Scalable Statistical Introgression Mapping Using Approximate Coalescent-Based Inference

Qiqige Wuyun
Department of Computer Science and Engineering
Michigan State University
East Lansing, Michigan
wuyunqiq@msu.edu

Nicholas W. VanKuren
Department of Ecology and Evolution
University of Chicago
Chicago, Illinois
nvankuren@uchicago.edu

Marcus Kronforst
Department of Ecology and Evolution
University of Chicago
Chicago, Illinois
mkronforst@uchicago.edu

Sean P. Mullen
Department of Biology & Center for Ecology and Conservation Biology
Boston University
Boston, Massachusetts
smullen@bu.edu

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

ABSTRACT

Recent advances in biomolecular sequencing have revealed the important role that interspecific gene flow has played in genome evolution throughout the Tree of Life. Current and future genomic studies will bring large amounts of genomic sequence to bear upon this topic, and scalable computational methodologies are needed to detect and analyze genomic signatures of interspecific introgression in large-scale datasets.

To address the methodological gap, we introduce a new computational framework known as PHiMM (or "fast PhyloNet + Hidden Markov Model"). PHiMM combines inference and learning under a combined model of genetic drift, substitutions, recombination, and gene flow with a coalescent-based approximation technique. We compare the performance of PHiMM against the state of the art using synthetic and empirical genomic sequence data. We find that PHiMM offers better computational runtime and main memory usage by multiple orders of magnitude, while returning comparable inference accuracy.

An open-source software implementation of the PHiMM framework and open data are publicly available at https://gitlab.msu.edu/liulab/phimm-dataset.

CCS CONCEPTS

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

ACM-BCB '19, September 07–10, 2019, Niagara Falls, NY

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

KEYWORDS

introgression, coalescent, phylogenetic network, gene flow, hidden Markov model, non-parametric resampling, mouse, butterfly

ACM Reference Format:
Qiqige Wuyun, Nicholas W. VanKuren, Marcus Kronforst, Sean P. Mullen, and Kevin J. Liu. 2019. Scalable Statistical Introgression Mapping Using Approximate Coalescent-Based Inference. In *ACM-BCB '19: ACM-BCB '19: ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, September 07–10, 2019, Niagara Falls, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Recent comparative genomic studies of interspecific gene flow in eukaryotes have sought to detect introgressed genetic variation and then understand the functional role of introgressed alleles. Examples include studies of house mice [23, 24, 33], butterflies [36], and humans and ancient hominins [9, 29]. A key challenge is inherent to the computational task of introgression mapping: eukaryotic genomes have evolved under complex evolutionary processes, including interspecific gene flow, genetic drift/incomplete lineage sorting (or ILS), recombination, and point mutations [4]. For example, one popular strategy for detecting introgression seeks to reconstruct local evolutionary histories along genomes, and then utilizes topological incongruence between gene trees as a pattern indicative of gene flow. However, deep coalescence caused by genetic drift/ILS can also result in local genealogical discordance. Thus, introgression detection remains a challenging problem, since developing a comprehensive approach for distinguishing genomic patterns left by interspecific introgression versus genetic drift/ILS (as well as other evolutionary processes) is sorely needed.

Recent methodological advances have sought to disentangle genomic signatures of gene flow from those left by the other evolutionary processes. A popular class of statistical introgression mapping methods directly analyzes historical introgression patterns from biomolecular sequence data using a combined model of population genetics and sequence evolution. The population genetic model component typically consists of an extended multi-species coalescent (MSC) model [20] which accounts for genetic drift, gene flow,

and recombination,where temporal probabilistic graphical models are used for a sequentially Markovian coalescent (SMC) approximation [26] to the full coalescent-with-recombination (CwR) model [11, 19]. The sequence evolution model component typically consists of a traditional Markovian model of point mutation processes [6]. Methods in this class include CoalHMM [14, 25] – an early method which was originally proposed for other population genetic inference problems but which has since been extended to address related tasks such as statistical introgression mapping. More recently, Liu et al. [23] introduced PhyloNet-HMM, an introgression mapping method that utilizes a statistical model which combines the multi-species network coalescent (MSNC) model [40] (for capturing genetic drift, incomplete lineage sorting, and gene flow), a finite site substitution model such as the general time-reversible (GTR) model [30], and a hidden Markov model (HMM) to capture intra-sequence dependence due to recombination (based on an SMC approximation to the full CwR model). Each MUL-tree encoded in a fixed species network is represented using a "row" of HMM states, and distinct gene tree topologies corresponding to local coalescent histories evolving within the MUL-tree are represented using distinct states within the row. HMM switching from one "row" of states to another is indicative of gene flow within a reticulation in the species network, depending on the MUL-trees involved; HMM switching between states in the same "row" is indicative of ILS and/or recombination. The combined model is coupled with a modified posterior decoding algorithm for statistical inference. Model parameters are learned using standard local optimization techniques. As far as we're aware, statistical methods in this class have not been used on datasets with more than a handful of genomic sequences. One contributing factor is algorithmic scalability. Increasing dataset size in terms of number of taxa and sampled genomes as well as greater evolutionary divergence can negatively impact computational runtime, memory usage, and inference/learning accuracy. Another popular class of introgression mapping methods applies statistical testing within sliding windows [3, 36], thus avoiding the computational burden of explicitly modeling coalescence, recombination, gene flow, and point mutation processes. Common simplifying assumptions made by this class of methods include the ad hoc nature of applying statistical tests within sliding windows across an input sequence alignment, the infinite sites model and its assumptions about sequence evolution, and others, which may result in relatively low inference/learning accuracy.

2 RELATED WORK

This study builds upon previous work including the PhyloNet-HMM and SERES frameworks. Here we review these related concepts and approaches.

2.1 PhyloNet-HMM algorithm

The PhyloNet-HMM algorithm employs a hidden Markov model for introgression detection. We utilize the model that is implemented in the recently released PhyloNet version 3.6[35, 38], which is a modification of the earlier model proposed by Liu et al. [23].

The HMM states include a trivial start state; every other HMM state corresponds to a distinct pair consisting of a MUL-tree and a gene tree. A MUL-tree is a type of multilabeled tree whose leaves can be labeled by the sampled alleles [15, 40]. The set of MUL-trees encoded by a species network $\mathcal{N}$ can be calculated using the NetworkToMulTree procedure described by [40]. Gene flow directionality is reflected in reticulation edge directionality in an explicit phylogenetic network. Let m and n be the number of MUL-trees and gene trees, respectively. As shown in Figure 1, the states can be represented by $s_{ij} = (T_i, G_j)$, where T_i is the i-th MUL-tree ($1 \leq i \leq m$) and G_j is the j-th gene tree ($1 \leq j \leq n$).

The stochastic behavior of the HMM is governed by transition probabilities, initial state probabilities, and emission probabilities. The transition from the start state to a state $s_{ij} = (T_i, G_j)$ can be calculated as follows:

$$t_{(T_i,G_j)} = \frac{z(s_{ij})}{\sum\limits_{k,l} z(s_{kl})}$$

where $z(s_{ij})$ is the probability of local gene tree G_j under MUL-tree T_i, which can be calculated using the approach of Yu et al. [40].

