Coal-Miner: A Statistical Method for GWA Studies of Quantitative Traits with Complex Evolutionary Origins

Hussein A. Hejase
Department of Computer Science and Engineering
Michigan State University
East Lansing, Michigan 48824
hijazihu@msu.edu

Natalie Vande Pol
Department of Plant, Soil and Microbial Sciences
Michigan State University
East Lansing, Michigan 48824
vandepo7@msu.edu

Gregory M. Bonito
Department of Plant, Soil and Microbial Sciences
Michigan State University
East Lansing, Michigan 48824
bonito@msu.edu

Patrick P. Edger
Department of Horticulture
Michigan State University
East Lansing, Michigan 48824
edgerpat@msu.edu

Kevin J. Liu*
Department of Computer Science and Engineering
Michigan State University
East Lansing, Michigan 48824
kjl@msu.edu

ABSTRACT

Association mapping (AM) methods are used in genome-wide association (GWA) studies to test for statistically significant associations between genotypic and phenotypic data. The genotypic and phenotypic data share common evolutionary origins – namely, the evolutionary history of sampled organisms – introducing covariance which must be distinguished from the covariance due to biological function that is of primary interest in GWA studies. A variety of methods have been introduced to perform AM while accounting for sample relatedness. However, the state of the art predominantly utilizes the simplifying assumption that sample relatedness is effectively fixed across the genome. In contrast, population genetic theory and empirical studies have shown that sample relatedness can vary greatly across different loci within a genome. This phenomenon – referred to as local genealogical variation – is commonly encountered in many genomic datasets. New AM methods are needed to better account for local variation in sample relatedness within genomes.

We address this gap by introducing Coal-Miner, a new statistical AM method. The Coal-Miner algorithm takes the form of a methodological pipeline. The initial stages of Coal-Miner seek to detect candidate loci, or loci which contain putatively associated markers. Subsequent stages of Coal-Miner perform test for association using a linear mixed model with multiple effects which account for sample relatedness locally within candidate loci and globally across the entire genome. Using synthetic and empirical datasets, we compare the statistical power and type I error control of Coal-Miner against state-of-the-art AM methods. The simulation conditions reflect a variety of genomic architectures for complex traits and incorporate a range of evolutionary scenarios, each with different evolutionary processes that can generate local genealogical variation. Across the datasets in our study, we find that Coal-Miner consistently offers comparable or typically better statistical power and type I error control compared to the state-of-the-art methods.

CCS CONCEPTS

•**Applied computing** → *Computational genomics; Computational biology; Molecular sequence analysis; Molecular evolution; Computational genomics; Systems biology; Bioinformatics; Population genetics;*

KEYWORDS

genome wide association study; GWAS; population stratification; genealogy; coalescent; Arabidopsis

1 INTRODUCTION

Genome-wide association (GWA) studies aim to pinpoint loci with genetic contributions to a phenotype by uncovering significant statistical associations between genomic markers and a phenotypic trait under study. We refer to the computational methods used in a GWA analysis as association mapping (AM) methods. Among the most widely studied organisms in GWA studies are natural human populations and laboratory strains of house mouse. Recently, GWA approaches have been applied to natural populations of other organisms sampled from across the Tree of Life. For example, the 1001 Genomes Consortium study [7] published whole genome sequences for over a thousand samples from globally distributed *Arabidopsis* populations. In combination with phenotypic data, the genomic sequence data was used in a GWA analysis to pinpoint genomic loci involved in flowering time at two different temperatures. Other recent GWA studies such as the study of Porter et al. [24] have focused on bacteria and other microbes (see [6] for a review of relevant literature).

Regardless of sampling strategy – from one or more closely related populations involving a single species to multiple populations from divergent species – it is well understood that sample relatedness can be a confounding factor in GWA analyses unless

*To whom correspondence should be addressed.

ACM-BCB'17, August 20-23, 2017, Boston, MA, USA.
 978-1-4503-4722-8/17/08.
DOI: http://dx.doi.org/10.1145/3107411.3107490

properly accounted for. Intuitively, the genotypes and phenotypes of present-day samples reflect their shared evolutionary history, or phylogeny. For this reason, covariance due to a functional relationship between genotypic markers and a phenotypic character must be distinguished from shared covariance due to common evolutionary origins. A number of AM methods have been developed to address this issue. EIGENSTRAT [26] is a popular AM method which accounts for sample relatedness as a fixed effect. Other statistical AM methods have utilized linear mixed models (LMMs) to capture sample relatedness using random effects; these include EMMAX [17] and GEMMA [31]. Local variation in functional covariance across the genome is a crucial signature that AM methods use to uncover putatively associated markers. In contrast, virtually all of the most widely used state-of-the-art AM methods assume that covariance due to sample relatedness does not vary appreciably across the genome. Sample relatedness is therefore evaluated "globally" across the genome, eliding over "local" genealogical variation across loci. The latter has been observed by many comparative genomic and phylogenomic studies (see [10] for a review of relevant literature). It is now well understood that local genealogical variation within genomes is pervasive across a range of evolutionary divergence – from structured populations within a single species to multiple species at various scales up to the Tree of Life, the evolutionary history of all living organisms on Earth. The evolutionary processes that can contribute to local genealogical variation include genetic drift and incomplete lineage sorting, recombination, gene flow, positive selection, and the combination of all of these processes (and others) [10].

Computational approaches for detecting local genealogical variation are broadly characterized by their modeling assumptions. One class of methods makes use of the Four-Gamete Test [14], which requires the simplifying assumption that sequence evolution can be described by the infinite sites model. The LRScan algorithm [29] belongs to this class of methods. Another class consists of parametric methods that make use of finite-sites models of sequence evolution. These include methods such as RecHMM [30]. More recently, coalescent-based methods such as PhyloNet-HMM [19] have been developed to infer local coalescent histories and explicitly ascribe local genealogical variation to different evolutionary processes.

Building upon these insights, we previously developed Coal-Map [12], an AM method that utilizes a fixed effects model to account for global sample relatedness and, depending upon whether the test marker is located within a locus containing putatively associated markers, local sample relatedness as well. The latter condition is evaluated using model selection criteria. Coal-Map requires local-phylogeny-switching breakpoints as input. We conducted a simulation study which demonstrated that Coal-Map's statistical power and type I error control was comparable or better than other state-of-the-art methods that account for global sample relatedness using fixed effects.

2 METHODS

2.1 Overview of Coal-Miner algorithm

In this study, we introduce Coal-Miner, a new statistical AM method which accounts for local variation of sample relatedness across genomic sequences as well as global sample relatedness. Coal-Miner's contributions relative to the state of the art (including Coal-Map) consist of the following. First, Coal-Miner utilizes an LMM with multiple effects to explicitly capture the genomic architecture of a phenotype, where both genotypic and phenotypic characters are the product of a complex evolutionary history which can cause sample relatedness to vary locally across genomic loci. The LMM captures global sample relatedness as a random effect, in contrast to the fixed-effect approach used by Coal-Map. Second, the pipeline-based design of Coal-Miner incorporates an intermediate stage to infer "candidate loci" for use in the new LMM, where a candidate locus is a locus that is inferred to contain one or more putatively associated SNPs.

We begin by introducing the high-level design of Coal-Miner. The input to the Coal-Miner algorithm consists of: (1) an $n \times k$ multi-locus sequence data matrix $\boldsymbol{X}$, (2) an $n \times 1$ vector $\boldsymbol{y}$ which represents a phenotypic character, and (3) ℓ^*, the number of candidate loci used during analysis. The output consists of an association score for each site $\boldsymbol{x} \in \boldsymbol{X}$.

Coal-Miner's statistical model captures the relationship between genotypic data $\boldsymbol{X}$ and the phenotypic character $\boldsymbol{y}$ in the form of a linear mixed model (LMM). The LMM incorporates multiple effects to capture the phenotypic contributions of and local genealogical variation among multiple candidate loci. A candidate locus is represented by a fixed effect, and a random effect is included to capture global sample relatedness as measured across all loci in $\boldsymbol{X}$. Ideally, the set of candidate loci identified during a Coal-Miner analysis is identical to the set of causal loci (i.e., loci containing causal SNPs) for the trait under study; in practice, the set of candidate loci are inferred as part of the Coal-Miner algorithm, which we discuss in greater detail below. The LMM takes the following form (based on the notation of Zhou and Stephens [31]):

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\alpha} + \boldsymbol{x}\beta + \boldsymbol{u} + \boldsymbol{\epsilon}$$
$$\boldsymbol{u} \sim \mathrm{MVN}_{\boldsymbol{n}}(0, \lambda\tau^{-1}\boldsymbol{K}_{\mathrm{global}})$$
$$\boldsymbol{\epsilon} \sim \mathrm{MVN}_{\boldsymbol{n}}(0, \tau^{-1}\boldsymbol{I}_{\boldsymbol{n}})$$

The fixed effects are represented by c covariates in the $n \times c$ matrix $\boldsymbol{W}$, which include covariates that capture local sample relatedness within each candidate locus, the $c \times 1$ vector $\boldsymbol{\alpha}$ of corresponding coefficients, and the test SNP is represented by the $n \times 1$ vector $\boldsymbol{x}$ with effect size β. Global sample relatedness (i.e., sample relatedness as measured across all loci in the genotypic data $\boldsymbol{X}$) is specified by the $n \times n$ relatedness matrix $\boldsymbol{K}_{\mathrm{global}}$ computed using $\boldsymbol{X}$, following the approach of state-of-the-art LMM-based AM methods (e.g., GEMMA [31]). The $n \times 1$ vectors $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$ represent random effects which account for global sample relatedness and residual error, respectively. Each of the two random effects follows an n-dimensional multivariate normal distribution (abbreviated "MVN") with mean 0. The random effects $\boldsymbol{u}$ have covariance $\lambda\tau^{-1}\boldsymbol{K}_{\mathrm{global}}$ and the random effects $\boldsymbol{\epsilon}$ have covariance $\tau^{-1}\boldsymbol{I}_{\boldsymbol{n}}$, where λ is the relative ratio between the two, $\boldsymbol{I}_{\boldsymbol{n}}$ is the $n \times n$ identity matrix, and the residual errors have variance τ^{-1}.

The design of the Coal-Miner algorithm takes the form of a methodological pipeline. We now discuss each pipeline stage in turn.

Stage one of Coal-Miner: inferring local-phylogeny-switching breakpoints. The input to the first stage of Coal-Miner is the genotypic data matrix $\boldsymbol{X}$. The output consists of a set of local-phylogeny-switching breakpoints $\boldsymbol{b}$ which partition the sites in $\boldsymbol{X}$ into loci $\{\boldsymbol{X_i}\}$, where $1 \leq i \leq \ell$ and ℓ is the number of loci. We require that $\ell^* \leq \ell$. (The ratio of ℓ^* and ℓ depends upon the genomic architecture of the trait corresponding to character $\boldsymbol{y}$.)

The general approach to address this computational problem is to infer local coalescent histories under an appropriate multi-species extension of the coalescent model [18], and then to assign breakpoints based upon gene tree discordance. Each pair of neighboring breakpoints delineates a locus for use in downstream stages of the Coal-Miner pipeline. The specific choice of model/method depends upon the relevant evolutionary processes involved in multi-locus sequence evolution, particularly regarding the source(s) of local genealogical discordance.

In this study, we use one of two different methods, depending upon assumptions about biomolecular sequence evolution. In the simulation study, the simulations make use of the infinite sites model. We therefore used the LRScan algorithm [29] to compute local-topology-switching breakpoints based upon the Four Gamete Test (FGT) [14]. In the empirical study, we did not make use of the infinite sites model and its assumptions about sequence evolution. Furthermore, multiple evolutionary processes were known to be involved in multi-locus sequence evolution, including genetic drift/incomplete lineage sorting (ILS), recombination/gene conversion, gene flow/horizontal gene transfer (HGT), and natural selection. Breakpoint inference under the corresponding extended coalescent model is suspected to be a computationally difficult problem. Existing methods for this problem (e.g., PhyloNet-HMM [19]) did not have sufficient scalability for the dataset sizes examined in our study. As a more feasible alternative, we inferred local-topology-switching breakpoints using Rec-HMM [30]. Rec-HMM performs fixed-species-phylogeny inference of local genealogies under a statistical model that combines a finite-sites substitution model and a hidden Markov model which is meant to capture intra-sequence dependence (such as arises from recombination).

Stage two of Coal-Miner: identifying candidate loci. The input to the second stage of Coal-Miner consists of the genotypic data matrix $\boldsymbol{X}$, the set of breakpoints $\boldsymbol{b}$ which partition $\boldsymbol{X}$ into loci $\{\boldsymbol{X_i}\}$, where $1 \leq i \leq \ell$ and ℓ is the number of loci, the phenotypic character $\boldsymbol{y}$, and ℓ^*, the number of candidate loci to identify. Note that the input $\boldsymbol{b}$ is an output of the preceding stage of Coal-Miner. The output is a set of candidate loci $\{\boldsymbol{X_j^*}\} \subseteq \{\boldsymbol{X_i}\}$ where $1 \leq j \leq \ell^*$.

Our general approach to this problem consists of a search among possible sets of candidate loci $\{X_j^*\}$ using optimization under a "null" version of Coal-Miner's LMM, where we do not consider a test SNP (i.e., $\beta = 0$ in Coal-Miner's LMM) and the phenotypic contributions from putatively associated SNPs in each candidate locus X_j^* is captured by covariates $\{\boldsymbol{w_j}\} \subseteq \boldsymbol{W}$. Since we compare fitted LMMs that may have varying fixed effects, our optimization criterion consists of the LMM log-likelihood $\mathcal{L}(\lambda, \tau, \boldsymbol{\alpha}, \beta) = \frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{H}| - \frac{1}{2}\tau(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\alpha}-\boldsymbol{x}\beta)^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\alpha}-\boldsymbol{x}\beta)$ where $\boldsymbol{H} = \lambda\boldsymbol{K}_{\text{global}} + \boldsymbol{I_n}$ (reproduced from equation (3) in [31]). Due to the computational difficulty of this optimization problem, numerical optimization procedures are typically used. We obtained estimates of λ in the range of $[10^{-5}, 1]$ using the optimization heuristic implemented in the GEMMA software library [31], which combines Brent's method [5] and the Newton-Raphson method.

For each candidate locus $\boldsymbol{X_j^*}$, local sample relatedness was evaluated using principal component analysis (PCA) [16] of $\boldsymbol{X_j^*}$ – similar to techniques that are widely used by AM methods to account for global sample relatedness as fixed effects [26]. The phenotypic contribution of candidate locus $\boldsymbol{X_j^*}$ was represented using covariates $\{\boldsymbol{w_j}\}$ which consisted of the top five principal components, where the zth principal component corresponds to the sample covariance matrix eigenvector with the zth largest eigenvalue and the number of covariates was based upon a design experiment in [12]. For added computational efficiency, we substituted the following search heuristic in place of set-based search among all possible ℓ^*-size sets of candidate loci. For each locus $\boldsymbol{X_i}$, we used MLE to fit an equivalent LMM, except that the covariates W included only the covariates $\{\boldsymbol{w_i}\}$ for locus $\boldsymbol{X_i}$ (as computed using the above PCA-based procedure). The output set of candidate loci consists of the top ℓ^* loci based upon fitted LMM likelihood.

Stage three of Coal-Miner: SNP-based association testing. The input to the third stage of Coal-Miner consists of the genotypic data matrix $\boldsymbol{X}$, the set of breakpoints $\boldsymbol{b}$ which partition $\boldsymbol{X}$ into loci $\{\boldsymbol{X_i}\}$, where $1 \leq i \leq \ell$ and ℓ is the number of loci, the phenotypic character $\boldsymbol{y}$, and the set of candidate loci $\{\boldsymbol{X_j^*}\}$. Note that the inputs $\boldsymbol{b}$ and $\{\boldsymbol{X_j^*}\}$ are outputs of stages one and two of Coal-Miner, respectively. The output of this stage is Coal-Miner's final output.

