Scalable and robust deep-learning methods power evolutionary-genetic studies of biobank-scale population genomic data

A Dissertation Presented

by

Ziyi Mo

to

The School of Biological Sciences

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Biological Sciences

Cold Spring Harbor Laboratory School of Biological Sciences

January 2024

Abstract of the Dissertation

Scalable and robust deep-learning methods power evolutionary-genetic studies of biobank-scale population genomic data

by

Ziyi Mo

Doctor of Philosophy

in

Biological Sciences

Cold Spring Harbor Laboratory School of Biological Sciences

2024

The advent of next-generation sequencing has brought forth an era where datasets containing genomic sequences for thousands of individuals are common. The key to leveraging rich datasets to generate impactful biomedical insights are high-quality computational tools for biological data analysis. The field of population genetics has a long tradition of using mathematical models to investigate how evolutionary forces shape the dynamics of genetic variants and their biological implications. More recently, artificial intelligence (AI) and machine learning (ML) methods have demonstrated state-of-the-art performance for a wide range of applications involving big data and are increasingly dominant in all areas of quantitative research. My thesis work addresses the unique promises and challenges of analyzing genomic data with AI/ML methods by pioneering rigorous, scalable and innovative deep learning models for population-genetic inference tasks, which ultimately open up broad opportunities for this emerging field of research.

A fundamental pursuit in evolutionary genetics is to identify beneficial mutations and measure the strength of their selective advantage, based on patterns of genetic variation. Studies of positive selection have led to new insights into the biological relevance of particular genomic elements, such as the discovery of mutations involved in immunity or adaptation to extreme environments. Despite many advances, major limitations remain in the sensitivity and accuracy of computational methods for identifying and characterizing selection. These limitations stem, in part, from the difficulty of estimating selective effects directly from DNA sequences. We developed a novel deep-learning method called **S**election **I**nference using the **A**RG (SIA), which makes use of a rich set of features extracted from a reconstructed ancestral recombination

graph (ARG) to make accurate inferences about selection from large-scale genomic data. The ARG can be thought of as a collection of local genealogies and therefore augments the raw sequences by encoding their complete evolutionary history. By exploiting both the richness of information in the ARG and the flexibility and scalability of deep-learning models, SIA offers notable improvements over a wide range of previous methods and therefore emerges as the state of the art for selection inference.

A defining feature of the new generation of AI/ML methods for applications in population genetics, including SIA, is that they generally rely on simulated data for supervised training. This simulate-and-train paradigm has the advantage of virtually unlimited and perfectly labeled training data, but the disadvantage that its performance depends strongly on simulation modeling assumptions. These methods can fail catastrophically when the simulations are mis-specified, such as when a demographic model fails to include a bottleneck event or migration between populations. To go beyond the current ad-hoc methods for handling this essential problem, we devised a domain-adaptive framework for deep-learning models trained on simulated population genetic data. This approach used domain adaptation – a specific form of transfer learning – to train models on one data distribution (simulated genomic data) that can perform well when applied to datasets drawn from a different distribution (real genomic data). This framework is the first to effectively address the critical problem of simulation mis-specification, which has hitherto been the major concern about current applications of AI/ML approaches in population genetics.

Our novel methodological frameworks mark a pivotal step to capitalize on hardware and software advancements for AI/ML, but only the beginning of AI/ML approaches to evolutionary modeling. Recently, large language models (LLMs) of protein and DNA have shown promising performance in a variety of problems in molecular biology such as protein structure or variant effect prediction. Similarly, large generative pre-trained evolutionary models based on genealogical embeddings of the ARG in the future have the potential to revolutionize population genetic research. Such models can be trained in a self-supervised manner with an incredibly wide range of simulations to learn the grammar and logic of how evolutionary processes manifest in different topologies of the ARG, much like the way LLMs "understand" natural languages. Generative models of evolution can be subsequently fine-tuned to perform diverse tasks such as inference of demography, population structure or admixture events. From this line of research that my thesis helped to pioneer, many more powerful AI/ML methods will emerge in the coming years to revolutionize population genetic research and other areas of genomics.

Contents

List of Figures

List of Tables

List of Acronyms

ABC	approximate Bayesian computation
AF	allele frequency
AI	artificial intelligence
ARG	ancestral recombination graph
AUPRC	area under the precision-recall curve
AUROC	area under the ROC curve
CNN	convolutional neural network
CV	computer vision
dadaSIA	**d**omain-**ada**ptive SIA
DAF	derived allele frequency
DFE	distribution of fitness effects
EM	expectation-maximization
FNN	feed-forward neural network
FP	false positive
FPR	false positive rate
GAN	generative adversarial network
GPU	graphics processing unit
GRL	gradient reversal layer
GWAS	genome-wide association study

HMM	hidden Markov model
LD	linkage disequilibrium
LLM	large language model
LSTM	long short-term memory
MAE	mean absolute error
MCMC	Markov chain Monte Carlo
ML	machine learning
MRCA	most recent common ancestor
NLP	natural language processing
PCA	principal component analysis
RBM	restricted Boltzmann machine
RMSE	root mean square error
RNN	recurrent neural network
ROC	receiver-operating characteristic
RTH	relative TMRCA half-life
SA/SC	Santa Ana and Santa Catalina
SFS	site frequency spectrum
SIA	**S**election **I**nference using the **A**RG
SNP	single nucleotide polymorphism
SPR	subtree prune-and-regraft
TMRCA	time to most recent common ancestry
TP	true positive
TPR	true positive rate
VAE	variational autoencoder

Acknowledgements

This endeavor would not have been possible without my thesis advisor, Dr. Adam Siepel, who has generously provided crucial guidance and feedback throughout my thesis project. I would also like to express my deepest gratitude to my thesis committee who provided their extensive knowledge and expertise. I would like to thank Dr. Jesse Gillis for spending the extra effort to keep my progress on track as the chair of my committee, Dr. David McCandlish for devoting long hours to explaining population genetics theory and giving me career advice as my academic mentor, Dr. Peter Koo for his comprehensive guidance on the world of deep learning, and Dr. Andy Kern for taking interest in my work and being an advocate at many conferences. My collaborators have also been indispensable to the success of my research projects. I would like to express my sincere gratitude to Drs. Leonardo Campagna, Nandita Garud, Hussein Hejase, Rob Martienssen, Patrick Reilly, Armin Scheben, Serena Tucci, Al Uy, Jeremiah Wander, as well as Mariana Harris and Josh Steinberg, In addition, I would like to thank the generous support from the Gladys & Roland Harriman Fellowship for funding my PhD at Cold Spring Harbor Laboratory.

I am genuinely grateful for the consistent support from everyone at the School of Biological Sciences. Thanks to Kim Creteur, Alex Gann, Kim Graham, Alyson Kass-Eisler, Zach Lippman, Monn Monn Myat, Victoria Panebianco and Catherine Perez for making a difficult PhD as smooth as possible.

I would like to extend my sincere gratitude to all current and past Siepel lab members. In particular, I am indebted to Hussein Hejase for his mentorship at the outset of my PhD, and grateful for the companionship from Katie Brenner, Noah Dukler, Ling Liu, Luiz Machado, Mehreen Mughal, Ritika Ramani, Armin Scheben and Xander Xue. Special thanks to Susan Fredericks for keeping the spirits of the office high, and Melissa Hubisz for her kind help on ARGweaver. I am also thankful for members of my cohort, Alexa, Amritha, Asad, Connor, Dani, Ilgin, Jenelys, Jonathan, Marie and Teri for going through some of the most difficult parts of this journey together. I'd like to

give a special shout-out to folks at the Thursday board game club, especially Cole and Michael, for providing a fun and thoughtful break from a busy work day.

Last, but by no means least, words cannot express my gratitude to my family for their unwavering support, both spiritually and financially. To my mom Dq, my dad Hui, aunt Daisy, aunt Jianying, grandpa John, grandma Lily, grandpa Aimin and grandma Guizhen, thank you for your unconditional love.

Chapter 1

Introduction

Population geneticist are historians telling the story of evolution. Mutation and recombination leave faithful historical records in the genome of every single organism on earth. The records have always been there, but over the past decades, the experimental and computational tools to decipher those records have improved dramatically. This introductory chapter surveys several key methodological trends that have transformed population genetic research, and finally delineates how these trends have built up the momentum for the original work presented in this thesis.

1.1 Genealogical modeling of evolution using the ancestral recombination graph

The genetic relationships between ancestors and descendants form the basis of all evolutionary genomics research. The simplest data structure that encodes ancestor-descendant relationships is a pedigree (light grey in Fig. 1.1A), commonly known as a "family tree". The pedigree is a graphical structure representing genealogical ancestry of individual organisms. During meiosis in sexually-reproducing diploid organisms, any given position in a haploid gamete is randomly sampled from either chromosome through meiotic recombination. Consequently, the pedigree alone cannot fully specify the genetic ancestry of every position in the genome. Since random shuffling of parental chromosomes through recombination creates a mosaic of genetic ancestry along the genome, different non-recombining segments of the genome have different paths of genetic inheritance in the pedigree. The collection of all paths (or lineages) along which inherited segments of the genome have been transmitted forms a complex graphical structure embedded in the pedigree known as an ancestral recombination graph (ARG) (dark grey in Fig. 1.1A, Griffiths et al., 1997).

The ARG is a *complete* record of the history of genetic inheritance for a set of sampled genomes (solid nodes Ⓐ, Ⓑ, Ⓒ and Ⓓ at the tips of the ARG in Fig. 1.1).

Bifurcating nodes in an ARG represent two types of events – coalescence and recombination. A node where two edges enter from the future but only a single edge exit to the past represents when two lineages find common ancestry and *coalesce* into a single lineage backward in time (e.g. grey coalescence nodes Ⓚ, Ⓟ, Ⓡ, Ⓦ and Ⓧ in Fig. 1.1). Forward in time, a coalescence event occurs through a parent providing the same copy of genomic segment to multiple descendants. Conversely, a node where a single edge enter from the future but two edges exit to the past represents a single lineage of a *recombinant* offspring from two parental lineages (e.g. red recombination nodes Ⓒ and Ⓠ in Fig. 1.1). Forward in time, a recombination node corresponds to a parent passing on a haploid gamete resulting from recombination between its two haploid genomes.

The ARG additionally records the age of each node (not labeled in Fig. 1.1) as well as the position of the recombination breakpoint (dashed red lines in Fig. 1.1) associated with each recombination node. Therefore, the full genealogy of every non-recombining genomic region can be constructed by traversing the the ARG backwards in time and following the lineage on the appropriate side of the recombination breakpoint. The correspondence between the full ARG and local genealogies naturally leads to an equivalent representation of an ARG as a series of genealogical trees along the genome with shared nodes and edges (Fig. 1.1B). Each local tree encodes the evolutionary history of a non-recombining genomic segment and can be transformed into the next one by removing a single edge and attaching it to a different node (arrows in Fig. 1.1B). This operation termed subtree prune-and-regraft (SPR) reflects the outcome of a recombination event manifested in local genealogies. The graphical representation of a full ARG can be recovered by sequentially combining the shared nodes and edges of each local tree while annotating each recombination node with its breakpoint position. In practice, the tree-sequence form of the ARG (see 1.2) is frequently used both as the output of inference algorithms and input for downstream applications (Lewanski et al., 2023), due to not only its tractability, but also the spatially local nature of many population genetic inference problems (such as identifying sites or region under selection).

The ARG constitutes the complete record of ancestral information among a set of genomes. Evolutionary processes such as selection, drift and gene flow all have a direct impact on the structure of an ARG. Consequently, many population and evolutionary genetic questions can be formulated as inquiries into the ARG (Rasmussen et al., 2014; Lewanski et al., 2023). For example, the rate

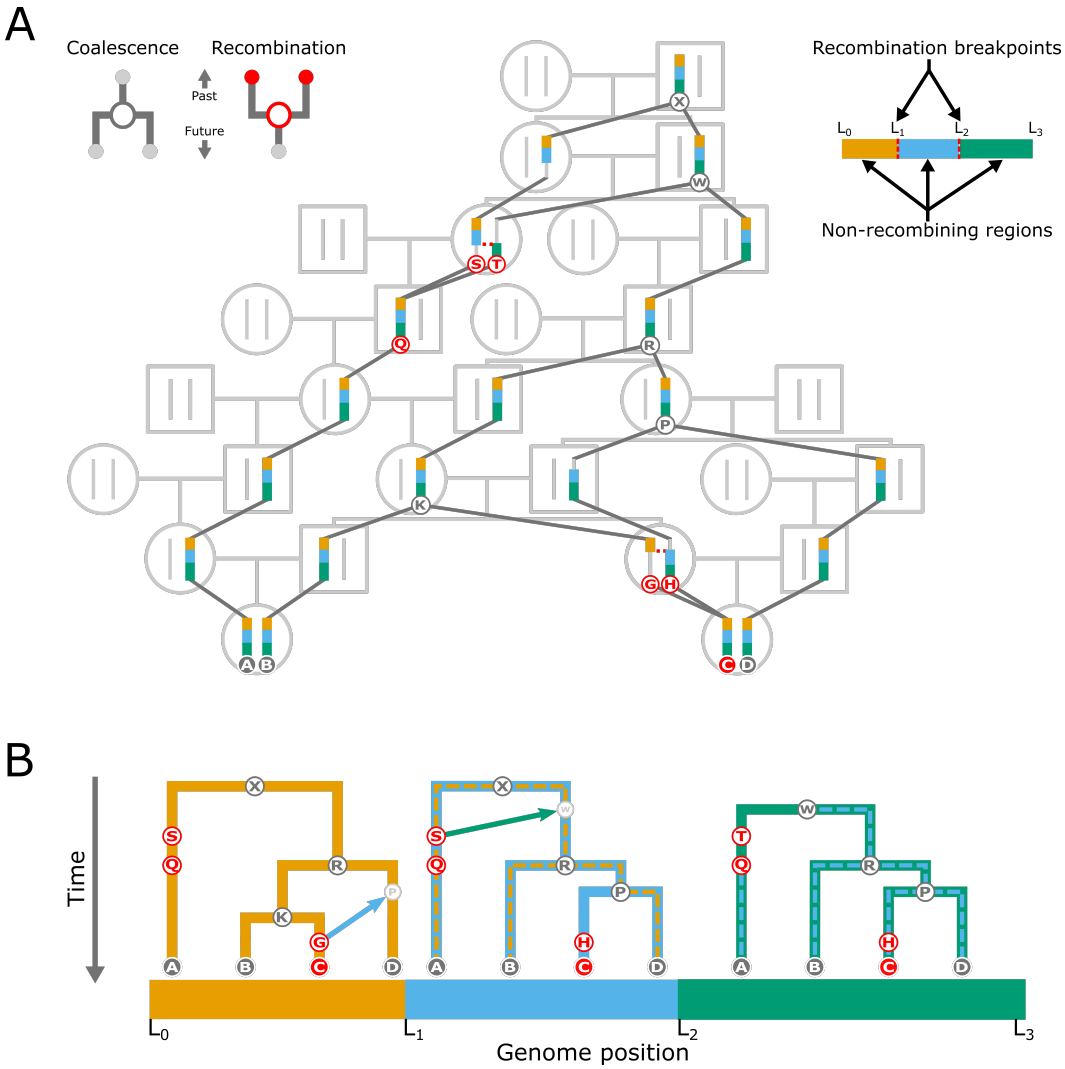


Figure 1.1: **A simple example of an ancestral recombination graph (ARG)**. (**A**) An ARG (dark grey) embedded in a pedigree (light grey). Each node of the pedigree corresponds to an individual organism, connected by edges representing parent-offspring relationships. Each node of the ARG corresponds to a haploid genome, connected by edges representing genetic inheritance between an ancestor and a descendant. Note that the example here assumes the samples are from the nuclear genome of sexually-reproducing diploid organisms, which is the most common scenario of interest. (**B**) An alternative representation of the ARG in (**A**) as a series of local genealogies that share nodes and edges. The arrows represent subtree prune-and-regraft (SPR) operations associated with recombination events that convert local genealogies to their rightward neighbor. The dashed lines highlight each tree's shared structure with its leftward neighbor. Figure adapted from Lewanski et al., 2023 under a CC BY 4.0 license.

of coalescence reflected by the ARG is informative of the effective population size over time, whereas the distribution of recombination breakpoints in the ARG is directly tied to recombination rate across the genome. Furthermore, under the infinite sites model, samples of genomic sequences are stochastic readout of the ARG through a Poisson process of mutations (Wakeley, 2005). Thus, any quantity or statistic derived from the genomic sequences (e.g. the SFS, F_{ST}, π, θ, heterozygosity etc.) is but a low-dimensional summary of the underlying ARG (Ralph et al., 2020). While these summary statistics have demonstrated great utility in providing meaningful evolutionary insights, the ARG holds much richer information that can be tapped into for evolutionary analyses.

Although the ARG is a powerful theoretical and conceptual tool to crack the code of evolution, in practice it must be inferred from population genomic data. ARG inference has historically been a very challenging problem (Rasmussen et al., 2014; I. Mathieson and Scally, 2020). The search space of all possible structures of an ARG grows rapidly with increasing genome and sample sizes. In addition, as mentioned previously, the observed genomic sequences are noisy readout of the true ARG. Mutation creates concordant patterns of genetic variation from which ARGs can be inferred, whereas recombination breaks up such patterns and reduces the amount of information per genealogy. The opposing forces of mutation and recombination impose a limit on ARG identifiability from genomic sequences (Hubisz and Siepel, 2020; Hayman et al., 2023), which in turn limits the utility of ARGs in downstream applications. Early methods aimed to built a parsimonious ARG that contains the minimal number of recombination events given a genotypic matrix (Wong et al., 2023), which is a NP-hard problem (L. Wang et al., 2001). These methods therefore rely on heuristics and are limited in scale of their applications. Recently, there has been great stride towards accurate ARG inference at a practical scale. ARGweaver (Rasmussen et al., 2014) and its extension ARGweaver-D (Hubisz et al., 2020) mark the inception of statistically rigorous genome-wide ARG inference. ARGweaver introduces a novel technique termed "threading" which adds an n-th sequence to an existing ARG of $n-1$ sequences under a likelihood model defined by a hidden Markov model (HMM). The state space of the HMM is simplified using approximations of the coalescent and discrete time to make the "threading" operation a computationally tractable sampling step from the posterior distribution of ARGs using Markov chain Monte Carlo (MCMC). ARGweaver can be applied to up to a hundred whole genomes and remains the state of the art in terms of accuracy (Brandt et al., 2022). With the rapid growth of modern biobank-scale genomic datasets, a number of methods that balances statistical rigor and computational efficiency

have been developed, such as Relate (Speidel et al., 2019) and tsinfer/tsdate (Kelleher et al., 2019; Wohns et al., 2022). These methods employ various heuristics and simplifications and consequently tend to underestimate recombination and only provide point estimates of the ARG (Lewanski et al., 2023; Wong et al., 2023), but have the remarkable ability to scale up to tens or even hundreds of thousands of genomic samples (see Brandt et al., 2022 for a benchmark and comparison of prevailing inference methods). The suite of ARG inference tools with a spectrum of accuracy-scalability tradeoffs has paved the way for a new generation of methods that tackle a variety of empirical questions in population genetics (Speidel et al., 2019; Stern et al., 2019; Stern et al., 2021).

1.2 Population-genetic simulations power large-scale *in silico* experiments of evolution

Without the ability to rewind time, we cannot observe evolution occur in real time in natural populations (barring some rare exceptions of feasible large-scale experimental evolutionary studies in species with short generation times). In most cases, *in silico* simulations provide the only way to replay various evolutionary scenarios and conduct experiments of evolution. Population simulators are therefore an indispensable tool across all aspects of population genetic research ranging from validating theoretical expectations, testing hypothesis to powering Monte Carlo methods.

The process of simulating the evolution of a population may seem straightforward. We can simulate the reproduction of every individual in the population according to some model of choice, most commonly the Wright-Fisher model (Wakeley, 2005; M. W. Hahn, 2018). The simulation runs forward in time for many generations while mutation and recombination processes are recorded in the genomes of extant individuals at each generation. In the end, to generate one simulated dataset, some number of individuals and their genomes are sampled from the latest generation. This *forward* simulation approach can be very flexible since it is facile to incorporate a wide range of evolutionary processes such as different forms of natural selection, complex demographic history or migration patterns into the simulations (M. W. Hahn, 2018). However, forward simulations can become unwieldy in practice. First, since each individual needs to be tracked every generation, the amount of computation scales with the population size N_e, which is not uncommon to be quite large. In addition, simulations need to run until the population reaches equilibrium, which usually takes a number of generations on the order of the long-term N_e

(Wakeley, 2005). Hence, forward simulations scale approximately quadratically with the population size.

Researchers have come up with two types of techniques to circumvent the computational challenge. At its heart, population genomic simulations are simulations of the ARG (see 1.1). Given a particular mutation model, neutral genetic variations in the genome are strictly governed by the ARG (Wakeley, 2005; M. W. Hahn, 2018). Therefore, it is not necessary to generate neutral mutations during forward simulation. They can be overlaid onto the genealogies under a mutation model of choice to generate genomic sequences post hoc. The forward simulator SLiM (Haller and Messer, 2019) employs this tree-sequence recording procedure (Haller et al., 2019, also see below) to improve computational efficiency. Another key observation of the forward simulation process is that many historical individuals do not turn out to be genetic ancestors to the contemporary population from which the samples are taken. Therefore, if we take a *backwards*-in-time approach, start from a sample of contemporary individuals and generate only the ancestors of these samples, we can avoid the computational burden of keeping track of the entire population every generation. This can be accomplished by sampling ancestral lineages under the coalescent process (Kingman, 1982a, 1982b; Tajima, 1983; Hudson, 1990). The coalescent is an approximation of a forward Wright-Fisher population under the key assumption $N_e >> n$, where N_e is the effective population size and n the sample size. This assumption holds for many realistic use cases and therefore coalescent simulators provide a highly efficient way to generate population samples at scale. Coalescent simulations trade flexibility for computational efficiency. Early coalescent simulators were limited to only neutrally evolving populations (M. W. Hahn, 2018; Lewanski et al., 2023). However, a new generation of coalescent simulators have emerged with the ability to handle more complex demographic scenarios and some form of selection. Among these, discoal (Kern and Schrider, 2016) and msprime (Kelleher et al., 2016) have been most widely adopted to simulate biologically realistic samples at a practical scale.

A key innovation that has further revolutionized evolutionary simulations is the introduction of a *hybrid* approach that exploits the strengths of both forward and backward (coalescent) simulators. The overall idea is to perform the more complex portion of the simulation where biological realism is paramount (e.g. population under various modes of selection) with a forward simulator and leave the simpler or less important portion (e.g. neutrally evolving population) to a coalescent simulator. This can be accomplished through either "recapitation" of uncoalesced lineages in the first generation of a forward simulation until a most recent common ancestor (MRCA) is found with

a coalescent simulator (Haller and Messer, 2019), or using a complete coalescent simulation to initialize a population and carry on the simulation with a forward simulator. This approach has been popularized thanks to the tskit python API (Kelleher et al., 2016; Kelleher et al., 2018), which provides an interoperable data structure (ts format) for the tree-sequence encoding of the ARG. The ts format is flexible, has a low storage footprint, and therefore has fostered an ecosystem of software tools along the simulation, inference and analysis pipeline of population genetic research.

Finally, it is worth mentioning that a community-driven effort to maintain a catalog of simulation models and a high-level API to simulate under these models out-of-the-box has emerged (Adrion, Cole, et al., 2020; Lauterbur et al., 2022). The stdpopsim project has made efficient large-scale simulations of complex evolutionary scenarios ever more accessible. The advancement of the simulation infrastructure has ultimately opened up opening up new opportunities and avenues for population genetic research, such as using simulations to generate perfectly labeled training data for supervised machine learning models.

1.3 Deep learning methods for population genetics

Population genetics is a century-old discipline that arose long before the era of genomics. Over the 20$^{\text{th}}$ century, a rich body of theory has emerged that aims to describe how the interplay among different evolutionary forces (such as mutation, recombination, selection, drift and gene flow) shapes the population dynamics of genetic variants. This tradition of probabilistic and statistical modeling led to the creation of many mathematical approaches to infer evolutionary parameters from observed patterns of genetic variation even before molecular genetic data became readily available (Korfmann et al., 2023). Traditional computational methods for population genetic inference (see Marjoram and Tavaré, 2006 for a detailed review) typically employ likelihood-based statistical approaches, such as maximum likelihood, Bayesian or Monte Carlo method to fit an evolutionary model, commonly derived from the Wright-Fisher model or the coalescent (see 1.2), to empirical data for parameter estimation. Traditional statistical approaches have dominated the field because of their interpretability. Since they are rooted in mechanistic or at least generative models of evolutionary processes, traditional methods can often be teased apart under population genetics theory to further our understanding of the molecular genetic mechanism of evolution.

The explosion of massive genomic datasets (e.g. 1000 Genomes Project [Auton et al., 2015], Allen Ancient DNA Resource [Mallick et al., 2023], UK Biobank [Sudlow et al., 2015], 1001 Genomes Project [Alonso-Blanco et al., 2016]) with the advent of high-throughput sequencing technologies has shifted the bottleneck of population genetic inference from the lack of data to the need for highly scalable and robust computational methods. Traditional model-based statistical methods rest on the assumption that the model sufficiently describes the data (Schrider and Kern, 2018). However, such probabilistic models more or less simplify complex evolutionary processes and sacrifice biological realism for computational tractability in practical applications. These simplifications may be appropriate when each evolutionary force is studied in isolation, but severely limit the utility of traditional statistical methods in deciphering complex evolutionary scenarios from the plethora of genomic data. In addition, the growing size and dimension of population genetic datasets pose a direct challenge to the computational efficiency of traditional methods. Popular algorithms for fitting probabilistic models to molecular genetic data such as Markov chain Monte Carlo (MCMC) and expectation-maximization (EM) scale poorly with sample size (Korfmann et al., 2023). Approximate Bayesian computation (ABC), a widely used method in population genetics for bypassing the calculation of intractable likelihood functions, suffers from "the curse of dimensionality" in that the approximation error increases with the growing number of summary statistics necessitated by high-dimensional data (Prangle, 2018). Therefore, while traditional inference methods established from population genetics theory have been successful in yielding valuable evolutionary insights from small-scale molecular markers, the emergence of modern biobank-scale genomic datasets has inevitably transformed population genetics from a theory-driven discipline to a data-driven one.

Machine learning (ML) has demonstrated state-of-the-art performance over a wide range of empirical applications involving big data and become increasingly dominant in all areas of quantitative research. ML methods are particularly successful in tackling problems for which large amount of data exist but no analytical solution is available or practical, such as computer vision (CV) and natural language processing (NLP) (Huang et al., 2023). ML algorithms are designed to automatically extract informative patterns in the data without explicit parametric models of the data and find solutions to specific inferential tasks. There has been several well-established applications of ML to population genetics. For example, hidden Markov models (HMMs) are widely used to segment genomic sequences and estimate ancestry (N. Li and Stephens, 2003) or other evolutionary parameters (Felsenstein and Churchill, 1996; Siepel et al., 2005; Boitard et al., 2009; Kern and Haussler, 2010) along chromosomes,

whereas principal component analysis (PCA) has become an essential tool to visualized high-dimensional genotypic matrices in low-dimensional clusters to inform relatedness among individuals (Schrider and Kern, 2018). These are examples of *unsupervised* ML where the model finds structure within the data without prior knowledge of labels provided by humans. *Supervised* ML, on the other hand, relies on training data with known labels to make predictions about new datapoints and does so by optimizing model parameters to maximize the prediction accuracy of a response variable from the input. The application of supervised ML is relatively new to the field of population genetics and was introduced by Pavlidis et al., 2010, where they addressed the lack of evolutionary ground truth (1.2) by utilizing synthetic training data generated via simulations.

