Coal-Miner: a coalescent-based 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 phenomena – 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 causal 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. The empirical benchmarks include a large-scale dataset that appeared in a recent high-profile publication. 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-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, population stratification, association mapping, genealogy, variation, discordance, coalescent, simulation study, Arabidopsis, Heliconius, Burkholdericeae

ACM Reference format:
Hussein A. Hejase, Natalie Vande Pol, Gregory M. Bonito, Patrick P. Edger, and Kevin J. Liu. 2017. Coal-Miner: a coalescent-based method for GWA studies of quantitative traits with complex evolutionary origins. In *Proceedings of ACM BCB, Boston, MA, 2017 (BCB),* 10 pages.
DOI: 10.475/123_4

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 study of Consortium [10] published whole genome sequences for over

[*] To whom correspondence should be addressed.

BCB, Boston, MA
© 2017 Copyright held by the owner/author(s). 123-4567-24-567/08/06…$15.00
DOI: 10.475/123_4

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. [47] have focused on bacteria and other microbes (see [9] 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 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. EIGENSTRAT [49] 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 EMMA [30], EMMAX [29], and GEMMA [62]. The question of whether sample relatedness is better modeled using the former or the latter – i.e., using fixed vs. random effects – is a matter of ongoing debate [50, 56].

Local variation in functional covariance across the genome is a crucial signature that AM methods use to uncover putatively causal 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 [16] for a review of relevant literature). It is well understood now 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. Topological incongruence can be severe: for example, within a range of evolutionary conditions referred to as the "anomaly zone", the topology of the most frequently observed local genealogy can be incongruent with the species phylogeny itself [11]. 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) [16, 18]. These observations are applicable across different GWA settings ranging from traditional studies involving closely related populations representing a single species to a comparative study involving multiple species. The latter typically involves relatively greater evolutionary divergence, which introduces added complexity in terms of accounting for sample relatedness.

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 [24], which requires the simplifying assumption that sequence evolution can be described by the infinite sites model. Methods in this class include the LRScan algorithm [59]. Another class of parametric methods make use of finite sites models of sequence evolution. One example is RecHMM [61], which applies a sequentially Markovian approximation [40] to the full coalescent-with-recombination model [21]. More recently, coalescent-based methods have been developed to infer local coalescent histories and explicitly ascribe local genealogical variation to different evolutionary processes. Examples include Coal-HMM [15, 20, 37, 38] and PhyloNet-HMM [35].

To address this methodological gap, we recently introduced Coal-Map, a new AM method that accounts for local genealogical variation across genomic sequences. Coal-Map performs statistical inference under a linear mixed model (LMM). The LMM utilizes fixed effects to account for global sample relatedness and, depending upon whether the test marker is located within a locus containing putatively causal markers, local sample relatedness as well. The latter condition is evaluated using model selection criteria. Coal-Map required local-phylogeny-switching breakpoints as input. We validated Coal-Map's performance using simulated and empirical data. Our performance study 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 sample relatedness using fixed effects.

2 METHODS

2.1 Overview of Coal-Miner algorithm

We begin by introducing the high-level design of Coal-Miner, our new algorithm for statistical AM which accounts for local variation of sample relatedness across genomic sequences. The input to the Coal-Miner algorithm consists of: (1) an n by k multi-locus sequence data matrix $\boldsymbol{X}$, (2) a phenotypic character $\boldsymbol{y}$, and (3) ℓ^*, the number of "candidate loci" used during analysis, where a candidate locus is a locus that is inferred to contain one or more putatively causal SNPs. The output consists of an association score for each site $x \in \boldsymbol{X}$.