Each test SNP $\boldsymbol{x}$ is tested for association under Coal-Miner's LMM. Variation in local sample relatedness across candidate loci $\{\boldsymbol{X_j^*}\}$ is captured by covariates in $\boldsymbol{W}$: specifically, if the test SNP $\boldsymbol{x}$ is located within a candidate locus $\boldsymbol{X_j^*}$, the covariates $\boldsymbol{W}$ include a corresponding covariate $\boldsymbol{w_j}$ which consists of the top principal component from PCA applied to $\boldsymbol{X_j^*}$ (see above discussion of previous stage), and otherwise not. (Stages two and three of the Coal-Miner pipeline utilize different covariates $\boldsymbol{W}$ due to the absence or presence of a test SNP effect in their respective LMMs.) The LMM is fitted using the likelihood-based numerical optimization procedures that are also used in stage two of Coal-Miner, and the association score is computed using a likelihood ratio test of the fitted model against a null model with no SNP effect.

2.2 Simulation study

Neutral simulations of multi-locus sequence data were based upon either tree-like or non-tree-like evolutionary scenarios. The evolutionary scenarios shared a species phylogeny that we used in a prior simulation study (Supplementary Figure S1 in the Supporting Online Materials (SOM)). We used ms [13] to simulate coalescent histories (and embedded gene trees) under an extension of the coalescent model [18] which allows instantaneous unidirectional admixture (IUA) [9]. Under this model, the parameterization of the model phylogeny includes an admixture proportion γ. Appropriate choices of γ allow us to explore the impact of tree-like and non-tree-like evolution in our simulation study, where we set γ to either 0.0 or 0.5, respectively. Each replicate dataset sampled 10 independently and identically distributed loci and 1000 individuals; taxa A, B, and C were represented by 250, 250, and 500 samples, respectively. Bi-allelic sequence evolution was simulated under the infinite sites model to obtain 250 bp per locus, resulting in total sequence length of 2.5 kb per replicate dataset.

As a means to investigate the impact of the genomic architecture of phenotypes, we simulated phenotypic characters using the approach from our previous work [12]. For each synthetic multi-locus sequence dataset in the neutral simulations, we randomly selected either 10%, 20%, or 30% of loci as causal. Twenty causal SNPs were then randomly selected from causal loci such that each causal locus contained at least one causal SNP and causal SNPs had minor allele frequency between 0.1 and 0.3. Given a set of causal SNPs δ, we sampled character y under an extension of the quantitative trait model used by Long and Langley [20]. The trait value for the ith individual is represented as $y_i = \pi \sum_{j \in \delta} \frac{Q_{i,j}}{|\delta|} + (1 - \pi)N(0, 0.01)$ where π specifies the ratio between the genotypic contribution and an environmental residual, Q is 1 if sample i has the derived allele at the jth causal SNP and 0 otherwise, and the environmental residual is normally distributed with mean 0 and standard deviation 0.01. Our simulations utilized a ratio π of 0.5.

Our simulation study also included non-neutral simulations that incorporated positive selection. We used msms [11] to conduct forward-time coalescent simulations of genotypic sequence evolution (in place of an otherwise equivalent neutral backward-time coalescent simulation using ms), where causal loci were evolved under deme-dependent positive selection with a finite sites mutation model and all other loci evolved neutrally (as discussed above in the neutral simulation procedure). We used a selection coefficient of $s = 0.56$, which is in line with estimates from prior studies of positive selection in natural *Mus* populations [28]. Quantitative traits were simulated using the above procedure.

The simulation study experiments involving quantitative traits with varying genomic architectures included 12 different model conditions in total. To recap, the model conditions differed in terms of the proportion of causal loci (either 10%, 20%, or 30%), model phylogeny (either tree-like or non-tree-like), and the presence or absence of positive selection. For each model condition, we repeated the simulation procedure to obtain 20 replicate datasets.

Performance evaluation. The other methods in our study consisted of Coal-Map, GEMMA, and EIGENSTRAT. We followed the procedure from [12] to obtain FGT-based local-phylogeny-switching breakpoints and run Coal-Map analyses. For consistency with the other LMM-based AM methods in our study, we ran GEMMA using an IBS kinship matrix as our measure of global sample relatedness and MLE and LRT to obtain association scores. EIGENSTRAT was run with default settings using the top ten principal components from the genotypic data matrix $\boldsymbol{X}$, following the recommendations of Price et al. [26]. Detailed software commands are listed in the SOM Appendix.

We evaluated performance based on statistical power, type I error, and AUROC. To compare AUROC, we performed DeLong *et al.* tests [8] using the Daim v. 1.1.0 package [25] in R [27]. Custom scripts were used to conduct the simulation study. All scripts are provided under an open source license. See SOM Appendix for details and download instructions.

2.3 Empirical study

We re-analyzed an *Arabidopsis* dataset which consists of whole genome sequence (WGS) data and phenotypic data for two quantitative traits: flowering time at 10 °C and 16 °C. A total of 1,135 samples from natural populations across the globe are represented. The phylogeny shown in SOM Supplementary Figure S16 depicts the geographic origins of and evolutionary relationships among the samples. The dataset was originally published and analyzed by the 1001 Genomes Consortium [7], and we obtained genomic sequences and quantitative trait data from the 1001 Genomes Project database (accessible at www.1001genomes.org); the former includes both assembled WGS data and variant calls for a total of 10,707,430 bi-allelic SNPs. (Details about sequencing, assembly, filtering, quality controls, and variant calling are described in the 1001 Genomes Consortium study [7].)

Stage one of the Coal-Miner pipeline made use of RecHMM [30] to infer local-phylogeny-switching breakpoints. For computational efficiency, the breakpoint inference utilized a subset of taxa rather than the full set of taxa. The subset was chosen to maximize evolutionary divergence and was comprised of one sample from each of the following geographic regions: Spain, Sweden, USA, and Russia. For chromosomes 1 through 5, the analysis in stage one resulted in 1876, 991, 783, 559, and 913 loci with an average locus length of 16 kb, 19 kb, 30 kb, 33 kb, and 29 kb, respectively.

Using the loci obtained in stage one as input, the second stage of Coal-Miner was run on both trait characters. The 10 °C analysis identified 179, 99, 108, 109, and 95 candidate loci in chromosomes 1 through 5, respectively. The 16 °C analysis identified 115, 42, 88, 65, and 89 candidate loci in chromosomes 1 through 5, respectively. Coal-Miner also requires that ℓ^*, the number of candidate loci, be provided as an input parameter. In practice, model selection approaches are typically used in this context. Our study utilized the following procedure to determine a suitable value for ℓ^*. We calculated the likelihood score of the fitted "null" LMM for each locus (see above), and we examined the distribution of likelihood scores (Supplementary Figure S15 in the SOM). We then assigned ℓ^* based on the distribution's inflection point.

The inputs to the third stage of Coal-Miner consisted of the set of candidate loci, a quantitative trait character (flowering time at either 10 °C or 16 °C), and the genotypic sequence data matrix which consisted of sites with minor allele frequency threshold of 0.03 (i.e., sites having a minor allele frequency less than or equals to 0.03 were removed). The third stage of Coal-Miner was run using the same settings as in the simulation study.

3 RESULTS

3.1 Simulation study

We conducted experiments that varied the proportion of causal loci as a means to investigate the impact of the genomic architecture of a trait on AM method performance. The model conditions utilized simulations with between 10% and 30% causal loci and either neutral or non-neutral evolution on either tree-like or non-tree-like model phylogenies. The methods under study included Coal-Miner, our new AM method, as well as representative methods from different classes of state-of-the-art methods: Coal-Map, an AM method that accounts for local and global sample relatedness as fixed effects, GEMMA, a LMM-based AM method that accounts for global sample relatedness as a random effect (but does not account for local sample relatedness), and EIGENSTRAT, an AM method that accounts for global sample relatedness as a fixed effect (but does not account for local sample relatedness). We compared the statistical power and type I error control of each method using receiver operating characteristic (ROC) curves (Supplementary figures S2 through S5 in the SOM), and Table 1 compares the area under ROC curve (AUROC) of each method.

Regardless of the proportion of causal loci and the evolutionary scenario explored in these model conditions, Coal-Miner's AUROC was significantly better than the next best method in our study (either Coal-Map or GEMMA) based upon the corrected test of DeLong et al. [8] (Table 1). A similar observation was made when measuring performance using true positive rate (TPR) at a false positive rate (FPR) of 0.1 (Supplementary Table S2 in the SOM), except that Coal-Miner's performance advantage over the next best method was even more pronounced. The TPR difference was 0.158 on average and ranged as high as 0.248. Across these model conditions, we observed a consistent ranking of AM methods by AUROC (with two minor exceptions): Coal-Miner first, Coal-Map second, GEMMA third, and EIGENSTRAT fourth. The minor exceptions involved the two lowest AUROC values on the neutral, non-tree-like model condition with 10% or 20% causal loci, where GEMMA and EIGENSTRAT swapped rankings. We noted that Coal-Map's AUROC was second best on model conditions with the smallest proportion of causal loci, but its performance tended to degrade as the proportion increased. Coal-Map's AUROC was only marginally better than GEMMA on model conditions with the highest proportion of causal loci.

The impact of varying the proportion of causal loci was similar for all methods: AUROC tended to degrade as the proportion of causal loci increased from 10% to 30%. However, Coal-Miner's performance advantage relative to the other AM methods was flat or improved as the proportion of causal loci increased.

The model conditions included different combinations of genetic drift/incomplete lineage sorting and/or gene flow – evolutionary processes which can generate local variation in sample relatedness. Note that model conditions with non-tree-like model phylogenies incorporated all of these evolutionary processes (including genetic drift/incomplete lineage sorting). The impact of the different evolutionary processes varied across the methods. Coal-Miner's AUROC tended to be larger on model conditions involving both drift/ILS and gene flow as sources of local genealogical variation, and Coal-Map's AUROC was similarly affected. On the other hand, GEMMA's AUROC was comparable (within 0.01) based on this comparison, with the exception of non-neutral model conditions involving 10% or 20% causal loci.

A comparison of model conditions that differed only with respect to neutral versus non-neutral evolution revealed the impact of positive selection on AM method performance. We note that, in our experiments, the evolution of causal loci differed from non-causal loci since positive selection acted only upon the former but not the latter. Coal-Miner and Coal-Map returned comparable AUROC (within 0.025) regardless of neutral versus non-neutral evolution. GEMMA and EIGENSTRAT performed similarly, although slightly greater variability (within 0.035) was observed. For LMM-based methods, there was no obvious trend in terms of direction of change when comparing neutral versus non-neutral experiment results. There was an apparent trend for EIGENSTRAT, however: positive selection tended to reduce EIGENSTRAT's AUROC, with one exception (model conditions with a tree-like model phylogeny and 10% causal loci).

3.2 Empirical study

We used Coal-Miner to re-analyze an *Arabidopsis* dataset which was originally studied by the 1001 Genomes Consortium [7]. The dataset includes samples from 1,135 high quality re-sequenced natural lines adapted to different environments with varying local climates [7]. The sampled data included whole genome sequences and quantitative trait data for two traits: flowering time under high and low temperature – 16 °C and 10 °C, respectively.

A key component of the 1001 Genomes Consortium study was a GWA analysis of the genomic sequences and quantitative trait data using EMMAX [17], another state-of-the-art statistical AM method (see [31] for a comparison of EMMAX and other state-of-the-art statistical AM methods). A major focus of the analysis was a set of five genes which are known to regulate flowering and contribute to flowering time variation at 10 °C in *Arabidopsis* [7]: FLOWERING LOCUS T (FT), SHORT VEGETATIVE PHASE (SVP), FLOWERING LOCUS C (FLC), DELAY OF GERMINATION 1 (DOG1), and VERNALIZATION INSENSITIVE 3 (VIN3). Plants rely on both endogenous and environmental (e.g. temperature and photoperiod) cues to initiate flowering [1, 2]. These five genes encode major components of the vernalization (exposure to the prolonged cold) and autonomous pathways known to regulate the initiation of flowering in *Arabidopsis*. Allelic and copy number variants (CNV) for many of these genes, including FLC, are known to serve important roles in generating novel variation in flowering time and permit plants to adapt to new climates [21–23]. DOG1 is known to be involved in determining seasonal timing of seed germination and influences flowering time in Arabidopsis [15].

Under a conservative Bonferroni-corrected threshold [4], Coal-Miner identified significant peaks associated with flowering time under high and low temperature (16 °C and 10 °C, respectively). In particular, Coal-Miner identified significantly associated markers in all five genes (FT, SVP, FLC, DOG1, and VIN3) for both the 16 °C dataset and the 10 °C dataset (Supplementary Figure S12 in the SOM). Within the five genes, Coal-Miner analyses returned peaks which largely agreed across the 10 °C and 16 °C datasets. Some differences involved association scores that were borderline significant in one dataset but not the other.

Table 2 compares the Coal-Miner analysis with similar analyses using two other state-of-the-art statistical AM methods. The EMMAX analysis in the 1001 Genomes Consortium study [7] identified significant associations for three of the genes at 10 °C, and association score peaks were marginally below a Bonferroni-corrected threshold in the other two genes (SVP and FLC). Furthermore, significant peaks were only detected in DOG1 at 16 °C, but no significant peaks were detected in the other four genes for this dataset. (See Figure 2 in the 1001 Genomes Consortium study [7] for the original Manhattan plot.) GEMMA's performance was qualitatively similar to EMMAX (Supplementary Figure S13 in the SOM). At 10 °C,

Model condition			AUROC				
Neutral vs. non-neutral	Model phylogeny	Percentage of causal loci (%)	Coal-Miner	Coal-Map	GEMMA	EIGENSTRAT	q-value
Neutral	Non-tree-like	10	**0.962 (0.009)**	0.939 (0.009)	0.866 (0.017)	0.871 (0.014)	< 0.00001
		20	**0.921 (0.010)**	0.899 (0.009)	0.849 (0.015)	0.859 (0.012)	< 0.00001
		30	**0.904 (0.013)**	0.882 (0.009)	0.847 (0.017)	0.832 (0.018)	< 0.00001
Neutral	Tree-like	10	**0.943 (0.014)**	0.922 (0.010)	0.870 (0.009)	0.833 (0.019)	0.0053
		20	**0.904 (0.016)**	0.847 (0.011)	0.843 (0.010)	0.813 (0.016)	< 0.00001
		30	**0.904 (0.013)**	0.853 (0.009)	0.844 (0.008)	0.799 (0.022)	0.00003
Non-neutral	Non-tree-like	10	**0.959 (0.009)**	0.933 (0.013)	0.896 (0.014)	0.836 (0.022)	< 0.00001
		20	**0.926 (0.009)**	0.897 (0.009)	0.856 (0.017)	0.847 (0.013)	< 0.00001
		30	**0.894 (0.015)**	0.863 (0.010)	0.832 (0.018)	0.816 (0.014)	< 0.00001
Non-neutral	Tree-like	10	**0.954 (0.014)**	0.922 (0.010)	0.856 (0.012)	0.841 (0.018)	< 0.00001
		20	**0.890 (0.015)**	0.850 (0.013)	0.832 (0.014)	0.796 (0.020)	0.00003
		30	**0.879 (0.014)**	0.836 (0.011)	0.830 (0.009)	0.783 (0.018)	0.0007

Table 1: The impact of the genomic architecture of a quantitative trait on the performance of Coal-Miner and the other AM methods. Multi-locus sequences were simulated under neutral or non-neutral evolution on tree-like or non-tree-like model phylogenies, and quantitative traits were simulated using causal markers sampled from 10%, 20%, or 30% of loci (see Methods section for more details). The performance of each AM method was evaluated based on the area under its receiver operating characteristic (ROC) curve, or AUROC. We report each method's AUROC as an average (and standard error in parentheses) across twenty replicate datasets for each model condition. Coal-Miner's AUROC is shown in bold where it significantly improved upon the AUROC of the most accurate of the other AM methods, based upon the test of DeLong et al. [8] ($n = 20$; $\alpha = 0.05$). We corrected for multiple tests using the approach of Benjamini and Hochberg [3], and corrected q-values are shown. (The corresponding ROC plots are shown in Supplementary Figures S2 through S5 in the SOM.)

