf/doi/10.1093/sysbio/syad053/52190589/syad053.pdf`.

[70] Le Corre, V. & Kremer, A. The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology* **21**, 1548–1566 (2012). URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2012.05479.x`. `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-294X.2012.05479.x`.

[71] Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLOS Genetics* **10**, 1–25 (2014). URL `https://doi.org/10.1371/journal.pgen.1004412`.

[72] Comeron, J. M., Williford, A. & Kliman, R. M. The Hill–Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. *Heredity* **100**, 19–31 (2008). URL `https://doi.org/10.1038/sj.hdy.6801059`.

[73] Speed, D., Hemani, G., Johnson, M. & Balding, D. Improved Heritability Estimation from Genome-wide SNPs. *The American Journal of Human Genetics* **91**, 1011–1021 (2012). URL `https://doi.org/10.1016/j.ajhg.2012.10.010`. Publisher: Elsevier.

[74] Simons, Y. B., Mostafavi, H., Smith, C. J., Pritchard, J. K. & Sella, G. Simple scaling laws control the genetic architectures of human complex traits. *bioRxiv* 2022.10.04.509926 (2022).

[75] Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76–82 (2011).

[76] Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* **157**, 1819–1829 (2001). URL https://doi.org/10.1093/genetics/157.4.1819. https://academic.oup.com/genetics/article-pdf/157/4/1819/42032331/genetics1819.pdf.

[77] Mrode, R. A. *Linear Models for the Prediction of Animal Breeding Values* (CABI Wallingford, Oxfordshire, UK, 2013), 3rd edn.

[78] Goddard, M., Meuwissen, T. & Daetwyler, H. Prediction of Phenotype from DNA Variants. In *Handbook of Statistical Genomics*, 799–20 (2019). URL https://doi.org/10.1002/9781119487845.ch28.

[79] Henderson, C. R. *Applications of Linear Models in Animal Breeding* (University of Guelph, 1984).

[80] Gianola, D. & Fernando, R. L. Bayesian Methods in Animal Breeding Theory. *Journal of Animal Science* **63**, 217–244 (1986). URL https://doi.org/10.2527/jas1986.631217x.

[81] Lynch, M., Walsh, B. *et al. Genetics and analysis of quantitative traits*, vol. 1 (Sinauer Sunderland, MA, 1998).

[82] Gillespie, J. H. *Population genetics: a concise guide* (JHU press, 2004).

[83] Henderson, C. Theoretical basis and computational methods for a number of different animal models. *Journal of Dairy Science* **71**, 1–16 (1988). URL https://www.sciencedirect.com/science/article/pii/S0022030288799749.

[84] Kruuk, L. E. B. Estimating genetic parameters in natural populations using the 'animal model'. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**, 873–890 (2004). URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2003.1437. https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2003.1437.

[85] de los Campos, G., Sorensen, D. & Gianola, D. Genomic Heritability: What Is It? *PLOS Genetics* **11**, e1005048 (2015). URL https://doi.org/10.1371/journal.pgen.1005048. Publisher: Public Library of Science.

[86] Ralph, P., Thornton, K. & Kelleher, J. Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes. *Genetics* **215**, 779–797 (2020). URL https://doi.org/10.1534/genetics.120.303253. https://academic.oup.com/genetics/article-pdf/215/3/779/42225422/genetics0779.pdf.

[87] McVean, G. A genealogical interpretation of principal components analysis. *PLoS genetics* **5**, e1000686 (2009).

[88] Fan, C., Mancuso, N. & Chiang, C. W. A genealogical estimate of genetic relationships. *The American Journal of Human Genetics* **109**, 812–824 (2022). URL https://doi.org/10.1016/j.ajhg.2022.03.016. Publisher: Elsevier.

[89] Wang, S. *et al.* Estimating Genetic Similarity Matrices Using Phylogenies. *Journal of Computational Biology* **28**, 587–600 (2021). URL https://doi.org/10.1089/cmb.2020.0375. Publisher: Mary Ann Liebert, Inc., publishers.