The advancement in deep learning has an outsized impact on the success of the supervised ML paradigm. Deep learning uses neural networks with nodes in multiple layers that are capable of many possible mathematical operations and can therefore learn a deep hierarchical representation of the data. During training, the connection weights between nodes are optimized to minimize the loss usually defined by the prediction error. Neural networks are universal function approximators that automatically learn complex features from raw data and thus provide exceptional performance on prediction tasks across a wide range of domains (LeCun et al., 2015). The rest of this section will briefly examine the major neural network architectures with a focus on supervised deep learning models for predictive tasks (other learning paradigms in population genetics are reviewed by Huang et al., 2023; Korfmann et al., 2023, see also 5.2 for discussion of the most recent developments in generative models).

Feed-forward neural networks (FNNs) are the archetypal architecture of deep learning models. A vector of input features enters the FNN via the input layer, which is connected to the output layer via a series of hidden layers. Each layer receives output from the previous layer and applies some form of non-linear transformation through activation functions before passing its output to the next layer. Under the Universal Approximation Theorem, FNNs can approximate any continuous function to any desired precision and therefore are suitable for many generic prediction tasks (Korfmann et al., 2023). A notable early adoption of FNNs in population genetics is the evoNet model that jointly infers selection and demography from summary statistics (Sheehan and Song, 2016). FNNs have subsequently been applied to various problems such as inferring mutation rates, population structure or admixture events (Huang et al., 2023).

Convolutional neural networks (CNNs) are uniquely suited for genomic ap-

plications as they are specifically designed to handle grid-like input. Originally proposed to process image data, CNNs are widely adopted for CV tasks such as image classification or segmentation (LeCun et al., 2015). CNNs consist of consecutive sets of convolutional and pooling layers. Each convolutional layer uses a set of kernels (or filters) to generate feature maps of its input through convolution operations, and the subsequent pooling layer reduces the size of the feature maps by combining information from adjacent regions. It is straightforward to treat genotype or haplotype sequences as images and CNNs have indeed been utilized in this manner for inferring demography (Flagel et al., 2019; Sanchez et al., 2021), local ancestry (Montserrat et al., 2020; Oriol Sabat et al., 2022), selection (Flagel et al., 2019; Torada et al., 2019), introgression (Blischak et al., 2021; Gower et al., 2021; Ray et al., 2023; Zhang et al., 2023) and recombination (Adrion, Galloway, et al., 2020).

Another deep learning architecture tailored to specific input format are recurrent neural networks (RNNs). RNNs incorporate a mechanism that allows the output of a layer to flow back to a previous or current layer and thereby are capable of retaining information from previous inputs. At each stage the current input and the previous output can be combined to produce the next output. The iterative nature of an RNN makes it suitable for handling sequential inputs such as speech or text, although novel architectures based on the transformer (Vaswani et al., 2017) have emerged as the state of the art for NLP tasks (see 5.2). In the context of population genetics, the input data could be spatially sequential along biological sequences or temporally sequential along evolutionary time. RNNs have been used for genome-wide estimation of coalescence rate (Khomutov et al., 2021) and recombination landscape (Adrion, Galloway, et al., 2020).

There are rich opportunities for deep learning approaches to drive profound scientific discoveries in the genomic era, especially in light of the recent success of generative AI models (discussed in 5.2). Within the supervised learning paradigm, two challenges for building accurate and robust deep learning models in population genetics stand out. First, deep learning models so far take either summary statistics or raw genotypic data as input. Given that ARGs are the ultimate record of evolution but difficult to model analytically (see 1.1), can we take advantage of deep learning models to perform inference from the ARG? Second, since most deep learning models in population genetics rely on synthetic data for training, inference on real data is prone to be biased by ill-defined simulation model. This "simulation mis-specification problem" remains a major concern about current applications of deep learning in population genetics (Korfmann et al., 2023) and calls for a timely solution.

1.4 Studies of selective sweeps generate key insights into adaptive evolution

The ability to accurately detect and quantify the influence of natural selection from genomic sequencing data is one of the main pursuits in population genetics. Studies of natural selection aim to tackle questions ranging from the genetic basis of historical evolutionary events to the functional and disease relevance of genetic variants in both human and non-human contexts (Henry and Nevo, 2014; Karlsson et al., 2014). Selection can be broadly categorized into different modes, including positive selection, where an allele is favored and increases in frequency, negative selection (or purifying selection), where deleterious alleles are removed from the genome, and balancing selection, where multiple alleles are actively maintained at an appreciable frequency (Vitti et al., 2013).

Adaptive evolution, at the molecular level, is driven by selection acting on alleles that enhance organismal fitness, thereby increasing their frequency in a population (Vitti et al., 2013). Alleles under positive selection often carry important phenotypic relevance and therefore are of particular interest. For example, genome-wide scans of positive selection in human have identified immune-related alleles (e.g. in genes encoding signal transducer of inflammatory response) implicated in pathogen resistance (Fumagalli et al., 2011). These immune-related variants may have therapeutic relevance for both infectious diseases and autoimmune diseases. Studies of selection in plants have also identified alleles associated with adaptation to biotic and abiotic stress, potentially leading to breeding strategies that help create crops resilient to adverse conditions caused by climate change (Henry and Nevo, 2014).

Population genetics methods identify positive selection through the detection of selective sweeps, which are genomic signatures of positive selection at the population level. A hard sweep is the classic case where a *de novo* beneficial allele rapidly increases in frequency due to selection. A soft sweep differs from a hard sweep in that selection acts on a standing genetic variant rather than a *de novo* mutation. More recent works have started to elucidate cases of polygenic selection in the human genome (Berg and Coop, 2014; Racimo et al., 2018; Edge and Coop, 2019), which is a more complex scenario of positive selection where selection of a trait acts simultaneously on many variants across the genome. Traditionally, methods to detect selective sweeps take advantage of the hitchhiking effect, where the neutral alleles around the site under selection that are linked to the beneficial allele "hitch-hike" to high frequency. The hitchhiking effect leads to a characteristic reduction of genetic diversity in the vicinity of the site under selection, manifested in both the site frequency

spectrum (SFS) and the haplotype structure. The hitchhiking effect can be captured by specifically designed summary statistics to make inferences about selection. Newer approaches to detect selection include likelihood-based statistical methods and more recently machine learning methods are increasingly adopted for selection inference. Despite many advances, major limitations remain in the sensitivity and scalability of computational methods for identifying and characterizing selection. These limitations stem, in part, from the difficulty of estimating selective effects directly from DNA sequences (Hejase, Dukler, et al., 2020).

1.5 Objectives and outline of thesis

This general introductory chapter 1 is followed by three stand-alone chapters, each presenting the entirety of a research project and featuring its own introduction, results, discussion and methods. These projects share the goal of capitalizing on the progress in AI/ML and genomic big data to drive scientific discoveries in population genetics. The concluding chapter 5 discusses the findings of the chapters as a whole and raises several promising avenues for future work with a focus on opportunities presented by advancements in generative AI.

Chapter 2 presents a novel deep learning method – SIA – to infer positive selection from reconstructed ARGs. This chapter describes in detail the methodological innovations and demonstrates notable improvements in performance over a wide range of previous methods using simulation experiments. The work on SIA ultimately serves to extend the frontier of analyzing ARGs with deep learning models.

Chapter 3 describes an application of the SIA method to uncover the genetic components that drive the speciation of *Monarcha* bird populations in the Solomon Islands. This chapter provides a blueprint for addressing various open evolutionary questions with SIA.

Chapter 4 introduces an original approach to train deep learning models on both simulated and real population genetic data with a domain-adaptive neural network architecture. This framework goes beyond the current ad-hoc methods for handling the simulation mis-specification problem and is the first to effectively address this major concern about current applications of supervised ML models for population genetics inference.

Chapter 2

A deep-learning approach for inference of selective sweeps from the ancestral recombination graph

Content of this chapter was previously uploaded to bioRxiv (2021) under the title "SIA: Selection Inference Using the Ancestral Recombination Graph" by Hussein A. Hejase, Ziyi Mo, Leonardo Campagna and Adam Siepel. The manuscript was published in Molecular Biology and Evolution (2021) under the title "A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph". H.H. and Z.M. contributed equally to this work.

2.1 Abstract

Detecting signals of selection from genomic data is a central problem in population genetics. Coupling the rich information in the ancestral recombination graph (ARG) with a powerful and scalable deep-learning framework, we developed a novel method to detect and quantify positive selection: **S**election **I**nference using the **A**RG (SIA). Built on a long short-term memory (LSTM) architecture, a particular type of a RNN, SIA can be trained to explicitly infer a full range of selection coefficients, as well as the allele frequency trajectory and time of selection onset. We benchmarked SIA extensively on simulations under a European human demographic model, and found that it performs as well or better as some of the best available methods, including state-of-the-art machine-learning and ARG-based methods. In addition, we used SIA to es-

timate selection coefficients at several loci associated with human phenotypes of interest. SIA detected novel signals of selection particular to the European (CEU) population at the *MC1R* and *ABCC11* loci. In addition, it recapitulated signals of selection at the *LCT* locus and several pigmentation-related genes. Finally, we reanalyzed polymorphism data of a collection of recently radiated southern capuchino seedeater taxa in the genus *Sporophila* to quantify the strength of selection and improved the power of our previous methods to detect partial soft sweeps. Overall, SIA uses deep learning to leverage the ARG and thereby provides new insight into how selective sweeps shape genomic diversity.

2.2 Introduction

The ability to accurately detect and quantify the influence of selection from genomic sequence data enables a wide variety of insights, ranging from understanding historical evolutionary events to characterizing the functional and disease relevance of observed or potential genetic variants. Adaptive evolution is driven by increases in frequency of alleles that enhance reproductive fitness. In addition, alleles experiencing such positive selection often provide insights into the functional or mechanistic basis of phenotypes of interest. Examples of genetic determinants of important phenotypic traits under selection in human populations include a family of mutations in the hemoglobin-β cluster, which confer resistance to malaria and are at high frequencies in many populations (Currat et al., 2002; Ohashi et al., 2004), loci controlling growth factor signaling pathways that contribute to short stature in Western Central African hunter-gatherer populations (Jarvis et al., 2012; Lachance et al., 2012), as well as mutations in several genes involved in immunity, hair follicle development, and skin pigmentation (Sabeti et al., 2007)(reviewed in Sabeti et al., 2006; Kelley and Swanson, 2008; Fu and Akey, 2013; Hejase, Dukler, et al., 2020).

Population genetic methods predominantly identify positive selection through the detection of selective sweeps. As the frequency of an advantageous allele increases, linked variants in the vicinity can "hitchhike" to high frequency, leading to local reductions in genetic diversity. Previous approaches to detecting selective sweeps (such as traditional summary statistics [Tajima, 1989], approximate likelihood and approximate Bayesian computation [ABC] methods [Peter et al., 2012], or supervised ML methods [Schrider and Kern, 2016; Kern and Schrider, 2018]) exploit the effect of genetic hitchhiking on the spatial haplotype structure and SFS. Summary statistics have the advantage of being fast and easy to compute, but may confound the effects of selection on genetic diversity with the effects of complex demographic histories includ-

ing bottlenecks, population expansions, and structured populations. Besides, they cannot easily be used to estimate the value of the selection coefficient. Approximate likelihood and ABC methods, on the other hand, can provide an estimate of the strength of selection by aggregating multiple summary statistics (Peter et al., 2012), but can be prohibitively computationally expensive when applied at a large scale. ML methods for inferring selection can be more scalable and can capture complex nonlinear relationships among features. With the exception of a handful of recently developed methods that operate on the multiple sequence alignment itself (Flagel et al., 2019; Torada et al., 2019), however, the majority of ML approaches to selection inference solely make use of traditional summary statistics as features for prediction. In short, previous methods (including ABC and most ML methods) predominantly rely on low-dimensional summary statistics, which, even in combination, capture only a small portion of the information in the sequence data.

Recently, a new generation of inference methods have made it possible to go beyond summary statistics and estimate or sample a full ARG (Hudson, 1990; Griffiths and Marjoram, 1996; Wiuf and Hein, 1999) for a collection of sequences of interest. The ARG is a complex data structure that summarizes the shared evolutionary history and recombination events that have occurred in a collection of DNA sequences, and therefore contains highly informative features that can potentially be leveraged to make accurate inferences about selection. The ARG representation is interchangeable with a sequence of local genealogies along the genome and the recombination events that transform each genealogy to the next. The influence of selection on each allele can be characterized from the ARG, based on departures from the patterns of coalescence and recombination expected under neutrality as reflected in the local genealogies. Traditional ARG inference methods (Hein, 1993; Song and Hein, 2005; Kuhner, 2006; Minichiello and Durbin, 2006; O'Fallon, 2013) were restricted in accuracy and scalability, limiting the practical application of ARGs. Recent advances (Rasmussen et al., 2014), however, have enabled scalable yet statistically rigorous genome-wide ARG inference with dozens of genomes. Moreover, methods such as Relate (Speidel et al., 2019) and tsinfer (Kelleher et al., 2019) have further dramatically improved the scalability of ARG inference to accommodate thousands or even hundreds of thousands of genomes. The latest progress in genealogical inference has paved the way for ARG-based methods to address many different questions in population genetics (Arenas, 2013; Rasmussen et al., 2014; Kelleher et al., 2019; Speidel et al., 2019).

One natural way to exploit the richness of the ARG representation in inference of selection would be to extract features from inferred ARGs and feed them into a modern supervised ML framework. Deep-learning methods,

in particular, have recently achieved unprecedented success on a variety of challenging problems, including image recognition, machine translation, and game-play (LeCun et al., 2015). Deep learning is also highly flexible, providing many opportunities for the design of novel model architectures motivated by biological knowledge. An ARG-guided deep-learning model could potentially provide new insight into how natural selection impacts the human genome, human diseases and other phenotypes, and human evolution.

With these goals in mind, we developed a new method, called **S**election **I**nference using the **A**RG (SIA), that uses an RNN (Hochreiter and Schmidhuber, 1997; Maas et al., 2011) to infer the selection coefficient and AF trajectory of a variant that maps to a gene tree embedded in an ARG. Rather than relying on traditional sequence-based summary statistics, SIA makes use of features based on the local genealogies extracted from the ARG. Based on these local topological features, SIA learns to infer the selection coefficient and AF trajectory of a beneficial variant (see Fig. 2.1). As described below, SIA performs well on benchmarks and is reasonably robust to model mis-specification. Applying SIA to data from the 1000 Genomes Northern and Western European (CEU) population, we identified new and known loci under positive selection that are associated with a variety of phenotypes and estimated selection coefficients at these loci. In addition, using SIA, we built on our previous work (Hejase, Salman-Minkov, et al., 2020) on a bird species-complex in the genus Sporophila by elucidating the strength and targets of selection at specific loci tied to a collection of rapid speciation events. Overall, SIA is the first method that couples ARG-based features with an ML approach for population genetic inference.

2.3 Results

2.3.1 Methodological overview

SIA is based on a RNN that is trained to predict selection at a genomic site from genealogical features at that site of interest and nearby sites (see Materials and methods for detailed descriptions; see Fig. 2.1 for a conceptual overview of SIA; and Fig. S1 in Appendix A for an illustration of ARG features and the RNN architecture). Based on the demography of a particular population of interest, training data including genomic regions under various strengths of selection are simulated. The ARG is then inferred from each simulated data set. ARG-level statistics are extracted at the site under selection (or a neutral site) as features to be used as input to the deep-learning model. Specifically, we use lineage counts at a set of discrete time points as a fixed-

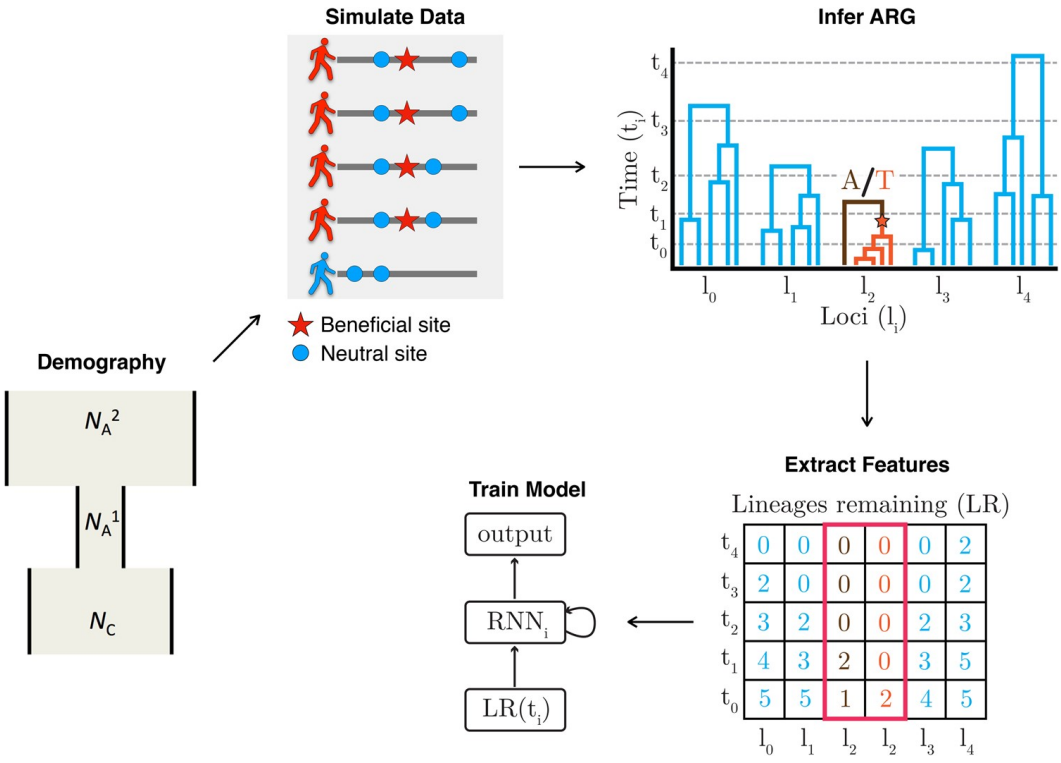


Figure 2.1: **A high-level framework for automating the detection of selective sweeps.** We first estimate the demographic history for the population of interest, then based on the estimated demographic history, we simulate neutral regions and sweeps using the discoal simulator (Kern and Schrider, 2016). We proceed with ARG inference and then extract ARG-level statistics from each simulated region. The ARG-level statistics are used as features for a deep-learning RNN model. Finally, the trained model is applied to the empirical data to infer sweeps, selection coefficients, and AF trajectories.

dimension encoding of a genealogy. The encoding of the genealogy at the focal site as well as similar encodings of flanking genealogies constitute the feature vector for that site. SIA uses a LSTM architecture, designed specifically to handle the temporal nature of the feature set. The LSTM unrolls temporally such that the lineage counts at each time point are fed to the network iteratively. Finally, the model trained on simulations is applied to ARGs inferred from empirical data to identify sweeps, infer selection coefficients, and AF trajectories.

2.3.2 Classification of sweeps

We first compared SIA with several existing methods, including the Tajima's D (Tajima, 1989) and H1 (Garud et al., 2015) summary statistics, iHS (Voight et al., 2006), a genealogy-based statistic (Speidel et al., 2019), and a summary-statistic-based ML method (Schrider and Kern, 2016; Kern and Schrider, 2018) (see Materials and methods), in the classification task of distinguishing hard sweeps from neutrally evolving regions. Our performance comparison was conducted across 16 combinations of selection coefficients and segregating allele frequencies such that the beneficial site was subjected to selection ranging from weak to strong, resulting in low to high derived allele frequencies (DAFs). Because a priori we expected sweep sites with lower selection coefficients and lower DAFs to be harder to detect, we performed a stratified analysis of SIA's performance by selection coefficient and DAF. Figure 2.2 reports the receiver-operating characteristic (ROC) curves using simulations based on the CEU demographic model (Tennessen et al., 2012) where inferred genealogies were used as input to SIA to account for gene tree uncertainty. As expected, all methods tended to perform better in a regime with higher selection coefficients and DAFs, as indicated by increasing values of the area under the ROC curve (AUROC) statistic from left to right (increasing selection) and from top to bottom (increasing DAF). SIA outperformed the other methods across model conditions, with a more pronounced performance advantage for sites under weaker selection and segregating at lower DAFs (Fig. 2.2). For each given selection coefficient, the AUROC of the Relate tree statistic (shown in red in Fig. 2.2), which measures how unlikely it is that the observed expansion of the derived lineages is purely due to genetic drift, did not substantially improve as the DAF increased. Alleles at higher frequency tend to be older and subjected to drift over longer periods, which may lead to reduced power for Relate to distinguish lineage expansion under selection from the neutral expectation. Consequently, although the ARG-based methods SIA and Relate both outperformed other methods at low DAFs, SIA was alone in maintaining this advantage at higher DAFs.

In addition, we validated the ability of SIA to classify genomic regions with additional test sets simulated under a demographic model for southern capuchinos, a group of songbirds in which we previously identified and characterized many examples of sweeps (Hejase, Salman-Minkov, et al., 2020), finding a predominance of "soft" rather than "hard" sweeps (meaning that they tend to be based on standing genetic variation rather than new mutations; see Materials and methods). Figure S2 in Appendix A reports the ROC curves for the task of distinguishing partial soft sweeps from neutral regions. Despite soft sweeps being harder to detect, the classifier achieved good perfor-

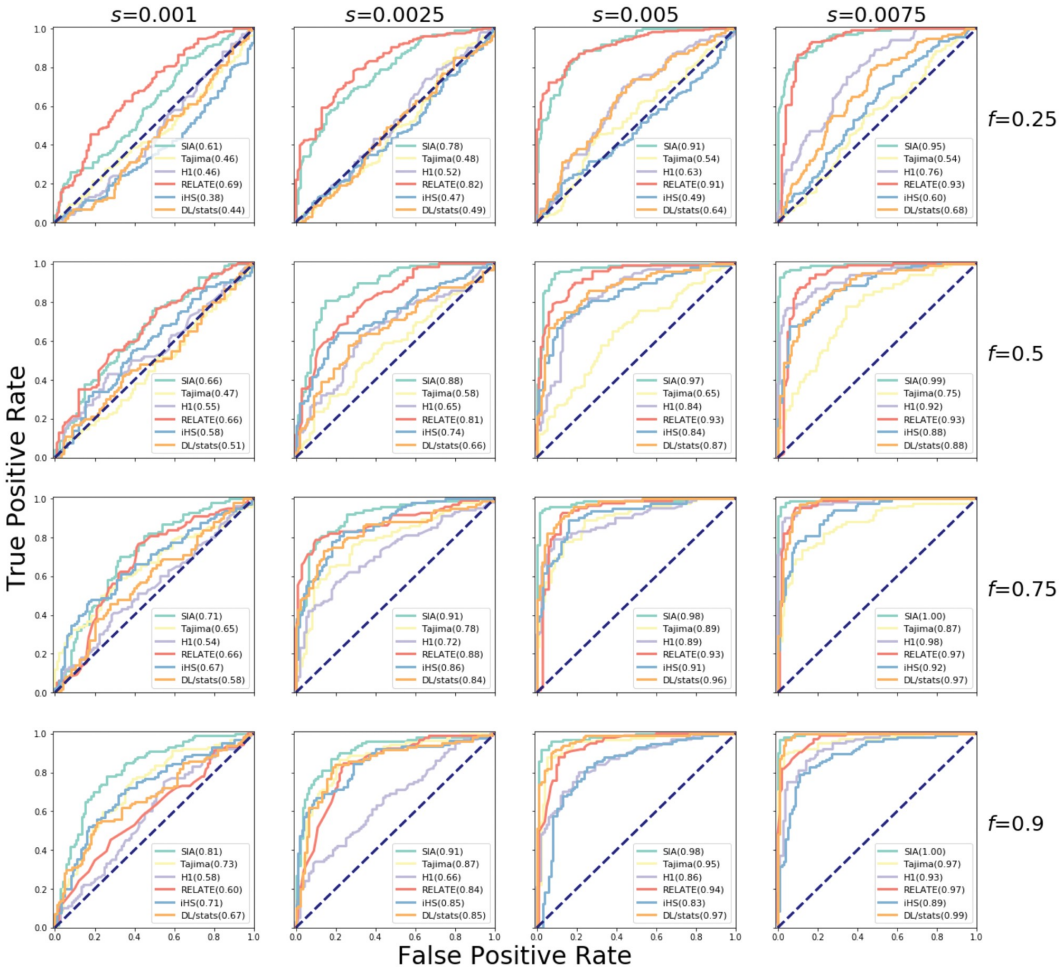


Figure 2.2: **Classification performance of SIA and other methods on simulated data.** Sequence data were simulated under a variety of selection regimes (s, shown horizontally) and DAFs for the beneficial mutation under selection (f, shown vertically) (see Materials and methods for more details). The prediction task distinguished neutral regions and sweeps. The methods were tested on a set of 200 regions per panel (100 per class), and the ROC curve records the true positive (TP) rate as a function of the false positive (FP) rate. The curve is obtained by varying the prediction threshold from 0 to 1 and recording for each threshold the number of regions correctly assigned (TPs) or misassigned (FPs) as positives (with prediction probability above the threshold). The performance of each method was evaluated based on the area under its ROC curve, or AUROC (shown in parenthesis in figure legend). Note that inferred genealogies were used as input to SIA.

mance in the moderate-to-strong selection regimes ($s = 0.005$ and $s = 0.0075$) where the accuracy ranged between 82% and 96%, a substantial improvement over the previous accuracy of 56% (Hejase, Salman-Minkov, et al., 2020). SIA performed particularly well in identifying partial soft sweeps when the site under selection was at a high segregating frequency. For example, at segregating frequencies of 0.75 and 0.9, the performance of SIA ranged between 80% and 96% across a variety of selection regimes ($s = 0.0025$, 0.005, and 0.0075). The performance of SIA degraded somewhat for weak selection ($s = 0.001$) with an accuracy ranging between 63% and 74%.

2.3.3 Selection coefficient inference using true gene trees

We assessed the performance of SIA in correctly predicting the selection coefficient and compared it with CLUES (Stern et al., 2019). Like SIA, CLUES uses local genealogies based on the ARG to infer a selection coefficient. However, CLUES calculates the likelihood of the genealogy analytically using a HMM, and does not rely on simulated training data. In addition, CLUES uses a single genealogy at the focal site, whereas SIA additionally considers flanking trees.

We began by supplying both methods with true genealogies, in order to later disentangle the error deriving from the ARG inference step from other sources of error (see Discussion). We found that SIA identified regions under neutrality with approximately no bias (median inferred $s = 7.5 \times 10^{-5}$; Fig. 2.3). Similarly, SIA correctly inferred the selection coefficient for regions under moderate to strong selection ($s \in \{0.0025, 0.005, 0.0075, 0.01\}$) with the median inferred s deviated from the true s by at most 3%. On the other hand, SIA somewhat underestimated the selection coefficient (median inferred $s = 0.00037$) for the weak selection regime (true $s = 0.001$), likely owing to limits in the training set within that selection regime (see Discussion). We further binned the results by segregating frequency and selection coefficient and found that, in general, the variance in estimates of s for SIA (as well as CLUES) tended to decrease as the segregating frequency of the beneficial allele increased (Fig. S3 in Appendix A).