GEMMA recovered significant associations in three of the genes but not in the remaining two (SVP and FLC); at 16 °C, no significant peaks were detected in three genes, a peak just above the threshold of significance was detected in FT, and another peak was detected in DOG1.

4 DISCUSSION

Simulation study. For the model conditions that varied the proportion of causal loci with neutral or non-neutral evolution on tree-like or non-tree-like model phylogenies, Coal-Miner had better performance than all of the other state-of-the-art methods in our study, as measured using AUROC and TPR at an FPR of 0.1. This suggests that Coal-Miner's performance advantage is robust to the specific proportion of causal loci that contribute genetic effects to a quantitative trait, which relates to trait architecture, as well as the evolutionary processes involved. We note that, as even more causal loci are added beyond the proportions explored in our study, the effects contributed by any individual locus becomes more diffuse, and global sample structure will become a more reasonable approximation of different causal loci with different local sample structures. In general, we found traits with "diffuse" genomic architecture (i.e., traits with a relatively high proportion of causal loci) to be challenging for all methods. Coal-Miner tended to cope better with the challenge relative to the other methods in our study, which we attribute to the design of the second stage in the Coal-Miner pipeline (i.e., candidate locus detection). Consistent performance trends were observed when comparing neutral versus non-neutral simulations. This suggests that, for the model conditions that we explored in our study, Coal-Miner's performance is robust to the presence or absence of positive selection. A similar outcome was observed when comparing IUA model-based experiments involving two different types of model phylogenies – tree-like and non-tree-like.

Taken together, the model conditions included multiple sources of local genealogical variation, including genetic drift/ILS, gene flow, positive selection, and combinations thereof. The specific evolutionary processes contributing to local genealogical variation did not seem to matter as much as the presence of local genealogical variation, and Coal-Miner's performance advantage was not necessarily predicated on specific evolutionary cause(s) of local genealogical discordance. These findings seem to suggest that Coal-Miner's model and algorithm may be generalized to other evolutionary scenarios, so long as the breakpoint inference method used in the Coal-Miner pipeline suitably accounts for evolutionary processes with first-order contributions to genome evolution. An additional consideration is that the simulations utilized minor allele frequencies of at least 0.1, and future work is needed to understand Coal-Miner's performance in GWA studies involving rare variants.

Empirical study. The empirical datasets in our study were more challenging than the simulated datasets because the former likely involved more complex evolutionary evolutionary scenarios compared to the latter. Additional evolutionary processes which may have played an important role include other types of natural selection and demographic events (e.g., fluctuations in effective population size).

For both of the *Arabidopsis* datasets, Coal-Miner was able to detect significant associations in all five positive control regions. In contrast, neither GEMMA nor EMMAX – the statistical AM method used in the 1001 Genomes Consortium study [7] – were able to do the same. The vernalization requirement for flowering in

		Significantly associated markers detected?		
Dataset	Positive control gene	Coal-Miner	EMMAX	GEMMA
10 °C	FLOWERING LOCUS T (FT)	Yes	Yes	Yes
10 °C	SHORT VEGETATIVE PHASE (SVP)	Yes	No*	No
10 °C	FLOWERING LOCUS C (FLC)	Yes	No*	No
10 °C	DELAY OF GERMINATION 1 (DOG1)	Yes	Yes	Yes
10 °C	VERNALIZATION INSENSITIVE 3 (VIN3)	Yes	Yes	Yes
16 °C	FLOWERING LOCUS T (FT)	Yes	No	Yes
16 °C	SHORT VEGETATIVE PHASE (SVP)	Yes	No	No
16 °C	FLOWERING LOCUS C (FLC)	Yes	No	No
16 °C	DELAY OF GERMINATION 1 (DOG1)	Yes	Yes	Yes
16 °C	VERNALIZATION INSENSITIVE 3 (VIN3)	Yes	No	No

Table 2: A comparison of Coal-Miner and two other state-of-the-art statistical AM methods based upon analyses of the two *Arabidopsis* datasets. The other AM methods are GEMMA and EMMAX, the statistical AM method used in the 1001 Genomes Consortium study [7]. We evaluated whether the three AM methods detected significantly associated markers in five genomic regions centered on positive control genes which are known to regulate flowering time in *Arabidopsis*. We used a Bonferroni-corrected threshold for significance. For two of the five genomic regions in the 10 °C dataset, EMMAX returned association scores that were near the threshold of significance (marked using an asterisk). The corresponding Manhattan plots for the Coal-Miner and GEMMA analyses are shown in Supplementary Figures S12 and S13 in the SOM, respectively. The corresponding Manhattan plot for the EMMAX analysis is shown as Figure 2 in the 1001 Genomes Consortium study [7].

Arabidopsis suggests that the flowering response at 16 °C presents a greater AM challenge than at 10 °C. Our findings were consistent with a need for more statistical power for the former as compared with the latter as well as the overall findings in the simulation study, which suggested that Coal-Miner offered improved statistical power relative to the state of the art. As noted above, the empirical datasets likely involved relatively complex evolutionary histories as compared to the synthetic datasets in our study, and an expanded simulation study would be needed to confirm our initial comparison of performance findings using synthetic and empirical data. Furthermore, Coal-Miner analysis of the *Arabidopsis* dataset identified putatively novel markers (i.e., markers which were not flagged using other AM methods). Additional comparative and functional analyses are needed to interpret these findings.

5 CONCLUSIONS

Across the range of genomic architectures and evolutionary scenarios explored in our study, Coal-Miner had comparable or typically improved statistical power and type I error control compared to state-of-the-art AM methods. The scenarios included different evolutionary processes such as genetic drift and ILS, positive selection, gene flow, and recombination – all of which can generate local genealogical variation that differs from the true species phylogeny. More work needs to be done to explore additional evolutionary processes which have first-order impacts on genome evolution (e.g., gene duplication and loss, other genome rearrangement events, etc.). As more divergent samples are included in a GWA study, more evolutionary processes potentially will become relevant to AM analysis. We fully expect that more algorithmic development will need to be done in this case, particularly regarding the breakpoint inference stage of Coal-Miner.

We conclude with our thoughts on future work. As an alternative to the pipeline-based design of Coal-Miner, simultaneous inference of local coalescent histories and AM model parameters will avoid error propagation across different stages of a pipeline-based algorithm. Furthermore, viewed through the lens of evolution, genotype and phenotype are arguably two sides of the same coin. The same could be said of "intermediate-scale" characters (e.g., interactomic characters). A combination of the extended coalescent models and LMMs could be used to capture evolutionary relatedness of and functional dependence between heterogeneous biological characters across multiple scales of complexity and at higher evolutionary divergences.

6 SUPPORTING ONLINE MATERIALS (SOM)

SOM files are located at https://doi.org/10.6084/m9.figshare.5165470.v1. These materials include: (1) an appendix with supplementary text, tables, and figures, (2) source code for software used in this study, and (3) datasets analyzed in this study. All materials are provided under open and free licenses.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the following support: National Science Foundation Grants No. CCF-1565719 and No. CCF-1714417 (to KJL), grants from the BEACON Center for the Study of Evolution in Action (NSF STC Cooperative Agreement DBI-093954) to KJL and GAB, and Michigan State University faculty startup funds (to KJL, to GAB, and to PPE). The authors would also like to thank the anonymous referees for their valuable feedback and suggestions.

REFERENCES

[1] Richard Amasino. 2010. Seasonal and developmental timing of flowering. *The Plant Journal* 61, 6 (2010), 1001–1013.
[2] Richard M Amasino and Scott D Michaels. 2010. The timing of flowering. *Plant Physiology* 154, 2 (2010), 516–520.
[3] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* 57, 1 (1995), 289–300.
[4] Carlo E Bonferroni. 1936. *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber.

[5] Richard P. Brent. 1973. *Algorithms for Minimization without Derivatives*. Dover Publications, Mineola, New York. 1–208 pages.
[6] Peter E Chen and B Jesse Shapiro. 2015. The advent of genome-wide association studies for bacteria. *Current Opinion in Microbiology* 25 (2015), 17–24.
[7] 1001 Genomes Consortium. 2016. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 2 (2016), 481–491.
[8] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 3 (1988), 837–845.
[9] Eric Y. Durand, Nick Patterson, David Reich, and Montgomery Slatkin. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* 28, 8 (2011), 2239–2252.
[10] Scott V Edwards. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1 (2009), 1–19.
[11] Gregory Ewing and Joachim Hermisson. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* 26, 16 (2010), 2064–2065.
[12] Hussein A Hejase and Kevin J Liu. 2016. Mapping the genomic architecture of adaptive traits with interspecific introgressive origin: a coalescent-based approach. *BMC Genomics* 17, 1 (2016), 41.
[13] Richard R. Hudson. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 2 (2002), 337–338.
[14] Richard R Hudson and Norman L Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111, 1 (1985), 147–164.
[15] Heqiang Huo, Shouhui Wei, and Kent J. Bradford. 2016. DELAY OF GERMINATION1 (DOG1) regulates both seed dormancy and flowering time through microRNA pathways. *Proceedings of the National Academy of Sciences* 113, 15 (2016), E2199–E2206.
[16] Ian Jolliffe. 2002. *Principal Component Analysis*. Wiley Online Library.
[17] Hyun Min Kang, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-yee Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* 42, 4 (Apr 2010), 348–354.
[18] John F. C. Kingman. 1982. On the genealogy of large populations. *Journal of Applied Probability* 19 (1982), pp. 27–43.
[19] Kevin J. Liu, Jingxuan Dai, Kathy Truong, Ying Song, Michael H. Kohn, and Luay Nakhleh. 2014. An HMM-based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Computational Biology* 10, 6 (06 2014), e1003649.
[20] Anthony D Long and Charles H Langley. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research* 9, 8 (1999), 720–731.
[21] Dustin Mayfield, Z Jeffrey Chen, and J Chris Pires. 2011. Epigenetic regulation of flowering time in polyploids. *Current Opinion in Plant Biology* 14, 2 (2011), 174 – 178.
[22] Belén Méndez-Vigo, F. Xavier Picó, Mercedes Ramiro, José M. Martínez-Zapater, and Carlos Alonso-Blanco. 2011. Altitudinal and climatic adaptation is mediated by flowering traits and FRI, FLC, and PHYC genes in *Arabidopsis*. *Plant Physiology* 157, 4 (12 2011), 1942–1955.
[23] J. Chris Pires, Jianwei Zhao, M. Eric Schranz, Enrique J. Leon, Pablo A. Quijada, Lewis N. Lukens, and Thomas C. Osborn. 2004. Flowering time divergence and genomic rearrangements in resynthesized *Brassica* polyploids (Brassicaceae). *Biological Journal of the Linnean Society* 82, 4 (2004), 675–688.
[24] Stephanie S Porter, Peter L Chang, Christopher A Conow, Joseph P Dunham, and Maren L Friesen. 2017. Association mapping reveals novel serpentine adaptation gene clusters in a population of symbiotic *Mesorhizobium*. *The ISME Journal* 11, 1 (2017), 248–262.
[25] Sergej Potapov, Werner Adler, Benjamin Hofner, and Berthold Lausen. 2013. *Daim: Diagnostic accuracy of classification models*. R package version 1.1.0.
[26] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38, 8 (2006), 904–909.
[27] R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
[28] Ying Song, Stefan Endepols, Nicole Klemann, Dania Richter, Franz-Rainer Matuschka, Ching-Hua Shih, Michael W. Nachman, and Michael H. Kohn. 2011. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology* 21, 15 (2011), 1296 – 1301.
[29] Jeremy Wang, Kyle J. Moore, Qi Zhang, Fernando Pardo-Manual de Villena, Wei Wang, and Leonard McMillan. 2010. Genome-wide compatible SNP intervals and their properties. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology (BCB '10)*. ACM, New York, NY, USA, 43–52.
[30] Oscar Westesson and Ian Holmes. 2009. Accurate detection of recombinant breakpoints in whole-genome alignments. *PLoS Computational Biology* 5, 3 (03 2009), e1000318.
[31] Xiang Zhou and Matthew Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44, 7 (2012), 821–824.

Appendix with Supplementary Materials

Hussein A. Hejase & Natalie VandePol & Gregory M. Bonito
& Patrick P. Edger & Kevin J. Liu

Commands and software options used in simulation study

Simulation study

Neutral with non-tree-like model phylogeny. The following ms command was used to generate a multiple sequence alignment for the neutral model conditions with non-tree-like model phylogenies that include drift/ILS and gene flow:

```
ms 1000 10 -t 4.0 -s 250 -T -I 4 250 250 Ca Cb -ej 3.0 2 1 -ej 2.0 3
1 -ej 2.0 4 2
```

where the number of taxa is 1000, the number of gene trees is 10, the -t switch represents the mutation parameter $4N_0\mu$ where N_0 is the diploid population size ($N_0 = 2.5 \text{ x } 10^5$) and μ is the neutral mutation rate for a locus ($\mu = 4 \text{ x } 10^{-6}$), the number of segregating sites is 250, the -T parameter outputs the gene trees, which represent the evolutionary history of the sampled taxa. The -I parameter is followed by the number of subpopulations (k = 4) and a list of integers ($n\_A = 250$, $n\_B = 250$, $n\_Ca = Ca$, $n\_Cb = Cb$) that represent the number of taxa sampled for each subpopulation. Ca and Cb vary across loci and are dependent on γ. The -ej switch (-ej t i j) moves all lineages from subpopulation i to subpopulation j at time t.

We further investigated the impact of different admixture times by simulating two more datasets with admixture occurring at $t_1 = 1.0$ and $t_1 = 2.9$. We used the following ms commands to generate the aforementioned simulations:

ms 1000 10 -t 4.0 -s 250 -T -I 4 250 250 Ca Cb -ej 3.0 2 1 -ej **1.0** 3 1 -ej **1.0** 4 2

ms 1000 10 -t 4.0 -s 250 -T -I 4 250 250 Ca Cb -ej 3.0 2 1 -ej **2.9** 3 1 -ej **2.9** 4 2

Neutral with tree-like model phylogeny. The following ms command was used to generate a multiple sequence alignment for the neutral model conditions with tree-like model phylogenies that include drift/ILS:

```
ms 1000 10 -t 4.0 -s 250 -T -I 3 250 250 500 -ej 2.0 3 2 -ej 3.0 2 1
```

We further investigated the impact of different split times by simulating two more datasets with divergence occurring at $t_1 = 1.0$ and $t_1 = 2.9$. We used the following ms commands to generate the aforementioned simulations:

ms 1000 10 -t 4.0 -s 250 -T -I 3 250 250 500 -ej **1.0** 3 2 -ej 3.0 2 1
ms 1000 10 -t 4.0 -s 250 -T -I 3 250 250 500 -ej **2.9** 3 2 -ej 3.0 2 1

Isolation with migration. ms [1] was used to simulate a multiple sequence alignment for the neutral model conditions with non-tree-like model phylogenies incorporating an isolation-with-migration (IM) model of gene flow:

```
ms 1000 10 -t 4.0 -s 250 -T -I 3 250 250 500 -ej 2.0 3 2 -ej 3.0 2 1
```

```
-em 1 1 3 1
```

where the -em switch (-em t i j x) sets $4N_0 m_{ij}$ ($m_{ij} = 10^{-6}$) to x at time t and m_{ij} is the fraction of subpopulation i in each generation which consist of migrants from subpopulation j. The migration rate used in this simulation is inline with previous studies [2].

Recombination. We further simulated a multiple sequence alignment under the coalescent model with uniform recombination rate across a locus. We used a total sequence length of 2.5 kb, and a p parameter of 0.35, which is $4N_0r$, where r is the probability of cross-over per generation between the ends of the locus. The per-generation crossover probability of $10^{-9.85}$ between adjacent sites was used. Therefore, the probability of cross-over between the ends of the locus is: $10^{-9.85}$ x $(2500 - 1) = 3.5$ x 10^{-7} and $p = 4$ x 2.5 x 10^5 x 3.5 x $10^{-7} = 0.35$. On average, we obtained 10 recombinant regions per replicate.