Let Δ_G be the probability of switching from a local gene tree to one having a different topology (i.e., switching between columns in Figure 1), while Δ_T is the probability of switching from a MUL-tree to one having a different topology (i.e., switching between rows in Figure 1).

A transition from $s_{ij} = (T_i, G_j)$ to $s_{kl} = (T_k, G_l)$ where $1 \leq i, k \leq m$ and $1 \leq j, l \leq n$ occurs with the following probability:

$$a_{(T_i,G_j)\rightarrow(T_k,G_l)} = \epsilon(T_i, T_k)\delta(G_j, G_l)\frac{z(s_{kl})}{\sum\limits_{i,j} z(s_{ij})}$$

where

$$\epsilon(T_i, T_k) = \begin{cases} 1-\Delta_T & if \;\; i = k \\ \frac{\Delta_T}{m-1} & if \;\; i \neq k \end{cases}, \; \delta(G_j, G_l) = \begin{cases} 1-\Delta_G & if \;\; j = l \\ \frac{\Delta_G}{n-1} & if \;\; j \neq l \end{cases}.$$

Given a hidden state $s_{ij} = (T_i, G_j)$, the emission probability can be calculated based on the observation sequence (i.e. the columns of the input alignment A). The observation sequences A can be defined as $\{A, C, T, G\}^{K\times L}$, where K is the number of taxa and L is the length of genomic sequence alignment. Emissions occur according to a substitution model ϕ, which was the generalized time-reversible (GTR) model [30] in our study. For each site of the observation sequence A, which we define as a_i ($1 \leq i \leq L$), the emission probability is $e_{s,\phi} = P[a_i|s, \phi] = P[a_i|\ell_T, \ell_G, \phi]$ where ℓ_T are the branch lengths of the MUL-tree and ℓ_G are the branch lengths of gene tree associated with state $s = (T, G)$.

Given the observation sequences A, the model parameters θ are learned under the maximum likelihood criterion $\underset{\theta}{\operatorname{argmax}}\, P(A|\theta)$ where the model parameters θ consist of:

- The set of MUL-trees (topologies and branch lengths);
- The set of local genealogies (topologies and branch lengths);
- DNA substitution model parameter ϕ;
- MUL-tree and gene tree switching probabilities Δ_T and Δ_G

While the model likelihood for a fixed θ can be calculated efficiently using dynamic programming [28], model likelihood optimization to learn θ is computationally difficult. For this reason, HMM learning is typically addressed using local search heuristics such as the Baum-Welch algorithm and the expectation-maximization algorithm [28]. The PhyloNet version 3.6 implementation of the PhyloNet-HMM framework utilizes the BOBYQA algorithm [27]

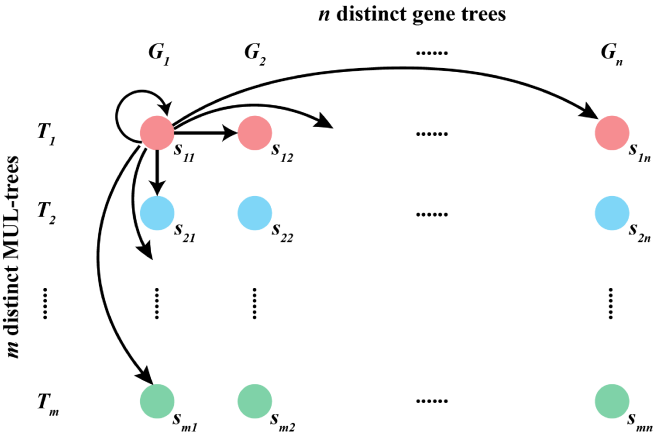


Figure 1: An illustration of the states in the PhyloNet-HMM framework. Note that only transitions outgoing from $s_{11} = (T_1, G_1)$ are shown to simplify the presentation.

to iteratively perform multi-variate optimization as part of a hill-climbing search (whereas the initial version 0.1 implementation of PhyloNet-HMM utilizes Brent's method for univariate optimization [1]).

Given an optimized model θ^*, the forward and backward algorithms are used to calculate the posterior decoding probability:

$$P(\pi_t = (T_i, G_j)|A, \theta^*) = \frac{f_t(i,j)b_t(i,j)}{P(A|\theta^*)}$$

where A is the observed multiple sequence alignment with L columns, and each aligned columns of A is $a_t \in A$ $(1 \leq t \leq L)$; π_t is the t-th state of path π; (T_i, G_j) represents all possible hidden states $(1 \leq i \leq m,\ 1 \leq j \leq n)$ as shown in Figure 1; the forward probability $f_t(i,j) = P(a_1, a_2, ..., a_t, \pi_t = (T_i, G_j)|\theta^*)$ is calculated using the forward algorithm; the backward probability $b_t(i,j) = P(a_{t+1}, a_{t+2}, ..., a_L|\pi_t = (T_i, G_j), \theta^*)$ is calculated using the backward algorithm; $P(A|\theta^*)$ is the probability of the alignment which can be computed by either the forward or backward algorithms.

Consistent with the study of Liu et al. [23], we used a modified posterior decoding approach for inference purposes. The modified posterior decoding probability that a column t $(1 \leq t \leq L)$ in the alignment A has introgressed origin is computed as follows:

$$p_t = \sum_{\substack{T_i \in \Omega_T \\ 1 \leq j \leq n}} P(\pi_t = (T_i, G_j)|A, \theta^*)$$

where Ω_T represents the set of MUL-trees having introgressive origins.

2.2 SERES resampling

Non-parametric resampling approaches allow researchers to use empirical data to build a distribution from which they can obtain critical values, calculate p-values, or construct confidence intervals. For the problem of introgression mapping, non-parametric resampling methods can be used to generate resampled replicates of a genome alignment. Inference/analysis can be performed and compared across replicates.

The SERES framework [37] consists of non-parametric or semi-parametric sequential resampling techniques that generalize the standard bootstrap method for non-parametric resampling [5] and the Heads-or-Tails method [21]. Critically, SERES resampling satisfies the "neighbor preservation property", which means the resampling preserves neighboring bases within the original sequences. The generalized procedure takes the form of random walks along either aligned or unaligned biomolecular sequences. In the former case, the random walk is conducted using the following procedure. A starting point and direction for the random walk are chosen uniformly at random across all sites. As the random walk proceeds, reversals occur with certainty at the start or end of the MSA; reversals can also occur during each step with probability γ. The random walk stops once the number of sampled characters is equal to the fixed MSA length.

3 METHODS

In this study, we address the methodological limitations of state-of-the-art approaches for statistical introgression mapping by augmenting the PhyloNet-HMM framework with two key algorithmic approaches: an MSNC model approximation and SERES-based resampling and re-estimation to "boost" statistical inference accuracy. We refer to the resulting model and method as the PHiMM (or "fast PhyloNet + Hidden Markov Model") framework. The PHiMM model is nested within the PhyloNet-HMM model, and the latter typically has many more parameters compared to the former.