CLUES performed roughly similarly to SIA in this experiment, but tended to slightly overestimate s for the neutral regions (i.e., true $s = 0$) and underestimate s for the moderate to high selection regimes (i.e., true $s = 0.005$, 0.0075, and 0.01). Under these conditions, SIA's median predictions of s were noticeably closer to the true values (Fig. 2.3A). At the same time, CLUES performed slightly better than SIA in weak selection regimes (i.e., true $s = 0.001$ and 0.0025) (Fig. 2.3). Overall, SIA (RMSE $= 9.52 \times 10^{-4}$) achieved a lower error in estimating s than CLUES (RMSE $= 1.44 \times 10^{-3}$), when true

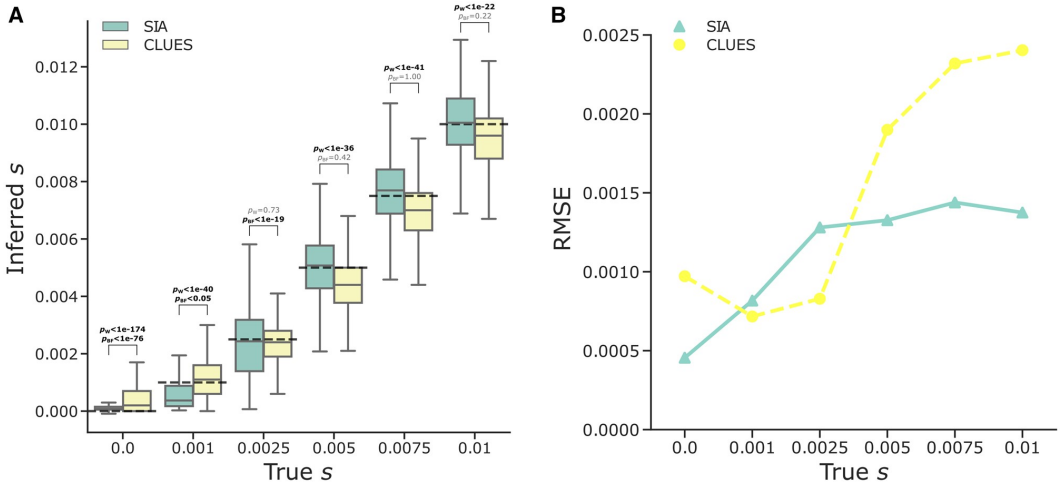


Figure 2.3: **Predictions of selection coefficients on simulated regions using SIA and CLUES based on true genealogies. (*A*)** The distribution of inferred selection coefficients for each method under each model condition are reported using a box plot. The box plot for each method reports these five statistics (from bottom to top): minimum, first quartile, median, third quartile, and maximum. The y-axis shows the inferred selection coefficient, whereas the x-axis shows the true selection coefficient. The dashed-black line indicates the true selection coefficient for each model condition. The simulations are based on the CEU demographic model and true genealogies were used as input to both methods. Each model condition (i.e., box plot) represents a set of 400 independent simulations. The mean ranks and variances of the distributions of inferred s were compared using the Wilcoxon signed-rank test (p_{W}) and the Brown–Forsythe test (p_{BF}), respectively. **(*B*)** The root mean square error (RMSE) for each method under each model condition evaluated on 400 independent simulations.

genealogies were used as input to both methods (Wilcoxon signed-rank test for difference in mean of squared error, $P = 1.25 \times 10^{-42}$). This finding potentially reflects the benefit of linkage information utilized by SIA through the additional flanking genealogies (see Discussion).

2.3.4 Selection coefficient inference using inferred gene trees

To account for gene-tree uncertainty, we next used ARGs inferred with Relate, which is scalable to the size of the training data set for SIA (see Materials and methods), as input to SIA and CLUES and compared their performance on

CEU simulations. Using a reduced sample size of 32 haplotypes, we additionally compared SIA with CLUES supplied with genealogies sampled using ARGweaver. Furthermore, we compared both methods with a supervised ML method, ImaGene (see Fig. S23 in Appendix A), that operates directly on an image of the alignment itself. ImaGene does not require gene trees as input and instead uses a CNN to perform dimensionality reduction of the sequence alignment, allowing for accurate and efficient classification and regression.

Overall, we found that SIA and ImaGene outperformed CLUES in these experiments (Fig. 2.4). CLUES tended to underestimate selection coefficients for the moderate-to-strong selection regimes, to a greater extent compared with the case where true genealogies were used for inference (Figs. 2.3A and 2.4A). This decrease in performance of CLUES evidently derives from error at the ARG reconstruction step. SIA, on the other hand, appeared to be more robust to the same ARG reconstruction error, and maintained an advantage even when CLUES was provided posterior samples of genealogies from ARGweaver (Fig. S5 in Appendix A). ImaGene performed remarkably similarly to SIA, given that it relies solely on the sequence alignment. SIA exhibited lower error at neutral sites and sites with low-to-moderate values of s, whereas ImaGene prevailed at sites under strong selection (Fig. 2.4B). Nevertheless, SIA showed a slightly smaller overall RMSE (2.75×10^{-3}) compared with ImaGene (2.91×10^{-3}) (Wilcoxon signed-rank test, $P = 6.18 \times 10^{-38}$), and in particular, SIA produces estimates of s much closer to 0 for neutral loci. Notably, in this case both SIA and ImaGene were trained with simulations under the same uniform distribution of s values (see Materials and methods). A different choice of training distribution could impact their performance across selection regimes (see Discussion). Furthermore, we binned the results of these methods by both the segregating frequency and the selection coefficient (see Fig. S4 in Appendix A) and again found that in general they exhibit higher variance under low segregating frequency of the beneficial allele. As before, we also tested our regression framework on true and inferred gene trees of test sets simulated under the *Sporophila hypoxantha* demographic model (see Fig. S6 in Appendix A). We found that SIA was approximately unbiased for the moderate ($s = 0.005$) and high ($s = 0.01$) selection regimes but appeared to overestimate the selection coefficient for regions under weak selection ($s = 0.001$ and 0.0025), when both true and inferred genealogies were used as input. Furthermore, SIA appeared to overestimate the selection coefficient for neutral regions when inferred gene trees were used as input, whereas it was approximately unbiased for true gene trees.

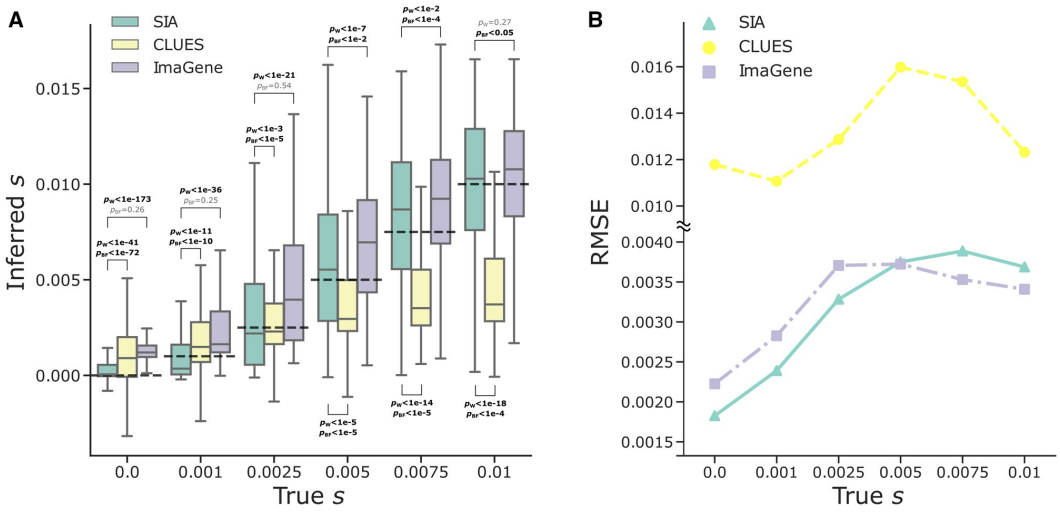


Figure 2.4: **Predictions of selection coefficient on simulated regions using SIA and CLUES based on inferred genealogies, and ImaGene. (*A*)** The distribution of inferred selection coefficients and **(*B*)** RMSE for each method under each model condition. The simulations are based on the CEU demographic model where inferred genealogies were used as input to SIA and CLUES, whereas sequence alignments were used as input to ImaGene. Figure layout and description are otherwise similar to Figure 2.3.

2.3.5 Performance on selection coefficient prediction with different sample sizes

To explore the tradeoffs associated with the use of larger data sets, we examined the performance of SIA under different sample sizes, assuming a constant-sized demographic model ($N_e = 10,000$). Figure S7 in Appendix A shows the error in selection coefficient inference on a held-out test set, stratified by the age of the allele (Fig. S7A and B in Appendix A) and present-day DAF (Fig. S7C and D in Appendix A) at the site of interest. We observed that sites with low frequency (AF < 0.33) and more recent (onset $< 0.2 \times 2N_e$ generations) alleles experience the most significant reduction in error as sample size increases. Notably, the performance of SIA on more ancient alleles (onset $> 0.2 \times 2N_e$ generations) had little to no improvement as the sample size increased from 32 to 254. These observations are in line with the expectation that having more samples improves the chance of capturing low-frequency alleles, but provides limited information about more ancient events. The reason for this age-dependency is that, looking backwards in time, most lineages coalesce rapidly and only a few survive to more ancient epochs, in a manner that

depends only weakly on the sample size. It may be useful to consider these observations when choosing the sample size for use in studying selection in a particular context (see Discussion).

2.3.6 Inference of AF trajectory

We further adapted the deep-learning architecture of SIA to model the AF trajectory at a site by retaining the output of the LSTM at each time point (Fig. S1 in Appendix A; see Materials and methods). We then evaluated the performance of SIA in the inference of the AF trajectory using simulations under the CEU demography across a range of selection coefficients and current DAFs. SIA was largely able to capture the expected trend of more rapidly increasing AF under stronger selection (Figs. S8 and S11 in Appendix A). In addition, AF estimates by SIA using both true and inferred genealogies were generally unbiased, although AF at more recent time points tended to be slightly underestimated when data was simulated under weaker selection. AF estimates also appeared to be more accurate in terms of variance for alleles under stronger selection (Figs. S9 and S12 in Appendix A). As expected, the variance of AF estimates tended to increase going further back in time (Figs. S9 and S12 in Appendix A). We also observed that overall SIA tended to produce more accurate AF estimates than CLUES (Figs. S9 and S10 in Appendix A).

2.3.7 Model performance on simulations with mis-specified demographic models

To evaluate the robustness of SIA to mismatches between the demographic parameters used for simulating training data and the true underlying demography of real data, we tested the method on the selection-coefficient inference task with data sets simulated under a range of alternative parameters. Each aspect of this model mis-specification was assessed independently of the others. In particular, the mis-specified data sets contained simulations under 1) combinations of population mutation (θ) and recombination (ρ) rates sampled beyond the range used for the training data (Figs. S13 and S16 in Appendix A); 2) various alternative demographic scenarios (Figs. S14, S17, and S19 in Appendix A); and 3) various effective population sizes (Figs. S15 and S18 in Appendix A). We compared the performance of SIA on these mis-specified data sets with that of CLUES (Stern et al., 2019), supplying both methods with the true genealogies. We consider CLUES the "silver standard" when it comes to robustness because it is unsupervised and therefore should not be

susceptible to mis-specified training data compared with supervised learning methods such as SIA. Overall, we found that both CLUES and SIA were reasonably robust to model mis-specification (Figs. S13–S15 in Appendix A), although the performance of both methods inevitably declined when tested on severely mis-specified data (Fig. S15 in Appendix A). Interestingly, SIA tended to overestimate selection coefficient when the true N_e was much smaller than that used for training, and underestimate it when the true N_e was much larger, whereas CLUES did the opposite (Fig. S15 in Appendix A). Because the CLUES likelihood model of AF transition is parameterized by the population-scaled selection coefficient ($\alpha = 2Ns$), a larger N_e likely appears to CLUES as equivalent to a higher s. On the other hand, features used by SIA capture broad information of coalescence and linkage in the ARG, and therefore can be distorted by mis-specified N_e in more subtle ways (see Discussion). Using the same mis-specified data set, we also ran SIA with Relate-inferred genealogies and compared its performance with that of the genotyped-based deep-learning model ImaGene (Flagel et al., 2019; Torada et al., 2019). In general, SIA appeared to be more robust to model mis-specifications, achieving an overall RMSE of 0.00362, 0.00318, and 0.00374 in the mis-specified θ/ρ, demography, and N_e experiments, respectively, compared with ImaGene, whose RMSE was 0.00416, 0.00330, and 0.00462 in the corresponding experiments (Figs. S16–S18 in Appendix A). The advantage of SIA was particularly noticeable in cases of mis-specified demographic parameters (Figs. S17 and S18 in Appendix A). Notably, SIA exhibited reduced bias when working with inferred genealogies compared with true genealogies, under conditions of extremely mismatched N_e (compare Figs. S15 and S18 in Appendix A).

2.3.8 Model prediction at genomic loci of interest in CEU population

We then applied the SIA model to identify selective sweeps and infer selection coefficients at selected genomic loci in the 1000 Genomes CEU population. These loci included the canonical example of selection at the *MCM6* gene, which regulates the neighboring *LCT* gene and contributes to the lactase persistence trait (Bersaglieri et al., 2004), the *ABCC11* gene regulating earwax production, several pigmentation-related genes, as well as genes associated with obesity, diabetes and addiction (Table 2.1).

For *LCT*, SIA detected a strong signal of selection at the nearby SNP that has been associated with the lactase persistence trait (rs4988235). At this SNP, SIA inferred a sweep probability close to 1 and a selection coefficient > 0.01, making this one of the strongest signals of selection in the human

Table 2.1: List of genomic loci of interest along with their derived allele frequencys (DAFs), sweep probabilities, and selection coefficients inferred by SIA in the 1000 Genomes CEU population.

Gene	**SNP ID**	**Chr**	**Position***	**DAF**	$P_{\mathbf{sweep}}$	**Selection coefficient (95% CI)**
LCT (Bersaglieri et al., 2004)	rs4988235	2	136608646	0.74	0.999	[0.01019, 0.01056]
OCA2 (Han et al., 2008; Sturm et al., 2008)	rs12913832	15	28365618	0.77	0.750	[0.00539, 0.00575]
MC1R (Sulem et al., 2007; Han et al., 2008)	rs1805007	16	89986117	0.12	0.949	[0.00362, 0.00384]
ABCC11 (Yoshiura et al., 2006)	rs17822931	16	48258198	0.13	0.620	[0.00034, 0.00036]
ASIP (Eriksson et al., 2010)	rs619865	20	33867697	0.12	0.777	[0.00172, 0.00197]
TYR (Sulem et al., 2007; Eriksson et al., 2010)	rs1393350	11	89011046	0.24	0.616	[0.00085, 0.00135]
KITLG (Sulem et al., 2007)	rs12821256	12	89328335	0.13	0.869	[0.00183, 0.002]
TYRP1 (Kenny et al., 2012)	rs13289810	9	12396731	0.37	0.144	[0.00004, 0.00006]
TTC3 (F. Liu et al., 2010)	rs1003719	21	38491095	0.62	0.011	[0, 0]
OCA2	rs7495174	15	28344238	0.94	0.013	[0, 0.00005]
TCF7L2 (Lyssenko et al., 2007)	rs7903146	10	114758349	0.69	0.035	[0, 0]
ANKK1 (Spellicy et al., 2014)	rs1800497	11	113270828	0.80	0.045	[0, 0]
FTO (Frayling et al., 2007)	rs9939609	16	53820527	0.56	0.011	[0, 0]

*Genomic coordinates in GRCh37 (hg19) assembly

genome. A close examination of the local genealogy at this site reveals a clear pattern indicative of a selective sweep—a burst of recent coalescence among the derived lineages (orange taxa are the lineages carrying the derived allele) is clearly visible from the tree (Fig. 2.5).

At a number of pigmentation genes (Sulem et al., 2007; Han et al., 2008; Sturm et al., 2008; F. Liu et al., 2010; Kenny et al., 2012), SIA detected signals of moderate selection, including *MC1R* (rs1805007, $P_{\text{sweep}} = 0.95$, $s \approx 0.0037$), *KITLG* (rs12821256, $P_{\text{sweep}} = 0.87$, $s \approx 0.0019$), *ASIP* (rs619865, $P_{\text{sweep}} = 0.78$, $s \approx 0.0019$), *OCA2* (rs12913832, $P_{\text{sweep}} = 0.75$, $s \approx 0.0056$), and *TYR* (rs1393350, $P_{\text{sweep}} = 0.62$, $s \approx 0.0011$). In addition, SIA identified a weak signal of selection at a SNP in the *ABCC11* gene (rs17822931), which influences earwax and sweat production (Yoshiura et al., 2006), with a selection coefficient of around 0.00035. There are few other estimates for these genes available for comparison, but, notably, our estimate for *LCT* of $s \approx 0.01$ is consistent with a previous estimate on the order of 0.01–0.1 (Bersaglieri et al., 2004), and with recent studies of ancient DNA samples (S. Mathieson and Mathieson, 2018; I. Mathieson, 2020) suggesting a value closer to 0.01. Our estimates suggest that selection at the pigmentation loci is considerably weaker than at *LCT*, in contrast to previous estimates for these loci, which covered a wide range but were generally considerably larger (ranging from 0.02 to 0.1) (Wilde et al., 2014). Interestingly, CLUES estimated s at the *OCA2* locus to be on the order of 0.001 (roughly similar to SIA's estimate of 0.0056), but s at the *KITLG*, *ASIP*, *TYR* loci to be > 0.01 (in comparison to SIA's considerably smaller estimates of 0.0019, 0.0019, and 0.0011) (Stern et al., 2019). The apparent discrepancy between the estimates may be partially due to the fact that the two methods used samples from two different populations (CEU for SIA and GBR/British for CLUES).

On the other hand, SIA did not detect significant evidence of positive selection at several disease-associated loci (rs7903146/*TCF7L2*, rs1800497/*ANKK1*, and rs9939609/*FTO*) or at several other pigmentation loci (rs13289810/*TYRP1*, rs1003719/*TTC3*, and rs7495174/*OCA2*) (Table 2.1). Notably, allele frequencies at these six loci tend to be similar in African and European populations (Marcus and Novembre, 2017), suggesting that they are not likely to be under strong environment-dependent positive selection, although it is possible that they have experienced very recent selective pressure that SIA lacks the power to detect (see Discussion). Notably, *TYRP1* and *TTC3* also lacked signals of selection in the CLUES analysis. Compared with the genealogies at sweep sites (Fig. 2.5), the trees at these putatively neutral loci lack the distinctive signature of recent bursts of coalescence among derived lineages (Fig. 2.6).

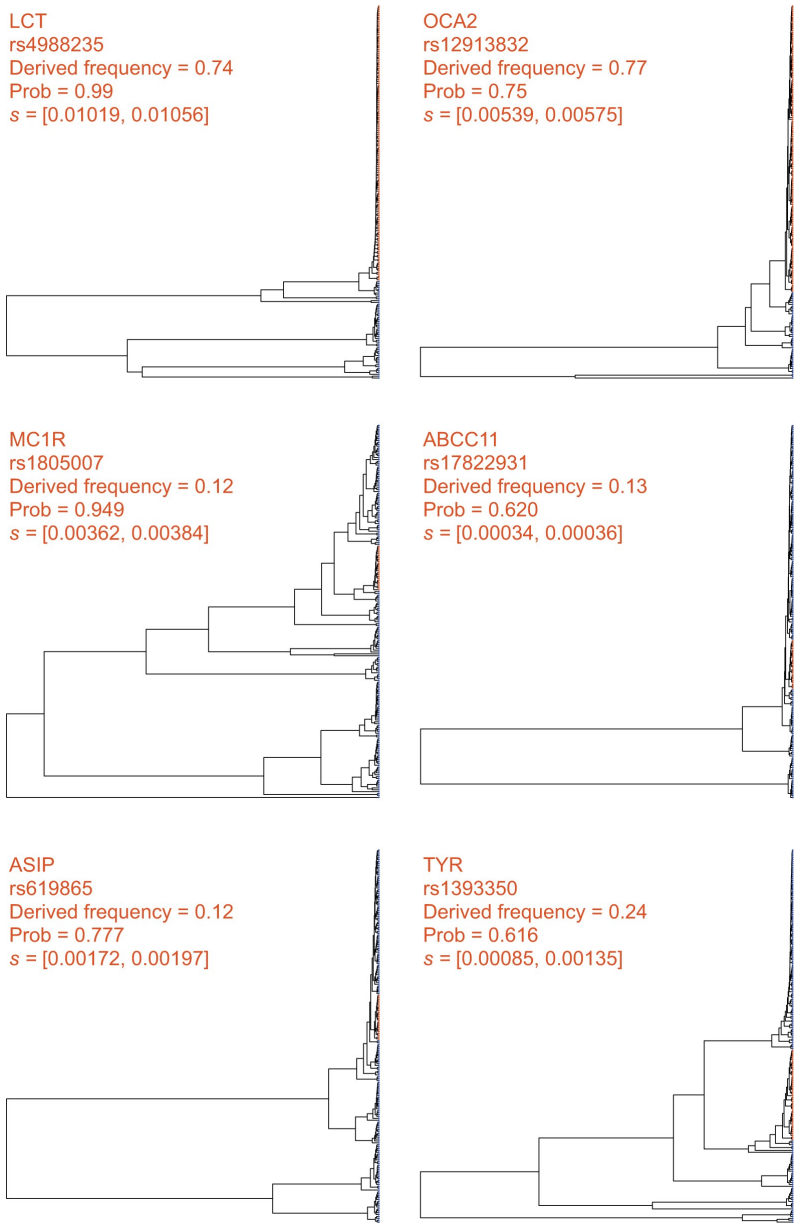


Figure 2.5: **Local genealogies at six loci inferred to be under positive selection in the 1000 Genomes CEU population.** Gene name, RefSNP number, derived AF, SIA-inferred sweep probability and SIA-inferred selection coefficient range for each locus are indicated at the top of each panel (see Table 2.1 for more details). Taxa carrying the ancestral and derived alleles are colored in blue and orange, respectively.

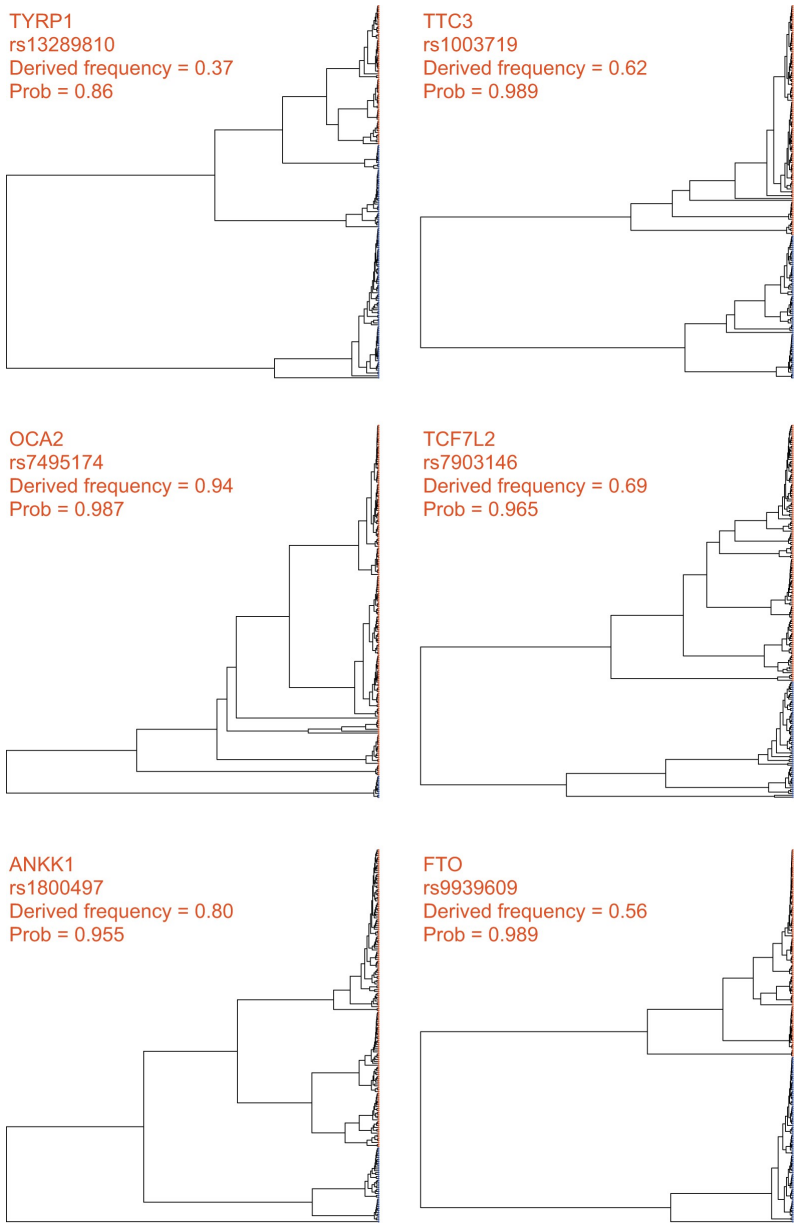


Figure 2.6: **Local genealogies at six loci lacking signal of positive selection in the 1000 Genomes CEU population.** Gene name, RefSNP number, derived AF and probability of neutrality inferred by SIA for each locus are indicated at the top of each panel (see Table 2.1 for more details). Taxa carrying the ancestral and derived alleles are colored in blue and orange, respectively.

2.3.9 Southern capuchino species analysis

Our previous study of southern capuchino seedeaters made use of the full ARG and ML to detect and characterize selective sweeps, and suggested that soft sweeps are the dominant mode of adaptation in these species (see Materials and methods for more details). To further characterize the targets and strengths of positive selection in these species, we applied SIA to polymorphism data (Turbek et al., 2021) for *S. hypoxantha*, and adopted a conservative approach by reporting only sites with DAF ≥ 0.5, SIA-inferred $s \geq 0.0025$, and SIA-inferred sweep probability $P_{\text{sweep}} \geq 0.99$ (see Materials and methods). In addition to loci near top F_{ST} peaks and known pigmentation-related genes (Table 2.2), we identified many more sites under positive selection located outside the previously scanned F_{ST} peaks, amounting to a total of 15,551 putative partial soft sweep sites across the 333 scanned scaffolds for *S. hypoxantha*. These sites can be prioritized for further evaluation and downstream analysis. Notably, SIA enabled us to distinguish between selection at regulatory and coding sequences, and we found that sweep loci near F_{ST} peaks and pigmentation genes fall mostly in noncoding regions (Table 2.2). We additionally surveyed all putative sweep sites identified by SIA and found that they are indeed enriched in noncoding regions (Fisher's exact test, $P = 6.80 \times 10^{-5}$), particularly noticeable in the "near-coding" regions (Fig. S22 in Appendix A). Consistent with the observation that the most highly differentiated SNPs among taxa are noncoding (Campagna et al., 2017; Turbek et al., 2021), our finding suggests that positive selection may act on *cis*-regulatory regions to drive differentiation and the subsequent speciation process. Furthermore, we examined many individual predictions in detail, considering the local trees inferred by Relate at these high-confidence predictions (Fig. 2.7). We found, in numerous cases, that these sweeps had distinct genealogical features, displaying evidence of a burst of coalescence events, corresponding to unusually large and young clades. Prominent examples include predictions near pigmentation-related genes *ASIP*, *KITL*, *SLC45A2*, and *TYRP1*.