The following ms command was used to generate a multiple sequence alignment for the neutral model conditions with tree-like model phylogenies incorporating recombination:

```
ms 1000 1 -t 4.0 -s 2500 -T -I 3 250 250 500 -ej 2.0 3 2 -ej 3.0 2 1
-r 0.35 2500
```

Non-neutral with non-tree-like model phylogeny. We used msms [3] to generate a forward-time simulation that explicitly modeled positive selection for the causal loci in the "neutral with non-tree-like model phylogeny" model conditions. The msms-based simulation utilized a sequence mutation model that allowed recurrent mutations between two alleles. Our forward-time coalescent simulation used a selection coefficient $s = 0.56$ which was based upon previously reported estimates from natural mouse populations that were involved in adaptive introgression linkage to emulate the genomic patterns of positive selection. The following msms command was used to generate a multiple sequence alignment for the non-neutral model conditions with non-tree-like model phylogenies that include drift/ILS, gene flow, and positive selection:

```
java -jar msms.jar 1000 <Number of causal loci> -t 4.0 -s 250 -T -I
4 250 250 Ca Cb 0 -ej 3.0 2 1 -ej 2.0 3 1 -ej 2.0 4 2 -SI 2.0 4 0 0
0 0 -Sc 0 4 11200 6272 0 -Sc 0 3 11200 6272 0 Smu 4.0 -N 10000
```

where the -SI switch (-SI t <number of populations> A B Ca Cb) sets the start of selection to time t forward in time from this point, the -Sc switch (-Sc t i α_{AA} α_{Aa} α_{aa}) sets the selection strength in population i pastward from time t to $2Ns$, the -Smu switch sets the forward mutation rate for the selected allele, and the -N switch is the effective population size.

Non-neutral with tree-like model phylogeny. The following msms command was used to generate a multiple sequence alignment for the non-neutral model conditions with tree-like model phylogenies that include drift/ILS and positive selection:

```
java -jar msms.jar 1000 <Number of causal loci> -t 4.0 -s 250 -T -I
3 250 250 500 0 -ej 2.0 3 2 -ej 3.0 2 1 -SI 2.0 3 0 0 0 -Sc 0 3 11200
6272 0 -Smu 4.0 -N 10000
```

EIGENSTRAT

EIGENSTRAT [4] utilizes a fixed effect model and uses Principal Component Analysis (PCA) to infer population structure in genetic data. From an n by m genotypic matrix X where n is the number of SNPs and m is the number of individuals, an m by m covariance matrix ϕ is computed. The top k principal components are defined as the top k eigenvectors of ϕ (e.g. k eigenvectors of the k largest eigenvalues). Using the top

k principal components as covariates, EIGENSTRAT corrects for population structure using the following:

$$X_{ij,adjusted} = X_{ij} - \alpha_i a_j \tag{1}$$

where i =1 to n, j =1 to m, α_i is the regression coefficient, and a_j is the axis of variation. After genetic and phenotypic adjustment based on the top principal components using equation (1), EIGENSTRAT applies a χ^2 association analysis between each genetic locus and the phenotype.

The following command was used to generate the principal components:

smartpca.perl -i example.geno -a example.snp -b example.ind -k 10 -q YES -o example.pca -p example.plot -e example.eval -l example.log -m 5 -t 2 -s 6

where the -i parameter specifies the genotype file, the -a parameter specifies the snp file, the -b parameter specifies the individual file, the k parameter specifies the number of principal components to output, the -q parameter specifies whether the phenotype is quantitative, the -o parameter specifies the output file of principal components, the -p parameter specifies the prefix of output plot files of top 2 principal components, the -e parameter specifies the output file of all eigenvalues, the -l parameter specifies the output log file, the -m parameter specifies the maximum number of outlier removal iterations, the -t parameter specifies the number of principal components along which to remove outliers, and the -s parameter specifies the number of standard deviations which an individual must exceed to be removed as an outlier.

The following command was used to apply the association analysis:

smarteigenstrat.perl -i example.geno -a example.snp -b example.ind -q YES -p example.pca -k 10 -o example.chisq -l example.log

where the -p parameter specifies the input file of principal components, the -k parameter specifies the number of principal components along which to correct for population structure, the -o parameter specifies the χ^2 association statistics, and the -l parameter specifies the standard output file.

GEMMA

We used GEMMA [5] which utilizes a linear mixed model to account for sample structure. GEMMA represents the phenotype Y as a function of fixed ($W\alpha + X\beta$) and random ($u + \epsilon$) effects:

$$\boldsymbol{y} = \boldsymbol{W\alpha} + x\beta + \boldsymbol{u} + \boldsymbol{\epsilon} \tag{2a}$$

$$\boldsymbol{u} \sim MVN_n(0, \lambda\tau^{-1}\boldsymbol{K}) \tag{2b}$$

$$\boldsymbol{\epsilon} \sim MVN_n(0, \tau^{-1}\boldsymbol{I_n}) \tag{2c}$$

where $\boldsymbol{y}$ is the phenotype vector, $\boldsymbol{W}$ includes the fixed effects, $\boldsymbol{\alpha}$ contains the coefficients of the covariates located in $\boldsymbol{W}$, x is the test locus, β is the effect size of x, $\boldsymbol{u}$ is a random effect that follows an n-dimensional multivariate normal distribution, $\boldsymbol{K}$ is a kinship matrix which is represented as a pairwise genotypic similarity between individuals, λ is the ratio between two variance components (genetic and environmental effects), τ is the variance of residual errors, $\boldsymbol{\epsilon}$ is a random effect that follows an n-dimensional multivariate normal distribution and is used to model any unexplained variation in $\boldsymbol{y}$, and $\boldsymbol{I_n}$ is an n by n identity matrix. The parameters $\hat{\boldsymbol{\alpha}}$, $\hat{\beta}$, $\hat{\tau}$, and $\hat{\lambda}$ are estimated using maximum likelihood where the association test statistics for $\boldsymbol{x_j}$ are generated using likelihood-ratio test between the fitted model against a null model with no SNP effect.

The following command was used to generate a kinship matrix:

gemma -g <specify input genotype file name> -p <specify input phenotype file name> -a <specify input SNPs annotation file name> -gk 1 <kinship/relatedness matrix type> -o <specify output file prefix>

The following command was used to run the association test:

gemma -g <specify input genotype file name> -p <specify input phenotype file name> -a <specify input SNPs annotation file name> -n 1 <specify phenotype column in the phenotype file> -maf 0 <specify minor allele frequency threshold> -r2 1 <specify r-squared threshold> -k <specify input kinship/relatedness matrix file name> -lmm 2 <specify frequentist analysis choice> -o <specify output file prefix>

Coal-Map

We applied Coal-Map [6] that models the local genealogical variation using a linear mixed model. Details on how to run Coal-Map are shown here https://gitlab.msu.edu/liulab/Coal-Map. We represented each of the global and local sample structures using five principal components.

Simulation study experiments involving alternative scenarios of neutral evolution

Multi-locus sequence evolution in our simulation study (see main manuscript) is impacted by genetic drift and incomplete lineage sorting, admixture, positive selection, and combinations of these processes. Our simulation study also included additional model conditions that involved alternative models of multi-locus sequence evolution. Each model condition was an extension of the above neutral model condition with 10% causal loci. One set of model conditions varied split time t_1 in the model tree shown in Figure S1 panel (a). Another set of model conditions varied admixture time t_1 in the model phylogeny network shown in Figure S1 panel (b), where $\gamma = 0.5$. The impact of recombination was explored in a model condition which made use of the coalescent-with-recombination model [7]. The simulations generated 2.5 kb alignments under a finite-sites model of recombination with per-generation crossover probability between adjacent sites of $10^{-9.85}$, which is 1-2 orders of magnitude smaller than estimates for mouse, rat and human [8]. We further explored the impact of gene flow using a model condition which substituted the isolation-with-migration model [9] in place of the IUA model.

Supplementary Table S1 shows an AUROC comparison of Coal-Miner and the other AM methods on the additional model conditions.

For model conditions that varied divergence time, involved recombination, or incorporated an isolation-with-migration (IM) model of gene flow, Coal-Miner returned significantly improved AUROC compared to the next best method based upon the test of DeLong *et al.* [10], and the other AM methods were ranked similarly to the experiments which varied the proportion of causal loci. A similar ranking was obtained when performance was measured using TPR at an FPR of 0.1 (Supplementary Table S3). Coal-Miner returned a comparable AUROC (within 0.027) as the divergence time t_1 increased from 1.0 to 2.9. The other methods performed similarly, except that the AUROC difference was larger (within 0.031). In the IM-based model condition, all methods returned AUROC that was comparable relative to experiments using the IUA model that were otherwise equivalent.

For IUA-based model conditions that varied the admixture time t_1, Coal-Map and Coal-Miner had comparable AUROC which was better than GEMMA and EIGENSTRAT. When comparing TPR at an FPR of 0.1, Coal-Miner returned a significant performance improvement relative to Coal-Map and the other AM methods (Supplementary Table S3). As seen in Supplementary Figures S8 and S9, Coal-Miner's TPR was better than Coal-Map when the false positive rate was 0.1 or less; the reverse was true only for large false positive rates (greater than around 0.15 for the $t_1 = 1.0$ model condition and greater than around 0.2 for the $t_1 = 2.9$ model condition). Among the AM methods in our study, Coal-Miner's AUROC was least impacted by the choice of admixture time and differed by at most 0.029 as the time t_1 increased from 1.0 to 2.9. The AUROC of the other AM methods became smaller as the admixture time became more ancient, and the AUROC difference was relatively greater than Coal-Miner (as much as 0.086).

Overall, Coal-Miner retained its performance advantage relative to the state-of-the-art, with one exception: Coal-Miner and Coal-Map had comparable AUROC on model conditions involving neutral evolution on non-tree-like model phylogenies and 10% causal loci, although Coal-Miner's TPR at an FPR of 0.1 was significantly better than Coal-Map's. These model conditions involved the smallest proportion of causal loci in our study. We note that Coal-Map's performance tended to degrade more rapidly than Coal-Miner as the proportion of causal loci increased, and the relative performance of the two methods may have changed for model conditions with higher proportions of causal loci that are otherwise equivalent.

Additional empirical datasets

To demonstrate the flexibility of the Coal-Miner framework, we conducted Coal-Miner analyses of three empirical datasets which spanned a range of GWAS settings. Each of the three datasets sampled taxa from a different kingdom and ranged from well-studied organisms to relatively novel organisms about which little is known. Specifically, the datasets sampled (1) natural populations of a single plant species (see main manuscript), (2) multiple closely related butterfly species where gene flow is a countervailing force versus genetic isolation, and (3) divergent bacterial species where horizontal gene transfer is suspected to be rampant. The datasets also varied in terms of the evolutionary processes with first-order impacts upon genome/phenotype evolution. The empirical analyses served two purposes: methodological validation using positive and negative controls based upon previous literature, and generation of new hypotheses for future study.

Heliconius erato dataset. We re-analyzed data from the study of Supple *et al.* [11]. The dataset includes 45 *H. erato* samples collected from four hybrid zones located in Peru, Ecuador, French Guiana, and Panama. Each sample exhibits one of two red phenotypes – postman and rayed – where 28 samples had the postman phenotype and 17 samples had the rayed phenotype. The genotypic data were sequenced from the 400 kb genomic region referred to as the D interval in *H. erato*. The D interval spans 56,862 biallelic SNPs and is known to modulate red phenotypic variation. Coal-Miner was run on the *H. erato* dataset using the same approach as in the *Arabidopsis* dataset analysis (see main manuscript). The first stage of Coal-Miner identified seven loci and the second stage inferred a single candidate locus.

Burkholderiaceae dataset. Bacteria belonging to the *Burkholderiaceae* are of interest given their importance in human and plant disease, but also given their role as plant and fungal endosymbionts and their metabolic capacity to degrade xenobiotics. Fully sequenced (closed) genomes belonging to *Burkholderiaceae* were selected and downloaded from the PATRIC web portal (www.patricbrc.org/) [12]. Supplementary Table S8 lists sampled species names along with other information (IDs, groups, and pathogenicity). We chose to maximize phylogenetic and ecological diversity in this sampling, so we included available genomes belonging to free-living, pathogenic, and endosymbiotic species spanning across the genera *Burkholderia*, *Ralstonia*, *Pandoraea*, *Cupriavidus*, *Mycoavidus*, and *Polynucleobacter*. A total of 57 samples were included, of which 52 samples were free-living and 5 were endosymbionts. Genomes ranged in size from 1.56 Mb to 9.70 Mb and spanned between 2,048 and 9,172 coding DNA sequences (CDS). The software package Proteinortho [13] was run using default parameters to detect single copy orthologs in the selected genomes. A total of 549 orthologs were recovered in the Proteinortho analysis. We analyzed a phenotype that identified each sample's status as either an animal pathogen or non-animal pathogen. Coal-Miner was used to analyze the genomic sequence data and phenotypic character using the same approach as in the other empirical analyses (see above). The initial stages of Coal-Miner identified 55 candidate loci. Genes with significant associations based upon the Coal-Miner analysis were further classified based upon their Gene Ontology [14] and KEGG [15] pathway assignments.

Coal-Miner re-analysis of the Heliconious erato dataset. Supplementary Figure S14 displays the Manhattan plot generated after applying Coal-Miner on the *H. erato* dataset across the D interval. We identified two significant peaks ranging from 502 kb to 592 kb and 658 kb to 682 kb, respectively. The second peak is located at the 3′ of the optix transcription factor, a gene previously shown to be behind the red phenotype variation in *Heliconius* [11]. The first peak is located in a noncoding region more distant from the 3′ of the optix transcription factor.

Coal-Miner analysis of the Burkholdericeae dataset. We applied Coal-Miner on an empirical dataset of complete genomes of bacteria belonging to the *Burkholderiaceae* and spanning a diversity of ecological states including animal and plant pathogens. Supplementary Table S7 shows the genes inferred by Coal-Miner to be associated with human pathogenicity, along with their inferred KEGG pathway and gene ontology assignments. In total, we identified 16 genes associated with human pathogenicity in *Burkholderia*. Four of these genes have been implicated in pathogenicity by others, and in some cases validated through gene knockout and experimental evolution experiments. For example, the cell division protein FtsK that Coal-Miner associated with human pathogenicity was found to be one of three genes under positive selection in *Burkholderia multivorans* during a 20-year cystic fibrosis infection [16]. Modifications of another gene identified by Coal-Miner, DNA gyrase subunit A, are well known to be implicated with virulence and antibiotic resistance to quinolone and ciprofloxacin in pathogenic *Burkholderia* [17, 18]. For example, Lieberman *et al.* [19] found that the DNA gyrase subunit A gene was under positive selection during a *Burkholderia dolosa* outbreak among multiple patients with cystic fibrosis [19]. Another gene identified by Coal-

Miner, Excinuclease ABC subunit A, has been shown to bind to previously published vaccine targets [20]. Coal-Miner also associated the protein dihydrofolate synthase with animal pathogenicity. Point mutations leading to nonsynonymous base changes in the dihydrofolate reductase gene have previously been demonstrated to be associated with trimethoprim resistance in cystic fibrosis patients infected by *Burkholderia cenocepacia* [21, 22].

Inferring local-phylogeny-switching break-points

The local partition breakpoint vector b for the simulated data required as input to Coal-Miner was inferred using the Four-Gamete Test [23], which identifies segregating sites that did not arise without either recombination or a repeat mutation. The Four-Gamete Test is an appropriate choice to detect breakpoints due to the simplifying assumptions of our simulation study (infinite sites model, free recombination between markers, and complete linkage within each marker).

For our empirical studies, we used RecHMM [24], an HMM-based method for computing local-phylogeny-switching breakpoints. The following command was used to run RecHMM:

> ./runTraining.py <FASTA input alignment> -lb -prefix <empty existing working directory> -k 2 <number of hidden states> -lt

Using 2 states for the -k option corresponds to two parental trees for the model network.

Sensitivity of breakpoint inference

We tested the sensitivity of inferring local-phylogeny-switching breakpoints using a variant of the LRScan algorithm [25] on the performance of Coal-Miner using neutral and non-tree-like model conditions that included 10% causal loci. The AUROC results for Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT were 0.962, 0.939, 0.866, and 0.871, respectively, with Coal-Miner significantly more accurate (q-value < 0.00001) than the next best method (Coal-Map). The Coal-Miner results are similar to the results obtained when using the LRScan algorithm for breakpoint inference, which suggests that Coal-Miner is robust to breakpoint inference.