[90] Lande, R. Natural selection and random genetic drift in phenotypic evolution. *Evolution* 314–334 (1976).

[91] Lynch, M. & Hill, W. G. Phenotypic evolution by neutral mutation. *Evolution* **40**, 915–935 (1986).

[92] Lynch, M. The evolutionary scaling of cellular traits imposed by the drift barrier. *Proceedings of the National Academy of Sciences* **117**, 10435–10444 (2020).

[93] Hansen, T. F. & Martins, E. P. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution* **50**, 1404–1417 (1996).

[94] Hansen, T. F., Pienaar, J. & Orzack, S. H. A comparative method for studying adaptation to a randomly evolving environment. *Evolution* **62**, 1965–1977 (2008).

[95] Harvey, P. H., Pagel, M. D. *et al.* *The comparative method in evolutionary biology*, vol. 239 (Oxford university press Oxford, 1991).

[96] Mendes, F. K., Fuentes-Gonzalez, J. A., Schraiber, J. G. & Hahn, M. W. A multispecies coalescent model for quantitative traits. *Elife* **7**, e36482 (2018).

[97] Hibbins, M. S., Breithaupt, L. C. & Hahn, M. W. Phylogenomic comparative methods: Accurate evolutionary inferences in the presence of gene tree discordance. *Proceedings of the National Academy of Sciences* **120**, e2220389120 (2023).

[98] Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459–463 (2010). URL https://doi.org/10.1038/nrg2813.

[99] Yao, Y. & Ochoa, A. Limitations of principal components in quantitative genetic association models for human studies. *eLife* **12**, e79238 (2023). URL https://doi.org/10.7554/eLife.79238.

[100] Caetano, D. S., O'Meara, B. C. & Beaulieu, J. M. Hidden state models improve state-dependent diversification approaches, including biogeographical models. *Evolution* **72**, 2308–2324 (2018).

[101] Diniz-Filho, J. A. F., Sant'Ana, C. E. R. d. & Bini, L. M. An eigenvector method for estimating phylogenetic inertia. *Evolution* **52**, 1247–1262 (1998).

[102] de Vienne, D. M., Aguileta, G. & Ollier, S. Euclidean nature of phylogenetic distance matrices. *Systematic biology* **60**, 826–832 (2011).

[103] Revell, L. J. Size-correction and principal components for interspecific comparative studies. *Evolution* **63**, 3258–3268 (2009).

[104] Uyeda, J. C., Caetano, D. S. & Pennell, M. W. Comparative analysis of principal components can be misleading. *Systematic Biology* **64**, 677–689 (2015).

[105] Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature genetics* **51**, 1749–1755 (2019).

[106] Freckleton, R. P., Harvey, P. H. & Pagel, M. Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist* **160**, 712–726 (2002).

[107] Revell, L. J. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution* **1**, 319–329 (2010).

[108] Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics* **9**, 255–266 (2008). URL `https://doi.org/10.1038/nrg2322`.

[109] Tung Ho, L. s. & Ané, C. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Systematic biology* **63**, 397–408 (2014).

[110] Wu, Y. & Sankararaman, S. A scalable estimator of snp heritability for biobank-scale data. *Bioinformatics* **34**, i187–i194 (2018).

[111] Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nature genetics* **50**, 906–908 (2018).

[112] Baumdicker, F. *et al.* Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* **220**, iyab229 (2022).

[113] Battey, C. J., Ralph, P. L. & Kern, A. D. Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data. *Genetics* **215**, 193–214 (2020). URL `https://doi.org/10.1534/genetics.120.303143`. `https://academic.oup.com/genetics/article-pdf/215/1/193/42100884/genetics0193.pdf`.

[114] Rohlf, F. J. Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution* **55**, 2143–2160 (2001).

[115] Freckleton, R. P., Cooper, N. & Jetz, W. Comparative methods as a statistical fix: the dangers of ignoring an evolutionary model. *The American Naturalist* **178**, E10–E17 (2011).