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 (in the notation of Zhou and Stephens [62]):

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

The fixed effects are represented by covariates $\boldsymbol{W}$ with coefficients $\boldsymbol{\alpha}$, which include covariates that capture local sample relatedness within each candidate locus, and the test SNP $\boldsymbol{x}$ 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 relatedness matrix $\boldsymbol{K}_{\text{global}}$. The random effects $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$ account for global sample relatedness and residual error, respectively. Each of the two random effects follows an $m - dimensional$ multivariate normal distribution (abbreviated "MVN") with mean 0. The random effects $\boldsymbol{u}$ have covariance $\lambda\tau^{-1}\boldsymbol{K}_{\text{global}}$ and the random effects $\boldsymbol{\epsilon}$ have covariance $\tau^{-1}\boldsymbol{I_n}$, where λ is the relative ratio between the two, $\boldsymbol{I_n}$ is the identity matrix, and the residual errors have variance τ^{-1}. $\boldsymbol{Z}$ is the design matrix corresponding to random effects $\boldsymbol{u}$.

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

2.2 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 $\{X_i\}$, 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 extension of the multi-species coalescent (MSC) model [22, 31, 58], 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 [59] to compute local-topology-switching breakpoints based upon the Four Gamete Test (FGT) [24]. 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 MSC model is suspected to be a computationally difficult problem. Existing methods for this problem (e.g., PhyloNet-HMM [35]) 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 [61]. 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 and other evolutionary processes).

2.3 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 $\{X_i\}$, 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. The output is a set of candidate loci $\{X_j^*\} \subseteq \{X_i\}$ 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 causal SNPs in each candidate locus X_j^* is captured by covariates $\{w_j\} \subseteq W$. Since we compare fitted LMMs that may have varying fixed effects, we use LMM log-likelihood as our optimization criterion (reproduced from equation (3) in [62]):

$$\mathcal{L}(\lambda, \tau, \boldsymbol{\alpha}, \beta) = \frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{H}| \\ - \frac{1}{2}\tau(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\alpha} - \boldsymbol{x}\beta)^T\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\alpha} - \boldsymbol{x}\beta)$$

where $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{K}_{\text{global}}\boldsymbol{Z}^T$ and $\boldsymbol{H} = \lambda\boldsymbol{G} + \boldsymbol{I_n}$. 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 [62], which combines Brent's method [8] and the Newton-Raphson method.

For each candidate locus X_j^*, sample relatedness was evaluated using principal component analysis (PCA) [27] of X_j^* – similar to techniques that are widely used by AM methods to account for global sample relatedness as fixed effects [49]. The phenotypic contribution of candidate locus X_j^* was represented using covariates $\{w_j\}$ which consisted of the top five principal components. (The zth principal component corresponds to the sample covariance matrix eigenvector with the zth largest eigenvalue.) 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 X_i, we used MLE to fit an equivalent LMM, except that the covariates W included only the covariates $\{w_i\}$ for locus 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.

2.4 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 $\{X_i\}$, where $1 \leq i \leq \ell$ and ℓ is the number of loci, the phenotypic character $\boldsymbol{y}$, and the set of candidate loci $\{X_j^*\}$. 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 $\{X_j^*\}$ is captured by covariates in $\boldsymbol{W}$: specifically, if the test SNP $\boldsymbol{x}$ is located within a candidate locus X_j^*, the covariates $\boldsymbol{W}$ include a corresponding covariate w_j which consists of the top principal component from PCA applied to 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.

2.5 Simulation study

Experiments involving quantitative traits with varying genomic architectures. 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 (shown in Supplementary Figure S1 panels (a) and (b)). We used ms [23] to simulate coalescent histories (and embedded gene trees) under an extension of the coalescent model [31] which allows instantaneous unidirectional admixture (IUA) [14]. 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 utilized a γ of 0.0 and 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 [19]. 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 [36] and Besenbacher et al. [6]. The trait value for the ith individual is represented as:

$$y_i = \pi \sum_{j \in \omega} \frac{Q_{i,j}}{|\delta|} + (1 - \pi)N(0, 0.01)$$

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 [17] 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 [54]. Quantitative traits with between one and three causal loci 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 number of causal loci (between one and three), 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.