Leaving-one-chromosome-out (LOCO) analysis

We demonstrate that Coal-Miner works better than an approach that performs standard linear mixed model association analysis, where the relatedness is controlled for all other loci when testing for each SNP. The performance advantage of Coal-Miner over the leaving-one-chromosome-out standard approach was significantly more accurate by 0.066 based on AUROC on model conditions that included neutral and non-tree-like model phylogenies with 10% causal loci.

Running time

We explored the running time of different stages of Coal-Miner using neutral and non-tree-like model conditions that included 10% causal loci. On average, the running time of the Coal-Miner pipeline across twenty replicates was 1.43 hours (with standard error of 0.02). The first stage of Coal-Miner, which involves inferring local-phylogeny-switching breakpoints, took approximately 70 minutes to complete while the other stages (stages two and three) took no more than 15 minutes to complete the analysis. These results suggest that stage one of Coal-Miner could pose a performance bottleneck as either the number of taxa or length of the multiple sequence alignment increases. We recommend performing sampling of taxa to mitigate the computational impact of this dimension of scale on the computational performance of stage one.

Impact of the number of loci

The impact of the number of loci on the performance of Coal-Miner was explored using a simulation study containing 20 loci (250 sites per locus) for neutral and non-tree-like model conditions that included 10% causal loci. The AUROC results for Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT were 0.965, 0.933, 0.841, and 0.866, respectively. Coal-Miner was significantly more accurate compared to the next most accurate method (Coal-Map) using a corrected test of DeLong *et al.* [10]. These results suggest that increasing the number of loci from 10 to 20 preserves the performance advantage of Coal-Miner over the other AM methods.

Additional trait model

We simulated a continuous additive trait using the following:

$$y = X\beta + \epsilon$$

where X is an n by p genotype matrix at p causal SNPs. Twenty causal SNPs were randomly selected from causal loci such that each causal locus contained at least one causal SNP and causal SNPs had minor allele frequency between 0.1 and 0.3. β follows a normal distribution with mean of zero and variance of $\frac{h^2}{p}$ (h is the heritability of the trait), and ϵ is the residual effect generated from a normal distribution with mean of 0 and variance of $(X\beta)$ x $(\frac{1}{h^2} - 1)$.

We explored the performance of Coal-Miner and the other methods on neutral and non-tree-like model conditions that included 10% causal loci using the above continuous additive trait model across a range of heritability values (h = 0.25, 0.5, and 0.75). For h = 0.5, the AUROC results for Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT were 0.952, 0.924, 0.865, and 0.869, respectively, with Coal-Miner significantly more accurate than the next most accurate method (Coal-Map) using a corrected test of DeLong *et al.* [10]. Furthermore, the TPR values at an FPR value of 0.1 for Coal-

Miner, Coal-Map, GEMMA, and EIGENSTRAT were 0.915, 0.799, 0.689, and 0.587, respectively. The results for the other heritability values are shown in Tables S5 and S6. These results suggest that Coal-Miner performs better than the other AM methods under a more complex continuous trait model with varying effect size. We compared the performance of Coal-Miner under the above trait model with $h = 0.5$ relative to $h \sim 0$. Results are shown in Figure S17, suggesting that Coal-Miner's performance is better relative to the null.

Bonferroni correction

The simulations used Benjamini-Hochberg procedure [26] to correct for multiple tests. We further used Bonferroni corrected by the number of independent loci, which is 10, for multiple tests correction. The results are shown in Table S4 and are consistent with the results shown in Table 1.

Vary recombination rate

A multiple sequence alignment under the coalescent model with uniform recombination rate across a locus, with a total sequence length of 10 kb and a p parameter of 1, was simulated using ms. The model condition included 10% causal loci. The AUROC results for Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT were 0.827, 0.772, 0.799, and 0.732, respectively, with Coal-Miner significantly more accurate than the next most accurate method (GEMMA) using a corrected test of DeLong *et al.* [10]. Furthermore, the TPR values at an FPR value of 0.1 for Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT were 0.485, 0.317, 0.446, and 0.296, respectively. These results suggest that increasing the recombination rate does not impact the performance advantage of Coal-Miner over the other AM methods.

Vary sequence length

We demonstrate the performance of Coal-Miner relative to the other AM methods by simulating larger sequence length datasets using neutral and non-tree-like model conditions that included 10% causal loci. Furthermore, the simulations contained 10 loci per replicate (0.1 Mb per locus), with a total sequence length of 1 Mb. The AUROC results for Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT were 0.946, 0.889, 0.862, and 0.851, respectively, with Coal-Miner significantly more accurate relative to the next most accurate method (Coal-Map) using a corrected test of DeLong *et al.* [10]. Furthermore, the TPR values at an FPR value of 0.1 for Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT were 0.923, 0.750, 0.739, and 0.513, respectively. These results suggest that increasing the sequence length preserves the performance advantage of Coal-Miner over the other AM methods.

SI Tables

	AUROC				
Model condition	Coal-Miner	Coal-Map	GEMMA	EIGENSTRAT	q-value
Non-tree-like model phylogeny with admixture time $t_1 = 1.0$	0.959	**0.963**	0.922	0.905	0.9959
Non-tree-like model phylogeny with admixture time $t_1 = 2.9$	**0.933**	0.922	0.836	0.843	< 0.00001
Tree-like model phylogeny with split time $t_1 = 1.0$	**0.959**	0.899	0.884	0.852	< 0.00001
Tree-like model phylogeny with split time $t_1 = 2.9$	**0.932**	0.895	0.853	0.849	< 0.00001
Coalescent-with-recombination	**0.841**	0.768	0.77	0.738	< 0.00001
Isolation-with-migration	**0.953**	0.931	0.881	0.868	< 0.00001

Supplementary Table S1: **Additional evolutionary scenarios exploring other evolutionary processes that can generate local genealogical variation.** The additional model conditions were variants of the model condition with neutral evolution on a tree-like model phylogeny and 10% causal loci (see Table 1 in the main manuscript). Each model condition incorporated an alternative evolutionary scenario (see Methods section for more details). Otherwise, table layout and description are identical to Table 1 in the main manuscript.

Model condition	Model phylogeny	Percentage of causal loci (%)	TPR				q-value
			Coal-Miner	Coal-Map	GEMMA	EIGENSTRAT	
Neutral	Non-tree-like	10	**0.934**	0.806	0.678	0.654	0.0102
		20	**0.823**	0.645	0.667	0.539	0.0003
		30	**0.785**	0.616	0.63	0.546	0.0005
Neutral	Tree-like	10	**0.782**	0.76	0.582	0.51	0.2299
		20	**0.804**	0.488	0.556	0.443	0.00003
		30	**0.766**	0.549	0.556	0.489	$< 10^{-5}$
Non-neutral	Non-tree-like	10	**0.934**	0.841	0.698	0.619	0.0062
		20	**0.841**	0.639	0.651	0.505	0.0002
		30	**0.788**	0.527	0.645	0.455	0.0016
Non-neutral	Tree-like	10	**0.93**	0.725	0.561	0.524	0.0033
		20	**0.714**	0.532	0.523	0.426	0.0009
		30	**0.643**	0.469	0.483	0.426	0.0006

Supplementary Table S2: **The true positive rate (TPR) values of Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT at false positive rate (FPR) value of 0.1 across different model conditions.** The most accurate method for each model condition is highlighted in bold. We report the corrected q-value of the performance advantage of Coal-Miner over the next most accurate method, which were compared using a pairwise t-test with Benjamini-Hochberg correction [26].

Model condition	Model phylogeny	Percentage of causal loci (%)	TPR				q-value
			Coal-Miner	Coal-Map	GEMMA	EIGENSTRAT	
Neutral	Non-tree-like with admixture-time $t_1 = 1.0$	10	**0.913**	0.894	0.822	0.771	0.5572
Neutral	Non-tree-like with admixture-time $t_1 = 2.9$	10	**0.885**	0.726	0.676	0.599	0.0002
Neutral	Tree-like with split-time $t_1 = 1.0$	10	**0.888**	0.726	0.633	0.55	0.0018
Neutral	Tree-like with split-time $t_1 = 2.9$	10	**0.827**	0.674	0.542	0.519	0.0463
Neutral	Tree-like with recombination	10	**0.493**	0.328	0.379	0.27	0.073
Neutral	Non-tree-like with isolation-with-migration	10	**0.89**	0.727	0.712	0.607	0.029

Supplementary Table S3: **The true positive rate (TPR) values of Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT at false positive rate (FPR) value of 0.1 across different model conditions.** Table layout and description are otherwise similar to Supplementary Table S2.

Model condition			q-value
Neutral vs. non-neutral	Model phylogeny	Percentage of causal loci (%)	
Neutral	Non-tree-like	10	< 0.00001
		20	< 0.00001
		30	< 0.00001
Neutral	Tree-like	10	0.0006
		20	< 0.00001
		30	0.00008
Non-neutral	Non-tree-like	10	< 0.00001
		20	< 0.00001
		30	< 0.00001
Non-neutral	Tree-like	10	< 0.00001
		20	0.00009
		30	0.002

Supplementary Table S4: **Coal-Miner's AUROC improvement upon the AUROC of the most accurate of the other AM methods, based upon the test of [10] ($n = 20$; $\alpha = 0.05$). We corrected for multiple tests using the approach of Bonferroni, and corrected q-values are shown.**

	AUROC				
Heritability value	Coal-Miner	Coal-Map	GEMMA	EIGENSTRAT	q-value
0.25	**0.949**	0.913	0.860	0.863	$< 10^{-5}$
0.5	**0.952**	0.924	0.865	0.869	$< 10^{-5}$
0.75	**0.940**	0.934	0.887	0.848	$< 10^{-5}$

Supplementary Table S5: **The AUROC values of Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT across different heritability trait values.** The most accurate method for each model condition is highlighted in bold. We report Coal-Miner's AUROC improvement upon the AUROC of the most accurate of the other AM methods, based upon the test of [10] ($n = 20$; $\alpha = 0.05$). We corrected for multiple tests using the approach of Benjamini-Hochberg [26], and corrected q-values are reported.

	TPR				
Heritability value	Coal-Miner	Coal-Map	GEMMA	EIGENSTRAT	q-value
0.25	**0.912**	0.788	0.694	0.576	0.0092
0.5	**0.915**	0.799	0.689	0.587	0.0022
0.75	0.817	**0.833**	0.682	0.580	0.726

Supplementary Table S6: **The true positive rate (TPR) values of Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT at false positive rate (FPR) value of 0.1 across different heritability trait values.** The most accurate method for each model condition is highlighted in bold. We report the corrected q-value of the performance advantage of Coal-Miner over the next most accurate method, which were compared using a pairwise t-test with Benjamini-Hochberg correction [26].

Supplementary Table S7: **Empirical study results involving bacteria belonging to the *Burkholderiaceae*.** Results are shown for proteins inferred to be associated with human pathogenicity along with their KEGG pathway and gene ontology assignments.

Proteins	Pathway Assignments	Gene Ontology Assignments
Dihydrofolate synthase	KEGG:00790 Folate biosynthesis	GO:0008841 dihydrofolate synthase activity, GO:0004326 tetrahydrofolylpolyglutamate synthase activity

Aspartokinase	KEGG:00260 Glycine, serine and threonine metabolism, KEGG:00270 Cysteine and methionine metabolism, KEGG:00300 Lysine biosynthesis	GO:0004072 aspartate kinase activity
NADH-ubiquinone oxidoreductase chain G	KEGG:00190 Oxidative phosphorylation, KEGG:00910 Nitrogen metabolism	GO:0008137 NADH dehydrogenase (ubiquinone) activity
Excinuclease ABC subunit A	-	GO:0005524 ATP binding, GO:0016887 ATPase activity
Carboxyl-terminal protease	-	-
Homoserine O-acetyltransferase	KEGG:00270 Cysteine and methionine metabolism, KEGG:00920 Sulfur metabolism	GO:0004414 homoserine O-acetyltransferase activity
Glutamate-ammonia-ligase adenylyltransferase	-	GO:0008882 [glutamate-ammonia-ligase] adenylyltransferase activity
Undecaprenyl-diphosphatase	KEGG:00550 Peptidoglycan biosynthesis	GO:0050380 undecaprenyl-diphosphatase activity
Cell division protein FtsK	-	-
DNA gyrase subunit A	-	GO:0003918 DNA topoisomerase (ATP-hydrolyzing) activity
Diaminohydroxyphosphoribosylaminopyrimidine deaminase	KEGG:00740 Riboflavin metabolism	GO:0008703 5-amino-6-(5-phosphoribosylamino)uracil reductase activity, GO:0008835 diaminohydroxyphosphoribosylaminopyrimidine deaminase activity
Ribonucleotide reductase of class Ia (aerobic), alpha subunit	KEGG:00230 Purine metabolism, KEGG:00240 Pyrimidine metabolism, KEGG:00480 Glutathione metabolism	GO:0004748 ribonucleoside-diphosphate reductase activity
DNA gyrase subunit B	-	GO:0003918 DNA topoisomerase (ATP-hydrolyzing) activity
Ketol-acid reductoisomerase	KEGG:00290 Valine, leucine and isoleucine biosynthesis, KEGG:00770 Pantothenate and CoA biosynthesis	GO:0004455 ketol-acid reductoisomerase activity
Phosphoribosylformylglycinamidine synthase, synthetase subunit	KEGG:00230 Purine metabolism	GO:0004642 phosphoribosylformylglycinamidine synthase activity
DNA polymerase I	KEGG:00230 Purine metabolism, KEGG:00240 Pyrimidine metabolism	GO:0003887 DNA-directed DNA polymerase activity

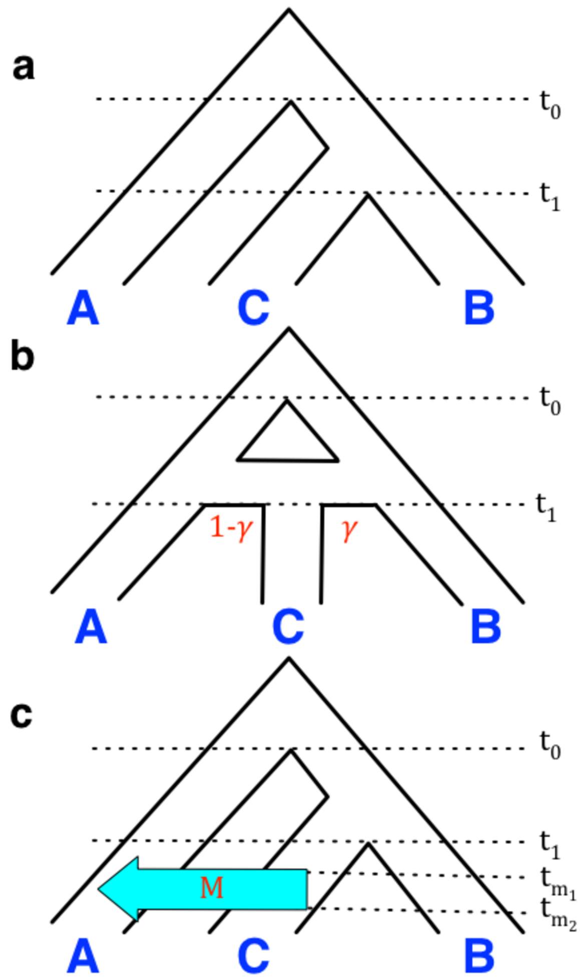
Supplementary Table S8: **Empirical study data involving bacteria belonging to the *Burkholderiaceae*.** The PATRIC accession numbers along with the species names, and group and pathogenicity categories are shown.