We first describe the computational problem addressed by the statistical introgression mapping methods under study. The problem input consists of a multiple sequence alignment A with K aligned sequences and L columns and a species phylogeny in the form of a rooted phylogenetic network N. The problem output is an annotation of each aligned character $a_t \in A$ with an introgression probability p_t for $1 \leq t \leq L$.

3.1 PHiMM algorithm

The PHiMM algorithm for statistical introgression mapping consists of a multi-stage pipeline. The pseudocode of our algorithm is given in Algorithm 1. The first stage consists of the following "truncation" algorithm: (a) Using $\mathcal{N}$ as a species network model, conduct Monte Carlo sampling of z local gene tree topologies from the gene tree topology distribution under the MSNC model [40]. We used $z = 1000$ in our simulation study experiments and empirical dataset analyses. The observed frequency distribution of local gene tree topologies is normalized to obtain an estimated probability distribution $\hat{f_{\mathcal{N}}}$. (b) Rank each topology in the domain of $\hat{f_{\mathcal{N}}}$ based on its estimated probability, and let Δ be the top k_n topologies based on the topology ranking. The PHiMM analyses in our study set k_n to 30. (c) Truncate the distribution $\hat{f_{\mathcal{N}}}$ such that the domain consists only of topologies in Δ. Normalize the truncated $\hat{f_{\mathcal{N}}}$ to obtain a truncated probability distribution $g_{\mathcal{N}}$. The second stage of the PHiMM algorithm constructs a hidden Markov model (HMM) in a manner similar to the PhyloNet-HMM algorithm, with a single modification. The set of MUL-trees T_i $(1 \leq i \leq m$, where m is the number of MUL-trees) in the MUL-tree representation of $\mathcal{N}$ [40]

is enumerated using the procedure NetworkToMulTree described by Yu et al. [40]. HMM state construction utilizes the truncated distribution g rather than the full theoretical distribution f, where a row of HMM states is instantiated for each distinct MUL-tree T_i and each state in a row corresponds to a distinct local gene tree topology G_j ($1 \leq j \leq k_n = |\Delta|$) in the domain of g_N. The final stage of the PHiMM algorithm performs model fitting and statistical inference under the fitted model using the same procedures as the PhyloNet-HMM algorithm.

The PHiMM framework can be augmented with SERES-based resampling and re-estimation to enhance inference and learning (Algorithm 1). First, the PHiMM algorithm is run with default settings to perform optimization-based learning on the original input MSA. SERES random walks are conducted to perform resampling; the number of SERES replicates was 10 and the reversal probability γ was set to 0.0003. The optimized model is then used to perform fixed-parameter-value inference on resampled SERES replicates. Finally, re-estimated posterior probability distributions are averaged across SERES replicate analyses to obtain a final inferred distribution.

3.2 Simulation study

We conducted a simulation study to evaluate the performance of PHiMM relative to the state of the art. Our simulation procedure generally followed that of Liu et al. [23], with one major change. We utilized msmove [8], a newer coalescent model simulation tool which annotates loci that evolved under gene flow, as opposed to ms [16], which does not provide such annotation. The two simulation studies also utilized comparable model conditions. The model conditions were originally chosen to reflect evolutionary scenarios involving adaptive interspecific introgression that have been recently described in eukaryotes [9, 24, 29]. Additionally, we also conducted additional experiments using the original ms-based simulation protocol from [23]. Our experiments indicated that there was not a large difference between the two simulation protocols, and we therefore focus on the msmove-based simulation protocol. (Detailed descriptions and results for the ms-based protocol are listed in the Appendix.)

The model phylogenies used in our simulation study were generated using the procedure of Hejase et al. [13], which consisted of the following steps. First, we used r8s [31] to generate a random rooted tree under a birth-death process, where the number of taxa $n \in [5, 10]$ and each tree had a total height of 1.0 coalescent unit. Next, a time-consistent level-r rooted tree-based phylogenetic networks was obtained in a manner similar to Leaché et al. [22]. Reticulations were topologically constrained to the class corresponding to paraphyletic gene flow based on Leaché et al. [22]'s classification scheme. A total of r reticulations were added by iterating the following steps: a time t_M between 0 and the tree height was selected uniformly at random, two tree edges for which corresponding ancestral populations existed during a time interval $[t_A, t_B]$ such that $t_M \in [t_A, t_B]$ were randomly selected, and a reticulation at time t_M was added to connect the pair of tree edges. An outgroup taxon was then added to the resulting network with divergence time at 10.0 coalescent units. Similar to Leaché et al. [22], we further classified model phylogenies based on whether gene flow was "deep" or "non-deep" based on topological placement of reticulations. A reticulation is non-deep when its placement involved two leaf edges, and all other reticulations are considered deep; non-deep model phylogenies include only non-deep reticulations, and all other model phylogenies are considered to be deep.

Given a model species network, local coalescent histories and gene trees were simulated under a multispecies network coalescent with recombination (MNSCwR) model. The MNSCwR simulations were performed using [8] since it can annotate a simulated dataset to indicate whether or not local coalescent histories involved historical instantaneous unidirectional admixture (IUA). Following the study of [3], reticulations were modeled as IUA events. Experiments examining impact of differing levels of admixture used model conditions with γ fixed to either 0.1 or 0.5; all other model conditions assigned γ to 0.5. Recombination was modeled using Hudson's finite-sites recombination model [16]. The crossover probability and simulated sequence lengths in our study reflect inferred recombination rates from the study of Jensen-Seaman et al. [17]. The simulation procedure incorporates multi-locus genomic sequence evolution, where k independent and identically distributed (i.i.d.) "query" loci are sampled from the same MNSCwR model and then embedded within a genome sequence. The choice of simulation protocol was practical, since msmove only provides gene flow annotation at the granularity of a whole locus. To implement this protocol, we simulated two different classes of loci: shorter "query" loci had sequence length of 100 bp, and longer "inter-locus regions" had sequence length of 1 kb. For each of the two classes of loci (i.e., query loci and inter-locus regions), the population recombination rate under the finite-sites recombination model was set to a value consistent with the overall recombination rate ρ: 1 and 10, respectively. Loci from the two classes were interleaved, with ten query loci and nine non-query loci sampled per dataset. The sampling design ensures that query loci are separated by sufficient sequence length for the purpose of assuming free recombination between query loci, based on linkage disequilibrium decay observed in previous empirical studies [34]. The sampling scheme also has an effect equivalent to introducing a recombination breakpoint between two adjoining loci. The total sequence length for each simulated dataset was 10 kb. The MNSCwR simulation outputs consist of a sequence of local coalescent histories and embedded gene trees – one for each recombination-free tract. Finally, DNA sequence evolution on each gene tree was simulated under the Jukes-Cantor model [18] with mutation rate $\theta = 2$. For each model condition, we repeated the simulation procedure 20 times to obtain 20 replicate datasets.

Our simulation study compared the performance of the PHiMM algorithm against the PhyloNet-HMM algorithm. PhyloNet-HMM analyses were run using default settings, i.e., the number of iterations for model parameter learning was 300, and the number of runs was set to 10. (Detailed commands are listed in the Appendix.)