Experiments involving alternative evolutionary scenarios. Multi-locus sequence evolution in the above simulations 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 above model tree (i.e., $\gamma = 0$). Another set of model conditions varied admixture time t_1 in the above model phylogeny where $\gamma = 0.5$. The impact of recombination was explored in a model condition which made use of the coalescent-with-recombination model [21]. 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 [26]. We further explored the impact of gene flow using a model condition which substituted the isolation-with-migration model [44] in place of the IUA model.

Open data access, methods, and performance evaluation. Detailed software commands and instructions for accessing simulation study datasets under an open license are provided in the SI.

The other methods in our study consisted of Coal-Map, GEMMA, and EIGENSTRAT. We followed the procedure from our original study [19] 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. [49]. Detailed software commands are listed in the SI.

We evaluated performance based on statistical power, type I error, and AUROC. To compare AUROC, we performed Delong *et al.* tests [12] using the Daim v. 1.1.0 package [48] in R [51]. Custom scripts were used to conduct the simulation study; all scripts are provided under an open source license (see SI for details and download instructions).

2.6 Empirical study

Arabidopsis dataset. The dataset 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 Supplementary Figure S16 depicts the geographic origins of and evolutionary relationships among the samples. The dataset was originally published and analyzed by Consortium [10], and we obtained genomic sequences and quantitative trait data from the 1001 Genomes Project database [10] (accessible at www.1001genomes.org); the former includes both assembled WGS data and variant calls for a total of 10,707,430 biallelic SNPs. (Details about sequencing, assembly, filtering, quality controls, and variant calling are described in [10].)

Stage one of the Coal-Miner pipeline made use of RecHMM [61] 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. We followed the general approach of Solís-Lemus and Ané [53] to determine a suitable value for ℓ^*. Specifically, 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). 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 0.3. The third stage of Coal-Miner was run using the same settings as in the simulation study.

Heliconius erato dataset. We re-analyzed data from the study of Supple et al. [57]. 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 above). 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/) [60]. Supplementary Table S5 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 [32] 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 [3] and KEGG [28] pathway assignments.

3 RESULTS

In this study, we introduce Coal-Miner, a new statistical AM method which advances the state of the art in terms of its statistical power and type I error control. Coal-Miner's performance advantage derives primarily from two factors. First, Coal-Miner utilizes a new 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. We validated the performance of Coal-Miner using two different types of datasets: synthetic datasets and empirical datasets sampled from natural populations of non-model organisms – each from a different kingdom. The synthetic datasets were simulated under a range of evolutionary scenarios that included multiple causes of local genealogical variation (see Methods for more details).

3.1 Simulation study

Experiments involving varying genomic architecture of a quantitative trait. 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, a LMM-based 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), and Table 1 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. [12] (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), 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 differed 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, causal loci evolved differentially compared to 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).

Experiments involving alternative evolutionary scenarios. Our simulation study also included additional experiments that explored other neutral evolutionary scenarios. These model conditions fixed the proportion of causal loci to 10%. 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. [12], 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).

3.2 Empirical study

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, (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.

Arabidopsis dataset. We used Coal-Miner to re-analyze an *Arabidopsis* dataset which the 1001 Genomes Consortium published in Cell this past summer [10]. The dataset includes samples from 1,135 high quality re-sequenced natural lines adapted to different environments with varying local climates [2, 10]. 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.