PATRIC Number	Species	Group	Pathogenicity
13373.18	Burkholderia-mallei	Ingroup	Human and animal pathogen
13373.19	Burkholderia-mallei	Ingroup	Human and animal pathogen
882378.3	Burkholderia-rhizoxinica	Ingroup	Fungal endosymbiont, plant pathogen
224135.3	Glomeribacter-endosymbiont-AG77	Ingroup	Fungal endosymbiont
452638.4	Polynucleobacter-necessarius-subsp-necessarius-STIR1	Outgroup	Freshwater bacterium, endosymbiont of protist
339670.8	Burkholderia-ambifaria	Ingroup	Opportunistic animal pathogen (cystic fibrosis)

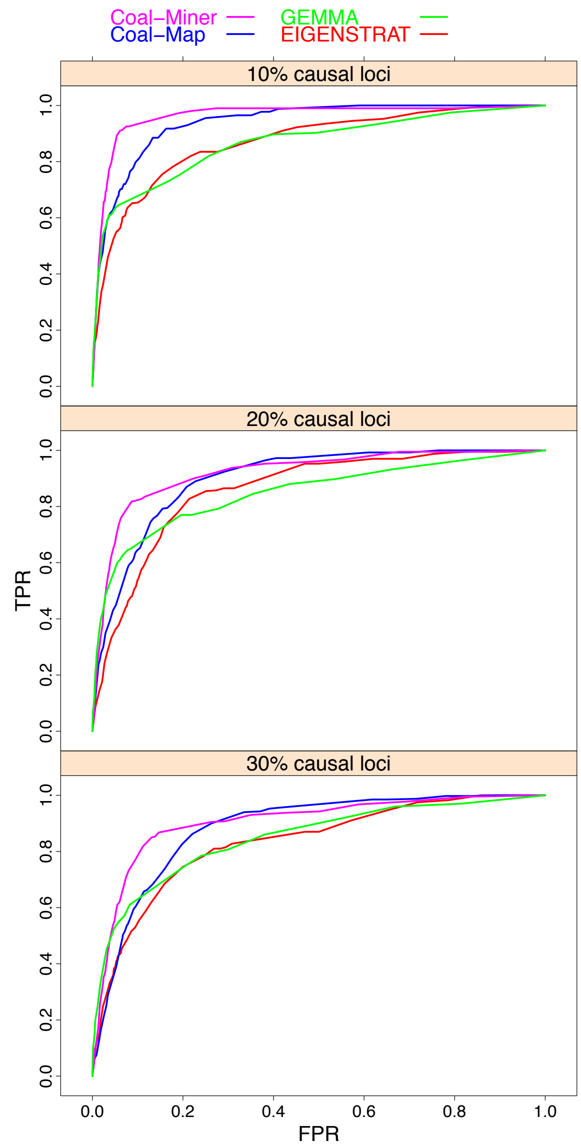
398577.6	Burkholderia-ambifaria-MC40-6	Ingroup	Opportunistic animal pathogen (cystic fibrosis)
331271.8	Burkholderia-cenocepacia	Ingroup	Opportunistic animal pathogen (cystic fibrosis), plant pathogen
95486.54	Burkholderia-cenocepacia	Ingroup	Opportunistic animal pathogen (cystic fibrosis), plant pathogen
1009846.3	Burkholderia-cepacia	Ingroup	Opportunistic animal pathogen (cystic fibrosis), plant pathogen
999541.3	Burkholderia-gladioli	Ingroup	Opportunistic animal pathogen, opportunistic plant pathogen, plant and fungal symbiont
626418.3	Burkholderia-glumae-BGR1	Ingroup	Plant pathogen
595500.3	Burkholderia-glumae-PG1	Ingroup	Plant pathogen
87883.36	Burkholderia-multivorans	Ingroup	Opportunistic animal pathogen (cystic fibrosis)
395019.8	Burkholderia-multivorans-ATCC-17616	Ingroup	Opportunistic animal pathogen (cystic fibrosis)
342113.3	Burkholderia-oklahomensis	Ingroup	Opportunistic human pathogen
391038.7	Burkholderia-phymatum-STM815	Ingroup	Plant symbiont (N-fixation)
398527.4	Burkholderia-phytofirmans-PsJN	Ingroup	Plant symbiont
1435365.3	Burkholderia-pseudomallei	Ingroup	Human and animal pathogen
28450.84	Burkholderia-pseudomallei	Ingroup	Human and animal pathogen
28450.87	Burkholderia-pseudomallei	Ingroup	Human and animal pathogen
1487955.3	Burkholderia-sp-BGK	Ingroup	Unknown
640510.4	Burkholderia-sp-CCGE1001	Ingroup	Unknown
640511.6	Burkholderia-sp-CCGE1002	Ingroup	Plant symbiont
640512.4	Burkholderia-sp-CCGE1003	Ingroup	Unknown
416344.3	Burkholderia-sp-KJ006	Ingroup	Plant-endophyte and symbiont (growth promoter)
758793.3	Burkholderia-sp-RPE64	Ingroup	Insect endosymbiont (acquired from soil)
758796.3	Burkholderia-sp-RPE67	Ingroup	Insect endosymbiont (acquired from soil)
1439853.3	Burkholderia-sp-TSV202	Ingroup	Unknown
1097668.3	Burkholderia-sp-YI23	Ingroup	soil (xenobiotic degrader)
1241582.3	Burkholderia-thailandensis	Ingroup	Plant-associate, opportunistic human pathogen
57975.4	Burkholderia-thailandensis	Ingroup	Plant-associate, opportunistic human pathogen
269482.11	Burkholderia-vietnamiensis	Ingroup	Opportunistic animal pathogen, opportunistic plant pathogen, plant symbiont
266265.5	Burkholderia-xenovorans	Ingroup	Soil bacterium (degrades xenobiotics)
36873.6	Burkholderia-xenovorans	Ingroup	Soil bacterium (degrades xenobiotics)
68895.5	Cupriavidus-basilensis-4G11	Outgroup	Unknown
266264.9	Cupriavidus-metallidurans-CH34	Outgroup	Soil bacterium (degrades xenobiotics; metal tolerant)
1042878.5	Cupriavidus-necator-N-1	Outgroup	Soil bacterium (metal tolerant); bacterium and fungal predator
164546.7	Cupriavidus-taiwanensis	Outgroup	Plant symbiont (N-fixation)
93218.7	Pandoraea-apista-TF81F4	Outgroup	Opportunistic animal pathogen (cystic fibrosis)
93220.9	Pandoraea-pnomenusa	Outgroup	Opportunistic animal pathogen (cystic fibrosis)
1416914.3	Pandoraea-pnomenusa-3kgm	Outgroup	Opportunistic animal pathogen (cystic fibrosis)
93221.4	Pandoraea-pulmonicola-DSM-16583	Outgroup	Opportunistic animal pathogen (cystic fibrosis)
1380774.3	Pandoraea-sp-RB-44	Outgroup	Soil bacterium
93222.6	Pandoraea-sputorum-DSM-21091	Outgroup	Opportunistic animal pathogen

312153.5	Polynucleobacter-necessarius-subsp-asymbioticus	Outgroup	Freshwater bacterium
381666.6	Ralstonia-eutropha-H16	Outgroup	Soil bacterium
264198.6	Ralstonia-eutropha-JMP134	Outgroup	Soil bacterium
428406.5	Ralstonia-pickettii-12D	Outgroup	Soil and freshwater bacterium, opportunistic human pathogen
1366050.3	Ralstonia-pickettii-DTP0602	Outgroup	Soil and freshwater bacterium, opportunistic human pathogen
859656.5	Ralstonia-solanacearum-CFBP2957	Outgroup	Plant pathogen
859655.3	Ralstonia-solanacearum-CMR15	Outgroup	Plant pathogen
1262456.3	Ralstonia-solanacearum-FQY_4	Outgroup	Plant pathogen
267608.8	Ralstonia-solanacearum-GMI1000	Outgroup	Plant pathogen
564065.5	Ralstonia-solanacearum-MolK2	Outgroup	Plant pathogen
1031711.3	Ralstonia-solanacearum-Po82	Outgroup	Plant pathogen
859657.5	Ralstonia-solanacearum-PSI07	Outgroup	Plant pathogen

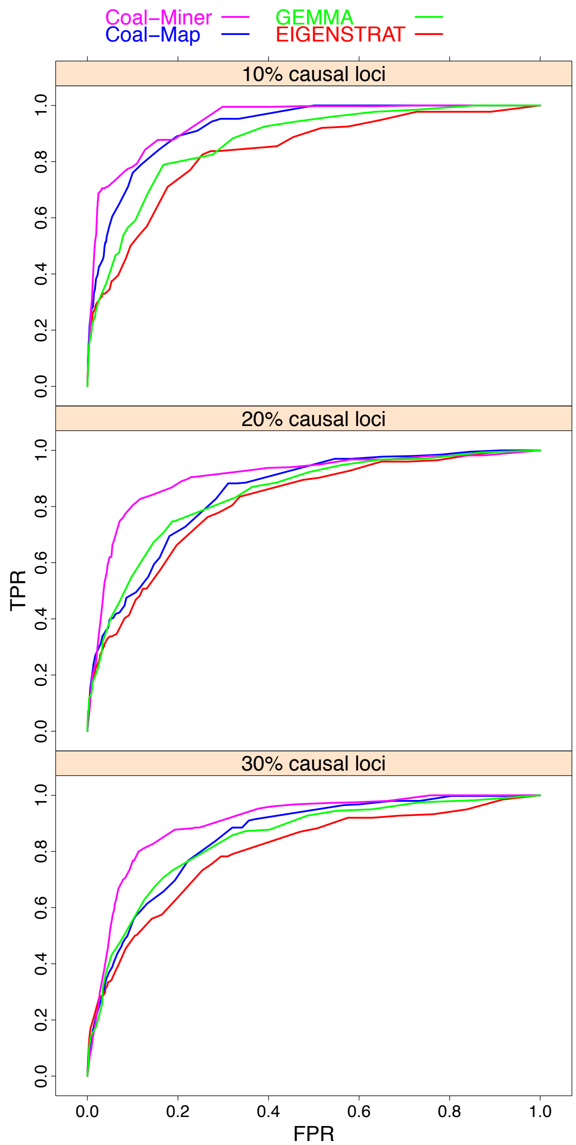
SI Figures



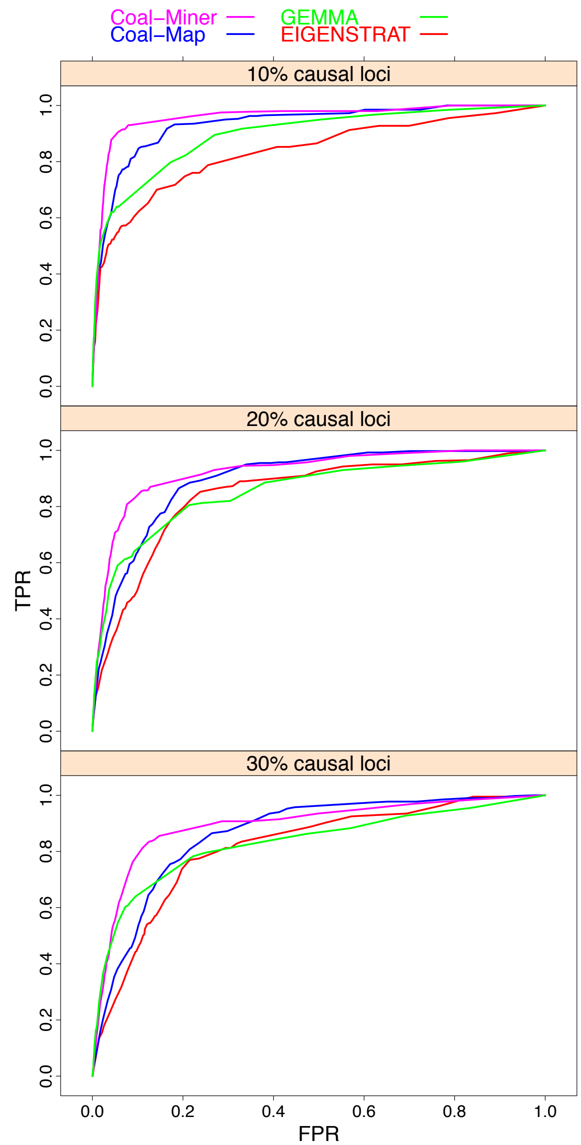
Supplementary Figure S1: **Model phylogenies used in the simulation study.** (a) Tree-like phylogeny, (b) Non-tree-like phylogeny with instantaneous unidirectional admixture (IUA), and (c) Non-tree-like phylogeny incorporating an isolation-with-migration (IM) model of gene flow.



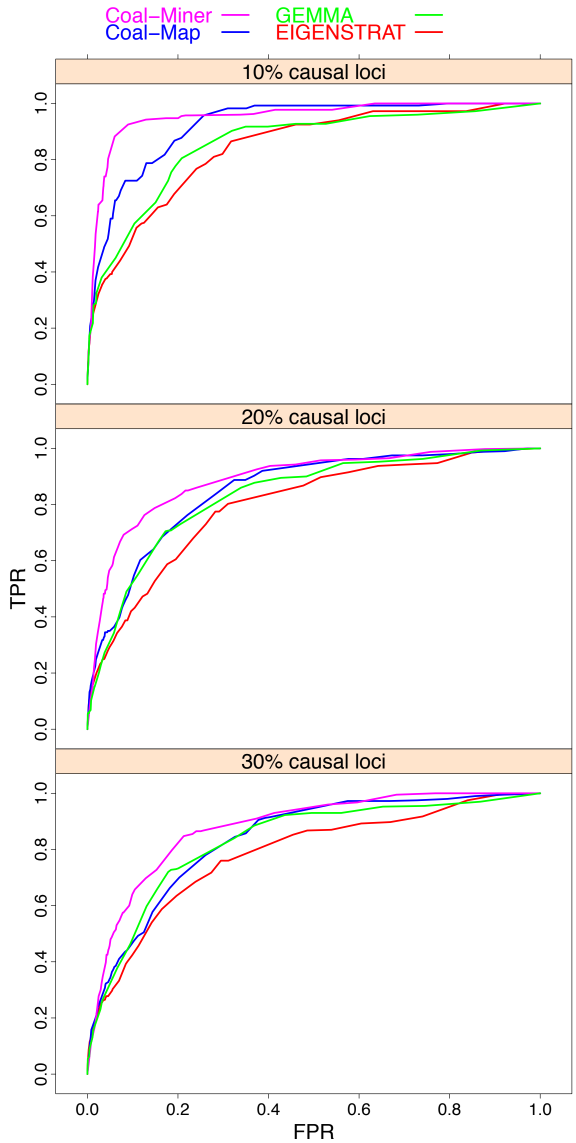
Supplementary Figure S2: **Results for neutral model conditions with non-tree-like model phylogenies that include drift/ILS and gene flow ($\gamma = 0.5$).** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. The receiver operating characteristic (ROC) curve shows the relationship between false positive rate (FPR) versus the true positive rate (TPR). Results are shown for three genomic architectures of quantitative traits with proportion of causal loci of 10%, 20%, and 30%, respectively.



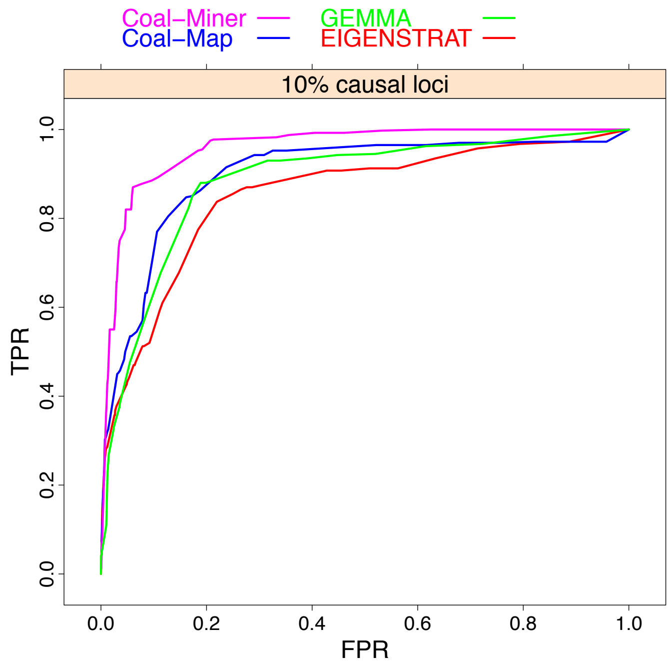
Supplementary Figure S3: **Results for neutral model conditions with tree-like model phylogenies that include drift/ILS ($\gamma = 0$).** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. Figure layout and description are otherwise similar to Supplementary Figure S2.