2.4 Discussion

The ARG is useful for addressing a wide variety of biological questions ranging from inferring demographic parameters to estimating allele ages. SIA exploits the particular utility of the ARG for accurate inference of positive selection in a way that makes use of the full data set, as opposed to traditional summary statistics, which necessarily discard substantial information. Direct use of the ARG improves upon traditional summary statistics in two key ways.

Table 2.2: The top 25 F_{ST} peaks identified in Hejase, Salman-Minkov, et al., 2020 along with the number of partial soft sites in *S. hypoxantha* identified for each scaffold using SIA.

Scaffold	**Start position (Mb)**	**End position (Mb)**	**Length (kb)**	**No. of partial soft sites***
59	5.74	5.86	120	11
118	7.16	7.22	60	5
252	0.40	0.54	140	3
257.1	21.24	21.78	540	26
257.2	24.40	24.84	440	43
257.3	28.66	28.96	300	10
257.4	31.30	31.38	80	8
257.5	5.78	6.20	420	25 (1)
263	0.00	0.58	580	31
308	0.04	0.20	160	0
404.1	5.04	5.84	800	115 (7)
404.2	10.76	10.96	200	30
412	3.38	3.62	240	15
430	10.98	11.10	120	24
567	2.50	2.80	300	0
637.1	6.00	6.32	320	2
637.2	6.84	6.92	80	4
762	1.65	1.73	80	30
766	1.98	2.10	120	1
791	9.90	9.98	80	15
1,717	0.92	0.98	60	7
3,622	0.96	1.36	400	8
1,635	3.71	3.75	40	4
1,954	2.8	2.9	100	17
579	0.1	0.16	60	0

Note: To avoid cases with limited power, we focused on sites with segregating frequency ≥ 0.5, SIA-inferred $s > 0.0025$, and SIA-inferred sweep probability $P_{\text{sweep}} \geq 0.99$.

* The number of sweep sites in coding regions is shown in parenthesis.

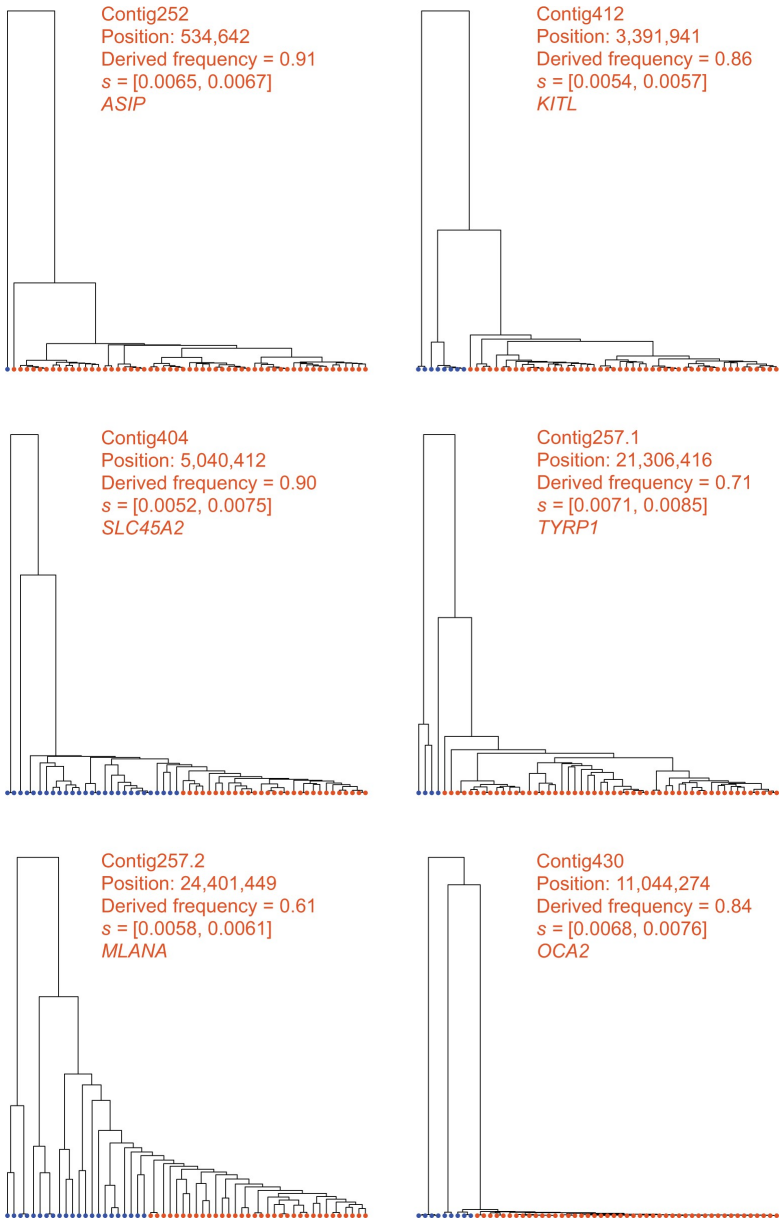


Figure 2.7: **Local genealogies at six loci inferred to be under positive selection in *S. hypoxantha*.** Contig name, position of SNP, DAF, SIA-inferred selection coefficient range, and the pigmentation gene closest to the locus in question are indicated at the top of each panel. Haploid genomes carrying the ancestral and derived alleles are colored in blue and orange, respectively.

First, it enables consideration of the temporal distribution of coalescence and recombination events in the history of the analyzed sequences, in contrast to traditional summary statistics that simply average over these coalescence and/or recombination events. In addition, ARG-based methods provide better spatial resolution by separately examining individual genealogies and the recombination breakpoints between them, rather than averaging across windows containing unknown numbers of genealogies. These detailed patterns of coalescences and linkage enable the ARG-based approaches to capture a more localized and fine-grained picture of selection (e.g., infer selection coefficient and AF trajectory) as well as to achieve a better classification performance. This performance advantage is particularly noticeable at lower DAFs and when selection is weak, a regime where previous methods for selection inference fall short (Fig. 2.2).

At the same time, the supervised ML approach sets SIA apart from another ARG-based method, CLUES, which approximates a full likelihood function for ARGs in the presence of selection using importance sampling and an HMM. Although the accuracy of both SIA and CLUES degraded when using inferred genealogies compared with true genealogies, reflecting the error and uncertainty at the ARG inference step, SIA appeared to be more robust to gene tree uncertainty (Figs. 2.3 and 2.4). One possible reason for this observation is that CLUES effectively assumes that the selection coefficient at the focal site is conditionally independent of the flanking trees given the focal tree. This assumption should hold in the presence of fully specified genealogies, but it may make CLUES more sensitive to errors in the inferred genealogies. In other words, through its use of supervised learning, SIA may be able to compensate for the effects of genealogy inference error on its estimation of the selection coefficient by also directly considering the flanking trees and LD-related patterns among them. Still, the drop in accuracy observed across methods underscores the dependency of ARG-based approaches on the ARG inference method. For this reason, we anticipate that SIA may benefit substantially from further improvement in ARG inference tools (see Hejase, Dukler, et al., 2020).

The ARG-based feature set distinguishes SIA from other supervised ML approaches for characterizing selective sweeps. SIA uses local topological features of the ARG, which are more informative than the SFS- or LD-based summary statistics employed by ML methods such as S/HIC, SFselect, and evolBoosting. Using simulations, we demonstrated that the SIA classifier outperformed a deep-learning method that aggregates these traditional summary statistics (Fig. 2.2). We also compared SIA with ImaGene, which represents another flavor of supervised learning methods, inspired by the recent rise of CNNs for image recognition. ImaGene encodes sequence alignments as images

for powerful population genetic inferences with CNNs and provides a state-of-the-art benchmark to compare against. We found that ImaGene performs remarkably well across a wide range of simulations, but SIA does appear to be somewhat less biased and more robust to model mis-specification than ImaGene. The evolutionary information in the ARG is implicit in the sequence alignment but some of this information may be difficult for a brute-force ML model to discover directly.

We demonstrated that utilizing the ARG granted SIA considerably improved performance over deep-learning models solely employing traditional summary statistics. However, a possible drawback of an ARG-based model is the potentially prohibitive computational overhead incurred by ARG inference, especially as sample size grows. Picking a sample size when running SIA involves a tradeoff between scalability (fewer samples, faster ARG inference) and performance (more samples, slower ARG inference). We have found that SIA can infer selection coefficients reasonably well with as few as 16 haplotypes. Including more samples did improve performance but with a sublinear reduction in error (Fig. S7 in Appendix A). Therefore, a sample size from a few dozen to a few hundreds—well within the capabilities of most modern ARG inference methods—strikes a good balance between performance and scalability. Moreover, we found that larger sample sizes improved prediction performance primarily for alleles at lower frequencies but had little impact on the performance for more ancient alleles (as most lineages would have already coalesced going further back in time) (Fig. S7 in Appendix A). This observation suggests that the choice of the sample size when applying SIA should be guided by the biological question of interest – ancient selection can be studied with just a handful of samples, whereas a larger sample size is better suited to detect more recent sweeps. Notably, the addition of ancient DNA samples could potentially enable selection to be inferred over much longer time scales. It should be possible to accommodate them with a relatively straightforward extension of the method.

Like other supervised learning methods, SIA relies on simulations to generate training data. In order to apply SIA in a particular population, a fresh set of training data tailored to that population needs to be simulated. Although it takes on the order of 100 CPU hours to simulate the training data compared with ten CPU hours to train the model (see Materials and methods), simulations can be easily distributed across multiple machines as each of them runs independently. Another potential drawback common to supervised methods is that they could be biased by subjective choices of simulation parameters. For example, SIA and ImaGene cannot make accurate predictions of selection coefficients outside the range represented in the training data (Fig. S20 in Ap-

pendix A), whereas unsupervised methods such as CLUES are not limited to a predefined range (Fig. S21 in Appendix A). This problem could be circumvented by training on an extended range of s. Similarly, the tendency of SIA to underestimate the selection coefficient for sites under weak selection (Figs. 2.3 and 2.4) could be mitigated by augmenting the training set with simulations densely sampled from the weak selection regime. A more subtle issue, however, arises when the underlying generative process of the real data does not match the assumptions made for the simulations of the training data, potentially compromising the accuracy of the method when applied to real data. Thus, we tested SIA on simulations with parameters mismatching those used in the training procedure. In general, we found that SIA was fairly robust to alternative parameter values, although, as expected, performance did degrade somewhat under severely mis-specified models. Notably, SIA achieved a similar level of robustness to model parameter mis-specification as the unsupervised (i.e., not relying on training data) likelihood method CLUES, yet outperformed the supervised deep-learning method ImaGene.

Applying SIA to the CEU panel from the 1000 Genomes Project yielded several noteworthy findings at loci with known ties to phenotypes of interest. In addition to confirming the canonical signal of selective sweep at the *LCT* locus, SIA detected a novel signal of selection at a GWAS SNP in the *MC1R* gene associated with red hair color, contrasting a previous study that could not find evidence of selection at *MC1R* in the European population (Harding et al., 2000). The derived allele at this locus segregates at around 10% in the CEU population but is nearly absent in non-European populations (Marcus and Novembre, 2017). In addition, at the *MC1R* locus the Relate test statistic for selection (Speidel et al., 2019), which tends to perform particularly well at low segregating frequencies (Fig. 2.2), falls slightly below the significance threshold of 0.05, supporting the evidence of positive selection at this locus. SIA also detected evidence of selection at a SNP in the *ABCC11* gene reported to be the determinant of wet versus dry earwax as well as sweat production, mirroring the signal of selection previously found in the East Asian population (Ohashi et al., 2011), although selection in the CEU population appeared to be much weaker. In addition, SIA identified selection at a few other pigmentation-related loci, yet determined previously identified SNPs in the *TYRP1* and *TTC3* genes to be largely free from selection (Table 2.1). These results were consistent with a previous study (Stern et al., 2019), which reported similar results for these pigmentation-related loci, albeit in a slightly different population (GBR). SIA notably did not detect positive selection at GWAS loci in the *TCF7L2* gene associated with type-2 diabetes, the *ANKK1* gene implicated in addictive behaviors, and the *FTO* gene associated with

obesity. Overall, this empirical study with the 1000 Genomes CEU population has illustrated how SIA can be applied to assess natural selection at the resolution of individual sites, suggesting that it may be useful in prioritizing GWAS variants for further scrutiny.

In our previous work on southern capuchino seedeaters (Hejase, Salman-Minkov, et al., 2020) (see Materials and methods), we applied newly developed statistical methods for ARG inference and ML for the prediction of selective sweeps. We found evidence suggesting that a substantial fraction of soft sweeps is partial but had limited power to identify them (i.e., average accuracy of 56%). SIA considerably improved our characterization of positive selection in the southern capuchino species in two key ways. The SIA framework performs inference of selection directly from genealogies instead of traditional summary statistics, and in doing so achieved an accuracy of up to 96% in detecting partial soft sweeps. Consequently, we found abundant evidence of soft sweeps beyond the previously scanned F_{ST} peaks, and additionally were able to estimate their selection coefficients. Importantly, SIA also took the analysis of selection beyond broad genomic windows containing sweeps to the identification of specific putative causal variants. We took advantage of this substantial improvement in genomic resolution and analyzed the distribution of these sweep sites, which revealed that positive selection on regions that likely contain *cis*-regulatory elements plays a role in driving the differentiation and speciation of southern capuchino seedeaters.

Although we believe SIA represents an important step forward in the use of the ARG for ML-based selection inference, there remain several possible avenues for improvement. For example, SIA currently uses a point-estimate of the ARG, rather than a distribution, and therefore does not explicitly take gene-tree uncertainty into account. Instead, the uncertainty of the inferred parameters is estimated with neural network dropouts (Gal and Ghahramani, 2016). The variance of parameter inference could alternatively be assessed from uncertainty in genealogy reconstruction by resampling coalescent times with Relate (Speidel et al., 2019), and moreover resampling trees from the posterior distribution of ARGs with ARGweaver (Rasmussen et al., 2014). Thus, it may be enlightening to compare these different approaches to analyzing uncertainty. Likewise, SIA will greatly benefit from better algorithms for ARG reconstruction that balance accuracy with scalability and can handle thousands of genomes. In addition, the SIA framework was applied in the context of single-locus selective sweeps, but could be extended to study polygenic selection, by making use of summary statistics from genome-wide association studies (as in Stern et al., 2021) and adapting the architecture of our neural network to account for selection acting at multiple sites. Finally, the robust-

ness of SIA to model mis-specifications can be further improved by ensuring the simulated data is generated under a distribution that is compatible with the real target data set. We anticipate that the continual advancement in ARG inference methods has the potential to open up many new applications for this flexible and powerful model of ARG-based deep learning in population genetics.

2.5 Materials and methods

2.5.1 Simulated data sets used for training and testing the SIA model

Training and testing data sets were generated using discoal (Kern and Schrider, 2016) by simulating 1,000,000 regions of length 100 kb for each model we considered (i.e., "neutral" or "hard sweep"). Aside from these regions, 2,000 were simulated for validation and 5,000 were simulated for testing. The number of sampled sequences was selected to match the number of individuals in the CEU population in the 1000 Genomes data set. Thus, a total of 198 haploid sequences were sampled. Simulations used a demographic model based on European demography (Tennessen et al., 2012). In non-neutral simulations, selection was applied to a single focal site located in the middle of the simulated region. We sampled each of the main demographic and selection parameters from a uniform distribution: 1) mutation rate $\mu \sim \mathcal{U}(1.25 \times 10^{-8}, 2.5 \times 10^{-8})$; 2) recombination rate $\rho \sim \mathcal{U}(1.25 \times 10^{-8}, 2.5 \times 10^{-8})$; 3) selection coefficient $s \sim \mathcal{U}(0.0001, 0.02)$; and 4) segregating frequency of the site under selection $f \sim \mathcal{U}(0.01, 0.99)$. The total storage footprint for the simulations was 1.6TB. The average cost of one simulation was 0.53 s, amounting to a total of 148 CPU hours to simulate the entire training set. The cost of simulation was mitigated by parallelization across multiple compute nodes.

2.5.2 ARG feature extraction

For each target variant, we extracted the corresponding gene tree from the ARG, then overlaid it with 100 discrete timepoints. These timepoints were fixed across all trees in an approximately log-uniform manner that resulted in finer discretization of more recent time scales (as in Rasmussen et al., 2014). We considered biallelic sites only and assumed no recurrent mutations; thus, each mutation was assumed to occur on the branch of the tree where the ancestral allele switches to the derived. For each timepoint, we calculated the number of active ancestral and derived lineages. Furthermore, we computed

the number of all active lineages (not distinguishing between ancestral and derived) at the same set of predefined timepoints in the two left- and right-flanking gene trees to account for linkage disequilibrium. We experimented with alternative numbers of flanking gene trees and found that the SIA model with two flanking gene trees (RMSE = 0.0027) outperforms a model with one (RMSE = 0.0029) or no (RMSE = 0.0030) flanking gene tree. Generally, more gene trees provide SIA with richer linkage information and thus improve its ability to estimate the effect of positive selection on a locus. The exact threshold of diminishing returns, however, can be computationally costly to establish. We therefore opted to include two flanking gene trees while noting that the user can control this hyperparameter when running SIA.

In the end, the ARG feature for each locus consisted of a 600-dimensional vector, which was then used as input to an RNN. The features for each simulated sweep region were extracted from the sweep site (by default at the center in all simulations) whereas the features for a simulated neutral region were extracted from a variant site (randomly chosen) with a predefined matched DAF. The features for each genomic locus of interest in the CEU population were extracted from all variant sites at that locus having a DAF of > 0.05.

2.5.3 Training a RNN to predict different modes of selection

An RNN was applied to the simulated training data sets to learn a classification or regression model for the task at hand. We used a LSTM, a particular form of RNN, to accommodate the temporal nature of our features, account for long-term dependencies, and tackle the vanishing gradient problem observed in traditional RNNs. Our model had 100 timepoints with the final target output depending on the use of classification or regression. For the classification task, the final target output is a binary class label predicting whether a region is under selection or neutrality. For the regression task, the final target output is a continuous value, representing the selection coefficient or the time of selection onset. We also took a many-to-many approach to model the AF trajectory for the site under selection. The `Keras` software was used to train and test the model. We used a two-stacked LSTM to account for greater model complexity where the number of units in each stack was set to 100 and the hyperbolic tangent (`tanh`) was used as an activation function. The `Adam` optimization method with its default operating parameters was used to update the network weights. For the classification task, the `Softmax` activation function was applied on the final dense layer and the `binary_crossentropy` was used to compute the cross-entropy loss between true labels and predicted

labels. For the regression task, the `linear` activation function was applied on the final dense layer and the `mean_squared_error` loss was used. The SIA deep-learning model took on average 7–10 h to train on a single GPU node with 32 GB memory and four threads, whereas applying the trained model for prediction took less than a minute.

2.5.4 Estimation of confidence intervals

To turn our single-valued regression model into one capable of returning a distribution of predictions of s, we reused the dropout technique that is typically used during training. Dropout enables a fraction of nodes to be randomly "turned off" in a certain layer, which assists in the regularization of the model and helps prevent overfitting. We applied dropout during inference, enabling us to sample a "thinned" network to generate a sample prediction. By repeatedly sampling thinned networks, we generated a distribution of predictions and then computed confidence intervals based on this distribution (Gal and Ghahramani, 2016).

2.5.5 ARG inference

Relate (Speidel et al., 2019) (`v1.0.17`) was used for inferring ARGs underlying simulated genomic samples as well as the CEU population in the 1000 Genomes data set. For simulations under the Tennessen et al., 2012 demography, Relate was run with the true simulation parameters (μ, ρ, and N_e) specified; whereas for genomic loci of the CEU population, Relate was run with a mutation rate of 2.5×10^{-8}/base/generation (`-m 2.5e-8`), a constant recombination map of 1.25×10^{-8}/base/generation and a diploid effective population size of 188,088 (`-N 376176`). The choice of mutation rate follows Stern et al., 2019 based on estimates from Nachman and Crowell, 2000. Although some more recent estimates have been lower (Scally and Durbin, 2012), these differences in mutation rate are unlikely to have a major effect on our selection inference because SIA appears to be fairly robust to mis-specification of mutation rate (Figs. S13 and S16 in Appendix A). For simulations and genomic loci of the *S. hypoxantha* population, Relate was run with $\mu = \rho = 1 \times 10^{-9}$/base/generation and a diploid N_e of 130,000. The branch lengths of Relate-inferred genealogies were estimated iteratively with the `EstimatePopulationSize.sh` script in the Relate package. Specifically, population size history was inferred from the ARG, the branch lengths are then updated for the estimated population size history and these steps are repeated until convergence. This was done for a default of five iterations (`-num_iter 5`).

2.5.6 Alternative methods for selection inference

To benchmark the performance of SIA for classification of sites under neutrality versus selective sweep, we ran the following methods: Tajima's D (Tajima, 1989), H1 (Garud et al., 2015), iHS (Voight et al., 2006), a summary statistics-based deep-learning model, and a tree-based statistic that is part of the Relate (Speidel et al., 2019) program. Tajima's D, H1, and iHS were calculated with the `scikit-allel` package. Haplotypes of the entire 100 kb simulated genomic segment were used for Tajima's D and H1 calculations. The unstandardized iHS was computed at every site with minor AF $> 5\%$, with respect to all other sites in the genomic segment (`min_maf = 0.05, include_edges = True`). iHS scores of all sites were then standardized in 50 AF bins. Finally, the iHS score of a genomic region was taken to be the mean of the iHS scores of all of its variant sites. For the summary statistics-based deep-learning model, we made use of the summary statistics used by S/HIC (Schrider and Kern, 2016; Kern and Schrider, 2018) as features for our deep-learning architecture. These included 11 sequence-based summary statistics (see Figure 3 in Schrider and Kern, 2018) which were used as features for our deep-learning model to distinguish among the two classes at hand (selective sweep vs. neutral drift). All statistics were collected along five consecutive 20-kb windows with the objective of identifying possible sweeps induced by a positively selected mutation in the third (middle) window. Some of these summary statistics corresponded to standard measures of diversity, such as ss (the number of segregating sites), π (Nei and Li, 1979), Tajima's D (Tajima, 1989), θ_{W} (Watterson, 1975), θ_{H} (Fay and Wu, 2000), the number of distinct haplotypes (Messer and Petrov, 2013), H1, H12, H2/H1 (Garud et al., 2015), Z_{nS} (Kelly, 1997), and maximum value of ω (Y. Kim and Nielsen, 2004). For each of these statistics, we computed an average value for each of the five 20 kb windows for the simulated population. Finally, each summary statistic was normalized by dividing the value recorded for a given window by the sum of values across all five windows. The Relate tree-based selection test was performed with an add-on module (`DetectSelection.sh`) using the inferred genealogy with calibrated branch lengths at a site of interest, yielding a $\log_{10} P$ value for each site.

We also compared the performance of SIA for selection coefficient inference with that of CLUES (Stern et al., 2019) and a genotype-based CNN framework (Flagel et al., 2019; Torada et al., 2019). Selection coefficient inference from true genealogies was performed with `clues-v0` (last accessed November 28, 2021). Transition probability matrices were built on a range of selection coefficients $[0, 0.05]$ at increments of 0.0001 and present-day allele frequencies $[0.01, 0.99]$ at increments of 0.01. Selection coefficient inference from Relate inferred genealogies was performed with CLUES (last accessed November 28,

2021). Branch lengths of the genealogy at the site of interest were resampled with Relate for 600 MCMC iterations, and CLUES was run with the following arguments: `-tCutoff 10000 -burnin 100 -thin 5`. For the genotype-based CNN model, each simulated genomic segment was preprocessed by first sorting the haplotypes and then converting the segment to a fixed-size genotype matrix. Haplotype sorting was performed by 1) calculating the pairwise Manhattan distances between haplotypes; 2) setting the haplotype with the smallest total distance to all other haplotypes as the first haplotype; and 3) sorting the remaining haplotypes in increasing distance to the first haplotype. To convert the sorted haplotypes to a fixed-size genotype matrix, centered on the middle variant of a simulated region, up to 180 variants on each side were retained. Variants beyond 180 were discarded and if there were fewer than 180, the missing variants were padded with zeros. Ancestral and derived alleles were coded with 0s and 1s, respectively. Consequently, each simulated genomic region was encoded as a (198×360) binary matrix, along with a real-valued vector encoding the genomic positions of the variants in the matrix. The CNN model had a branched architecture – one branch with five 1D convolution layers taking the genotype matrix as input and another branch with a fully connected layer taking the vector of variant positions as input. The output of the two branches was flattened, concatenated and fed into three fully connected layers, followed by a linear output layer to predict selection coefficient (Fig. S23 in Appendix A).

2.5.7 Evaluation metrics

To evaluate the performance of SIA's classification model and alternative methods, we computed an ROC curve for the binary class at hand ("neutral" or "sweep"), to provide a more complete summary of the behavior of different types of errors. We further assessed the performance of SIA and alternative methods in terms of correctly predicting the selection coefficient numerically using mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (r^2), and visually using a box plot that compares the simulated ground truth against the predictions by the method at hand.

2.5.8 Robustness study

We carried out an extensive analysis of the robustness of our approach, considering not only alternative demographic parameters (such as population size), but also alternative parameters for recombination rate, mutation rate, time of selection onset, and selection coefficients. In all cases, we took care to test our prediction methods under parameters well outside the range used in training.

2.5.9 Analysis of CEU population in 1000 Genomes data

We applied SIA to infer selection coefficients and AF trajectories in the 1000 Genomes (Auton et al., 2015) CEU population at 13 genomic loci with known association to phenotypes, some of which were previously identified as likely targets of positive selection (Table 2.1). For each gene of interest, the ARG was inferred with Relate from SNPs within a 2-Mb window centered at the gene. Once the ARG was inferred, only SNPs with valid ancestral allele ("`AA`" `INFO` field in the `vcf` file) were retained for downstream analysis. Following the aforementioned protocol (see ARG feature extraction), features at all variant sites in the 2 Mb window above a DAF threshold of 0.05 were extracted. Lastly, the SIA model was applied to classify neutrality versus selection, and infer selection coefficient and AF trajectory at each site.

2.5.10 Localizing sweeps in southern capuchino seedeaters

We recently applied a combination of ARG inference and ML methods for identifying selective sweeps to study previously identified "islands of differentiation" in southern capuchino seedeaters and distinguish among possible evolutionary scenarios leading to their formation (Hejase, Salman-Minkov, et al., 2020). Taking advantage of its improved power and genomic resolution, we applied SIA to sequence data for the species for which we have the most samples, *S.hypoxantha*. We simulated training (250,000 neutral; 250,000 soft sweeps), validation (1000 neutral; 1000 soft sweeps), and testing (2,500 neutral; 2,500 soft sweeps) data sets for SIA under a demographic model inferred by G-PhoCS (Campagna et al., 2015). Simulations were performed using discoal with the following parameters: 1) mutation rate $\mu = 1 \times 10^{-9}$; 2) recombination rate $\rho = 1 \times 10^{-9}$; 3) derived $N_e = 130,000$; 4) root divergence time = 1,850,000 generations ago; 5) root $N_e = 1,450,000$; 6) ancestral divergence time = 44,000 generations ago; 7) ancestral $N_e = 14,380,000$; 8) selection coefficient $s \sim \mathcal{U}(0.001, 0.02)$; 9) initial frequency at which selection starts acting on the allele $f_{\text{init}} \sim \mathcal{U}(0.01, 0.05)$; and 10) segregating frequency of the site under selection $f \sim \mathcal{U}(0.25, 0.99)$. A total of 56 haploid sequences were sampled from each simulation, matching the number of *S. hypoxantha* individuals (28) in the real data. The SIA model for *S. hypoxantha* was built, trained and evaluated in an otherwise similar fashion to that for the CEU population as outlined above.