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**	0.939	0.866	0.871	< 0.00001
		20	**0.921**	0.899	0.849	0.859	< 0.00001
		30	**0.904**	0.882	0.847	0.832	< 0.00001
Neutral	Tree-like	10	**0.943**	0.922	0.87	0.833	0.0053
		20	**0.904**	0.847	0.843	0.813	< 0.00001
		30	**0.904**	0.853	0.844	0.799	0.00003
Non-neutral	Non-tree-like	10	**0.959**	0.933	0.896	0.836	< 0.00001
		20	**0.926**	0.897	0.856	0.847	< 0.00001
		30	**0.894**	0.863	0.832	0.816	< 0.00001
Non-neutral	Tree-like	10	**0.954**	0.922	0.856	0.841	< 0.00001
		20	**0.89**	0.85	0.832	0.796	0.00003
		30	**0.879**	0.836	0.83	0.783	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 average AUROC 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. [12] ($n = 20$; $\alpha = 0.05$). We corrected for multiple tests using the approach of Benjamini and Hochberg [5], and corrected q-values are shown. (The corresponding ROC plots are shown in Supplementary Figures S2 through S5.)

A key component of the study of the 1001 Genomes Consortium was a GWA analysis of the genomic sequences and quantitative trait data using EMMAX [29], another state-of-the-art statistical AM method (see [62] for a comparison of EMMAX and other state-of-the-art statistical AM methods examined in our study). 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* [2, 25, 41]: 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 [39, 42, 45].

Under a conservative Bonferroni-corrected threshold [7], 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). 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 study of Consortium [10] 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. DOG1 is known to be involved in determining seasonal timing of seed germination and influences flowering time in Arabidopsis [25]. (See Figure 2 in [10] for the original Manhattan plot.) GEMMA's performance was qualitatively similar to EMMAX (Supplementary Figure S13). At 10 °C, 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.

Heliconious 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* [57]. The first peak is located in a noncoding region more distant from the 3′ of the optix transcription factor.

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 S4 shows the genes

		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 study of Consortium [10]. We evaluate 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 use 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, respectively. The corresponding Manhattan plot for the EMMAX analysis is shown as Figure 2 in [10].**

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 [52]. 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* [4, 55]. For example, Lieberman et al. [34] found that the DNA gyrase subunit A gene was under positive selection during a *Burkholderia dolosa* outbreak among multiple patients with cystic fibrosis [34]. Another gene identified by Coal-Miner, Excinuclease ABC subunit A, has been shown to bind to previously published vaccine targets [43]. 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* [13, 33].

4 DISCUSSION

4.1 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 selection coefficients 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.

The other model conditions in our simulation study explore alternative evolutionary scenarios where the proportion of causal loci was fixed. In these model conditions, 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 a 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.

Taken together, the model conditions included multiple sources of local genealogical variation, including genetic drift/ILS, gene flow, recombination, positive selection, and combinations thereof. We note that gene flow was not a necessary prerequisite for Coal-Miner's performance advantage, so long as the other processes were involved (e.g., drift/ILS). 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 stage one of the Coal-Miner pipeline suitable accounts for evolutionary processes with first-order contributions to genome evolution.

4.2 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 other 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 by Consortium [10] – were able to do the same. The vernalization requirement for flowering in *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. Coal-Miner also correctly analyzed positive and negative controls in the other empirical datasets. Furthermore, Coal-Miner analyses of the *Arabidopsis* and *Burkholdericeae* datasets 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.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the following support: National Science Foundation Grant CCF-1565719 (to KJL), a grant 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] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet* 25, 1 (May 2000), 25–29.

[4] Alejandro Beceiro, María Tomás, and Germán Bou. 2013. Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? *Clinical Microbiology Reviews* 26, 2 (April 2013), 185–230.

[5] Y. Benjamini and Y. 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.

[6] Søren Besenbacher, Thomas Mailund, and Mikkel H Schierup. 2009. Local phylogeny mapping of quantitative traits: higher accuracy and better ranking than single-marker association in genomewide scans. *Genetics* 181, 2 (2009), 747–753.

[7] Carlo E Bonferroni. 1936. *Teoria statistica delle classi e calcolo delle probabilita.* Libreria internazionale Seeber.

[8] R. P. Brent. 1973. *Algorithms for Minimization without Derivatives.* Dover Publications, Mineola, New York. 1–208 pages.

[9] 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.