The methods under study were evaluated based on several different performance measures. First, we assessed inference accuracy for each method, where the inference annotates each input alignment column with an introgression probability based on the modified posterior decoding calculation in [23]. For each query locus in a simulation replicate, each method's inference accuracy was assessed based on whether or not any sites were flagged as introgressed under a given posterior decoding probability threshold. Varying

Algorithm 1 PHiMM

procedure PHiMM(N, A)
  $\mathcal{N} \leftarrow GetSpeciesNetworkModel(N)$ ▷ N: Phylogenetic network; $\mathcal{N}$: Species network model
  $\Delta_z \leftarrow \emptyset$ ▷ Δ_z: Sampled gene tree topologies
  int $i \leftarrow 1$
  while $i \leq z$ **do** ▷ z: Sampling size
    $\Delta_z \leftarrow \Delta_z + GeneTreeMonteCarloSampling(N, \mathcal{N})$
    $i \leftarrow i + 1$
  $\Delta_d \leftarrow GetDistinctGeneTrees(\Delta_z)$ ▷ Δ_d: Distinct gene tree topologies in Δ_z
  $\hat{f_{\mathcal{N}}} \leftarrow EstimateProbability(\Delta_d)$ ▷ $\hat{f_{\mathcal{N}}}$: Estimated probability distribution of Δ_d
  $\hat{f_{\mathcal{N}}} \leftarrow RankTopology(\hat{f_{\mathcal{N}}})$
  $\Delta \leftarrow Truncate(\hat{f_{\mathcal{N}}}, \Delta_d, k_n)$ ▷ k_n: Truncation size; Δ: Selected gene tree topologies
  $g_{\hat{\mathcal{N}}} \leftarrow EstimateTruncatedProbability(\Delta)$ ▷ $g_{\hat{\mathcal{N}}}$: Estimated probability distribution of Δ

  $\theta \leftarrow InitializeModelParameters(N, \Delta, g_{\hat{\mathcal{N}}})$ ▷ θ: Model parameters
  while Not reaching the convergence criteria **do**
    $\theta \leftarrow HeuristicLearning(\theta, A)$ ▷ A: Input multiple sequence alignment with K aligned sequences and L columns
  $\{p_t\}_{1 \leq t \leq L} \leftarrow ModifiedPosteriorDecoding(\theta, N, A)$ ▷ p_t: Introgression probability for each aligned site t ($1 \leq t \leq L$)
  Return $\{p_t\}_{1 \leq t \leq L}$

procedure SERES-based PHiMM(N, A, r)
  $(\{p_t\}_{1 \leq t \leq L}, \theta) \leftarrow PHiMM(N, A)$
  int $i \leftarrow 1$
  while $i \leq r$ **do** ▷ r: Number of SERES resampling
    $startsite^{(i)}, direction^{(i)} \leftarrow SelectStartSite(A)$ ▷ $startsite^{(i)}, direction^{(i)}$: Starting point and direction for SERES random walk
    $A^{(i)}, mapping^{(i)} \leftarrow \emptyset, \emptyset$ ▷ $A^{(i)}, mapping^{(i)}$: Resampled alignment and mapping
    while Length of $A^{(i)} \leq L$ **do**
      $A^{(i)}, mapping^{(i)} \leftarrow A^{(i)}, mapping^{(i)} + RandomWalk(A, startsite^{(i)}, direction^{(i)}, \gamma)$ ▷ γ: Reversal rate
    $\{p_t^{(i)}\}_{1 \leq t \leq L} \leftarrow PHiMM(N, A^{(i)}, \theta)$ ▷ Run PHiMM with fixed model parameters
    $i \leftarrow i + 1$
  $\{\bar{p_t}\} \leftarrow AverageProbability(\{p_t\}, \{p_t^{(1)}\}, mapping^{(1)} ..., \{p_t^{(r)}\}, mapping^{(r)})$
  Return $\{\bar{p_t}\}_{1 \leq t \leq L}$

threshold values trades off between type I and type II errors. The tradeoff is commonly visualized using receiver operating characteristic (ROC) curves, which plot true positive rate ($\frac{TP}{TP+FN}$) vs. false positive rate ($\frac{FP}{FP+TN}$), or precision-recall (PR) curves, which plot precision ($\frac{TP}{TP+FP}$) vs. recall ($\frac{TP}{TP+FN}$) (where we abbreviate "FP" for false positive, "FN" for false negative, "TP" for true positive, and "TN" for true negative). We report area under curve (AUC) for both (AUROC and PR-AUC, respectively). Second, we assessed computational runtime and main memory usage for the methods under study.

3.3 Empirical study

Our performance study included empirical genomic sequence datasets with positive and negative control loci. The datasets were sampled from wild-derived and wild mouse samples from *Mus musculus domesticus* and *M. spretus*. For comparison purposes, we reproduced a subset of the PhyloNet-HMM analyses from [24], which utilized genomic sequence data from [2]. We briefly review relevant methodological details here (see [24] for more details). The data were sequenced using a SNP array designed by [39]; raw reads from the array were genotyped using MouseDivGeno software [2]. The genotypic sequence data was phased into haploid genomic sequences using fastPHASE [32]. Each dataset consisted of phased genomic sequences for three *M. m. domesticus* samples – one from the region of sympatry between *M. m. domesticus* and *M. spretus*, and two from far outside the region of sympatry, one *M. spretus* sample from the region of sympatry, and one outgroup sequence (*Rattus norvegicus* genome (RGSC Rnor_5.0/rn5)). Each dataset included 4 in-group taxa and 1 out-group taxon.

Our study also included larger "extended" datasets with taxon sampling that was a strict superset of the datasets from the Liu et al. [24] study. The larger size of the extended datasets necessitated the use of PHiMM for introgression mapping purposes. Other than larger set of taxa in new datasets, all other aspects of empirical data were the same (i.e., genotyping, phasing, etc.). The extended datasets included one additional *M. m. domesticus* sample from outside the region of sympatry between *M. m. domesticus* and *M. spretus*. PHiMM was run on the extended datasets with settings identical to the simulation study, with two exceptions. First, PHiMM analyses set the number of iterations for model parameter learning

to 1000 (rather than 300). Second, substitution model and MUL-tree branch lengths were optimized using chromosome 7 from the extended Spain-Arenal dataset.

We also re-analyzed the Limenitis dataset from Gallant et al. [7] and part of the reference genome assembly that is in preparation. We ran PhyloNet-HMM and PHiMM on the dataset of *Limenitis* AC scaffold containing the WntA gene. The dataset includes 4 in-group taxa, *Limenitis arthemis arizonensis*, *Limenitis arthemis arthemis*, *Limenitis arthemis astyanax* and *Limenitis archippus floridenesis*, where we assume that the *Limenitis arthemis arthemis* and *Limenitis arthemis astyanax* were first coalesced, and then their ancestors were coalesced with *Limenitis arthemis arizonensis* and finally with *Limenitis archippus floridenesis*. A reticulation from *Limenitis arthemis arizonensis* to *Limenitis arthemis astyanax* was postulated for the 4-taxa network. Due to the scalability limitation of PhyloNet-HMM, it was run on the 4-taxon dataset, while PHiMM utilized the extended dataset with an additional *Limenitis arthemis arthemis* sample.