Using a subset of polymorphism data from Turbek et al., 2021 of 28 *S. hypoxantha* and 2 *S. minuta* individuals, we applied our trained model to localize selective sweeps in *S. hypoxantha* on 19 scaffolds that contain top F_{ST} peaks in at least one pairwise species comparison (Campagna et al., 2017)

and/or harbor known pigmentation-related genes such as *ASIP* (located on scaffold 252; induces melanocytes to synthesize pheomelanin instead of eumelanin), *KITL* (located on scaffold 412; stimulates melanocyte proliferation), *SLC45A2* (located on scaffold 404; transports substances needed for melanin synthesis), and *CAMK2D* (located on scaffold 1717; cell communication), as well as 316 scaffolds that 1) are longer than 100 kb; 2) contain more than 1,000 variants; and 3) where more than 95% of sites have a consensus ancestral allele, as determined by four identical haplotypes for two individuals from the outgroup species *S. minuta*. The ARG was inferred with Relate for each scaffold independently. Once the ARG was inferred, the SIA model was applied to sites with consensus ancestral allele for classification and selection coefficient inference.

Chapter 3

Selective sweeps on different pigmentation genes mediate convergent evolution of island melanism in two incipient bird species

Content of this chapter was published in PLoS Genetics (2022) under the title "Selective sweeps on different pigmentation genes mediate convergent evolution of island melanism in two incipient bird species" by Leonardo Campagna, Ziyi Mo, Adam Siepel and J. Albert C. Uy. L.C. conceptualized the study, curated data, performed formal analyses, developed methodology and wrote the manuscript. Z.M. characterized selective sweeps in the bird populations using the SIA method, developed methodology and edited the manuscript. A.S. developed methodology and edited the manuscript. J.A.C.U. conceptualized the study, curated data, performed formal analyses, developed methodology and wrote the manuscript.

3.1 Abstract

Insular organisms often evolve predictable phenotypes, like flightlessness, extreme body sizes, or increased melanin deposition. The evolutionary forces and molecular targets mediating these patterns remain mostly unknown. Here we study the Chestnut-bellied Monarch (*Monarcha castaneiventris*) from the Solomon Islands, a complex of closely related subspecies in the early stages of speciation. On the large island of Makira *M. c. megarhynchus* has a chestnut

belly, whereas on the small satellite islands of Ugi, and SA/SC *M. c. ugiensis* is entirely iridescent blue-black (i.e., melanic). Melanism has likely evolved twice, as the Ugi and SA/SC populations were established independently. To investigate the genetic basis of melanism on each island we generated whole genome sequence data from all three populations. Non-synonymous mutations at the *MC1R* pigmentation gene are associated with melanism on SA/SC, while *ASIP*, an antagonistic ligand of *MC1R*, is associated with melanism on Ugi. Both genes show evidence of selective sweeps in traditional summary statistics and statistics derived from the ancestral recombination graph (ARG). Using the ARG in combination with machine learning, we inferred selection strength, timing of onset and allele frequency trajectories. *MC1R* shows evidence of a recent, strong, soft selective sweep. The region including *ASIP* shows more complex signatures; however, we find evidence for sweeps in mutations near *ASIP*, which are comparatively older than those on *MC1R* and have been under relatively strong selection. Overall, our study shows convergent melanism results from selective sweeps at independent molecular targets, evolving in taxa where coloration likely mediates reproductive isolation with the neighboring chestnut-bellied subspecies.

3.2 Introduction

The extent to which evolutionary change can be predicted has been a long-standing matter of debate in evolutionary biology (Gould, 1989; Grant and Grant, 2002; Blount et al., 2018). Instances of convergent evolution support the argument that evolutionary change can be deterministic, yet stochastic historical events can lead to divergent outcomes from recently split taxa. A better understanding of the eco-evolutionary forces and genetic mechanisms behind evolutionary changes will shed light on the conditions under which deterministic or stochastic outcomes can occur. Some examples of convergent evolution occurred deep in the tree of life, like the independent origins of wings in birds, bats and insects (Blount et al., 2018), while other cases represent more recent (and potentially ongoing) phenomena like the repeated radiations of ecomorphs in Caribbean lizards (Mahler et al., 2013), the loss of flight associated to insularity in insects and birds (Roff, 1994; Wright et al., 2016) or the evolution of island melanism (Mundy, 2005). These recent classic examples of phenotypic convergence can be leveraged to study the evolutionary forces and molecular mechanisms behind phenotypic change. Here we focus on island melanism in birds, a phenotype that involves the increased deposition of eumelanin, which leads to entirely black plumage coloration (Theron et al., 2001; Uy et al., 2016; Walsh et al., 2021).

The Chestnut-bellied Monarch (*Monarcha castaneiventris*) from the Solomon Islands represents a complex of closely related subspecies which are in the early stages of speciation and vary in plumage color, song, and body size (Mayr, 1999; Mayr and Diamond, 2001; Uy, Moyle, and Filardi, 2009). One of these subspecies, *M. c. ugiensis*, has entirely iridescent blue-black plumage, and is found on the small satellite islands to the north and southeast of the larger island of Makira (Fig. 3.1A). In contrast, the endemic subspecies on Makira is *M. c. megarhynchus* and has a chestnut belly and iridescent blue-black upper parts. Phylogenetic analyses using reduced-representation genomic data show that *M. c. ugiensis* individuals from the satellite islands of Ugi, and Santa Ana and Santa Catalina (SA/SC) are independently derived from the chestnut-bellied Makira population, suggesting that *M. c. ugiensis* is polyphyletic and melanism has evolved repeatedly and convergently (Cooper and Uy, 2017). A candidate gene study suggested that the molecular basis of increased melanin deposition differs between the Ugi and SA/SC populations (Uy et al., 2016). Melanism on each of the satellite islands is associated with mutations that affect the coding sequence of the *MC1R*/*ASIP* receptor and ligand pair, two molecules that regulate the balance between the production of eumelanin (a pigment conferring black/gray coloration) and pheomelanin (a pigment which leads to brown/yellow coloration). While the melanic individuals from SA/SC carry a derived non-synonymous mutation on the *MC1R* receptor, their counterparts from Ugi possess a non-synonymous mutation on the *ASIP* ligand, and heterozygotes at either mutation display an intermediate coloration phenotype (Fig. 3.1B; Uy et al., 2016). Finally, it is likely that changes in plumage color mediated by these mutations generate prezygotic reproductive isolation between the melanic populations on the satellite islands and the chestnut-bellied population on nearby Makira, as territorial males discriminate individuals by their phenotype, and respond predominantly to simulated territorial intrusions of males with the local plumage (and song) traits (Uy, Moyle, Filardi, and Cheviron, 2009; Uy and Safran, 2013). Convergent melanism, therefore, may result in repeated speciation between the chestnut-bellied population of Makira, and each of the two melanic populations of Ugi and SA/SC.

Here we generate a reference genome for the Chestnut-bellied Monarch and obtain high coverage whole-genome data for a sample of individuals from Makira and its satellite islands. Our study aims to uncover the molecular targets and evolutionary forces that shape convergent evolution of adaptive traits that can contribute to generating prezygotic reproductive isolation. We use these data to quantify differentiation, reconstruct phylogenetic affinities, and infer the demographic history of these populations. We then use a genome-

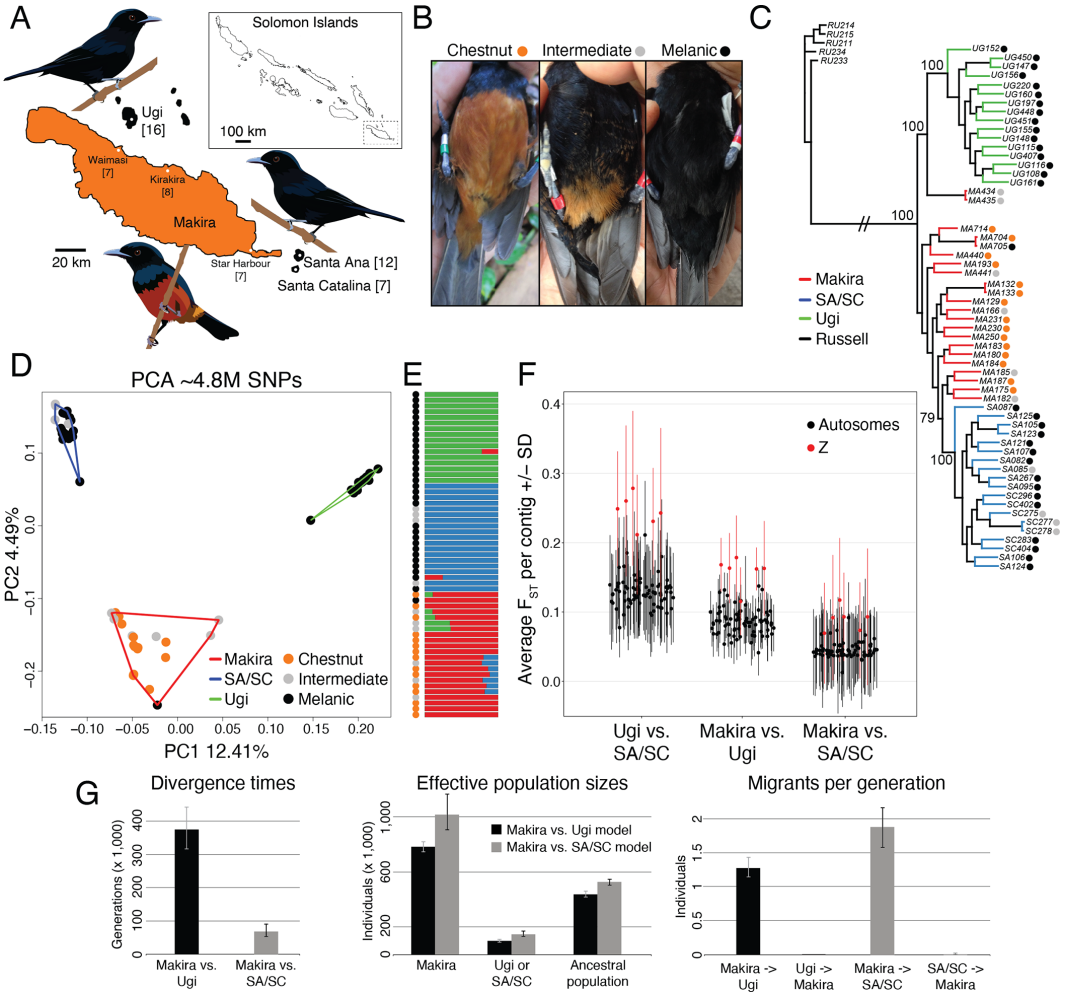


Figure 3.1: **Genetic differentiation and demography of *M. c. megarhynchus* and *M. c. ugiensis*.** **A.** Study area, sample sizes and predominant phenotype on each island. The map was downloaded and modified from www.diva-gis.org. **B.** Representative pictures of chestnut-bellied, intermediate and melanic individuals (color key used throughout Chapter 3). Maximum Likelihood tree (**C**) and PCA (**D**) indicating the origin and coloration phenotype of each individual. **E.** Admixture plot showing the proportion of ancestry for each individual belonging to three different genetic clusters. Each cluster is color-coded by the island from which samples originated and the phenotype is shown by color-coded circles on the left of the plot. **F.** Pairwise F_{ST} estimates summarized by contig. **G.** Demographic reconstructions indicating estimates of divergence times, effective population sizes, and migrants per generation.

wide approach to identify variants associated with melanic plumage. Finally, we infer the evolutionary processes that have shaped these phenotypes on each of the satellite islands of Ugi and SA/SC, estimate when mutations arose and the timing of these selective events.

3.3 Results

3.3.1 Melanic populations are independently derived from a chestnut-bellied ancestor

Birds grouped together by island, irrespective of their coloration phenotype (Fig. 3.1C&D). Individuals from the satellite islands of SA/SC (which are primarily melanic, yet show a low prevalence of the intermediate coloration phenotype) and Ugi (which are exclusively melanic) formed island-specific clades which were embedded among clades containing primarily chestnut-bellied individuals from the larger island of Makira (Fig. 3.1C). The relationships among birds from the three islands could not be resolved using mtDNA, as individuals from every locality share haplotypes (Fig. S1 in Appendix B). Consequently, melanism on the two satellite islands likely originated twice, independently from a chestnut-bellied ancestor (Uy et al., 2016; Cooper and Uy, 2017). We did not observe clear evidence of early generation inter-island hybrids in the genome-wide PCA (Fig. 3.1D), however the two individuals from Makira which form a clade with the individuals from Ugi (Fig. 3.1C) showed intermediate coloration and were sampled in the locality which is closest to Ugi (Waimasi), suggesting the possibility of either incomplete lineage sorting or gene flow. Furthermore, we observed Makira ancestry in one individual of each of the satellite islands, and SA/SC or Ugi ancestry in a few individuals on Makira (Fig. 3.1E). The admixed individuals on Makira were from the localities closest to the satellite island with which they shared ancestry (Waimasi for Ugi and Star Harbour for SA/SC). The levels of differentiation among populations were largest between Ugi and SA/SC, intermediate between Ugi and Makira, and smallest between SA/SC and Makira (Fig. 3.1F). The contigs showing the highest differentiation for each pairwise population comparison were in all cases Z-linked. The difference in the magnitude of genetic differentiation between populations could be due to variation in a combination of demographic parameters (i.e, the splitting time, the degree of gene flow experienced after this split, or the intensity of genetic drift due to differences in effective population sizes). We therefore used sequence data to conduct a demographic reconstruction with G-PhoCS, which suggested that the main reason for the observed difference in the levels of differentiation between pop-

ulations was that Ugi split from Makira approximately six times earlier than SA/SC branched from Makira (Fig. 3.1G). Additionally, the effect of genetic drift is likely to be slightly stronger in Ugi, as its effective population size was inferred to be nearly two thirds of that of SA/SC (and approximately one eighth of Makira's). Finally, G-PhoCS inferred significant levels of gene flow from Makira into each of the satellite islands (higher into SA/SC) and not in the reverse direction Fig. 3.1G, suggesting that the admixture observed on Makira (Fig. 3.1E) may be due to the retention of ancestral polymorphisms in this larger population.

3.3.2 Melanism on each satellite island associates with mutations in different genes

To test if the convergent melanic phenotype on each of the satellite islands was also convergent at the molecular level, we conducted two genome-wide association studies (GWAS) while controlling for population structure by including an inter-individual relatedness matrix as a covariate. The first included individuals from Makira and SA/SC and revealed a single peak on contig 400 (corresponding to chromosome 11) composed of 61 SNPs with association values above the significance threshold (Fig. 3.2A). This region contained 15 annotated genes, including the coloration gene *MC1R* (Fig. 3.2B and S1 Table in Appendix B). The second GWAS, derived from individuals from Ugi and Makira showed seven association peaks with 83 annotated genes (Fig. 3.2C and S1 Table in Appendix B), suggesting a larger number of genes could mediate melanism on Ugi. One of these association peaks, on contig 947 (located on chromosome 20), contained four of the six strongest hits in the GWAS. The *MC1R* antagonist *ASIP* was one of the 14 genes in this region (Fig. 3.2D). The variants within the seven association peaks were in high linkage disequilibrium (LD) (average intrachromosomal $R^2 = 0.84$; average interchromosomal $R^2 = 0.79$; Fig. S2 in Appendix B). We did not find other known coloration genes within the remaining association peaks (S1 Table in Appendix B), suggesting these genes have unknown functions in melanism or mediate other differences between the Ugi and Makira populations which may covary with changes in coloration (e.g., Ugi individuals are larger than those from Makira). Three of the seven association peaks were on the Z sex chromosome, which is consistent with this chromosome evolving faster than the autosomes in birds (Irwin, 2018). Additionally, when comparing within the region encompassed by the association peaks containing *MC1R* and *ASIP* and outside of this region (for contig 400 and 947 separately), we observed higher levels of differentiation between Makira and each of the satellite islands (Fig.

S3A&B in Appendix B).

The melanic individuals from SA/SC had two haplotypes in the region which contained the 61 association hits on contig 400, which were different from the most prevalent haplotype on Makira and Ugi (Figs. S4 and S5 in Appendix B). Three variants fell within the coding region of *MC1R*; two of these positions involved synonymous changes and one coded for an *Asp119Asn* substitution. Similarly, all the individuals from Ugi possessed two haplotypes that were different from the main one present in SA/SC and Makira individuals in the association region around *ASIP* (38 SNPs; Figs. S5 and S6 in Appendix B). A single position fell within the coding region of *ASIP* and involved a non-synonymous *Ile55Thr* substitution, with all Ugi individuals carrying the *Thr55* allele. In conclusion, melanic individuals always carried two copies of the coding *MC1R* mutation (*Asn119*) observed on SA/SC or of the coding *ASIP* mutation (*Thr55*) observed on Ugi.

3.3.3 The regions of the genome containing *MC1R* and *ASIP* show signatures of selective sweeps

We first searched for signatures of selection by calculating summary statistics from the focal contigs containing coloration genes. The region which includes the *MC1R* gene produced negative values of Tajima's D and low nucleotide diversity in the SA/SC population (Fig. 3.3A), as expected for a selective sweep, and high H12 and intermediate H2/H1 values, which are consistent with a relatively soft selective sweep (Fig. 3.3B). However, because of the windowed nature of this analysis we are cautious in interpreting the specific type of sweep that affected the *MC1R* gene. We observed windows within the peak on contig 947 for the Ugi population that showed an overall similar pattern to the one seen for *MC1R* on SA/SC (Fig. 3.3C&D). The positive value of Tajima's D for the window containing *ASIP* on contig 947 (1.4) may be consistent with balancing selection, yet represents an average for a 5 kb window which only included a single SNP (out of 12) from the gene region. In fact, when we calculate Tajima's D for 500 bp windows, the one which includes *ASIP* has a value close to zero (0.25; calculated from 3 SNPs in that window). We opted to present our results for 5 kb windows as these contain an average of 25 SNPs per window (vs. an average of 3 SNPs for 500 bp windows) and therefore represent more robust values of the summary statistics. Finally, we note that genome-wide values of Tajima's D tend to be close to zero for the three populations (-0.6 for Makira and 0.1 for both SA/SC and Ugi), which suggests this statistic hasn't been strongly impacted by demographic trends.

We next searched for signatures of selection on the focal contigs by calcu-

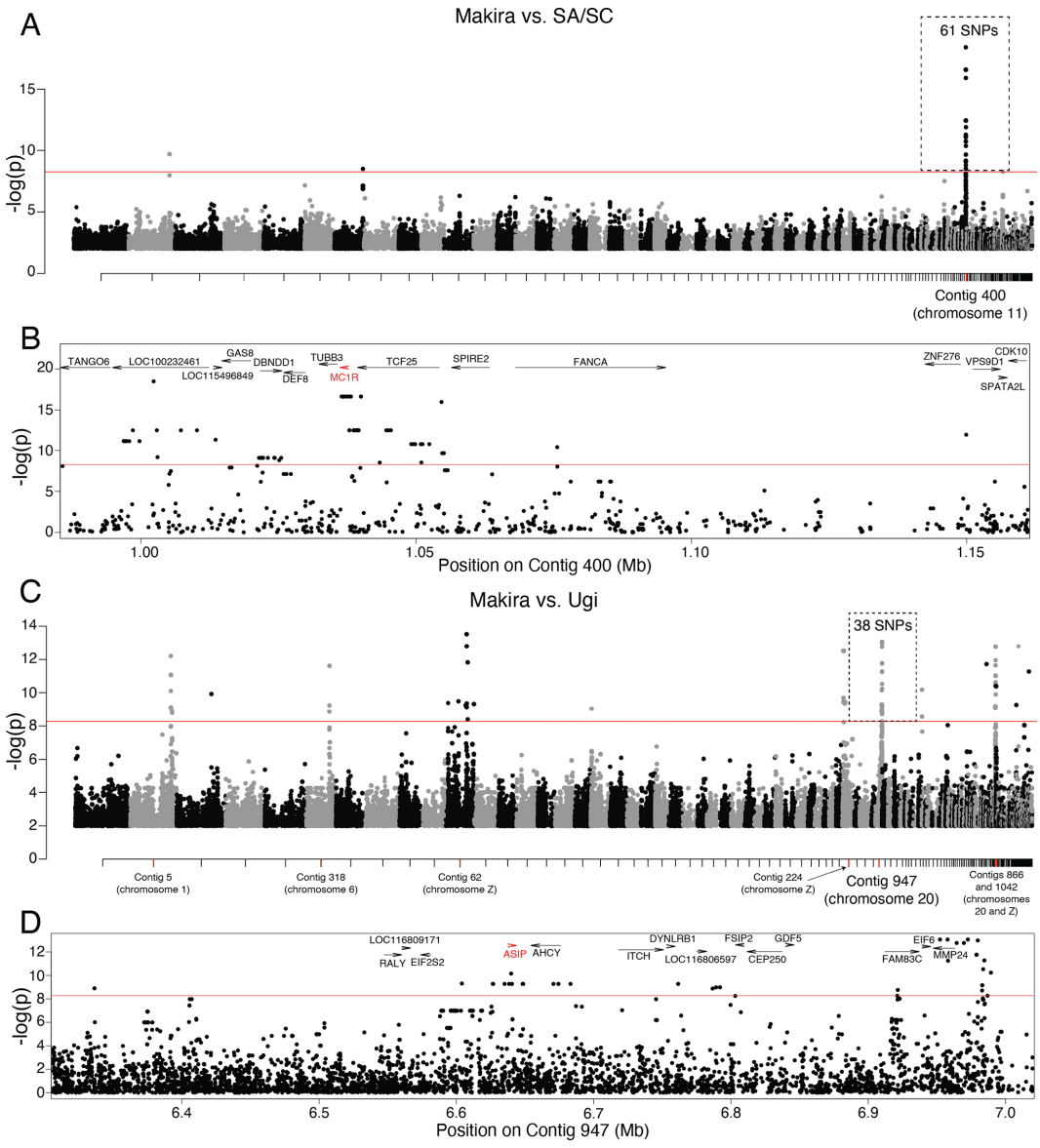


Figure 3.2: **Genome wide association study comparing individuals from subspecies of *Monarcha castaneiventris*. A.** Manhattan plot obtained from the GWAS comparing individuals from Makira and SA/SC. **B.** Zoom-in to the association peak in **A** indicating gene annotations within this region with *MC1R* in red. Equivalent plots for the GWAS obtained with individuals from Makira and Ugi (**C**, **D**).

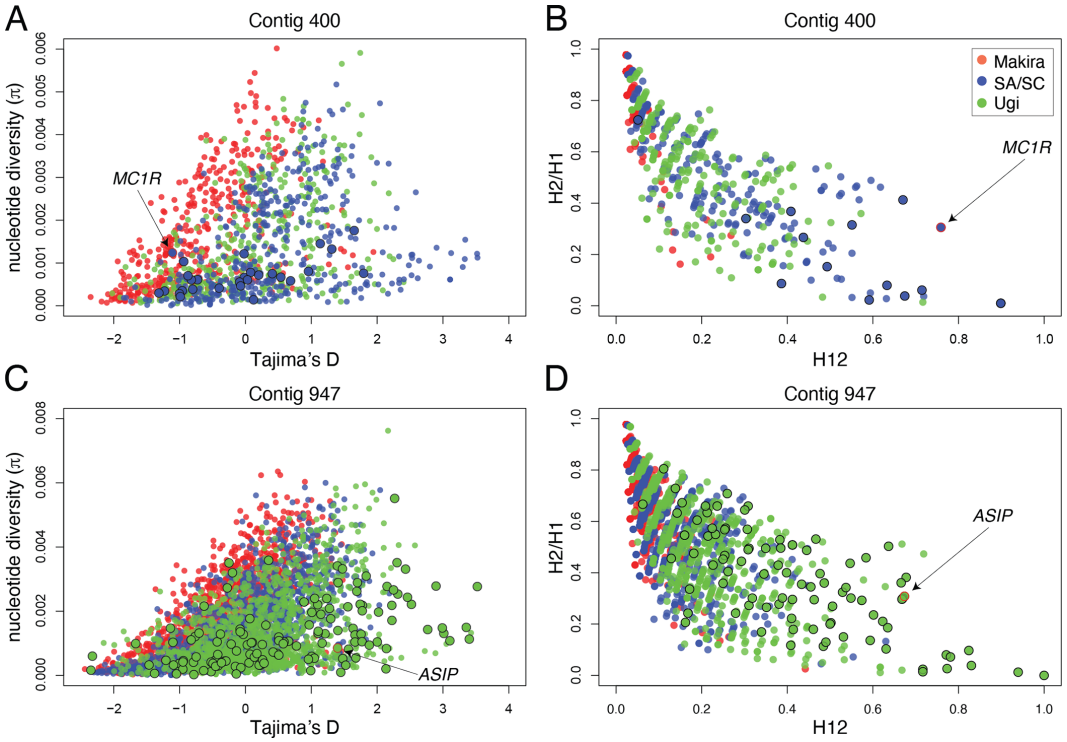


Figure 3.3: **Evidence of a selective sweep in *MC1R* for the SA/SC population and on *ASIP* in the Ugi population.** Biplots of Tajima's D vs. nucleotide diversity for contig 400 and contig 947 (**A**, **C**). Biplots of H12 vs. H2/H1 for the same contigs as above (**B**, **D**). Dots represent statistics derived from 5 kb windows (Tajima's D vs. nucleotide diversity) or 100-SNP windows (H12 vs. H2/H1), and are color coded based on the population of origin. Larger dots denote windows that belong to the outlier peak region, and those that have a red outline include the focal gene indicated by the arrow.

lating two statistics derived from the ancestral recombination graph (ARG): a species (or population) enrichment score and a measure of normalized time to most recent common ancestry (TMRCA) called the relative TMRCA half-life (RTH'; Rasmussen et al., 2014; Hejase, Salman-Minkov, et al., 2020, Fig. S7A in Appendix B). Species enrichment scores measure the probability of observing subtrees of different sizes containing individuals from a certain species. RTH' is the TMRCA of half of the haploid samples of a species divided by the age of the youngest subtree containing half of all the haploid samples, and measures the age of coalescence events independently of the overall coalescence rate. We reasoned that areas of the genome that have undergone a selective sweep in a given population should show shallow genealogies (low RTH' val-

ues) comprising most individuals of that population (high species enrichment score) (Hejase, Salman-Minkov, et al., 2020). We averaged these statistics across 20 kb windows and for each statistic we established population-specific thresholds based on the distribution of values obtained from control windows. In the *MC1R* region of contig 400, we observed a statistically significant elevation of the SA/SC enrichment score, which coincided with a dip in RTH' ($p < 0.005$ in both cases; Figs. 3.4A and 3.4B). In the *ASIP* region of contig 947, we observed a similar pattern for the Ugi population (enrichment: $p < 0.005$, RTH' $p < 0.01$; Figs. 3.4C and 3.4D). These statistical outliers were generated from trees with large and shallow population-specific clades (Fig. S8 in Appendix B). The statistics for the remaining populations in each of the focal contigs did not surpass the thresholds of statistical significance and resembled the values observed for the control contigs (see an example in Fig. S9 in Appendix B). Finally, we also observed clades with extreme enrichment scores on trees obtained from each of the gene regions themselves (Fig. 3.4E; *MC1R* enriched for SA/SC and *ASIP* enriched for Ugi individuals).