[10] 1001 Genomes Consortium. 2016. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 2 (2016), 481–491.

[11] James H Degnan and Noah A Rosenberg. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet* 2, 5 (2006), e68.

[12] 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.

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

[14] 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.

[15] Julien Y. Dutheil, Ganesh Ganapathy, Asger Hobolth, Thomas Mailund, Marcy K. Uyenoyama, and Mikkel H. Schierup. 2009. Ancestral Population Genomics: The Coalescent Hidden Markov Model Approach. *Genetics* 183, 1 (2009), 259–274.

[16] Scott V Edwards. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1 (2009), 1–19.

[17] 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.

[18] Jotun Hein, Mikkel Schierup, and Carsten Wiuf. 2004. *Gene Genealogies, Variation and Evolution: a Primer in Coalescent Theory.* Oxford University Press, Oxford.

[19] 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.

[20] Asger Hobolth, Ole F Christensen, Thomas Mailund, and Mikkel H Schierup. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genetics* 3, 2 (2007), e7.

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

[22] Richard R Hudson. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* (1983), 203–217.

[23] Richard R. Hudson. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 2 (2002), 337–338.

[24] 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.

[25] 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.

[26] Michael I Jensen-Seaman, Terrence S Furey, Bret A Payseur, Yontao Lu, Krishna M Roskin, Chin-Fu Chen, Michael A Thomas, David Haussler, and Howard J Jacob. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome research* 14, 4 (2004), 528–538.

[27] Ian Jolliffe. 2002. *Principal component analysis*. Wiley Online Library.

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

[29] 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.

[30] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. 2008. Efficient control of population structure in model organism association mapping. *Genetics* 178, 3 (2008), 1709–1723.

[31] J. F. C. Kingman. 1982. On the Genealogy of Large Populations. *Journal of Applied Probability* 19 (1982), pp. 27–43.

[32] Marcus Lechner, Sven Findeiß, Lydia Steiner, Manja Marz, Peter F Stadler, and Sonja J Prohaska. 2011. Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC bioinformatics* 12, 1 (2011), 1.

[33] Matthew D Lefebre and Miguel A Valvano. 2002. Construction and evaluation of plasmid vectors optimized for constitutive and regulated gene expression in *Burkholderia cepacia* complex isolates. *Applied and Environmental Microbiology* 68, 12 (Dec. 2002), 5956–5964.

[34] Tami D Lieberman, Jean-Baptiste Michel, Mythili Aingaran, Gail Potter-Bynoe, Damien Roux, Michael R Davis, Jr, David Skurnik, Nicholas Leiby, John J LiPuma, Joanna B Goldberg, Alexander J McAdam, Gregory P Priebe, and Roy Kishony. 2011. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature Genetics* 43, 12 (13 Nov. 2011), 1275–1280.

[35] 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.

[36] 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.

[37] Thomas Mailund, Julien Y. Dutheil, Asger Hobolth, Gerton Lunter, and Mikkel H. Schierup. 2011. Estimating Divergence Time and Ancestral Effective Population Size of Bornean and Sumatran Orangutan Subspecies Using a Coalescent Hidden Markov Model. *PLoS Genetics* 7, 3 (03 2011), e1001319.

[38] Thomas Mailund, Anders E. Halager, Michael Westergaard, Julien Y. Dutheil, Kasper Munch, Lars N. Andersen, Gerton Lunter, Kay Prüfer, Aylwyn Scally, Asger Hobolth, and Mikkel H. Schierup. 2012. A New Isolation with Migration Model along Complete Genomes Infers Very Different Divergence Processes among Closely Related Great Ape Species. *PLoS Genet* 8, 12 (12 2012), e1003125.

[39] 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.

[40] Gilean AT McVean and Niall J Cardin. 2005. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360, 1459 (2005), 1387–1393.