[116] Adams, D. C. & Church, J. O. The evolution of large-scale body size clines in plethodon salamanders: evidence of heat-balance or species-specific artifact? *Ecography* **34**, 1067–1075 (2011).

[117] Legendre, P. & Legendre, L. *Numerical ecology* (Elsevier, 2012).

[118] Cope, A. L., O'Meara, B. C. & Gilchrist, M. A. Gene expression of functionally-related genes coevolves across fungal species: detecting coevolution of gene expression using phylogenetic comparative methods. *BMC genomics* **21**, 1–17 (2020).

[119] Szklarczyk, D. *et al.* String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research* **47**, D607–D613 (2019).

[120] Jiang, D., Cope, A. L., Zhang, J. & Pennell, M. Decoupling of evolutionary changes in mrna and protein levels. *Molecular Biology and Evolution* **40**, msad169 (2023).

[121] Dimayacyac, J. R., Wu, S., Jiang, D. & Pennell, M. Evaluating the performance of widely used phylogenetic models for gene expression evolution. *Genome Biology and Evolution* **15**, evad211 (2023).

[122] Harmon, L. Phylogenetic comparative methods: learning from trees (2019).

[123] Beaulieu, J. M., O'Meara, B. C. & Donoghue, M. J. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. *Systematic biology* **62**, 725–737 (2013).

[124] Barton, N. H., Etheridge, A. M. & Véber, A. The infinitesimal model: Definition, derivation, and implications. *Theoretical population biology* **118**, 50–73 (2017).

[125] Taylor, M. B. & Ehrenreich, I. M. Higher-order genetic interactions and their contribution to complex traits. *Trends in Genetics* **31**, 34–40 (2015). URL `https://doi.org/10.1016/j.tig.2014.09.001`. Publisher: Elsevier.

[126] Campbell, R. F., McGrath, P. T. & Paaby, A. B. Analysis of Epistasis in Natural Traits Using Model Organisms. *Trends in Genetics* **34**, 883–898 (2018). URL `https://doi.org/10.1016/j.tig.2018.08.002`. Publisher: Elsevier.

[127] Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS genetics* **4**, e1000008 (2008).

[128] Crow, J. F. On epistasis: why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 1241–1244 (2010).

[129] Mäki-Tanila, A. & Hill, W. G. Influence of gene interaction on complex trait variation with multilocus models. *Genetics* **198**, 355–367 (2014).

[130] Hansen, T. F. Why epistasis is important for selection and adaptation. *Evolution* **67**, 3501–3511 (2013).

[131] Lynch, M. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. *The American Naturalist* **136**, 727–741 (1990).

[132] Estes, S. & Arnold, S. J. Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales. *The american naturalist* **169**, 227–244 (2007).

[133] Houle, D., Bolstad, G. H., van der Linde, K. & Hansen, T. F. Mutation predicts 40 million years of fly wing evolution. *Nature* **548**, 447–450 (2017).

[134] Arnold, S. J., Pfrender, M. E. & Jones, A. G. The adaptive landscape as a conceptual bridge between micro-and macroevolution. *Microevolution rate, pattern, process* 9–32 (2001).

[135] Krone, S. M. & Neuhauser, C. Ancestral processes with selection. *Theoretical population biology* **51**, 210–237 (1997).

[136] Wright, S. The results of crosses between inbred strains of guinea pigs, differing in number of digits. *Genetics* **19**, 537 (1934).

[137] Felsenstein, J. Quantitative characters, phylogenies, and morphometrics. *Systematics Association Special Volume* **64**, 27–44 (2002).

[138] Felsenstein, J. A comparative method for both discrete and continuous characters using the threshold model. *The American Naturalist* **179**, 145–156 (2012).

[139] Hadfield, J. D. Increasing the efficiency of mcmc for hierarchical phylogenetic models of categorical traits using reduced mixed models. *Methods in Ecology and Evolution* **6**, 706–714 (2015).