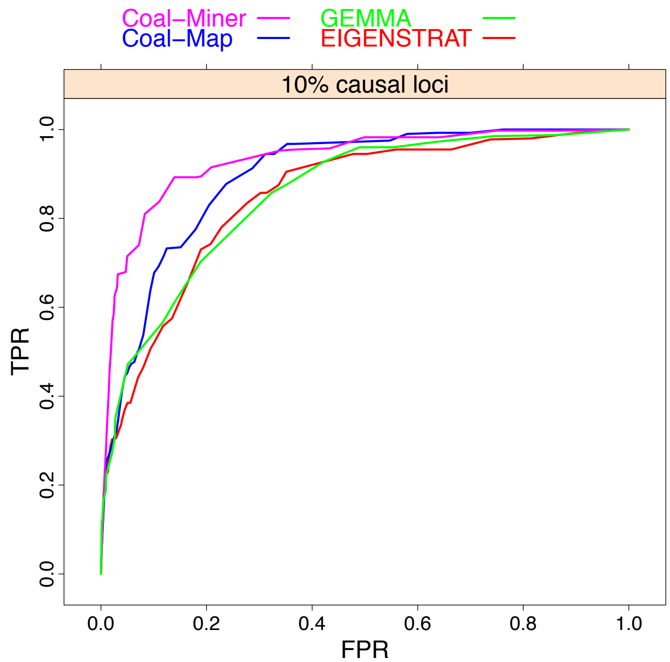
Supplementary Figure S4: **Results for non-neutral model conditions with non-tree-like model phylogenies that include drift/ILS, gene flow, and positive selection.** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT and GEMMA. Figure layout and description are otherwise similar to Supplementary Figure S2.



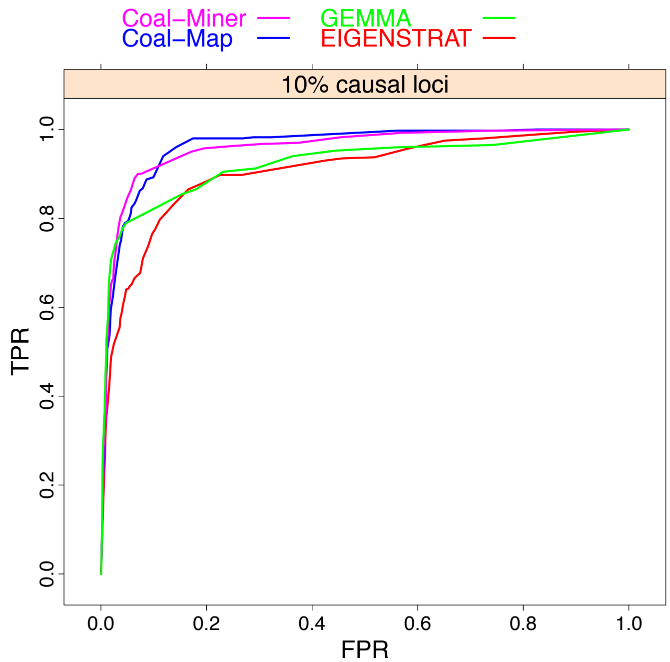
Supplementary Figure S5: **Results for non-neutral model conditions with tree-like model phylogenies that include drift/ILS and positive selection.** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGEN-STRAT, and GEMMA. Figure layout and description are otherwise similar to Supplementary Figure S2.



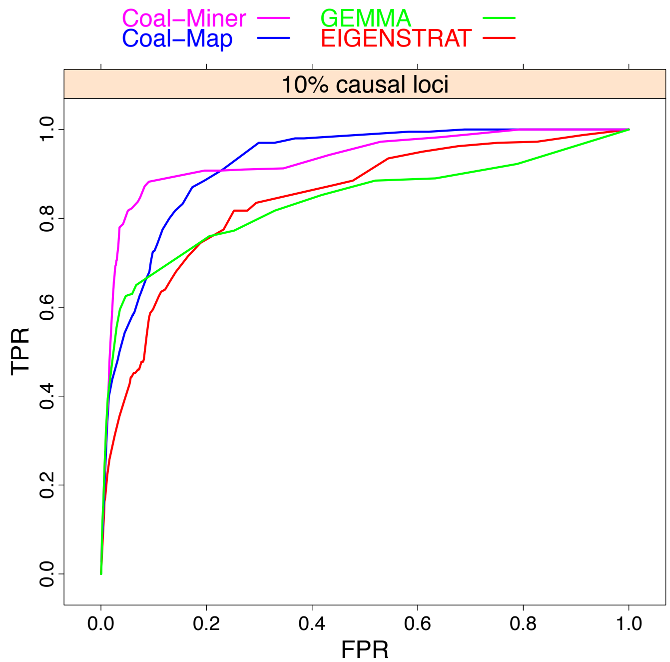
Supplementary Figure S6: **Results for neutral model conditions with tree-like model phylogenies that include drift/ILS ($\gamma = 0$) and divergence time t_1 = 1.0 (in coalescent units).** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. Figure layout and description are otherwise similar to Supplementary Figure S2.



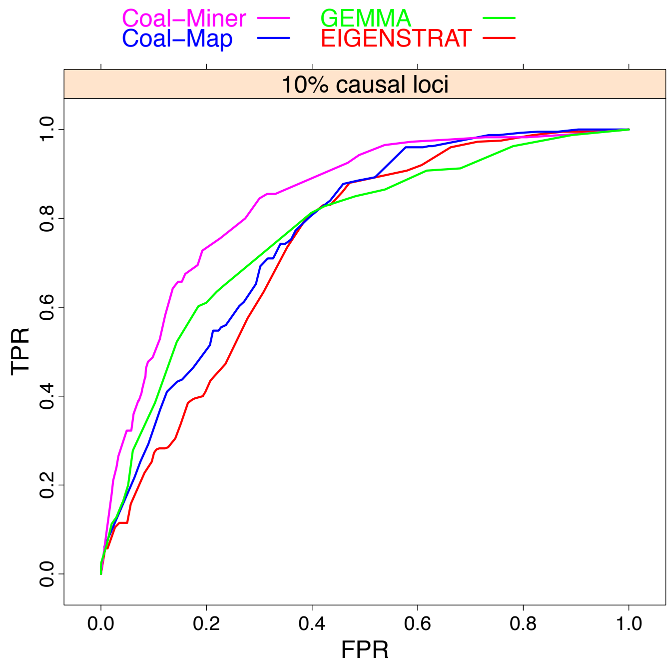
Supplementary Figure S7: **Results for neutral model conditions with tree-like model phylogenies that include drift/ILS ($\gamma = 0$) and divergence time $t_1 =$ 2.9 (in coalescent units).** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. Figure layout and description are otherwise similar to Supplementary Figure S2.



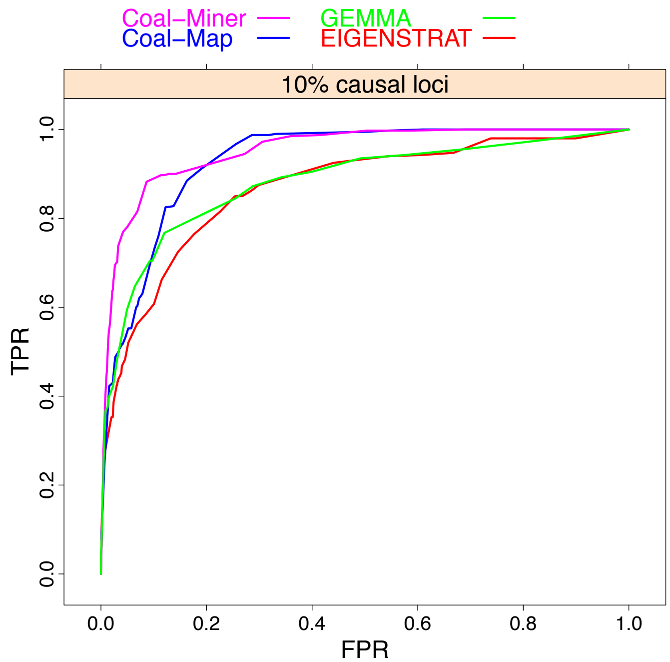
Supplementary Figure S8: **Results for neutral model conditions with non-tree-like model phylogenies that include drift/ILS and gene flow ($\gamma = 0.5$), and admixture time $t_1 = 1$ (in coalescent units).** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. Figure layout and description are otherwise similar to Supplementary Figure S2.



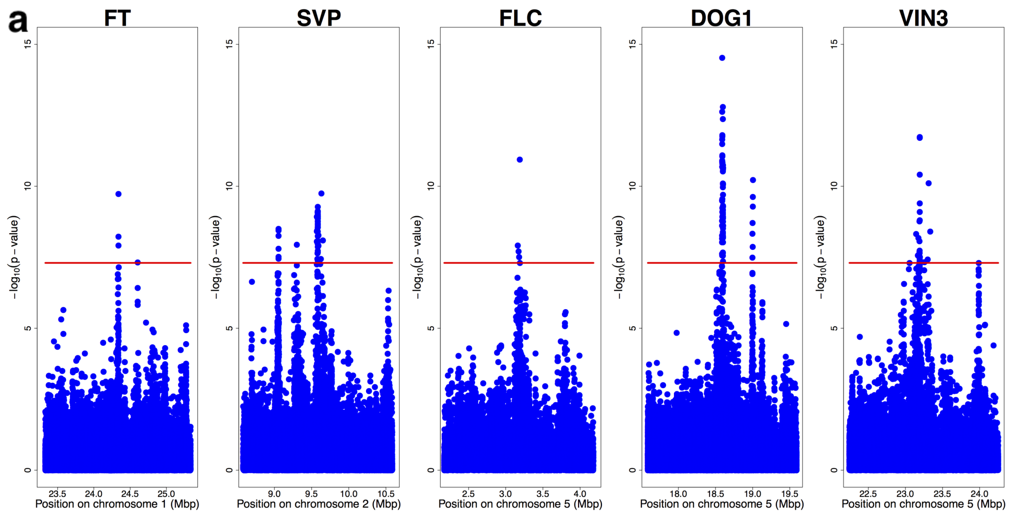
Supplementary Figure S9: **Results for neutral model conditions with non-tree-like model phylogenies that include drift/ILS and gene flow** ($\gamma = 0.5$)**, and admixture time** t_1 **= 2.9 (in coalescent units).** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. Figure layout and description are otherwise similar to Supplementary Figure S2.



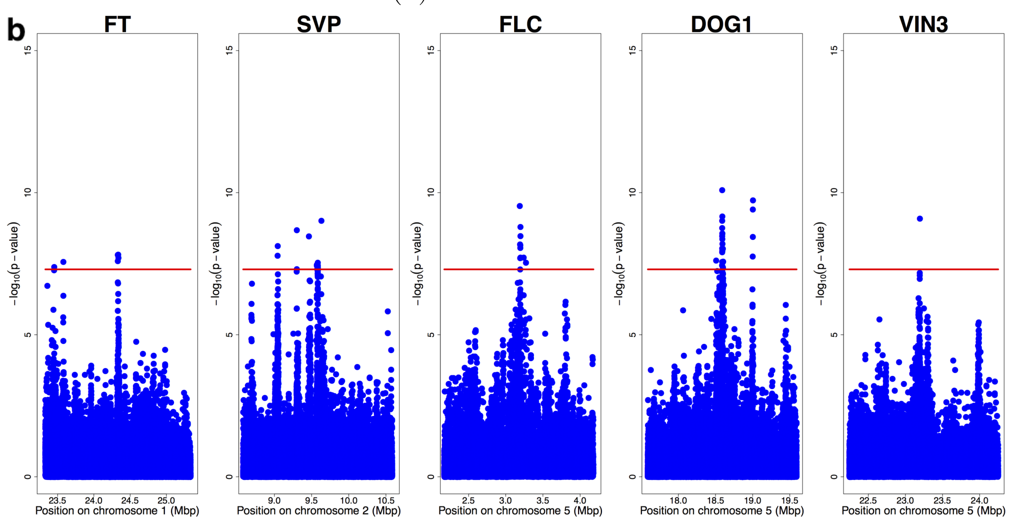
Supplementary Figure S10: **Results for neutral model conditions with tree-like model phylogenies incorporating recombination.** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. Figure layout and description are otherwise similar to Supplementary Figure S2.



Supplementary Figure S11: **Results for neutral model conditions with non-tree-like model phylogenies incorporating an isolation-with-migration (IM) model of gene flow.** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. Figure layout and description are otherwise similar to Supplementary Figure S2.