3.4 Software and data

Our software implementation of the PHiMM algorithm includes a custom implementation of the MSNC-based Monte Carlo algorithm as well as custom modifications of the PhyloNet software package [35]. Open-source software and open data for all study datasets are publicly accessible at https://gitlab.msu.edu/liulab/phimm-dataset.

4 RESULTS

4.1 Simulation study

We begin by describing the performance comparison of PHiMM versus PhyloNet-HMM. Area under receiver operating characteristic curve (AUROC), computational runtime, and memory usage comparisons are shown in Table 1.

Given up to a week of runtime and at most a TiB of main memory, PhyloNet-HMM was only able to complete analyses of 16 out of 20 of the smallest datasets in our study – those involving the 5-taxon model condition with a single non-deep reticulation; PhyloNet-HMM failed to complete analysis of the other 4 replicates due to excessive main memory requirements. On datasets with 6 or more taxa and model conditions involving deep reticulations or multiple non-deep reticulations, PhyloNet-HMM failed to complete analysis for the same reason. In contrast, PHiMM completed analyses of all of the simulated datasets in at most several hours – even the 10-taxon datasets – and main memory usage was also at most a few GiB (Figure 2).

PhyloNet-HMM's scalability constraints limited comparison of the two methods to the smallest 5-taxon datasets involving a single non-deep reticulation which PhyloNet-HMM successfully analyzed. On the 5-taxon model condition, PHiMM returned runtime and memory usage improvements that amounted to around two orders of magnitude compared to PhyloNet-HMM. On average, PhyloNet-HMM required around 41 hours and 319 GiB of main memory to complete analysis; in contrast, PHiMM required around 8 minutes and 2 GiB of main memory. As measured by AUROC and PR-AUC, PHiMM's inference was comparable to PhyloNet-HMM.

The remainder of the simulation study experiments focus on PHiMM due to PhyloNet-HMM's scalability limitations. On model conditions involving a single non-deep reticulation, PHiMM's runtime and memory requirements increased as the number of taxa increased from 5 to 10 (Figure 2 (a)). However, PHiMM's runtime and memory requirements on the largest datasets were still orders of magnitude smaller than PhyloNet-HMM on the smallest datasets in our study. On model conditions with between 5 and 10 taxa, average AUROC remained between 0.75 and 0.85.

Table 1: A performance comparison of PHiMM and PhyloNet-HMM on the 5-taxon model condition with a single non-deep reticulation. Performance was evaluated based on area under receiver operating characteristic curve ("AUROC"), area under precision-recall curve ("PR-AUC"), computational runtime, and main memory usage. Given a week of computational runtime and 1 TiB of main memory, PhyloNet-HMM successfully completed analysis of 16 out of 20 experimental replicates, but failed on the remaining 4 replicates due to excessive main memory requirements. In contrast, PHiMM completed analysis of all replicates. We report results on replicates on which PhyloNet-HMM ran to completion. Averages and standard errors are reported ($n = 16$).

		Methods	
		PhyloNet-HMM	**PHiMM**
AUROC	Average	0.7806	0.7653
	Standard Error	0.0534	0.0523
PR-AUC	Average	0.7197	0.7305
	Standard Error	0.0871	0.0726
Run time (h)	Average	40.8968	0.1291
	Standard Error	0.5607	0.0035
Memory (GiB)	Average	318.6493	2.3527
	Standard Error	4.3099	0.2271

Similar trends were seen on model conditions involving two non-deep reticulations (Figure 2 (b)). Compared to model conditions with a single non-deep reticulation, runtime and memory usage tended to be slightly larger on the two non-deep reticulation model conditions, but remained on the order of a few hours and GiB, respectively. AUROC on the latter model conditions was between 0.85 and 0.9.

On model conditions involving a single deep reticulation, runtime and memory usage were similar to other model conditions in our study (Figure 2 (c)). Unlike model conditions involving non-deep reticulations, PHiMM's AUROC on model conditions involving deep reticulations was impacted by increasing dataset size in terms of number of taxa: on single-deep-reticulation model conditions involving 5 or 6 taxa, AUROC was around 0.8 – comparable to its performance on equivalent single-non-deep-reticulation model conditions – but AUROC dropped as the number of taxa increased to 10. Experimental variability (as measured by standard error across experimental replicates) also tended to be larger on model conditions involving deep reticulations relative to those involving non-deep reticulations.

Furthermore, we compare the performance of SERES-based PHiMM and PHiMM based on the model condition with a single non-deep reticulation. As shown in Figure 3, the former method returned a

relatively small AUROC improvement relative to the latter on the larger model conditions with 9 and 10 taxa.

4.2 Empirical study

We also compared the performance of PHiMM and PhyloNet-HMM on mouse genomic sequence datasets that were originally analyzed in [24] and [2]. Due to PhyloNet-HMM's scalability limitations, we ran PhyloNet-HMM on a smaller 5-genome dataset that was a proper subset of the larger dataset used for PHiMM's analysis. The latter dataset includes more *Mus musculus domesticus* samples compared to the former, but are otherwise identical. (See Methods for details.)

Previous studies [24, 33] have reported adaptive interspecific introgression involving the chromosome 7 region surrounding the *Vkorc1* gene (i.e., the chromosome 7 region between coordinates 123 Mb and 134 Mb). As shown in Figure 4, both methods infer multi-megabase-long introgressed tracts that appear in all eight samples from Spain and Germany, except the sample from Arenal, Spain. Thus, both methods detect interspecific introgression for this positive control. The *Vkorc1*-containing genomic region contains the longest introgressed tracts that were detected in the mouse genome. Within this genomic region, we note that the total sequence length of introgressed tracts inferred by PHiMM is greater than that inferred by PhyloNet-HMM.

A similar situation was observed in other genomic regions where Liu et al. [24] detected introgressed tracts with hundreds of kilobases of sequence length or more. PHiMM and PhyloNet-HMM inferences were qualitatively similar in that they both detected introgression in these regions. When examining local inference patterns, two types of differences were noted: local differences in the pattern of introgressed and non-introgressed tracts (e.g., the chromosome 7 region between coordinates 102 Mb and 108 Mb and the chromosome 17 region between coordinates 4 Mb and 54 Mb), and longer and more numerous introgressed tracts inferred by PHiMM as compared to PhyloNet-HMM (e.g., chromosome 10, 12, and 15).

We further analyzed the performance of PHiMM and PhyloNet-HMM on the Limenitis sequence datasets, where PHiMM was run on the 5-genome dataset and PhyloNet-HMM was run on the 4-genome dataset (see Methods for details). As shown in Figure 5, the longest introgression tract detected by PHiMM for larger dataset was approximately similar to those inferred by PhyloNet-HMM. Furthermore, both methods detected introgression within the WntA gene region (from coordinates 27Kb to 101Kb), especially for 60Kb to 100Kb. The results were generally consistent with the Figure 4b of Gallant et al. [7].