[41] Belén Méndez-Vigo, José M. Martínez-Zapater, and Carlos Alonso-Blanco. 2013. The Flowering Repressor SVP Underlies a Novel *Arabidopsis thaliana* QTL Interacting with the Genetic Background. *PLoS Genetics* 9, 1 (01 2013), 1–10.

[42] Belén Méndez-Vigo, F Xavier Picó, Mercedes Ramiro, JoséM 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.

[43] Manne Munikumar, I Vani Priyadarshini, Dibyabhaba Pradhan, Amineni Umamaheswari, and Bhuma Vengamma. 2013. Computational approaches to identify common subunit vaccine candidates against bacterial meningitis. *Interdisciplinary Sciences* 5, 2 (June 2013), 155–164.

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

[45] 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.

[46] Stephanie S Porter, Peter L Chang, Christopher A Conow, Joseph P Dunham, and Maren L Friesen. 2016. Association mapping reveals novel serpentine adaptation gene clusters in a population of symbiotic Mesorhizobium. *The ISME Journal* (2016).

[47] 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.

[48] Sergej Potapov, Werner Adler, Benjamin Hofner, and Berthold Lausen. 2013. *Daim: Diagnostic accuracy of classification models.* https://CRAN.R-project.org/package=Daim R package version 1.1.0.

[49] 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.

[50] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. 2013. Response to Sul and Eskin. *Nature Reviews Genetics* 14, 4 (2013), 300–300.

[51] R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[52] Inês N Silva, Pedro M Santos, Mário R Santos, James E A Zlosnik, David P Speert, Sean W Buskirk, Eric L Bruger, Christopher M Waters, Vaughn S Cooper, and Leonilde M Moreira. 2016. Long-Term Evolution of *Burkholderia multivorans* during a Chronic Cystic Fibrosis Infection Reveals Shifting Forces of Selection. *mSystems* 1, 3 (May 2016).

[53] Claudia Solís-Lemus and Cécile Ané. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. *PLoS Genet* 12, 3 (03 2016), 1–21.

[54] 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.

[55] Sílvia A Sousa, Joana R Feliciano, Tiago Pita, Soraia I Guerreiro, and Jorge H Leitão. 2017. *Burkholderia cepacia* Complex Regulation of Virulence Gene Expression: A Review. *Genes* 8, 1 (19 Jan. 2017).

[56] Jae Hoon Sul and Eleazar Eskin. 2013. Mixed models can correct for population structure for genomic regions under selection. *Nature Reviews Genetics* 14, 4 (2013), 300–300.

[57] Megan A Supple, Heather M Hines, Kanchon K Dasmahapatra, James J Lewis, Dahlia M Nielsen, Christine Lavoie, David A Ray, Camilo Salazar, W Owen McMillan, and Brian A Counterman. 2013. Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. *Genome Research* (2013), gr–150615.

[58] Fumio Tajima. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105, 2 (1983), 437–460.

[59] 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.

[60] Alice R. Wattam, David Abraham, Oral Dalay, Terry L. Disz, Timothy Driscoll, Joseph L. Gabbard, Joseph J. Gillespie, Roger Gough, Deborah Hix, Ronald Kenyon, Dustin Machi, Chunhong Mao, Eric K. Nordberg, Robert Olson, Ross Overbeek, Gordon D. Pusch, Maulik Shukla, Julie Schulman, Rick L. Stevens, Daniel E. Sullivan, Veronika Vonstein, Andrew Warren, Rebecca Will, Meredith J.C. Wilson, Hyun Seung Yoo, Chengdong Zhang, Yan Zhang, and Bruno W. Sobral. 2013. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research* (2013).

[61] Oscar Westesson and Ian Holmes. 2009. Accurate Detection of Recombinant Breakpoints in Whole-Genome Alignments. *PLoS Comput Biol* 5, 3 (03 2009), e1000318.

[62] Xiang Zhou and Matthew Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44, 7 (2012), 821–824.

[63] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* 109, 4 (2012), 1193–1198.