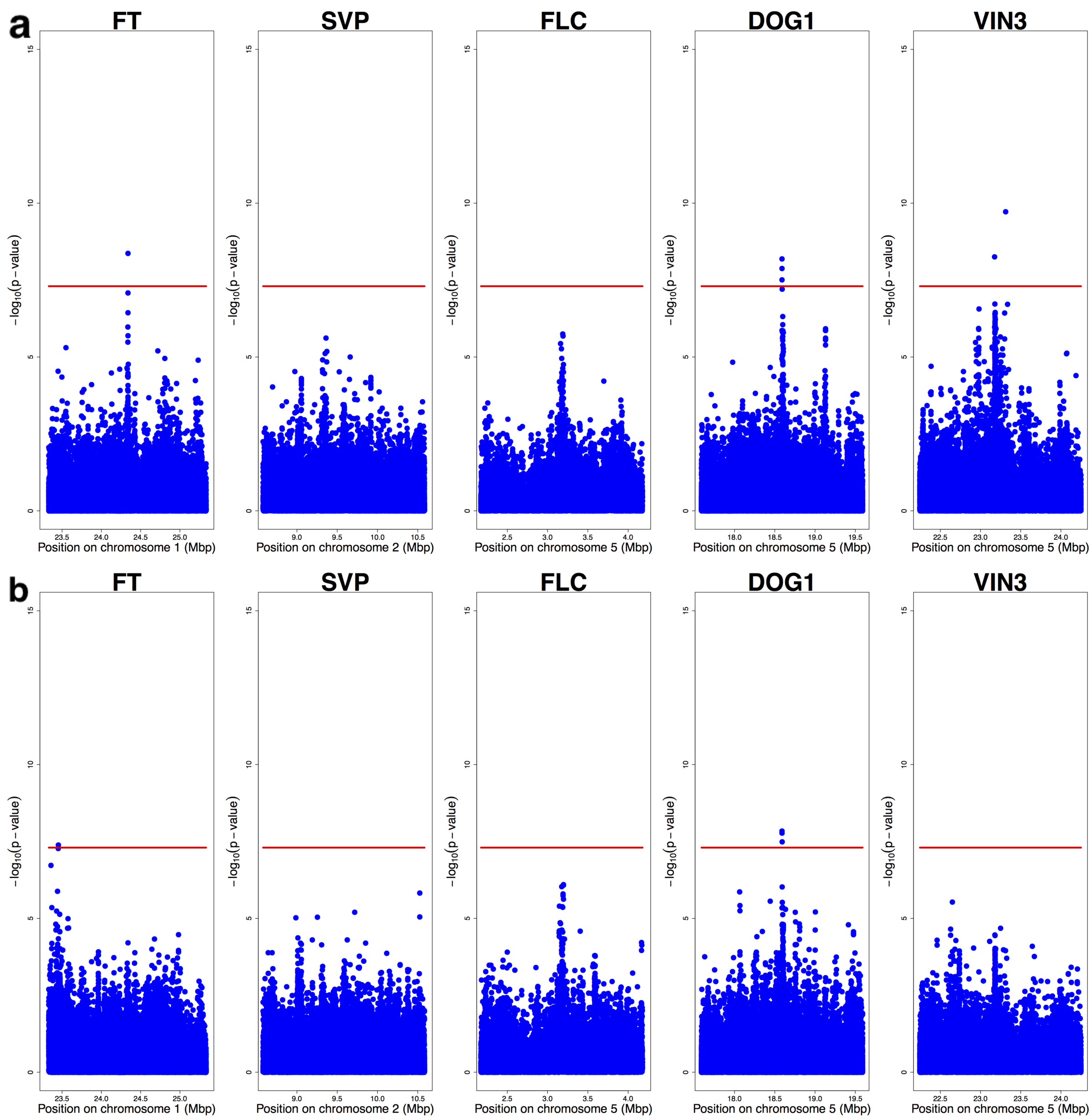


(a) 10 °C dataset

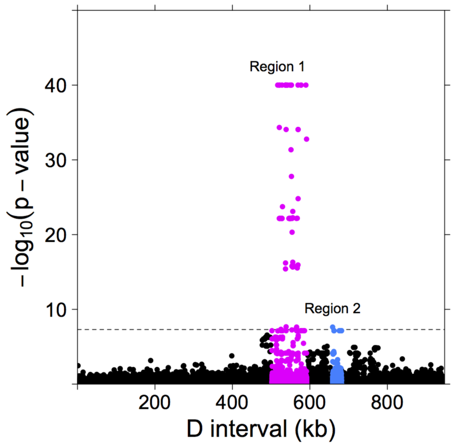


(b) 16 °C

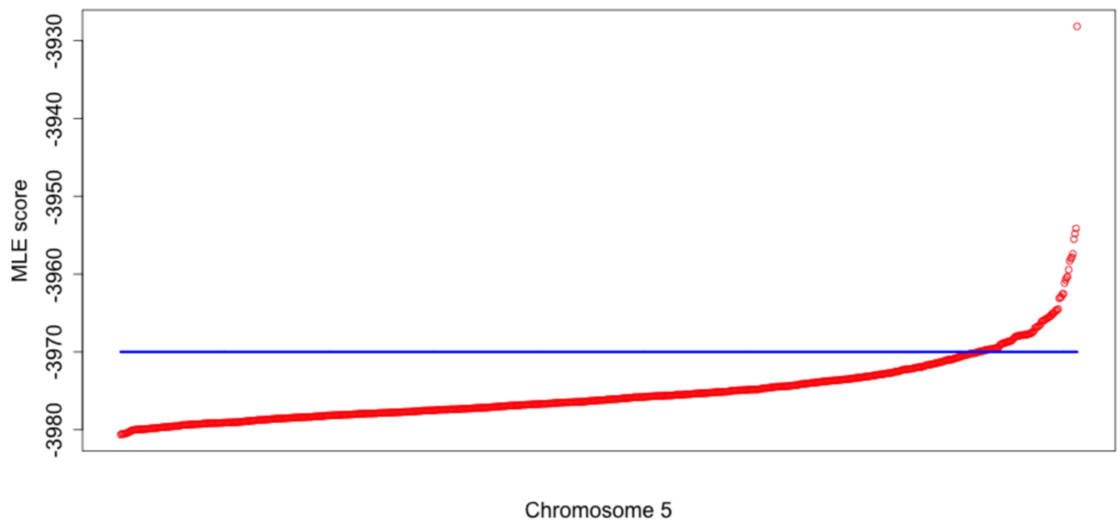
Supplementary Figure S12: **Selected Manhattan plots showing Coal-Miner analyses of two *Arabidopsis* datasets.** Results are shown for the (a) 10 °C dataset and (b) 16 °C dataset. Each Manhattan plot shows Coal-Miner association scores (blue dots) and a Bonferroni-corrected threshold of significance (red line) for selected regions of the *Arabidopsis* genome. The genomic regions are centered on five genes which are known to regulate flowering in *Arabidopsis* and were the focus of similar AM analyses in the study of [27]: FLOWERING LOCUS T (FT), SHORT VEGETATIVE PHASE (SVP), FLOWERING LOCUS C (FLC), DELAY OF GERMINATION 1 (DOG1), and VERNALIZATION INSENSITIVE 3 (VIN3).



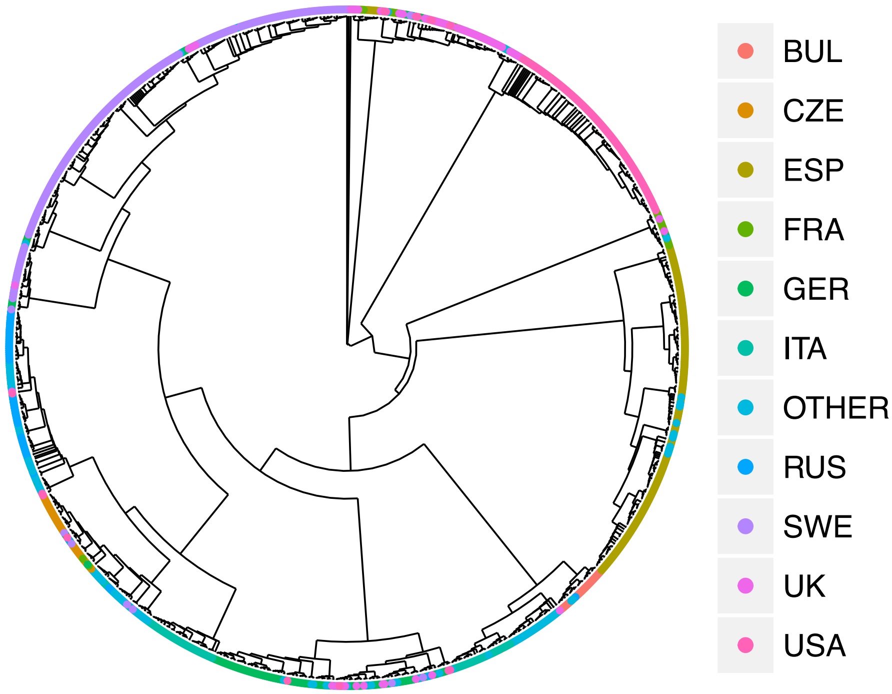
Supplementary Figure S13: **Results showing the Manhattan plots after applying GEMMA on the *Arabidopsis* dataset using two model conditions: flowering time at 10 °C (top panel a) and flowering time at 16 °C (bottom panel b).** The x axis represents the genomic position, and the y axis shows the $-log_{10}$ p-value for all SNPs. The genome-wide significant threshold (p-value = 5 x 10^{-8}) is indicated by the red line. Each sub-panel represents a gene, which was inferred by Coal-Miner to be significantly associated with flowering, and its nearby region. Minor allele frequency of 0.03 was used in the analysis.



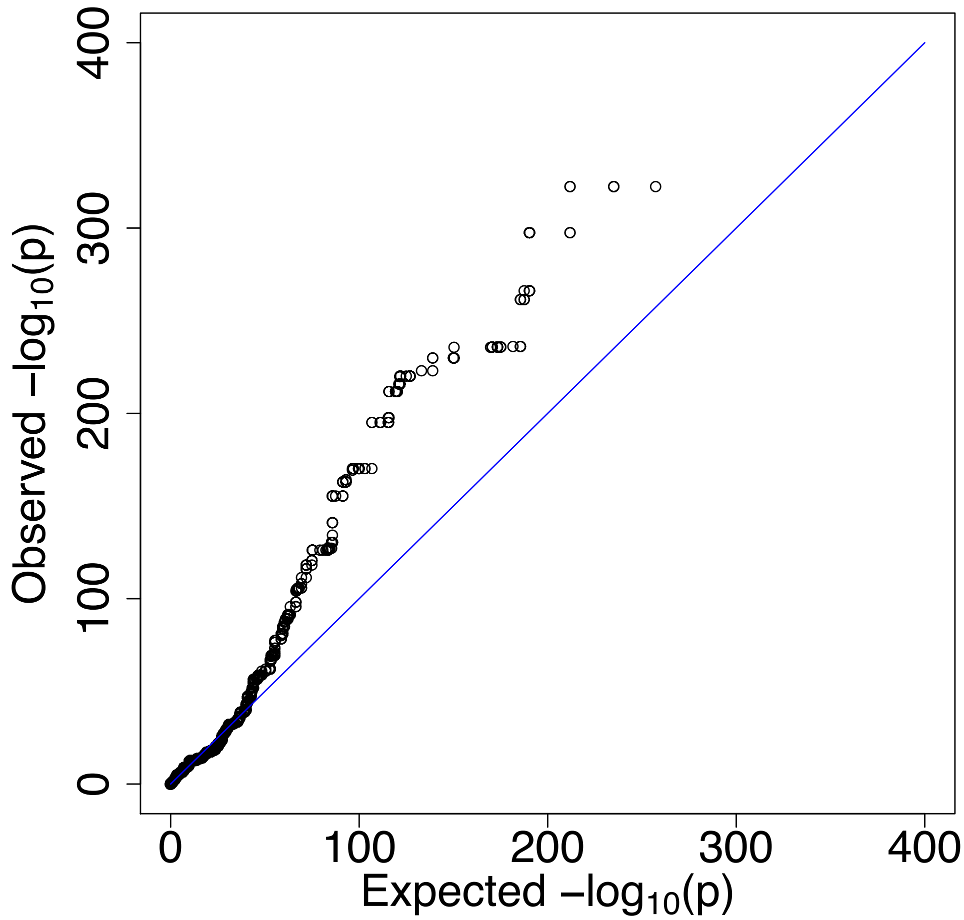
Supplementary Figure S14: **Manhattan plot showing the empirical study results involving *Heliconius erato* butterflies across the D interval.** The x axis represents the genomic position across the D interval, and the y axis shows the $-log_{10}$ p-value for all SNPs. The genome-wide significant threshold (p-value = 5 x 10^{-8}) is indicated by the dotted black line. The dots indicate genotype by phenotype association calculated for biallelic SNPs using Coal-Miner for four hybrid zones: Peru, Ecuador, French Guiana, and Panama (number of postman = 28; number of rayed = 17). The magenta and blue regions represent the two significant peaks identified by Coal-Miner.



Supplementary Figure S15: **Distribution of likelihood scores in stage two of Coal-Miner for loci in chromosome 5 (*Arabidopsis* dataset).** The x axis represents chromosome 5, and the y axis represents the likelihood scores. Results are shown for the flowering time at 10 °C model condition. Each circle represents a genomic locus. The blue line represents the threshold, which is the point of inflection in the distribution, that was used to detect candidate loci (i.e. loci that contain putatively causal SNPs). Any circles located above the threshold are considered candidate loci.



Supplementary Figure S16: **The phylogeny inferred from the 1,135** ***Arabidopsis*** **strains using RAxML.** Each tip in the phylogeny is colored according to its country code. The legend represents the different countries in the analysis (BUL: Bulgaria, CZE: Czech Republic, ESP: Spain, FRA: France, GER: Germany, ITA: Italy, OTHER: Other countries, RUS: Russia, SWE: Sweden, UK: United Kingdom, USA: United States of America)



Supplementary Figure S17: **A qq-plot showing the expected distribution of p-values compared to the observed p-values.** The observed distribution of p-values (y-axis) is obtained by simulating a trait under h = 0.5. The expected distribution of p-values (x-axis) is obtained by simulating a trait under h $\sim$ 0. Test statistic scores are reported as $-log_{10}$ p-value.

References

[1] Hudson, R.R.: Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**(2), 337–338 (2002). doi:10.1093/bioinformatics/18.2.337. http://bioinformatics.oxfordjournals.org/content/18/2/337.full.pdf+html

[2] Hejase, H.A., Liu, K.J.: A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. BMC Bioinformatics **17**(1), 422 (2016)

[3] Ewing, G., Hermisson, J.: MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**(16), 2064–2065 (2010)

[4] Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D.: Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics **38**(8), 904–909 (2006)

[5] Zhou, X., Stephens, M.: Genome-wide efficient mixed-model analysis for association studies. Nature Genetics **44**(7), 821–824 (2012)

[6] Hejase, H.A., Liu, K.J.: Mapping the genomic architecture of adaptive traits with interspecific introgressive origin: a coalescent-based approach. BMC Genomics **17**(1), 41 (2016)

[7] Hudson, R.R.: Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology **23**(2), 183–201 (1983)

[8] Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C.-F., Thomas, M.A., Haussler, D., Jacob, H.J.: Comparative recombination rates in the rat, mouse, and human genomes. Genome research **14**(4), 528–538 (2004)

[9] Notohara, M.: The coalescent and the genealogical process in geographically structured population. Journal of mathematical biology **29**(1), 59–75 (1990)

[10] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics **44**(3), 837–845 (1988)

[11] Supple, M.A., Hines, H.M., Dasmahapatra, K.K., Lewis, J.J., Nielsen, D.M., Lavoie, C., Ray, D.A., Salazar, C., McMillan, W.O., Counterman, B.A.: Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. Genome Research, 150615 (2013)

[12] Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E.K., Olson, R., Overbeek, R., Pusch, G.D., Shukla, M., Schulman, J., Stevens, R.L., Sullivan, D.E., Vonstein, V., Warren, A., Will, R., Wilson, M.J.C., Yoo, H.S., Zhang, C., Zhang, Y., Sobral, B.W.: PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Research (2013). doi:10.1093/nar/gkt1099. http://nar.oxfordjournals.org/content/early/2013/11/12/nar.gkt1099.full.pdf+html

[13] Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P.F., Prohaska, S.J.: Proteinortho: detection of (co-) orthologs in large-scale analysis. BMC bioinformatics **12**(1), 1 (2011)

[14] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: tool for the unification of biology. Nat Genet **25**(1), 25–29 (2000). doi:10.1038/75556

[15] Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research **28**(1), 27–30 (2000)

[16] Silva, I.N., Santos, P.M., Santos, M.R., Zlosnik, J.E.A., Speert, D.P., Buskirk, S.W., Bruger, E.L., Waters, C.M., Cooper, V.S., Moreira, L.M.: Long-Term evolution of *Burkholderia multivorans* during a chronic cystic fibrosis infection reveals shifting forces of selection. mSystems **1**(3) (2016)

[17] Beceiro, A., Tomás, M., Bou, G.: Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? Clinical Microbiology Reviews **26**(2), 185–230 (2013)

[18] Sousa, S.A., Feliciano, J.R., Pita, T., Guerreiro, S.I., Leitão, J.H.: *Burkholderia cepacia* complex regulation of virulence gene expression: A review. Genes **8**(1) (2017)

[19] Lieberman, T.D., Michel, J.-B., Aingaran, M., Potter-Bynoe, G., Roux, D., Davis, M.R. Jr, Skurnik, D., Leiby, N., LiPuma, J.J., Goldberg, J.B., McAdam, A.J., Priebe, G.P., Kishony, R.: Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nature Genetics **43**(12), 1275–1280 (2011)

[20] Munikumar, M., Priyadarshini, I.V., Pradhan, D., Umamaheswari, A., Vengamma, B.: Computational approaches to identify common subunit vaccine candidates against bacterial meningitis. Interdisciplinary Sciences **5**(2), 155–164 (2013)

[21] Drevinek, P., Mahenthiralingam, E.: *Burkholderia cenocepacia* in cystic fibrosis: epidemiology and molecular mechanisms of virulence. Clinical Microbiology and Infection **16**(7), 821–830 (2010)

[22] Lefebre, M.D., Valvano, M.A.: Construction and evaluation of plasmid vectors optimized for constitutive and regulated gene expression in *Burkholderia cepa-*

cia complex isolates. Applied and Environmental Microbiology **68**(12), 5956–5964 (2002)

[23] Hudson, R.R., Kaplan, N.L.: Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111**(1), 147–164 (1985)

[24] Westesson, O., Holmes, I.: Accurate detection of recombinant breakpoints in whole-genome alignments. PLoS Comput Biol **5**(3), 1000318 (2009). doi:10.1371/journal.pcbi.1000318

[25] Wang, J., Moore, K.J., Zhang, Q., de Villena, F.P.-M., Wang, W., McMillan, L.: Genome-wide compatible SNP intervals and their properties. In: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology. BCB '10, pp. 43–52. ACM, New York, NY, USA (2010). doi:10.1145/1854776.1854788. http://doi.acm.org/10.1145/1854776.1854788

[26] 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**(1), 289–300 (1995)

[27] Consortium, .G.: 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell **166**(2), 481–491 (2016)