5 DISCUSSION

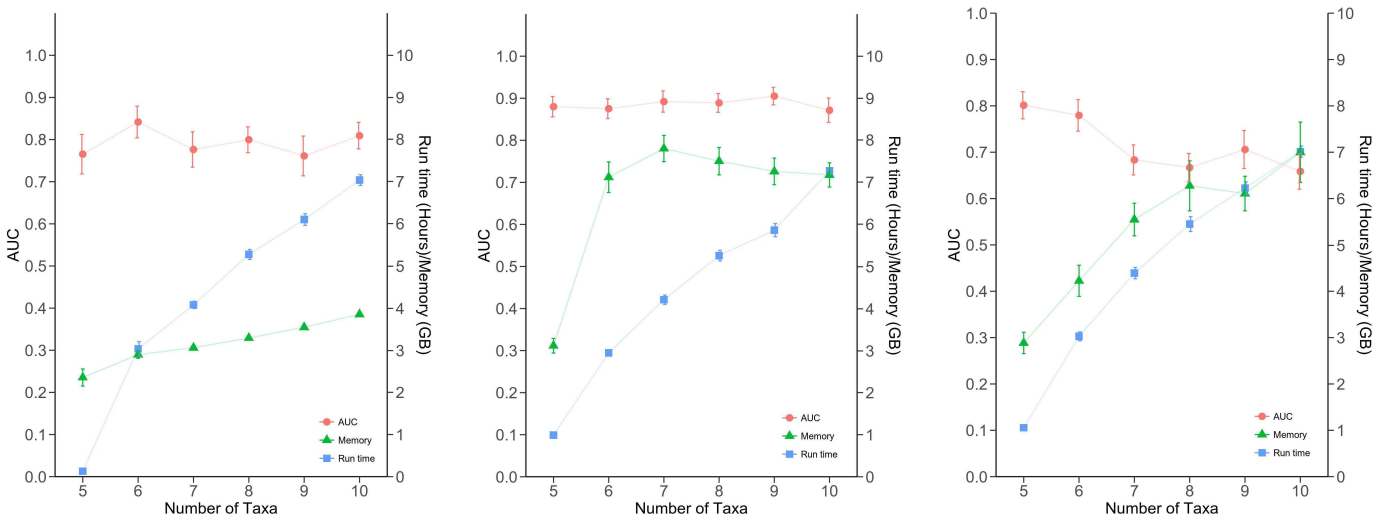
Our simulation study revealed that PHiMM's runtime and memory usage improved upon PhyloNet-HMM by multiple orders of magnitude. The scalability enhancements were primarily due to model approximation enabled by PHiMM's truncation algorithm. Model approximations can impose a penalty in terms of inference accuracy, but that was not the case in our study: in fact, PHiMM's AUROC and PR-AUC were similar to PhyloNet-HMM's. One explanation may be that PHiMM's truncation approach may curb model complexity without imposing much of a penalty in terms of model fit to observed data. Furthermore, a statistical model with fewer parameters may be better suited to the local optimization techniques that are traditionally used for computationally difficult statistical learning problems – as is the case for the methods under study.

PHiMM's AUROC performance was largely robust to increasing dataset sizes and the number and placement of reticulations in the model network, although some AUROC impact was seen on larger datasets involving deep reticulations. Previous studies suggest that deep gene flow (and ancient evolutionary events in general) may present a greater challenge to phylogenetic inference as compared to more recent evolutionary events [12, 13, 22]. While largely unaffected by the number and placement of reticulations, PHiMM's computational requirements increased as dataset sizes increased, but remained well within the capabilities of modern high-performance computing hardware.

SERES-based PHiMM returned comparable performance compared to standalone PHiMM on the smaller model conditions in our study. As the dataset size increased, the former began to return performance improvements relative to the latter, which suggests that the SERES resampling and re-estimation has the potential to "boost" PHiMM's inference accuracy.

On the empirical datasets, PHiMM and PhyloNet-HMM returned qualitatively similar inferences in terms of introgressed genomic regions. The findings are generally consistent with the molecular hypotheses proposed by [24], which identified candidate "driver" genes in these genomic regions that may play a causative role similar to *Vkorc1*. The pattern of local inferences were different between the two methods. In some genomic regions (e.g., the *Vkorc1*-containing genomic region in chromosome 7), PHiMM returned longer and more numerous introgressed tracts as compared to PhyloNet-HMM. Furthermore, the distribution of introgressed tract lengths tended to differ between the two methods. There were more introgressed tracts detected by PHiMM compared to PhyloNet-HMM, and PHiMM's histogram revealed clearer "separation" between two classes of tracts: megabases-long tracts – a few dozen in all – and shorter tracts which were more numerous. The former "long" class of tracts would be consistent with a hypothesis of adaptive introgression, where neutral recurrent back-crossing tends to shorten introgressed tracts over time but positive selection and genetic hitchhiking provides an opposite and countervailing effect [24]. The latter "short" class of tracts would be consistent with Liu et al. [24]'s hypothesis about more ancient bouts of adaptive interspecific introgression; sympatry between *M. musculus* and *M. spretus* is understood to have predated the recent introduction of pesticides [10, 24]. Consistent with the simulation study's performance comparison, we ascribe the observed differences in our empirical study to two factors: PHiMM's competitive statistical power and type I error control relative to PhyloNet-HMM, and denser allele sampling enabled by PHiMM's improved scalability relative to PhyloNet-HMM.

On the Limenitis empirical datasets, PhyloNet-HMM and PHiMM detected the similar introgression tracts that overlapped throughout much of the WntA-containing genomic region. The findings are generally consistent with the experiments given by Gallant et al. [7].



(a) Single non-deep reticulation (b) Two non-deep reticulations (c) Single deep reticulation

Figure 2: PHiMM's performance on model conditions with (a) a single non-deep reticulation, (b) two non-deep reticulations and (c) a single deep reticulation. Performance was evaluated based on area under receiver operating characteristic curve ("AUC"), computational runtime, and main memory usage. Averages and standard error bars are shown (n = 20).

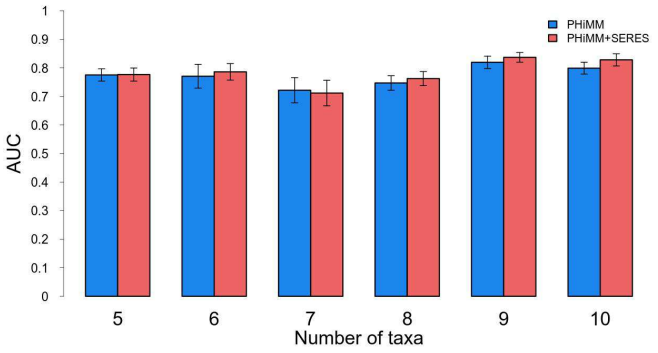


Figure 3: The accuracy comparison between PHiMM and SERES-based PHiMM on 5 to 10 taxa. The results are based on newly simulated dataset under the model condition with a single non-deep reticulation. Performance was evaluated based on area under receiver operating characteristic curve ("AUC"). Averages and standard error bars are shown (n = 20).