[140] Lande, R. & Arnold, S. J. The measurement of selection on correlated characters. *Evolution* 1210–1226 (1983).

[141] Kingsolver, J. G. *et al.* The strength of phenotypic selection in natural populations. *The American Naturalist* **157**, 245–261 (2001).

[142] Sanjak, J. S., Sidorenko, J., Robinson, M. R., Thornton, K. R. & Visscher, P. M. Evidence of directional and stabilizing selection in contemporary humans. *Proceedings of the National Academy of Sciences* **115**, 151–156 (2018).

[143] Stroud, J. T., Moore, M. P., Langerhans, R. B. & Losos, J. B. Fluctuating selection maintains distinct species phenotypes in an ecological community in the wild. *Proceedings of the National Academy of Sciences* **120**, e2222071120 (2023).

[144] Araki, H., Berejikian, B. A., Ford, M. J. & Blouin, M. S. Fitness of hatchery-reared salmonids in the wild. *Evolutionary applications* **1**, 342–355 (2008).

[145] Colautti, R. I. & Barrett, S. C. Rapid adaptation to climate facilitates range expansion of an invasive plant. *science* **342**, 364–366 (2013).

[146] Siepielski, A. M., DiBattista, J. D. & Carlson, S. M. It's about time: the temporal dynamics of phenotypic selection in the wild. *Ecology letters* **12**, 1261–1276 (2009).

[147] De Villemereuil, P. *et al.* Fluctuating optimum and temporally variable selection on breeding date in birds and mammals. *Proceedings of the National Academy of Sciences* **117**, 31969–31978 (2020).

[148] Butler, M. A. & King, A. A. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist* **164**, 683–695 (2004).

[149] Beaulieu, J. M., Jhwueng, D.-C., Boettiger, C. & O'Meara, B. C. Modeling stabilizing selection: expanding the ornstein–uhlenbeck model of adaptive evolution. *Evolution* **66**, 2369–2383 (2012).

[150] Bartoszek, K., Pienaar, J., Mostad, P., Andersson, S. & Hansen, T. F. A phylogenetic comparative method for studying multivariate adaptation. *Journal of Theoretical Biology* **314**, 204–215 (2012).

[151] Clavel, J., Escarguel, G. & Merceron, G. mvmorph: an r package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution* **6**, 1311–1319 (2015).

[152] Uyeda, J. C. & Harmon, L. J. A novel bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Systematic biology* **63**, 902–918 (2014).

[153] Khabbazian, M., Kriebel, R., Rohe, K. & Ané, C. Fast and accurate detection of evolutionary shifts in ornstein–uhlenbeck models. *Methods in Ecology and Evolution* **7**, 811–824 (2016).

[154] Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nature genetics* **50**, 746–753 (2018).

[155] Schoech, A. P. *et al.* Quantification of frequency-dependent genetic architectures in 25 uk biobank traits reveals action of negative selection. *Nature communications* **10**, 790 (2019).

[156] Simons, Y. B., Bullaughey, K., Hudson, R. R. & Sella, G. A population genetic interpretation of gwas findings for human quantitative traits. *PLoS biology* **16**, e2002985 (2018).

[157] Koch, E. M. & Sunyaev, S. R. Maintenance of complex trait variation: classic theory and modern data. *Frontiers in genetics* 2198 (2021).

[158] Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).

[159] Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20**, 467–484 (2019).

[160] Abdellaoui, A., Yengo, L., Verweij, K. J. & Visscher, P. M. 15 years of gwas discovery: Realizing the promise. *The American Journal of Human Genetics* **110**, 179–194 (2023). URL `https://www.sciencedirect.com/science/article/pii/S0002929722005456`.

[161] Del-Aguila, J. L. *et al.* Alzheimer's disease: rare variants with large effect sizes. *Current opinion in genetics & development* **33**, 49–55 (2015).

[162] Park, J.-H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics* **42**, 570–575 (2010).

[163] Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the japanese population. *Nature communications* **10**, 4393 (2019).