We next used SIA (Hejase et al., 2022), a supervised deep-learning method, to infer the strength and time of onset of selection on individual variants within the candidate regions associated with melanism on SA/SC and Ugi. Our models performed well on data simulated under the demographic parameters inferred by G-PhoCS (using msprime and SLiM), distinguishing neutral sites from those under selection, and were able to distinguish soft from hard sweeps in most cases (Figs. S10A and S10B in Appendix B). For this task, we assigned the class with the highest probability as the predicted class, which according to the benchmark with simulated data, resulted in a false positive rate (FPR) of 6–8% when distinguishing neutral regions from those under selection (Fig. S10A in Appendix B). We note, however, that a more stringent probability cutoff could be applied to specifically reduce the FPR for exploratory analyses such as whole-genome selection scans. When applied to the real data, SIA found evidence for soft selective sweeps on multiple variants in the peak region of contig 400, including sites associated with melanism in our GWAS (Fig. 3.5A). We observed the strongest selection ($s \approx 0.02$) on the variants within and around *MC1R* (which were found on the same haplotype), with the timing of selection onset inferred to be ∼500 generations before present on mutations that were ∼78K generations old (Fig. 3.5B). Although our models tended to overestimate selection coefficients when the true s was small, the overestimated values of s were typically below 0.01 (Fig. S10 in Appendix B), which is not the case for *MC1R*. On contig 947, the sites associated with melanism in our GWAS did not vary (i.e., are fixed) in the Ugi population (Fig. S6 in Appendix B), which may hinder our ability to detect signatures of selection

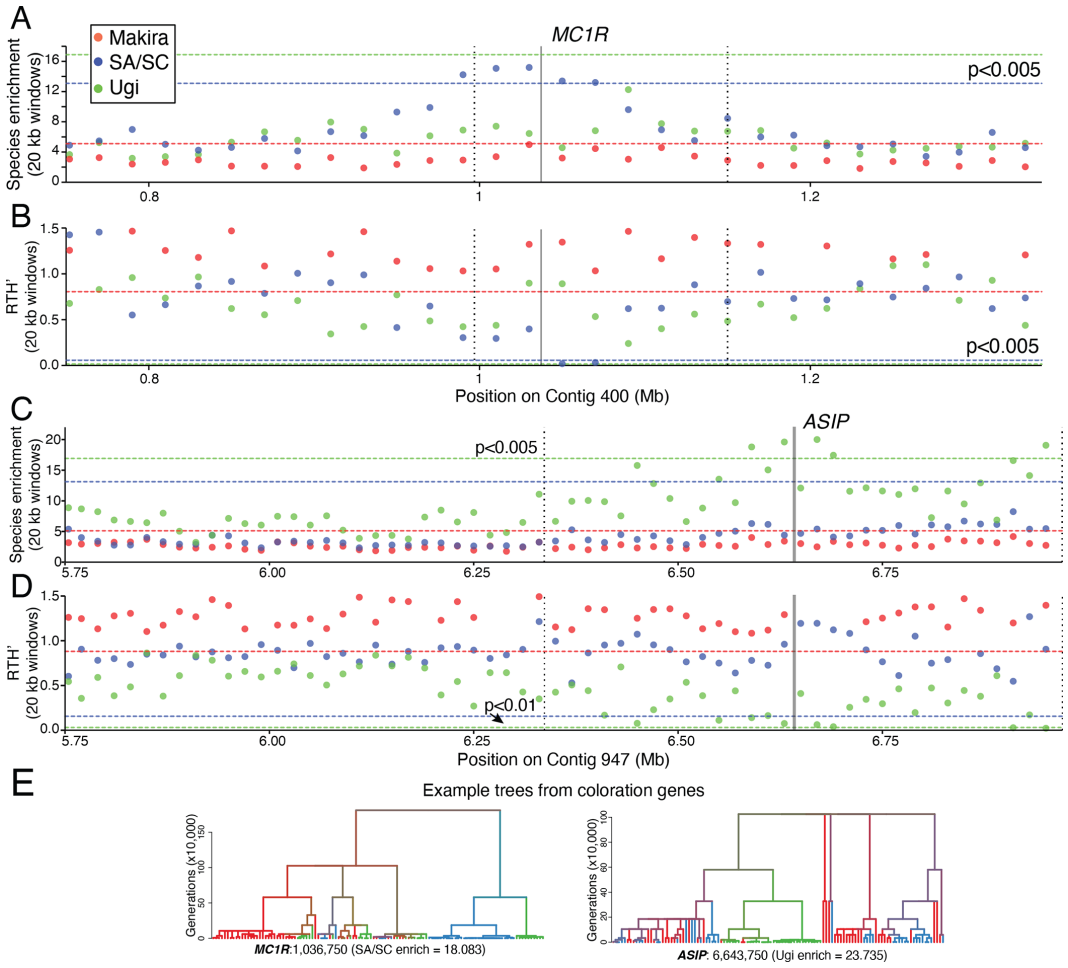


Figure 3.4: **Signatures of selective sweeps in ARG-based statistics on the focal contigs with coloration genes.** Plots showing species enrichment (**A**, **C**) and RTH' (**B**, **D**) in 20 kb windows along contig 400 and 947. Horizontal lines show species-specific levels of statistical significance, dashed vertical lines define the regions of the association peaks, and solid vertical lines show the position of coloration genes (*MC1R* and *ASIP*). **E**. Outlier values of species-enrichment within the *MC1R* and *ASIP* genes (the position on the contig from which each topology is derived is shown on the bottom of each tree). The terminal branches are color-coded by island and the color of internal branches represents an average over all offspring branches.

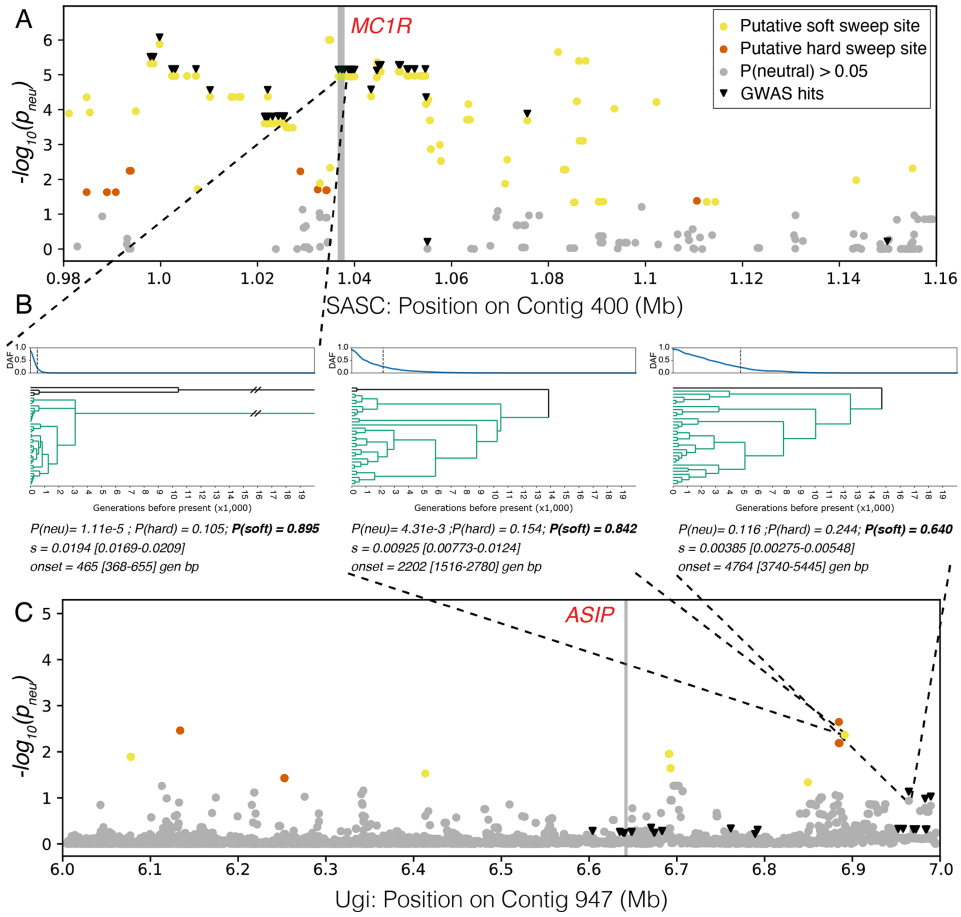


Figure 3.5: **Estimation of positive selection at *MC1R* in the SA/SC population and at *ASIP* in the Ugi population using SIA. A**. Negative log probability of neutrality ($-\log(p_{\text{neu}})$) in the SA/SC population at candidate sites near the *MC1R* gene. GWAS hits are highlighted with inverted triangles. Significantly non-neutral sites ($p_{\text{neu}} < 0.05$) are colored by the predicted class with the highest probability. **B**. Inferred derived allele frequency (DAF) trajectories, local genealogies and selection parameters at three loci of interest. 95% confidence intervals of inferred selection parameters are shown in square brackets. These were estimated using random dropouts at inference time (see Materials and methods). The vertical dashed line on the DAF plot indicates selection onset inferred by SIA. Derived lineages are colored in aquamarine. Highlighted here are a GWAS locus in the *MC1R* coding sequence, the locus inferred to be under the strongest selection near the *ASIP* gene and a GWAS locus near the *ASIP* gene. The location of these loci is projected onto panels (**A**) and (**C**) by dashed gray lines. **C**. Negative log probability of neutrality in the Ugi population at candidate sites near the *ASIP* gene. Details of the figure are otherwise similar to panel (**A**).

when applying SIA to this population. Despite observing several signals of selection in the association peak in other analyses (e.g., high species enrichment or low RTH'), SIA inferred these sites as neutral (Fig. 3.5C). Although among the sites identified by the GWAS we observed some towards the end of the contig which showed the highest probability of having undergone soft sweeps, they also had a $p_{\text{neu}} > 0.05$ and were therefore conservatively classified as neutral. We did however observe non-neutral sites close to *ASIP*. The site with the highest assignment probability to a given class has undergone a soft sweep ($P(\text{soft}) = 0.842$), is 12–14 k generations old, and was under selection ($s \approx 0.01$) for ∼2,200 generations. Overall, *MC1R* was among the strongest and most recent targets of selection in the genome of individuals from the SA/SC population (Fig. S11 in Appendix B).

3.4 Discussion

Our findings show how melanism originated twice in the polyphyletic *M. c. ugiensis* from a chestnut-bellied ancestor: once on Ugi and a second time on SA/SC (Uy et al., 2016; Cooper and Uy, 2017). Moreover, the molecular basis of this convergent phenotype is likely to be different on each island. Our study is novel in identifying how selection has shaped the phenotype on either island, and by being able to time these events.

Black plumage on SA/SC likely originated under strong and recent selection (in the order of 1,000 years before present, assuming a generation time of 2 years) on a series of standing mutations (i.e., a soft sweep) in and around the *MC1R* gene. Selection on *MC1R* is comparable in strength to what was found for the same gene in pocket mice (Hoekstra et al., 2004) or on the *LCT* gene (associated with the lactase persistence trait) in European human populations (Bersaglieri et al., 2004; Hejase et al., 2022). Our ARG-based analysis dates the origin of these mutations to ∼78 K generations before present (Fig. 3.5B), which is older than the inferred split between SA/SC and Makira (∼68 K generations), suggesting they could have originated in the larger Makira population and existed at low frequencies in SA/SC until the mating preference for melanic males reached a frequency threshold that triggered the recent selective sweep (e.g., Kokko et al., 2002). Consistent with this scenario, these derived mutations are present at low frequencies on Makira and Ugi (Fig. S4 in Appendix B). We found among the strongest signatures of selection in the genome on the coding *Asp119Asn* mutation in this gene, a substitution that has been observed independently in other taxa and is known to constitutively activate *MC1R* (Lu et al., 1998), leading to melanism in some domestic animals (Kijas et al., 1998; Våge et al., 1999; Mundy, 2005). Despite these

lines of evidence, we can't rule out that other nearby mutations (perhaps cis-regulatory) also contribute to shaping the coloration phenotype in the SA/SC population.

Various lines of evidence suggest selection in the genomic region containing *ASIP* in the melanic Ugi population, including high species enrichment, higher levels of differentiation (F_{ST}), low RTH', or low nucleotide diversity. However, these statistics were calculated as windowed averages, making it hard to precisely determine the variants under selection. The coding *Ile55Thr* substitution on *ASIP* (and the other sites identified by the GWAS) was fixed on Ugi and only found in five additional individuals from the other islands (mostly in heterozygosity; Fig. S6 in Appendix B). This association, together with the fact that mutations on the N-terminal portion of *ASIP* (where this substitution occurs) can disrupt binding and lead to melanism in other taxa (Hiragaki et al., 2008; Kingsley et al., 2009), suggest a causal role. However, SIA did not infer this mutation to be under selection (or any of those identified by our GWAS analysis) and instead found other positions close to *ASIP* to be under selection. It is possible that these variants were not identified by our GWAS because of their patterns of segregation on Makira and, conversely, that SIA did not find the GWAS hits to be under selection because of their lack of variation in the Ugi population or because those events may have been too old. The estimated age of the *Ile55Thr* substitution is in the order of 174 K generations before present, and SIA was not trained to detect selection older than 20 K generations, and rarely identified selection older than 10 K generations (Fig. S10 in Appendix B). There are a few alternative interpretations of these complex signatures of selection on contig 947. It is possible that the coding position on *ASIP* is an example of an old and completed sweep, and that SIA has detected selection on additional, perhaps cis-regulatory, and more recent mutations (dated to 12-14K generations before present with an estimated selection onset in the order of 5–10 times older compared to what was estimated for *MC1R*). These cis-regulatory mutations could modify the direct (i.e., plumage color) or potential pleiotropic (e.g., stress response, food intake) effects of the *Ile55Thr* substitution on *ASIP* (Ducrest et al., 2008). Alternatively, the GWAS and SIA may have identified sites towards the end of contig 947 (between positions 6.9 and 7) that are independent of selection on *ASIP* and could contribute or be unrelated to differences in coloration. Similarly, additional mutations in other association peaks may contribute to melanism on Ugi, although we did not identify coloration genes in those genomic regions, suggesting they mediate additional phenotypes in which the Ugi and Makira populations differ. Finally, recombination in the flanking regions of hard sweeps can lead to the erroneous identification of soft sweeps

(or "soft shoulders"; Schrider et al., 2015). SIA could have identified the soft shoulder of a hard sweep on *ASIP* (Fig. 3.5C), yet we consider this scenario to be unlikely since this erroneous classification was uncommon when using a similar approach on a different study system (Hejase, Salman-Minkov, et al., 2020). Overall, results from multiple population genetic approaches suggest selective sweeps occurred in the genomic region containing *ASIP* in the Ugi population.

The probability of gene reuse in parallel phenotypic evolution has been estimated to be particularly high when populations are young and closely related (Conte et al., 2012), as is the case for SA/SC and Ugi. It is therefore surprising that the older *ASIP* mutations do not also mediate melanism on SA/SC, especially since we observed one SA/SC and a few Makira individuals carrying Ugi haplotypes from the *ASIP* region (Fig. S6 in Appendix B). One possible explanation is that gene flow between satellite islands is sufficiently low that the *MC1R* mutation swept before the *ASIP* mutations reached SA/SC.

Our findings highlight how independent selective sweeps on a receptor/ligand pair can lead to melanism on two island populations. In *M. c. ugiensis* this trait has been repeatedly favored by selection, and it remains to be determined if the same is true for other instances of island melanism (Theron et al., 2001; Uy and Vargas-Castro, 2015; Walsh et al., 2021). There are several hypothesized benefits of darker plumage coloration, including abrasion resistance, protection from UV radiation, thermoregulation, crypsis, and parasite resistance (e.g., Jacquin et al., 2011; Marcondes et al., 2021). Furthermore, avian coloration is known to mediate reproductive isolation (Price, 2007), especially in the early stages of speciation, and numerous incipient species have been found to differ primarily in melanin-based coloration traits (Poelstra et al., 2014; Bourgeois et al., 2017; Uy et al., 2018; Semenov et al., 2021; Turbek et al., 2021). Field experiments have shown that species recognition is mediated by plumage color in melanic and chestnut-bellied birds from Santa Ana and Makira, respectively (Uy, Moyle, Filardi, and Cheviron, 2009; Uy and Safran, 2013). Therefore, the strong selective pressures we observe in and near pigmentation genes may be the combined product of the advantages of melanic plumage and sexual selection driven by female choice. Overall, our study shows how independent mutations on individual coloration genes can lead to the convergent evolution of a phenotype that is favored on small islands, which, in turn, could promote reproductive isolation and the repeated evolution of incipient species.

3.5 Materials and methods

3.5.1 Ethics statement

All birds were caught with mist nets then measured, tagged, blood-sampled and released as part of a long-term study. Permission to collect samples and work in the Solomon Islands was granted by the Ministry of Environment, Climate Change, Disaster Management & Meteorology (BR/2014/002). Research was approved by the University of Miami Institutional Animal Care and Use Committee (IACUC) protocols number 11–116, 14–097 and 17–071.

3.5.2 Sampling and dataset

A total of 57 individuals from the island of Makira and the neighboring islands of Ugi and Santa Ana/Santa Catalina (hereafter SA/SC) were included in this study from samples collected between 2006 and 2018 (S2 Table in Appendix B). Twenty-two Chestnut-bellied Monarch (*Monarcha castaneiventris megarhynchus*) birds were sampled from three sites on Makira: 7 birds from Waimasi directly across Ugi, 8 birds from Kirakira along the northern coast of Makira, and 7 birds from Star Harbour across from SA/SC. Thirty-five *M. c. ugiensis* birds were included from two satellite island groups: 16 birds from the island of Ugi, and 19 birds from SA/SC. Ten birds of intermediate plumage color (partial chestnut) were sampled, 6 from Makira and 4 from SA/SC. The remaining birds from SA/SC and all of those from Ugi were melanic, while those from Makira were all chestnut-bellied except for one melanic individual. Finally, a sample from a chestnut-bellied *M. c. obscurior* bird caught in the Russell Islands in 2013 was used to sequence and assemble the reference genome, and five additional individuals caught in 2012 were re-sequenced as outgroups for phylogenetic analysis.

3.5.3 Reference genome assembly and annotation

We assembled and annotated a genome from a male Chestnut-bellied Monarch (*Monarcha castaneiventris obscurior*) sampled in the Russell Islands (Solomon Islands; individual RU430). We obtained both short-read Illumina data and long-read Pacific Biosciences (PacBio) data from the same individual, and all sequencing was conducted by Novogene Co. A fragment library was prepared using the NEBNext DNA Library Prep Kit with an insert size of 350 bp, and paired-end sequenced on an Illumina Novaseq 6000 machine, producing 168 gb of raw data (approximately 140x coverage). PacBio SMRTbell libraries were prepared and sequenced on 5 flow cells of the Sequel platform, generat-

ing 71.4 gb of data (approximately 60x coverage) with 7 million subreads with an average N50 of 15.5 kb. We used Samtools version 1.11 (H. Li et al., 2009) to merge the subreads from the five flow cells, filter out subreads that were shorter than 4.5 kb (retaining 67.6% of the subreads), and to convert the file to fasta format. We performed the genome assembly with MaSuRCA version 3.3.3 (Zimin et al., 2013), an assembler which can incorporate both Illumina and PacBio data by being able to use reads of variable lengths. We produced assembly statistics with Quast version 5.0.2 (Gurevich et al., 2013), obtaining a total assembly length of 1.08 gb distributed in 899 contigs, with an N50 of 20.2 mb and 2.1 Ns per mb. We assessed the completeness of our reference assembly by searching for the Passeriformes set of 10,844 single copy orthologs using BUSCO version 5.1.2 (Simão et al., 2015). Our reference genome contained a complete copy of 95.7% of the orthologs in this gene set, 95.4% were found as single copy genes and 0.3% were duplicated. There was a total of 3.4% of these genes that were missing from our assembly and an additional 0.9% were found fragmented. We estimated the chromosomal location of the 899 contigs in our assembly by aligning them to the chromosome level Zebra Finch genome (bTaeGut2.pat.W.v2 downloaded from NCBI) with the Chromosemble function from the Satsuma version 3.1 pipeline (Grabherr et al., 2010), and assigning contigs to the chromosome with the top hit. This function also provides a version of the reference genome with contigs aligned and oriented into pseudochromosomes, assuming synteny between the Chestnut-bellied Monarch and the Zebra Finch. We conducted downstream analyses with both versions of the reference genome and obtained equivalent results (e.g., the same association peaks in our GWAS), so we decided to present those based on the version of the genome that does not assume synteny with the distantly related Zebra Finch.

We annotated the reference genome by first generating a library of the repetitive sequences with RepeatModeler version 2.01 (Flynn et al., 2020). These simple and complex (e.g., transposible elements) repeats can be subsequently masked to avoid being incorreclty annotated as genes from the organism of interest. We then ran two iterations of the MAKER pipeline version 3.01 (Cantarel et al., 2008) to produce gene models. The first iteration generated gene models by training algorithms with data from Zebra Finch transcript and protein databases (downloaded from the bTaeGut2.pat.W.v2 assembly). The models are subsequently refined during a second iteration of the pipeline that uses the output of the first MAKER run as input. In total the pipeline produced 15,226 gene models (72.3% of the 21,049 genes annotated for the Zebra Finch).

3.5.4 Population level genome sequencing and variant discovery

We sequenced the genomes of 57 individuals, 22 belonging to *M. c. megarhynchus* sampled on the island of Makira and 35 to individuals of *M. c. ugiensis*, 16 of which were sampled on the island of Ugi and 19 sampled on the islands of SA/SC. We extracted DNA from blood samples using the DNEasy blood and tissue kit (Qiagen, CA, USA) and libraries were prepared by Novogene Co with the NEBNext DNA Library Prep Kit, with an inset size of 350 bp. Sequencing was performed on an Illumina Novaseq 6000 machine by Novogene Co, obtaining 5,967 million paired end, 150 bp reads. Based on the number of raw (pre-filtering) reads, we expected the depth of coverage to range across all individuals from between 21.5 and 36.8x (average of 26.2x).

We first assessed the quality of individual libraries using fastqc version 0.11.8 and performed quality filtering and trimming, adapter removal and merged overlapping paired end reads with AdapterRemoval version 2.1.1 (Schubert et al., 2016). Once reads were filtered we proceeded to align them to the reference genome using Bowtie2 version 2.4.3 (Langmead and Salzberg, 2012) using the very sensitive local option, which resulted in an average alignment rate of 99.4%. We manipulated the alignment files using Samtools version 1.11 (H. Li et al., 2009), converting `sam` files into `bam` format and sorted and indexed them. We used Picard Tools version 2.8.2 to mark PCR duplicates, GATK version 3.8.1 (Van der Auwera and O'Connor, 2020) to realign around indels, and finally Picard Tools to fix mate-pairs. We obtained an average depth of coverage of 26.3 +/- 4.5x and an average duplication rate of 21.4 +/- 1.6 by computing alignment statistics using qualimap version 2.2.1 (Garcia-Alcalde et al., 2012).

Our genotyping pipeline started by producing individual genomic variant call files for each sample with the "Haplotypecaller" module from GATK, and we subsequently used the "GenotypeGVCFs" module to summarize variants into a single variant file for the entire dataset. We selected SNPs with the "SelectVariants" module of GATK and retained those that satisfied the following filters: QD < 2, FS > 60.0, MQ < 30.0, ReadPosRankSum $<$ -8.0. Finally, we used VCFtools version 0.1.16 (Danecek et al., 2011) to retain 4,799,460 variant sites present in at least 80% of individuals, with mean depth of coverage between 2 and 50 and a minor allele count of at least 8 (equivalent to a minimum of four homozygote individuals, which represents 25% of the population with the smallest sample size). We used this dataset for downstream analyses unless otherwise stated.

3.5.5 Population structure, genetic differentiation and summary statistics

We assessed population structure and admixture among individuals in our sample by conducting a PCA, constructing an admixture plot and building a Maximum Likelihood tree. We also quantified differentiation by calculating F_{ST} values among the populations from the three sampled islands. The PCA was conducted in R version 4.0.2 (R Core Team, 2021) with the package SNPRelate version 3.3 (Zheng et al., 2012). We assessed structure and admixture using the program Admixture version 1.3.0 (Alexander et al., 2009). For this analysis we first thinned the dataset to avoid including linked SNPs with VCFtools, retaining 101,076 SNPs that were at least 10 kb apart. We manipulated the vcf file in VCFtools and plink version 1.9 (Purcell et al., 2007) to convert it to bed format and ran Admixture with a K of three populations. We also ran Admixture analyses exclusively for the two focal contigs with association peaks, in 100 kb sliding windows. We ran the analysis separately for Makira vs. SA/SC individuals on contig 400 and Makira vs. Ugi on contig 947 (i.e., $K = 2$). We plotted these values by using a smoothing line in ggplot2 (Wickham, 2016). To build a tree we first re-ran the pipeline described in the previous section using identical parameters, but including five outgroup *M. c. obscurior* individuals sampled in the Russel Islands (> 330 km away). This iteration of the pipeline produced 5,811,866 SNPs, 5,094,873 of which (those that had the minor allele in homozygosity in at least one individual) could be used to build a tree using RAxML version 8.2.4 (Stamatakis, 2014). We implemented the "ASC_GTRGAMMA" model in combination with the Lewis correction for ascertainment bias, and carried out 200 bootstrap replicates. We used VCFtools to calculate F_{ST} values for non-overlapping 5 kb and 25 kb windows, and subsequently obtained average values and standard deviations for each contig/population comparison in R. We also calculated Tajima's D and nucleotide diversity (π) in non-overlapping, 5 kb windows, with VCFtools (independently for each population) using a dataset without the minor allele frequency filter (see the section on Demographic reconstruction). Additionally, we calculated the haplotype-based statistics H1, H2, H12 and H2/H1, which are designed to distinguish between soft and hard sweeps, using the package SelectionHapStats (Garud et al., 2015). We obtained these statistics for non-overlapping windows of 100 SNPs, merging haplotypes with only one difference (`-distanceThreshold 1`) and using the dataset without the minor allele frequency filter. Finally, we used the information of the chromosomal location of each contig (based on the results from *Chromosemble*, see above) to plot F_{ST} estimates obtained from autosomes and the Z chromosome separately, as values from the latter chromosome tended to be higher. We only plotted

values for contigs that were at least 150 kb (six non-overlapping windows).