6 CONCLUSIONS

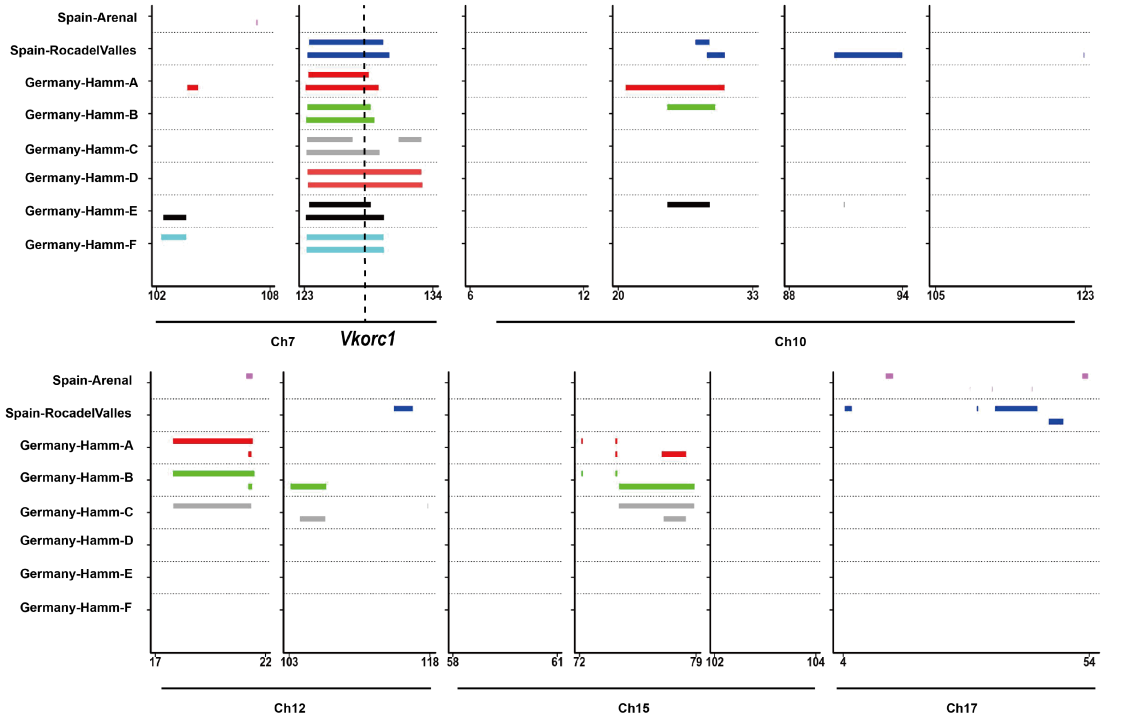
We have introduced PHiMM, a new computational framework for coalescent-based introgression mapping of genomic sequence datasets. Relative to the state of the art, PHiMM offers improved scalability that is better suited to the size and evolutionary scope of current phylogenomic studies. We evaluated the performance of PHiMM and another state-of-the-art method using simulations and empirical genomic sequence datasets.

ACKNOWLEDGEMENTS

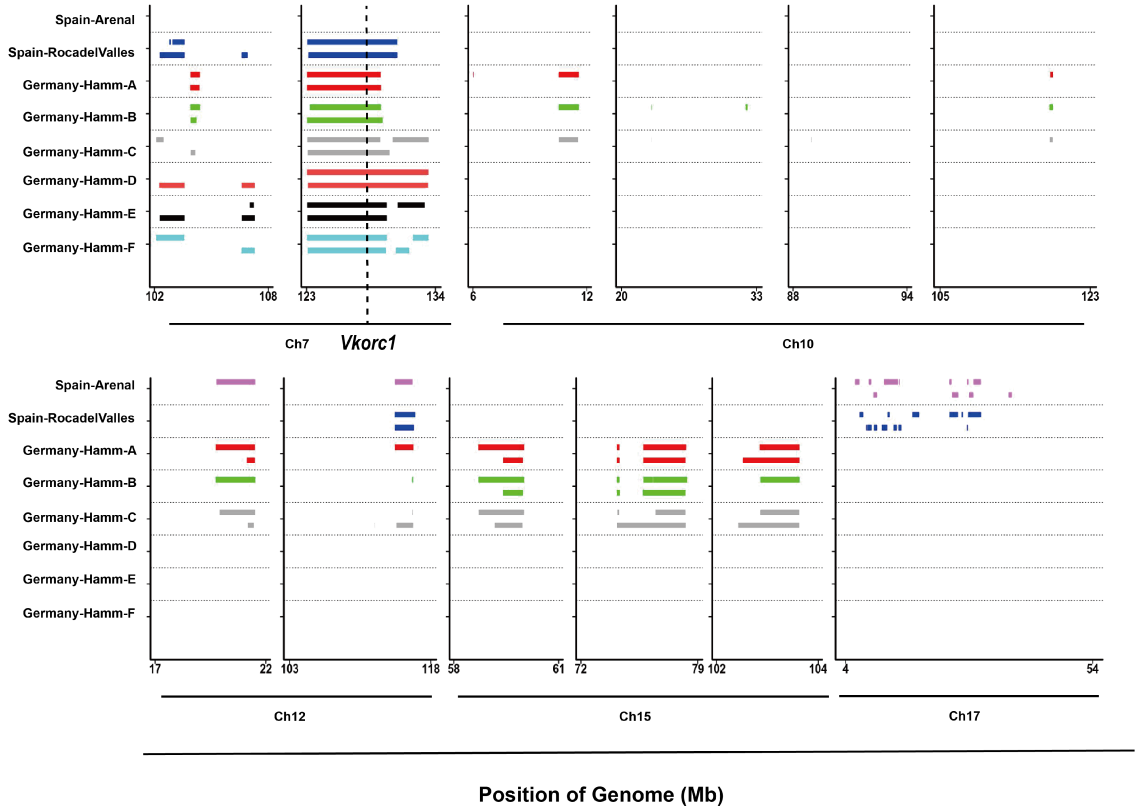
We gratefully acknowledge the support of the National Science Foundation (grants no. 1565719, 1714417, and 1737898), the BEACON Center for the Study of Evolution in Action (NSF STC Cooperative Agreement 093954), and the MSU High Performance Computing Center (HPCC).