[164] Park, J.-H. *et al.* Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences* **108**, 18026–18031 (2011).

[165] Dudley, J. T., Chen, R., Sanderford, M., Butte, A. J. & Kumar, S. Evolutionary meta-analysis of association studies reveals ancient constraints affecting disease marker discovery. *Molecular biology and evolution* **29**, 2087–2094 (2012).

[166] Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R. & Amos, C. I. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *The American Journal of Human Genetics* **82**, 100–112 (2008).

[167] Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).

[168] Gorlova, O. Y., Xiao, X., Tsavachidis, S., Amos, C. I. & Gorlov, I. P. Snp characteristics and validation success in genome wide association studies. *Human Genetics* **141**, 229–238 (2022).

[169] Benegas, G., Batra, S. S. & Song, Y. S. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences* **120**, e2311219120 (2023).

[170] Wright, S. The analysis of variance and the correlations between relatives with respect to deviations from an optimum. *Journal of Genetics* **30**, 243–256 (1935).

[171] Robertson, A. The effect of selection against extreme deviants based on deviation or on homozygosis: With two text-figures. *Journal of Genetics* **54**, 236–248 (1956).

[172] Keightley, P. D. & Hill, W. G. Quantitative genetic variability maintained by mutation-stabilizing selection balance in finite populations. *Genetics Research* **52**, 33–43 (1988).

[173] Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

[174] Kowalczyk, A. *et al.* Rerconverge: an r package for associating evolutionary rates with convergent traits. *Bioinformatics* **35**, 4815–4817 (2019).

[175] Benegas, G., Albors, C., Aw, A. J., Ye, C. & Song, Y. S. Gpn-msa: an alignment-based dna language model for genome-wide variant effect prediction. *bioRxiv* .

[176] Racimo, F. & Schraiber, J. G. Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. *PLoS genetics* **10**, e1004697 (2014).

[177] Plassais, J. *et al.* Natural and human-driven selection of a single non-coding body size variant in ancient and modern canids. *Current Biology* **32**, 889–897 (2022).

[178] Pease, J. B., Haak, D. C., Hahn, M. W. & Moyle, L. C. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS biology* **14**, e1002379 (2016).

[179] Natarajan, C. *et al.* Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. *Science* **354**, 336–339 (2016).

[180] Hu, Z., Sackton, T. B., Edwards, S. V. & Liu, J. S. Bayesian detection of convergent rate changes of conserved noncoding elements on phylogenetic trees. *Molecular biology and evolution* **36**, 1086–1100 (2019).

[181] Sackton, T. B. *et al.* Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* **364**, 74–78 (2019).

[182] Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annual review of genetics* **44**, 293–308 (2010).

[183] Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82–93 (2011).

[184] Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome medicine* **7**, 1–11 (2015).

[185] Guerrero, R. F. & Hahn, M. W. Quantifying the risk of hemiplasy in phylogenetic inference. *Proceedings of the National Academy of Sciences* **115**, 12787–12792 (2018).

[186] Hibbins, M. S., Gibson, M. J. & Hahn, M. W. Determining the probability of hemiplasy in the presence of incomplete lineage sorting and introgression. *Elife* **9**, e63753 (2020).

[187] Maddison, W. P. Gene trees in species trees. *Systematic biology* **46**, 523–536 (1997).

[188] Avise, J. C. & Robinson, T. J. Hemiplasy: a new term in the lexicon of phylogenetics. *Systematic biology* **57**, 503–507 (2008).

[189] Yan, H. *et al.* Phyloacc-gt: A bayesian method for inferring patterns of substitution rate shifts on targeted lineages accounting for gene tree discordance. *Molecular Biology and Evolution* **40**, msad195 (2023).

[190] Revell, L. J. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution* 217–223 (2012).

[191] Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

[192] Stadler, T. *TreeSim: Simulating Phylogenetic Trees* (2019). URL `https://CRAN.R-project.org/package=TreeSim`. R package version 2.4.

[193] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).