We also built minimum spanning networks in PopART 1.7 (Bandelt et al., 1999; Leigh and Bryant, 2015) from mitochondrial genomes. We first assembled mtDNA genomes from our filtered reads with MITObim 1.9.1 (C. Hahn et al., 2013), using the "quick" option and up to 40 iterations with the full mitochondrial genome from the Hooded Crow as a template (*Corvus cornix cornix*, GenBank number NC_024698.1). We subsequently aligned the 57 individual sequences in Geneious version 10.2.6 (Kearse et al., 2012) and imported the alignments into PopART 1.7. We repeated this process restricting the analysis to the COI gene alone, which is commonly used for species identification (Hebert et al., 2003).

3.5.6 Demographic reconstruction

We conducted demographic reconstructions using G-PhoCS version 1.3 (Gronau et al., 2011) which implements an isolation-with-migration model, obtaining estimates of effective population sizes, splitting times and bi-directional migration rates. Because of the computationally intensive nature of this analysis, we conducted two separate demographic reconstructions, one including individuals from Makira and Ugi and the second with individuals from Makira and SA/SC. We also subsampled our dataset, retaining 7 individuals per island (we did not include individuals with intermediate coloration or the melanic individual from Makira). We re-exported 11,537,213 SNPs without a minor allele frequency filter to avoid biasing our analysis by only using data including alleles segregating at higher frequencies, and used these SNPs to generate sequence files for each individual with the "FastaAlternateReferenceMaker" module in GATK. We subsequently sampled for each individual 1,700, 1 kb sequences at intervals of at least 100 kb from autosomal contigs that were larger than 1 Mb. We ran the multi-threaded version of the program for 2 million iterations, discarding the initial 100,000 as burn-in, and estimated 6 demographic parameters in each of our two models (three effective population sizes, one splitting time, and two migration rates). We checked that the traces from the different parameter estimates were stationary and that the effective sample sizes were large (range: 228–9020) using the coda package in R (Plummer et al., 2006). To convert median and 95% Bayesian credible intervals for each parameter from mutation scale to generations or individuals we used an approximate mutation rate estimate of 10^{-9} per bp per generation (Smeds et al., 2016). We note that the assumption of mutation rate will impact the absolute estimates of population sizes and divergence times produced by the model, however we try to focus our interpretations on relative comparisons which are independent of the assumed mutation rate. The number of migrants

per generation, which is independent of the assumption of mutation rate, was calculated as the mutation scaled per generation migration rate times a fourth of the theta parameter for the receiving population ($m_{a>b} \times \theta_b/4$).

3.5.7 Genome wide association analysis (GWAS) and identification of genes in divergent regions

We conducted a phenotype-genotype association analysis using the Wald test implemented in Gemma version 0.98.4 (Zhou and Stephens, 2014). We generated a phenotypic variable in which chestnut individuals were scored as 1, fully melanic individuals were scored as 2, and intermediate individuals as 1.5. The GWAS tests the association between this phenotypic variable and SNP genotypes by fitting univariate linear mixed models, which account for population structure by calculating and including an inter-individual relatedness matrix among all samples as a covariate. We conducted two analyses, one including individuals from Makira and SA/SC and a second with individuals from Makira and Ugi, as we had previous evidence indicating that each island had a different origin of melanism (Uy et al., 2016). We did not conduct a GWAS comparing SA/SC and Ugi individuals as these two populations are not sister and have pronounced population structure. We corrected for multiple tests by using the total number of comparisons conducted across both GWAS (Makira vs. SA/SC and Makira vs. Ugi), and used this conservative α threshold to assess significance ($\alpha = 0.05/(2 * 4.7)$ M SNPs $\approx 5.3 \times 10^{-9}$). We subsequently visualized our results by log-transforming the p-values, changing their sign, and building Manhattan plots with the R package qqman (Turner, 2018). SNPs showing statistically significant associations tended to cluster together in groups (generally more than 5 SNPs) which we defined as association peaks. In other cases, we also observed single or at most a couple of isolated SNPs beyond the level of statistical association which we did not treat as association peaks. We searched for the genes contained in the association peaks by inspecting these regions in the annotation file using Geneious version 10.2.6 (Kearse et al., 2012) and compiled a list of gene models within each region. We subsequently obtained information on these annotations of interest from the NCBI database. We explored the relationship between genotypes at different loci within each association peak by phasing and imputing missing data using BEAGLE version 3.3.2 (Browning and Browning, 2007). This resulted in two haplotypes per individual for each peak with which we calculated a distance matrix in the R package vegan (Oksanen, 2022) and plotted it with the function `phylo.heatmap()` from the R package phytools (Revell, 2012). We also calculated LD between different sites by computing R2 values in VCFtools.

3.5.8 ARG inference and derivation of ARG-based statistics

We generated statistics derived from ARGs as described in detail in Hejase, Salman-Minkov, et al., 2020 using scripts deposited in GitHub. We first inferred ARGs for two contigs with association peaks (contig 400 and contig 947; total of 9 Mb) and 20 similarly-sized contigs (ranging from 1.2 to 11 Mb; total of ∼65 Mb) that did not contain association peaks. We inferred ARGs using the `arg-sample` module from ARGweaver version 1 (Rasmussen et al., 2014), which estimates a local tree for each position along the contig. We ran the software independently on each of the 22 contigs indicating that the data were unphased and assuming a mutation and a recombination rate of 10^{-9}/bp/gen (Smeds et al., 2016; Hejase, Salman-Minkov, et al., 2020). We set the effective population size to 500,000 individuals and the following options for the remaining parameters required by the software: `-c 5 -ntimes 20 -maxtime 1e7 -delta 0.005 -resample-window-iters 1 -resample-window 10000 -n 1000`. We sampled the last of 1,000 MCMC iterations and used it to extract a local tree at intervals of 500 bp, discarding the edges of each ARG block (initial and final 50 kb) where there is uncertainty in the inferred topologies. We subsequently calculated two statistics from each tree for downstream analyses: species enrichment scores and RTH' (see Rasmussen et al., 2014 for RTH and Hejase, Salman-Minkov, et al., 2020 for a modification in how we normalize TMRCA to obtain RTH' or Fig. S7A in Appendix B).

Species enrichment scores are defined as the probability of observing a subtree with n leaves for which k are mapped to a certain species or population, assuming a hypergeometric distribution. Therefore, if a local tree contains a large clade composed of individuals from the same species this will be reflected in a high enrichment score for that species. Because any given tree contains various subtrees, the score for each species at each site was defined as the highest score obtained from all the possible subtrees. RTH' was calculated by dividing the time to the most recent common ancestor of half the haploid samples for a given species (TMRCAH; $n_{\text{Makira}} = 22$, $n_{\text{SA/SC}} = 19$, $n_{\text{Ugi}} = 16$) by the age of the youngest subtree that contained at least half of all haploid samples ($n = 57$). The benefits of this normalization are that it is sensitive to various types of selective sweeps (e.g., partial sweeps, those shared by multiple species or complete species-specific sweeps) and that it is independent of the variation in coalescent times that is observed along the genome (Hejase, Salman-Minkov, et al., 2020). We obtained 20 kb nonoverlapping window values for each statistic, derived from averaging statistics obtained from 40 individual trees. For each species, this process produced windowed averages for species enrichment and RTH', for the two contigs with association peaks

and the sample of 20 contigs that did not contain association peaks.

We assessed statistical significance by generating empirical distributions for each parameter from the total of 3,682 20 kb windows. We established species and parameter specific significance thresholds by finding the cutoff value that represented the top (species enrichment) or bottom (RTH') 0.01, 0.005 and 0.001 of the distribution. Cutoff values that defined different slices of the distribution were used to establish statistical significance at different P values. Windows that fell beyond or below a threshold (e.g., $P < 0.001$) were considered to come from a region of the genome with clades that are statistically significantly enriched in a given species, or to have statistically significantly shallow clades (RTH'), respectively. Finally, we exported randomly selected individual topologies or trees which illustrated extreme enrichment or RTH' values for particular species, for the regions containing association peaks or for specific genes.

3.5.9 Inference of positive selection

The analyses of species differentiation and cross-species ARG statistics are useful tools to detect signals of positive selection in genomic windows. To further localize the target of selection and infer parameters of positive selection such as the selection coefficient, time of selection onset and allele frequency trajectories, we employed the machine learning method implemented in SIA (Hejase et al., 2022). SIA uses a RNN to leverage features of single-population genealogies. Selection in a population leaves characteristic signals in its genealogy that can be picked up by SIA to make inferences of selection parameters for individual variants that map to gene trees embedded in an ARG (Fig. S7B in Appendix B).

We simulated data for training and benchmarking the SIA model by initializing neutral simulation in msprime (Kelleher et al., 2016) and continuing simulation of positive selection in SLiM (Haller et al., 2019; Haller and Messer, 2019), to maximize computational efficiency. We ran coalescent simulation in msprime up to the generation of selection onset (or in the case of neutral simulations, a randomly sampled generation), saved the progress in tree sequence format, and loaded the tree sequence in SLiM to carry on with forward simulation. We simulated separate datasets for the SA/SC/Makira and Ugi/Makira population pairs, each under a two-population, 5-parameter demographic model inferred by G-PhoCS (see above and Fig. 3.1G), with effective population size scaled down by 10-fold. Scaling down the population sizes reduces the running time of the simulations but requires scaling other parameters accordingly. Because the migration rates from each of the satellite islands into Makira were inferred by G-PhoCS to be negligible, these were

ignored in the simulations. However, we simulated gene flow from Makira into each of the satellite island populations because it was inferred to be much higher (Fig. 3.1G) and can have a non-trivial effect on selection inference. For example, a completed hard sweep followed by subsequent introgression and recombination of the ancestral haplotype could be mis-classified as soft by a model trained without simulations of such a scenario. For sweep simulations, selection coefficients (s) were sampled between 0.001 and 0.025 (scaled up to 0.01–0.25 for simulations) from an equal mixture of a uniform distribution and a log-uniform distribution. We kept only sweep simulations where the current derived allele frequencies at the sweep site was greater than 0.2 and allowed for alleles that are "recently fixed". This sampling scheme corresponded roughly to a range of selection onset from 250 to 20,000 generations before present, a regime in which SIA would be trained to detect positive selection. For soft sweep simulations, the allele frequency threshold (f_{init}) above which selection acts on the allele was sampled uniformly by $f_{\text{init}} \sim \mathcal{U}(0.01, 0.1)$. To simulate a soft sweep, at the generation of selection onset, we picked a random clade of the satellite island population whose size matches the sampled f_{init}). We then added a mutation to the branch leading to the MRCA of this clade before turning on selection at this variant. This particular MRCA could be a native (such that the mutation occurred on the satellite island), or alternatively a migrant from Makira (such that the allele came from standing variation in the Makira population). Nevertheless, since SIA uses features of single-population genealogies of the satellite island population, it is agnostic to the two scenarios. Each dataset consists of 1,500,000 neutral, soft and hard sweep simulations of 100kb regions (equal split among the three categories). For sweep simulations, the sweep site was at the center of the region. The datasets were used to train and benchmark two separate SIA models for the SA/SC and Ugi populations following a train-val-test split of 85%-5%-10%. The ARG inference process, genealogical feature extraction and deep learning architecture used for building the SIA model are described in detail in Hejase et al., 2022. A cartoon illustration of the genealogical features is provided in Fig. S7B in Appendix B. We applied the SIA model to detect signals of positive selection in the SA/SC and Ugi populations using the dataset without the minor allele frequency filter, and for putative sweep sites, we inferred selection coefficients and the time of selection onset. To gauge the uncertainty of the parameter estimates, we applied dropout to the trained SIA model at inference time (Gal and Ghahramani, 2016). We ran the model 1,000 times, each with random sets of dropout nodes, to obtain 1,000 samples of the model prediction from which a 95% confidence interval was derived. We conducted these predictions across the 190 scaffolds that were longer than 100kb and had at least 1,000

called variants (without applying a minor allele frequency filter). Sites with a probability of being neutral greater than 0.05 were considered to be neutral. In addition, for particular sites of interest, we applied the model to infer allele frequency trajectories. Finally, we dated the origin of several mutations by estimating the midpoint age of the branch in which they first appeared.

Chapter 4

Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data

Content of this chapter was previously uploaded to bioRxiv (2023) under the title "Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data" by Ziyi Mo and Adam Siepel. The manuscript was published in PLoS Genetics (2023) under the same title.

4.1 Abstract

Investigators have recently introduced powerful methods for population genetic inference that rely on supervised machine learning from simulated data. Despite their performance advantages, these methods can fail when the simulated training data does not adequately resemble data from the real world. Here, we show that this "simulation mis-specification" problem can be framed as a "domain adaptation" problem, where a model learned from one data distribution is applied to a dataset drawn from a different distribution. By applying an established domain-adaptation technique based on a gradient reversal layer (GRL), originally introduced for image classification, we show that the effects of simulation mis-specification can be substantially mitigated. We focus our analysis on two state-of-the-art deep-learning population genetic methods—SIA, which infers positive selection from features of the ancestral

recombination graph (ARG), and ReLERNN, which infers recombination rates from genotype matrices. In the case of SIA, the domain adaptive framework also compensates for ARG inference error. Using the dadaSIA model, we estimate improved selection coefficients at selected loci in the 1000 Genomes CEU population. We anticipate that domain adaptation will prove to be widely applicable in the growing use of supervised machine learning in population genetics.

4.2 Introduction

Advances in genome sequencing have allowed population genetic analyses to be applied to many thousands of individual genome sequences (Auton et al., 2015; Sudlow et al., 2015; Karczewski et al., 2020). Given adequately rigorous and scalable computational tools for analysis, these rich catalogs of genetic variation provide opportunities for addressing many important questions in areas such as human evolution, plant genetics, and the ecology of non-model organisms. Deep-learning methods, already well-established in other application areas (LeCun et al., 2015), have proven to be good matches for these analytical tasks and have recently been successfully applied to many problems in population genetics (Sheehan and Song, 2016; Kern and Schrider, 2018; Schrider and Kern, 2018; Flagel et al., 2019; Torada et al., 2019; Adrion, Galloway, et al., 2020; Caldas et al., 2022; Hejase et al., 2022; Huang et al., 2023; Korfmann et al., 2023).

The key to the success of deep learning in population genetics has been the use of large amounts of simulated data for training. Under simplifying, yet largely realistic, assumptions, evolution plays by relatively straightforward rules. By exploiting these rules and advances in computing power, a new generation of computational simulators has made it possible to efficiently produce large quantities of perfectly labeled synthetic data across a wide range of evolutionary scenarios (Haller et al., 2019; Haller and Messer, 2019; Baumdicker et al., 2022). At the same time, programming libraries such as stdpopsim have made these simulators accessible to a broad community of researchers while improving the reproducibility of simulation workflows (Adrion, Cole, et al., 2020; Lauterbur et al., 2022). The facility of generating synthetic training data serves as the foundation of the new simulate-and-train paradigm of supervised machine learning for population genetics inference (Fig. 4.1A; Schrider and Kern, 2018; Huang et al., 2023; Korfmann et al., 2023).

At the same time, this paradigm is highly dependent on well-specified models for simulation (Korfmann et al., 2023). If the simulation assumptions do not match the underlying generative process of the real data – that is, in

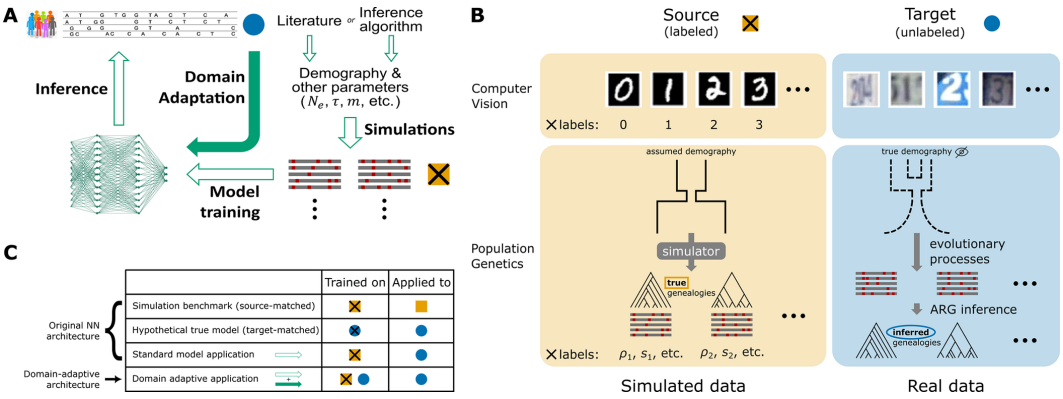


Figure 4.1: **Unsupervised domain adaptation in the context of population genetic inference. A)** A high-level overview of the supervised machine-learning approach for population genetic inference and how domain adaptation fits into the paradigm. **B)** Example formulations of the unsupervised domain adaptation problem with application to computer vision and population genetics. Note that in the specific case of SIA, which uses features of the ARG, the source domain data always consist of true genealogies generated in simulations, whereas the target domain data always consist of inferred genealogies reconstructed from observed sequence data. **C)** Four benchmarking scenarios considered in this study. The original model was both trained and tested on source domain data (simulation benchmark), both trained and tested on target domain data (hypothetical true model), or trained on source domain data but applied to target domain data (standard model application). These three cases contextualize the performance of the domain-adaptive model (see Methods for details). Gold squares represent source domain data, blue circles represent target domain data and crosses ($\times$) represent labels.

the presence of *simulation mis-specification* – the trained deep-learning model may reflect the biases in the simulated data and perform poorly on real data. Indeed, previous studies have shown that, despite being robust to mild to moderate levels of mis-specification, performance inevitably degrades when the mismatch becomes severe (Adrion, Galloway, et al., 2020; Hejase et al., 2022).

In a typical workflow, key simulation parameters such as the mutation rate, recombination rate, and parameters of the demographic model are either estimated from the data or obtained from the literature (Fig. 4.1A; Adrion, Cole, et al., 2020; Lauterbur et al., 2022). Sometimes these parameters are allowed to vary during simulation, and sometimes investigators evaluate the sensitivity of predictions to departures from the assumed range, but there is typically no

way to ensure that the ranges considered are adequately large. Moreover, these benchmarks do not usually account for under-parameterization of the demographic model. Particularly in the case of non-model organisms, the quality of the estimates can be further limited by the availability of data. Overall, some degree of mis-specification in the simulated training data is impossible to avoid.

One way to mitigate the effects of simulation mis-specification would be to engineer a simulator to force the simulated data to be compatible with real data. For example, one could simulate from an overdispersed distribution of parameters followed by a rejection sampling step (based on summary statistics) as in approximate Bayesian computation (ABC) methods, or one could use a generative adversarial network (GAN) (Z. Wang et al., 2021) to mimic the real data. These methods tend to be costly, however. For example, ABC methods scale poorly with the dimensionality of the parameter space, and GANs are notoriously hard to train.

Here we consider the alternative approach of adopting a deep-learning model that is explicitly designed to account for and mitigate the mismatch between simulated and real data (Fig. 4.1A). A standard machine learning model aims to make accurate predictions on data following the same probability distribution as the training instances. In contrast, the task of building well-performing models for a target dataset that has a *different* distribution from the training dataset is termed "domain adaptation" in the machine-learning literature (Csurka, 2017; Wilson and Cook, 2020). A typical setting of interest for domain adaptation is image classification (Fig. 4.1B). For example, suppose a digit-recognition model is needed for the Street View House Numbers (SVHN) dataset (the "target domain"), but abundant labeled training data is only available from the MNIST dataset of handwritten digits (the "source domain"). In this case, a method needs to train on one dataset and perform well on another, despite systematic differences between the two data distributions.

Various strategies for domain adaptation have been introduced. Prior to the advent of deep learning, early methods focused on reweighting training instances according to their likelihoods of being a source or target example (Shimodaira, 2000; Dai et al., 2007) or explicitly manipulating a feature space through augmentation (Daumé III, 2009), alignment (Fernando et al., 2013; Sun et al., 2016) or transformation (Pan et al., 2011). Recently, specialized neural network architectures have been developed for deep domain adaptation. Most model architectures of this kind share the common goal of learning a "domain-invariant" representation of the data through a feature extractor neural network, for example, by minimizing domain divergence (Rozantsev

et al., 2019), by adversarial training (Ganin and Lempitsky, 2014; M.-Y. Liu and Tuzel, 2016) or through an auxiliary reconstruction task (Ghifary et al., 2016). Domain adaptation so far has been most widely applied in the fields of computer vision (e.g., using stock photos for semantic segmentation of real photos) and natural language processing (e.g., using Amazon product reviews for sentiment analysis of movies and TV shows) where large, heterogeneous datasets are common but producing labeled training examples can be labor intensive (Wilson and Cook, 2020). More recently, deep domain adaptation has been used in regulatory genomics to enable cross-species transcription-factor-binding-site prediction (Cochran et al., 2022).

In this work, we reframe the simulation mis-specification problem in population genetics as an unsupervised domain adaptation problem – unsupervised in the sense that data from the target domain is not labeled (Fig. 4.1B). In particular, we use population-genetic simulations to obtain large amounts of perfectly labeled training data in the source domain. We then seek to apply the trained model to unlabeled real data in the target domain. We use domain adaptation techniques to explicitly account for the mismatch between these two domains when training the model.

To demonstrate the feasibility of this approach, we incorporated a domain-adaptive neural network architecture into two published deep learning models for population genetic inference: 1) SIA (Hejase et al., 2022), which identifies selective sweeps based on the ancestral recombination graph (ARG), and 2) ReLERNN (Adrion, Galloway, et al., 2020), which infers recombination rates from raw genotypic data. Through extensive simulation studies, we demonstrated that the domain adaptive versions of the models significantly outperformed the standard versions under realistic scenarios of simulation mis-specification. Our domain-adaptive framework for utilizing mis-specified synthetic data for supervised learning opens the door to many more applications in population genetics.

4.3 Results

4.3.1 Experimental design

We created domain-adaptive versions of the SIA and ReLERNN models, each of which employed a gradient reversal layer (GRL) (Ganin and Lempitsky, 2014) (Fig. 4.2A&B). As noted, the goal of domain adaptation is to establish a "domain-invariant" representation of the data (Fig. 4.1A). Our neural networks consist of two major components: the original networks ("feature extractor" in green and "label predictor" in blue in Fig. 4.2A&B), which are

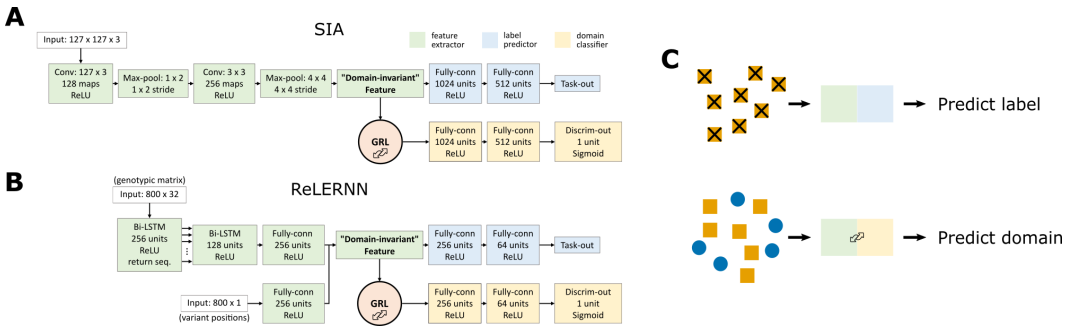


Figure 4.2: **Neural network architecture for domain adaptation.** The model architectures incorporating gradient reversal layers (GRLs) for **A)** SIA and **B)** ReLERNN. The feature extractor of SIA contains 1.49×10^5 trainable parameters, whereas the label predictor and domain classifier contains 1.22×10^8 each. The feature extractor of ReLERNN contains 1.52×10^6 trainable parameters, whereas the label predictor and domain classifier contains 1.49×10^5 each. Note that the total number of trainable parameters includes those in batch normalization layers. **C)** When training the networks, each minibatch of training data consists of two components: (1) labeled data from the source domain fed through the feature extractor and the label predictor; and (2) a mixture of unlabeled data from both the source and target domains fed through the feature extractor and the domain classifier. The first component trains the model to perform its designated task. However, the GRL inverts the loss function for the second component, discouraging the model from differentiating the two domains and leading to the extraction of "domain-invariant" features.

applied only to labeled examples from the "source" (simulated) domain; and alternative branches ("domain classifier" in yellow in Fig. 4.2A&B), which use the same feature-extraction portions of the first networks but have the distinct goal of distinguishing data from the "source" (simulated) and "target" (real) domains (they are applied to both). When the neural network is trained by back-propagation, the GRL reverses the sign of the gradient for the feature extractor with respect to the domain-classifier loss. By doing so, the GRL systematically undermines this secondary goal of distinguishing the two domains (Fig. 4.2, see Methods for details), and therefore promotes domain invariance in feature extraction.

We designed two sets of benchmark experiments to assess the performance of the domain-adaptive models relative to the standard models. In both cases, we tested the methods using "real" data in the target domain that was actually generated by simulation, but included features not considered by the simpler simulator used for the source domain. In the first set of experiments,

background selection was present in the target domain but not the source domain. In the second set of experiments, the demographic model used for the source-domain simulations was estimated from "real" data generated under a more complex demographic model and was therefore somewhat mis-specified (as detailed below). Below we refer to these as the "background selection" and "demography mis-specification" experiments.

4.3.2 Performance of domain-adaptive SIA model

We compared the performance of the **d**omain-**ada**ptive SIA (dadaSIA) model to that of the standard SIA model on held-out "real" data, considering both a classification (distinguishing selective sweeps from neutrality) and a regression (inferring selection coefficients) task. In all cases, we focused on a comparison of the domain-adaptive model to the standard case where a model is simply trained on data from the source domain and then applied to the target domain ("standard model"; Fig. 4.1C). Note that the version of SIA used by both the domain-adaptive and standard models includes a variety of minor improvements that led to modest gains in performance over the previously published version (see *Updates to genealogical features and deep learning architecture for the SIA model* in Methods and Fig. S1B&C in Appendix C). The codebase of the original SIA model has been updated accordingly.

For additional context, we also considered the two cases where the training and testing domains matched ("source-matched" or "target-matched"; Fig. 4.1C)—although we note that these cases are not achievable with real data and provide only hypothetical upper bounds on performance. Notably, in the source-matched (or "simulation benchmark") case, the standard model is both trained and tested with true genealogies from source-domain simulations. By contrast, in the target-matched (or "hypothetical true model") case, the standard model is trained as if target-domain data with ground-truth selection coefficient labels were available. Since genealogies need to be inferred in the target domain (Fig. 4.1B), the hypothetical true model is both trained and tested with inferred genealogies (see *Setup of benchmarking experiments* in Methods for details).

As noted, we considered two types of mis-specification, background selection and demographic mis-specification. In the background selection experiments, the target domain experienced selection in a central "genic" region (following a DFE from Boyko et al., 2008), leading to background selection in flanking regions. This genic region was omitted in the source domain. In the demographic mis-specification experiments, the demographic model for source-domain simulations was inferred from "real" data using G-PhoCS (Gronau et al., 2011). Both the real (target domain) and inferred (source domain) models

assumed three populations with migration, but the inferred model was under-parameterized and its parameters differed substantially from the real model (Fig. S1A in Appendix C) (see Methods for details).

In both the background selection and demography mis-specification experiments, and in both the classification and regression tasks, the domain-adaptive SIA model substantially improved on the standard model (Fig. 4.3). Indeed, in all cases, the domain-adaptive model (turquoise lines in Fig. 4.3A&C) nearly achieved the upper bound of the hypothetical true model (dashed gray lines) and clearly outperformed the standard model (gold lines), suggesting that domain adaptation had largely "rescued" SIA from the effects of simulation mis-specification (see also Fig. S2C&D in Appendix C). The standard model performed particularly poorly on the regression task (Fig. 4.3B&D), but the domain-adaptive model achieved substantial improvements, reducing both the absolute error as well as the upward bias of the estimation (Fig. S2C&D in Appendix C).