REFERENCES

[1] R. P. Brent. 1973. *Algorithms for Minimization without Derivatives.* Dover Publications, Mineola, New York. 1–208 pages.
[2] John Didion, Hyuna Yang, Keith Sheppard, Chen-Ping Fu, Leonard McMillan, Fernando de Villena, and Gary Churchill. 2012. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC Genomics* 13, 1 (2012), 34.
[3] 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.
[4] Scott V Edwards. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1 (2009), 1–19.
[5] B. Efron. 1979. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* 7, 1 (01 1979), 1–26.
[6] Joseph Felsenstein. 2004. *Inferring Phylogenies.* Sinauer Assoc., Sunderland, MA.
[7] Jason R Gallant, Vance E Imhoff, Arnaud Martin, Wesley K Savage, Nicola L Chamberlain, Ben L Pote, Chelsea Peterson, Gabriella E Smith, Benjamin Evans, Robert D Reed, et al. 2014. Ancient homology underlies adaptive mimetic diversity across butterflies. *Nature communications* 5 (2014), 4817.
[8] D Garrigan and AJ Geneva. 2014. msmove: A modified version of Hudson's coalescent simulator ms allowing for finer control and tracking of migrant genealogies.
[9] Richard E. Green, Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F. Hansen, Eric Y. Durand, Anna-Sapfo Malaspinas, Jeffrey D. Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer, Hernán A. Burbano, Jeffrey M. Good, Rigo Schultz, Ayinuer Aximu-Petri, Anne Butthof, Barbara Höber, Barbara Höffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum, Eric S. Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm, Christine Verna, Pavao Rudan, Dejana Brajkovic, Željko Kucan, Ivan Gušic, Vladimir B. Doronichev, Liubov V. Golovanova, Carles Lalueza-Fox, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Ralf W. Schmitz, Philip L. F. Johnson, Evan E. Eichler, Daniel Falush, Ewan Birney, James C. Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich, and Svante



(a) PhyloNet-HMM



(b) PHiMM

Figure 4: A comparison of genomic patterns of interspecific introgression based on (a) PhyloNet-HMM and (b) PHiMM analyses. Results are reported for *Mus musculus domesticus* samples from the region of sympatry between *M. m. domesticus* and *M. spretus* (two Spanish mice and six German mice). Genomic regions involving megabases of introgressed tract length are shown. Due to PhyloNet-HMM's scalability limitations, each PhyloNet-HMM analysis examined a subset of the sequence data for a corresponding PHiMM analysis. Panel layout is adapted from Figure 4a in [24].

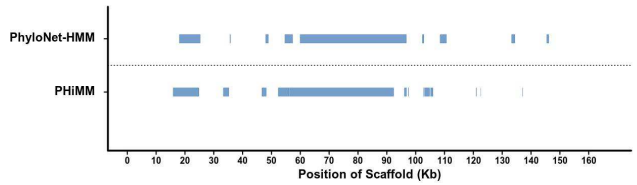


Figure 5: A comparison of genomic patterns of interspecific introgression based on (a) PhyloNet-HMM and (b) PHiMM analyses. Results are reported for *Limenitis arthemis arizonensis* samples from the region of sympatry between *Limenitis arthemis arthemis* and *Limenitis arthemis astyanax*. Inferred introgressed regions are shown in blue. Due to PhyloNet-HMM's scalability limitations, each PhyloNet-HMM analysis examined a subset of the sequence data for a corresponding PHiMM analysis.

Pääbo. 2010. A Draft Sequence of the Neandertal Genome. *Science* 328, 5979 (2010), 710–722.

[10] Bettina Harr, Emre Karakoc, Rafik Neme, Meike Teschke, Christine Pfeifle, Željka Pezer, Hiba Babiker, Miriam Linnenbrink, Inka Montero, Rick Scavetta, Mohammad Reza Abai, Marta Puente Molins, Mathias Schlegel, Rainer G. Ulrich, Janine Altmüller, Marek Franitza, Anna Büntge, Sven Künzel, and Diethard Tautz. 2016. Genomic resources for wild populations of the house mouse, *Mus musculus* and its close relative *Mus spretus*. *Scientific Data* 3 (2016), 160075 EP –.

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

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

[13] Hussein A. Hejase, Natalie VandePol, Gregory M. Bonito, and Kevin J. Liu. 2018. FastNet: Fast and Accurate Statistical Inference of Phylogenetic Networks Using Large-Scale Genomic Sequence Data. In *Comparative Genomics*, Mathieu Blanchette and Aïda Ouangraoua (Eds.). Springer International Publishing, Cham, 242–259.

[14] 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.

[15] Katharina T Huber, Bengt Oxelman, Martin Lott, and Vincent Moulton. 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. *Molecular Biology and Evolution* 23, 9 (2006), 1784–1791.

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

[17] 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.

[18] T.H. Jukes and C.R. Cantor. 1969. *Evolution of Protein Molecules*. Academic Press, New York, NY, USA, 21–132.

[19] John Frank Charles Kingman. 1982. The coalescent. *Stochastic Processes and their Applications* 13, 3 (1982), 235–248.

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

[21] Giddy Landan and Dan Graur. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. *Molecular biology and evolution* 24, 6 (2007), 1380–1383.

[22] Adam D Leaché, Rebecca B Harris, Bruce Rannala, and Ziheng Yang. 2014. The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology* 63, 1 (2014), 17–30.

[23] 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.

[24] K. J. Liu, E. Steinberg, A. Yozzo, Y. Song, M. H. Kohn, and L. Nakhleh. 2015. Interspecific introgressive origin of genomic diversity in the house mouse. *Proceedings of the National Academy of Sciences* 112, 1 (2015), 196–201.

[25] 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.

[26] 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.

[27] Michael JD Powell. 2009. The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge* (2009), 26–46.

[28] Lawrence R Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 2 (1989), 257–286.

[29] David Reich, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Tomislav Maricic, Jeffrey M. Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E. Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V. Shunkov, Anatoli P. Derevianko, Jean-Jacques Hublin, Janet Kelso, Montgomery Slatkin, and Svante Paabo. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468, 7327 (2010), 1053–1060.

[30] F. Rodriguez, J.L. Oliver, A. Marin, and J.R. Medina. 1990. The general stochastic model of nucleotide substitution. *J. Theoretical Biology* 142 (1990), 485–501.

[31] M. J. Sanderson. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 2 (2003), 301–302.

[32] Paul Scheet and Matthew Stephens. 2006. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics* 78, 4 (2006), 629 – 644.

[33] 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.

[34] Fabian Staubach, Anna Lorenc, Philipp W. Messer, Kun Tang, Dmitri A. Petrov, and Diethard Tautz. 2012. Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (*Mus musculus*). *PLoS Genetics* 8, 8 (2012), e1002891.

[35] Cuong Than, Derek Ruths, and Luay Nakhleh. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9, 1 (2008), 322.

[36] The Heliconious Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487, 7405 (2012), 94–98.

[37] Wei Wang, Jack Smith, Hussein A Hejase, and Kevin J Liu. 2018. Non-parametric and semi-parametric support estimation using SEquential RESampling random walks on biomolecular sequences. In *RECOMB International conference on Comparative Genomics*. Springer, 294–308.

[38] Dingqiao Wen, Yun Yu, Jiafan Zhu, and Luay Nakhleh. 2018. Inferring phylogenetic networks using PhyloNet. *Systematic biology* 67, 4 (2018), 735–740.

[39] Hyuna Yang, Jeremy R. Wang, John P. Didion, Ryan J. Buus, Timothy A. Bell, Catherine E. Welsh, Francois Bonhomme, Alex Hon-Tsen Yu, Michael W. Nachman, Jaroslav Pialek, Priscilla Tucker, Pierre Boursot, Leonard McMillan, Gary A. Churchill, and Fernando Pardo-Manuel de Villena. 2011. Subspecific origin and haplotype diversity in the laboratory mouse. *Nature Genetics* 43, 7 (Jul 2011), 648–655.

[40] Yun Yu, James H. Degnan, and Luay Nakhleh. 2012. The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection. *PLoS Genetics* 8, 4 (04 2012), e1002660.