The comparisons with the simulation benchmark and hypothetical true model were also informative in other ways. Notice that performance in the simulation benchmark case was considerably better than that in all other cases, including the hypothetical true model. For SIA in particular, the ARG is "known" (fixed in simulation) in the source domain, whereas in the target domain it must be inferred (Fig. 4.1B). Thus, the difference between the simulation benchmark (source-matched) and hypothetical true model (target-matched) cases represents a rough measure of the importance of ARG inference error (see Discussion). In addition, note that in many studies, benchmarking of population-genetic models is performed using the same, or similar, simulations as those used for training, as with our hypothetical true model. Thus, the difference between the hypothetical true model and the standard model is representative of the degree to which benchmarks of this kind may be overly optimistic about performance, depending on the degree to which the simulations are mis-specified.

We further investigated the effect of imbalanced training data from the target domain on the performance of the domain-adaptive model in the context of sweep classification. Despite the ability to simulate perfectly class-balanced labeled data in the source domain, in practice we have no control over whether real data are balanced. Using simulations for the background selection mis-specification experiments, we tested the performance of the domain-adaptive SIA model classifying sweeps when trained with unlabeled "real" data under different proportions of sweep vs. neutral examples. While a balanced dataset yielded the best performance, significantly skewed datasets (20% or 80% sweep examples) still provided the domain-adaptive model with reasonable improve-

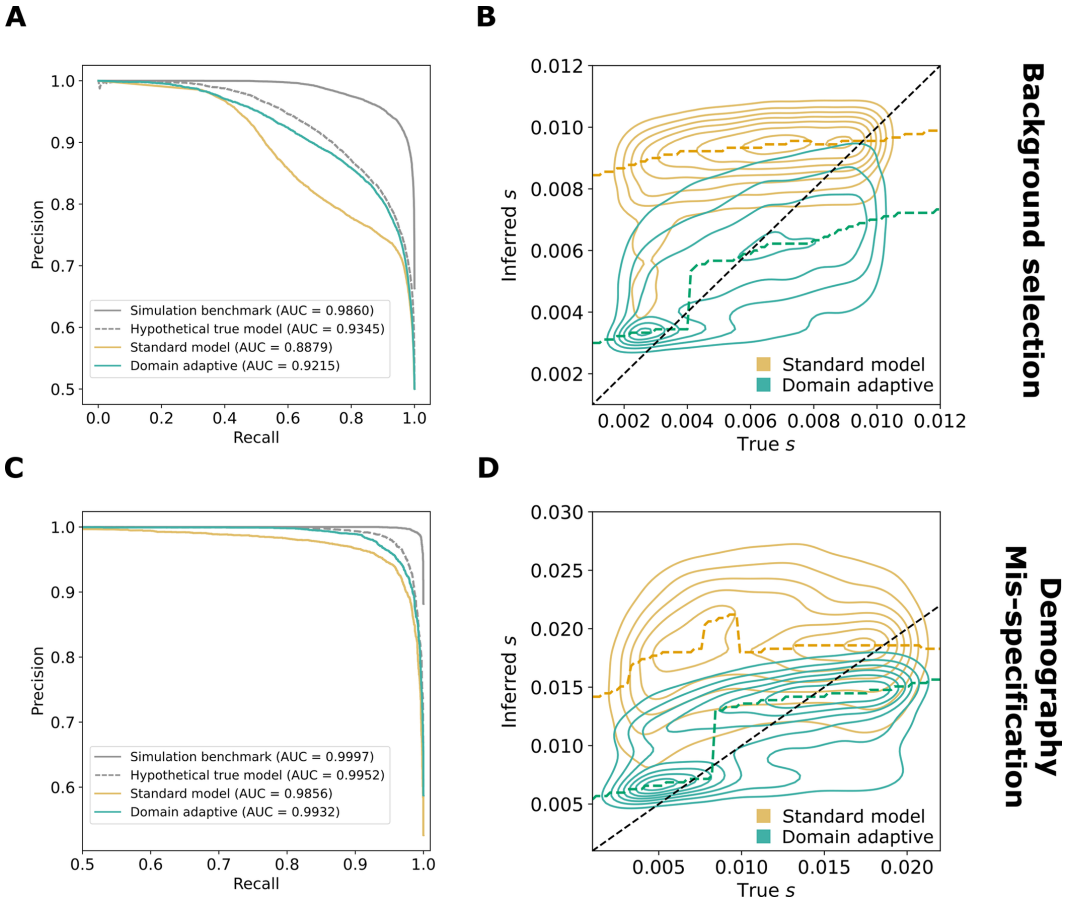


Figure 4.3: **Performance of domain-adaptive SIA models.** Results are shown from (**A**, **B**) the background-selection and (**C**, **D**) the demography-mis-specification experiments. (**A**, **C**) Precision-recall curves for sweep classification. (**B**, **D**) Contour plots summarizing true (horizontal axis) vs. inferred (vertical axis) selection coefficients (s) for the standard (gold) and domain adaptive (turquoise) models as evaluated on the held-out test dataset. The ridge along the horizontal axis of each contour is traced by a dashed line, representing the mode of the inferred value for each true value of s. Raw data underlying the contour plots are presented in Fig. S2 in Appendix C. See Fig. 4.1C for definition of the model labels.

ment upon the standard model (Fig. S3A&B in Appendix C). The exception appeared to be when the target domain data consisted entirely of sweep examples (100% sweep). Although highly unrealistic, this scenario demonstrates that the domain-adaptive model can underperform the standard model when the target domain data follow a radically different distribution.

Another type of imbalance arises if only a limited amount of target domain data is available to train the domain-adaptive model. Using the same set of simulations for the background selection mis-specification experiments, we tested the performance of the domain-adaptive SIA model when trained with less target domain data. With the target domain data at only 10% of the source domain data (source:target ratio = 10:1), the model suffered a noticeable drop in performance yet still maintained a clear advantage over the standard model (Fig. S3C-E in Appendix C). We did not examine the case where there is more target domain than source domain data, since one could always simulate additional source domain data to match the size of the target domain. In summary, our experiments suggest that domain adaptation can accommodate reduced or imbalanced data for the target domain but there is a cost in performance if the reduction or imbalance is extreme.

4.3.3 Performance of domain-adaptive ReLERNN model

We performed a parallel set of experiments with a domain-adaptive version of ReLERNN. In this case, the background selection experiment was essentially the same as for SIA, but we used a simpler design for the demography mis-specification experiment, following Adrion, Galloway, et al., 2020. Briefly, the "real" (target domain) data was generated according to the out-of-Africa European demographic model estimated by Tennessen et al., 2012. By contrast, the simulated data for the source domain simply assumed a constant-sized panmictic population at equilibrium with $N_e = \frac{\hat{\theta}_W}{4\mu}$, where $\hat{\theta}_W$ is the Watterson estimator obtained from the "real" data (see Methods for details).

Similar to our results for SIA, the domain-adaptive ReLERNN model both reduced the MAE and corrected for the downward bias in recombination-rate estimates compared to the standard model (Figs. 4.4 and S6 in Appendix C). In the background-selection experiment, the standard ReLERNN model performed quite well (Figs. 4.4A and S6A in Appendix C, MAE = 5.60×10^{-9}), but the domain-adaptive ReLERNN model nonetheless further reduced the MAE to 4.41×10^{-9} (Fig. S6C in Appendix C, Welch's t-test: $n = 25,000$, $t = 31.0$, $p < 10^{-208}$). The advantage of the domain-adaptive model was more apparent in the demography-mis-specification experiment (Figs. 4.4B and S6B in Appendix C), where it reduced the MAE from 8.06×10^{-9} to

5.45×10^{-9} (Fig. S6D in Appendix C, Welch's t-test, $n = 25,000$, $t = 72.4$, $p < 10^{-323}$). Notably, our results for the standard model in the demography-mis-specification experiment were highly similar to those reported by Adrion, Galloway, et al., 2020, including the approximate mean and range of the raw error (compare Fig. 4A from Adrion, Galloway, et al., 2020 and Fig. S6D in Appendix C), as well as the downward bias.

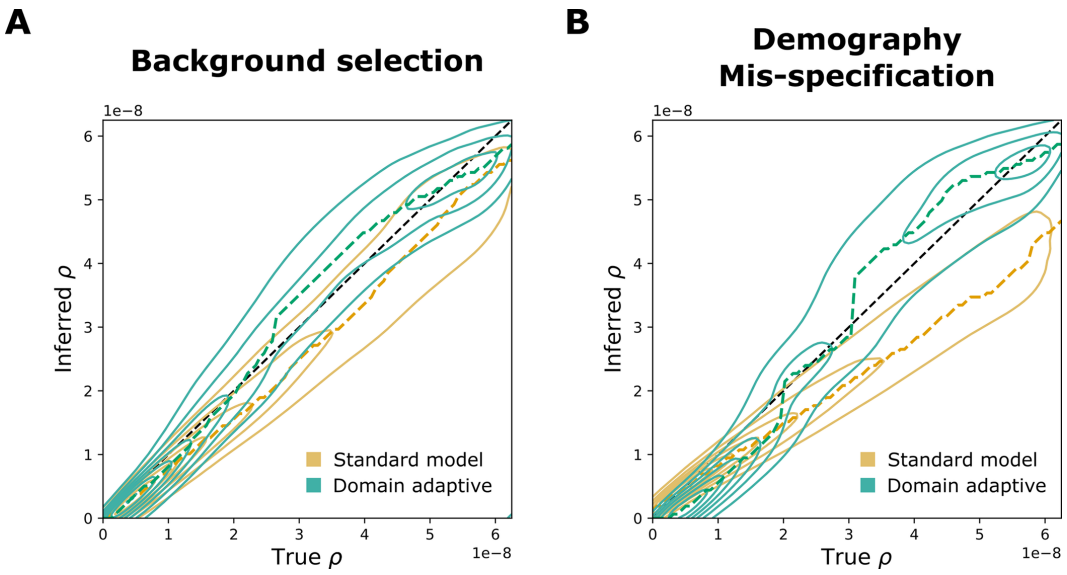


Figure 4.4: **Performance of domain-adaptive ReLERNN models.** Results are shown from (**A**) the background-selection and (**B**) the demography-mis-specification experiments. Each contour plot summarizes true (horizontal axis) vs. inferred (vertical axis) recombination rates (ρ) for the standard (gold) and domain adaptive (turquoise) models as evaluated on the held-out test dataset. The ridge along the horizontal axis of each contour is traced by a dashed line, representing the mode of the inferred value for each true value of ρ. Raw data underlying the contour plots are presented in Fig. S6 in Appendix C.

Interestingly, Adrion, Galloway, et al., 2020 observed that ReLERNN was sometimes more strongly influenced by demographic mis-specification than unsupervised methods such as LDhelmet, even though it still performed better in terms of absolute error. The addition of domain adaptation appears to considerably mitigate this susceptibility to demographic mis-specification, making an excellent method even stronger.

4.3.4 Efficacy of domain adaptation under various degrees of simulation mis-specification

So far, we have examined scenarios of relatively modest simulation mis-specifica -tion, likely to be encountered in real applications. While domain adaptation appeared to be effective in these cases, we expect a limit to its capability when mis-specification is extreme. We therefore carried out a series of experiments to probe the performance of the dadaSIA model under increasingly severe simulation mis-specification (Fig. S4 in Appendix C, also see Methods).

We found that dadaSIA exhibited good performance when mis-specification was caused by genealogy inference alone or by light to moderate bottlenecks. As the bottleneck became more severe, its performance deteriorated, but even with a 5% bottleneck, dadaSIA still outperformed the standard model (Fig. 4.5). To examine the limits of the method, we tested an extreme scenario with the 5% bottleneck, background selection and an 8-fold mis-specification of recombination rate. In this case, the model performed poorly, having virtually no power to classify sweeps and large errors in its selection coefficient estimates (Fig. 4.5). This example demonstrates that, while domain adaptation is useful over a broad range of mis-specification levels, it eventually does fail when mis-specification becomes extreme.

Does domain adaptation compromise performance at the opposite extreme, where there is little or no simulation mis-specification? To address this question, we tested the standard and domain-adaptive ReLERNN models in a setting without any simulation mis-specification. We focused here on ReLERNN, which directly uses raw genotypic data, as opposed to SIA, which always has some mis-specification due to genealogy inference error. We observed that the standard and domain-adaptive ReLERNN models performed nearly identically when no mis-specification was present, with only minor decreases in performance (Fig. S7 in Appendix C). Thus, there is perhaps some cost in using domain adaptation when it is not needed, but, at least in our case, that cost appears to be slight.

4.3.5 Application of domain-adaptive SIA to real data

In applications to real data, the true selection coefficient is not known, so it is impossible to perform a definitive comparison of methods. Nevertheless, it can be informative to evaluate the degree to which alternative methods are concordant, especially with consideration of their relative performance in simulation studies.

Toward this end, we re-applied our **d**omain-**ada**ptive SIA (dadaSIA) model to several loci in the human genome that we previously analyzed with SIA (He-

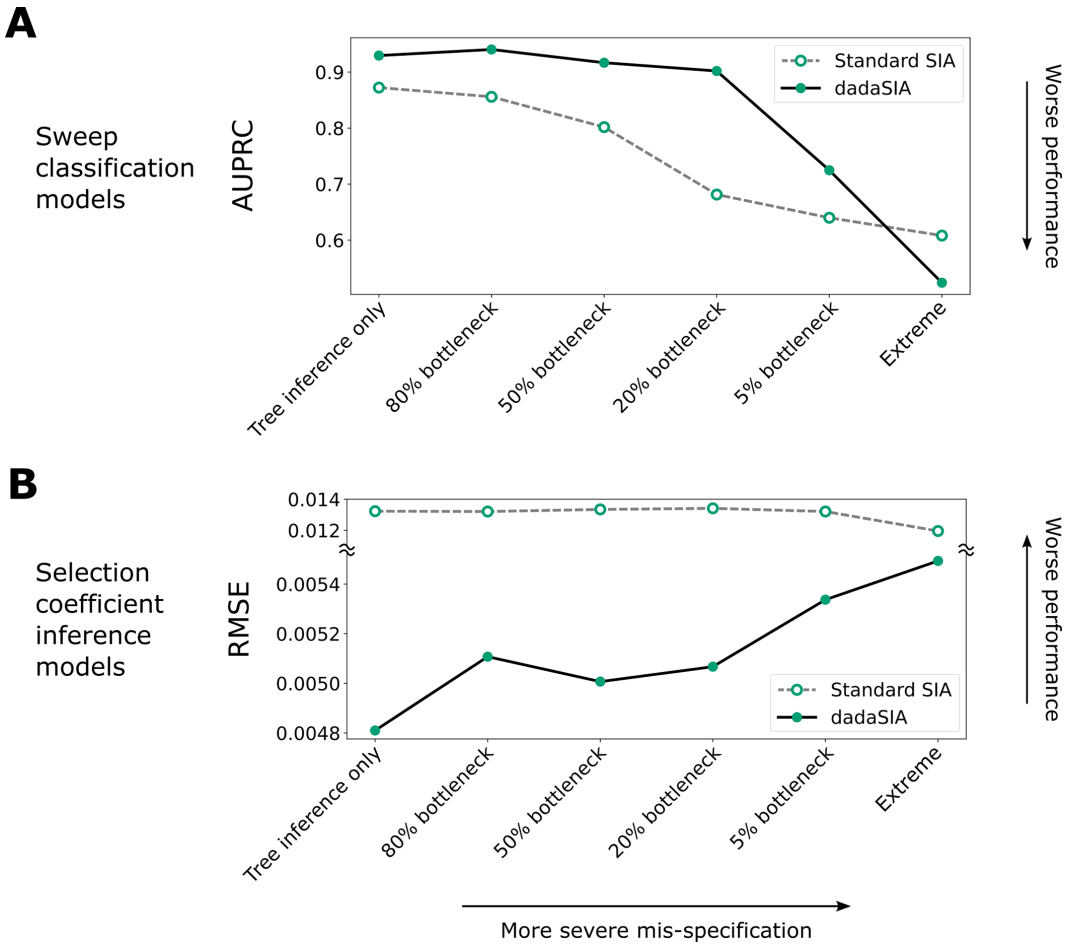


Figure 4.5: **Performance of domain-adaptive SIA (dadaSIA) model with different degrees of mis-specification.** The performance of the model on the sweep classification task is quantified by the AUPRC (**A**). Performance on the selection-coefficient inference task is quantified by RMSE (**B**). In the "tree inference only" case, there is no mis-specification other than that caused by error in genealogy inference. In the "extreme" case, mis-specification consists of a 5% bottleneck, background selection and an 8-fold mis-specification in recombination rate. See Fig. S4 in Appendix C for illustrations of the different bottlenecks and Methods for details.

jase et al., 2022), using whole-genome sequence data from the 1000 Genomes CEU population (Auton et al., 2015). For the target domain, we sampled genealogies from genome-wide ARGs inferred from the individual sequences (see Methods). The putative causal loci analyzed included single nucleotide polymorphisms (SNPs) at the *LCT* gene (Bersaglieri et al., 2004), one of the best-studied cases of selective sweeps in the human genome; at the disease-associated genes *TCF7L2* (Lyssenko et al., 2007), *ANKK1* (Spellicy et al., 2014) and *FTO* (Frayling et al., 2007); at the pigmentation genes *KITLG* (Sulem et al., 2007), *ASIP* (Eriksson et al., 2010), *TYR* (Sulem et al., 2007; Eriksson et al., 2010), *OCA2* (Han et al., 2008; Sturm et al., 2008), *TYRP1* (Kenny et al., 2012) and *TTC3* (F. Liu et al., 2010), which were also analyzed by Stern et al., 2019; and at the genes *MC1R* (Sulem et al., 2007; Han et al., 2008) and *ABCC11* (Yoshiura et al., 2006), where SIA reported novel signals of selection.

We found that dadaSIA generally made similar predictions to SIA at these SNPs, but there were some notable differences. The seven loci predicted by SIA to be sweeps were also predicted by dadaSIA to be sweeps (Table 4.1), although dadaSIA always reported higher confidence in these predictions (with probability of neutrality, $P_{\text{neu}} < 10^{-2}$ in all cases) than did SIA (P_{neu} up to 0.384 for *TYR*). The five loci predicted by SIA not to be sweeps were also predicted by dadaSIA not to be sweeps ($P_{\text{neu}} > 0.5$). At *LCT*, the strongest sweep considered, the selection coefficient (s) estimated by dadaSIA remained very close to SIA's previous estimate of $s = 0.01$ and also close to several prior estimates (Bersaglieri et al., 2004; S. Mathieson and Mathieson, 2018; I. Mathieson, 2020). In all other cases, the estimate from SIA was somewhat revised by dadaSIA, generally by factors of about 2-3. Importantly, in all cases, the estimates from dadaSIA remained much closer to those from SIA than to estimates by other methods (Table 4.1). Together, these observations suggest that the addition of domain adaptation does not radically alter SIA's predictions for real data but may in some cases improve them (see Discussion).

4.4 Discussion

Standard approaches to supervised machine learning rest on the assumption that the data they are used to analyze follow essentially the same distribution as the data used for training. In applications in population genetics, the training data are typically generated by simulation, leading to concerns about potential biases from simulation mis-specification when supervised machine-learning methods are used in place of more traditional summary-statistic- or model-based methods (Caldas et al., 2022; Korfmann et al., 2023). In this ar-

Table 4.1: Selection coefficients in the European population estimated by domain-adaptive SIA compared to previous estimates.

Gene	**SNP**	**Estimates of selection coefficient**		
		Domain-adaptive SIA	**SIA***	**Previous estimates**
KITLG	rs12821256	0.0035	0.0019	0.0161 (Stern et al., 2019)
ASIP	rs619865	0.0057	0.0019	0.0974 (Stern et al., 2019)
TYR	rs1393350	0.0028	0.0011	0.0112 (Stern et al., 2019)
OCA2	rs12913832	0.0093	0.0056	0.002 (Stern et al., 2019); 0.036 (Wilde et al., 2014)
MC1R	rs1805007	0.0027	0.0037	No selection (Harding et al., 2000)
ABCC11	rs17822931	0.0020	0.00035	≈ 0.01 in East Asian (Ohashi et al., 2011)
LCT	rs4988235	0.0097	0.010	≈ 0.01 (Bersaglieri et al., 2004; S. Mathieson and Mathieson, 2018; I. Mathieson, 2020)
TYRP1	rs13289810	$P_{\text{neu}} > 0.5$	$P_{\text{neu}} > 0.5$	No selection (Stern et al., 2019)
TTC3	rs1003719	$P_{\text{neu}} > 0.5$	$P_{\text{neu}} > 0.5$	No selection (Stern et al., 2019)
TCF7L2	rs7903146	$P_{\text{neu}} > 0.5$	$P_{\text{neu}} > 0.5$	N/A
ANKK1	rs1800497	$P_{\text{neu}} > 0.5$	$P_{\text{neu}} > 0.5$	N/A
FTO	rs9939609	$P_{\text{neu}} > 0.5$	$P_{\text{neu}} > 0.5$	N/A

*The original SIA model in Hejase et al., 2022 uses genealogies *inferred* from simulations for training, despite the availability of ground truth genealogies.

ticle, we have shown that techniques from the "domain adaptation" literature can effectively be used to address this problem. In particular, we showed that the addition of a GRL to two recently developed deep-learning methods for population genetic analysis – SIA and ReLERNN – led to clear improvements in performance on "real" data that differed in subtle but important ways from the data used to train the models. These improvements were observed both when the demographic models were mis-specified and when background selection was included in the simulations of "real" data but un-modeled in the training data.

While we observed performance improvements in all of our experiments, they were especially pronounced in the case where SIA was used to predict specific selection coefficients, rather than simply to identify sweeps. The standard model (with training on simulated data and testing on "real" data) performed particularly poorly in this regression setting and domain adaptation produced striking improvements (Fig. 4.3B&D). This selection-coefficient inference problem appears to be a harder task than either sweep classification or recombination-rate inference, and the performance in this case proves to be more sensitive to simulation mis-specification (cf. Fig. 4.3A&C). In general, we anticipate considerable differences across population-genetic applications in the value of domain adaptation, with some applications being more sensitive to simulation mis-specification and therefore more apt to benefit from domain adaptation, and others being less so.

We also observed some interesting differences in the ways SIA and ReLERNN responded to domain adaptation. For example, the performance gap between the "simulation benchmark" (trained and tested on simulated data) and "hypothetical true" (trained and tested on real data) models was considerably greater for SIA than for ReLERNN (Figs. S2C&D, S6C&D in Appendix C). This difference appears to be driven by ARG inference, which is required by SIA in the hypothetical true case but not the simulation benchmark case, and for which no analog exists for ReLERNN. For SIA, the uncertainty about genealogies given sequence data makes the prediction task fundamentally harder in the real world (target domain) than in simulation (source domain) (Fig. 4.1B). By contrast, ReLERNN does not depend on a similar inference task, and therefore the target and source domains are more or less symmetric. This same factor contributed to the much more dramatic drop in performance for SIA than ReLERNN under the "standard model," where the model is trained on simulated data and naively applied to "real" data (Figs. 4.3B&D, 4.4). It is, of course, also conceivable that simulation mis-specification has more impact on selection inference than recombination rate inference, rendering the standard SIA model less robust than the standard ReLERNN model. Re-

gardless of the exact cause, the result is more potential for improvement from domain adaptation with SIA than with ReLERNN (Figs. 4.3, 4.4, S2, S6 in Appendix C). In effect, in SIA, domain adaptation not only mitigates simulation mis-specification but also compensates for ARG inference error, as directly evidenced by the observation that domain adaptation improves model performance when mis-specification is due to genealogy inference alone (Fig. 4.5, "Tree inference only"). More broadly, we expect domain adaptation to be especially effective in applications that depend not only on the simulated data itself but also on nontrivial inferences of latent quantities that are known for simulated but not real data.

In addition, we performed a series of experiments to probe the limits of domain adaptation. As expected, the dadaSIA model gradually lost its power as simulation mis-specification became more severe. In an extreme case where mis-specification involved demography, selection and recombination rate, the dadaSIA model had virtually no power to classify sweeps and exhibited high error of selection coefficient inference (Fig. 4.5). In practice, simulation models themselves are inferred from real data. With high quality data, state-of-the-art inference tools are unlikely to fail completely (e.g., by missing a 5% bottleneck completely, or under-estimating recombination rate by an order of magnitude). We thus expect the most extreme scenario tested here to be fairly uncommon. Nevertheless, this experiment demonstrated that there are reasonable limits to the efficacy of domain adaptation. Consequently, it is important in real-world applications to begin with the best possible simulation model, before using domain adaptation to further optimize performance.

Because the accuracy of the simulation model is typically not known a priori, it is tempting to apply domain adaptation in all cases, regardless of the true degree of mis-specification. Indeed, we found that the domain-adaptive model performed very similarly to the standard model in the absence of mis-specification (Fig. S7 in Appendix C), suggesting little risk in applying the approach liberally. When the target domain is mis-specified, the domain classifier appears to "unlearn" the mis-specification, with its loss increasing steadily before plateauing where the source and target domains are no longer distinguishable. In contrast, when there is no mis-specification, the domain classifier starts with a high loss and this loss remains high (Figs. 4.2B, S8 in Appendix C). In this case, because the source and target domains are effectively indistinguishable, the domain classifier can never do much better than randomly guessing, leading to near-zero gradients along the domain classifier branch. In effect, the training process ignores the domain-classifier branch in this case, and improves only the feature-extractor and label-predictor portions of the model. For this reason, the domain-adaptive model behaves nearly identically

to the standard model in the absence of mis-specification.

The accuracy of even the best current selection-coefficient inference methods appears limited (Flagel et al., 2019; Stern et al., 2019; Torada et al., 2019; Hejase et al., 2022). More work is needed on models and methods for inference as well as on the problem of simulation mis-specification. Nevertheless, current methods can still be valuable in approximately characterizing the strength of selection. In our re-analysis of several loci in the 1000 Genomes CEU population, we found that dadaSIA made similar predictions to SIA, but it tended to exhibit higher confidence in its predictions (Table 4.1). Considering the extensive previous work on demography inference for the CEU population, we expect that simulation mis-specification is limited in severity for this analysis, but that some mis-specification is inevitable. Given the similar performance on benchmarks of SIA and other leading methods such as CLUES, their similar sensitivity to moderate levels of simulation mis-specification (Hejase et al., 2022), and the improvements offered by domain adaptation that are demonstrated in this work, we find it likely that dadaSIA improves on previous estimates of selection coefficients in this setting.

In a typical application of domain adaptation, the distribution shift between the source and target domains is treated as a nuisance. However, for certain population genetic questions, the gap between the simulated and real data could in principle help to reveal unmodeled evolutionary processes. We observed that the domain classifier generally tended to start with a lower loss and took more epochs to train when the mis-specification is more severe (Fig. S9 in Appendix C). It might be worthwhile, as a future endeavor, to try to identify the features driving this loss, understand their evolutionary significance, and, perhaps, incorporate them into a new set of simulations. In such a way, domain adaptation could be used to discover evolutionary processes and improve the models used for simulation.

Although our experiments were limited to background selection and demographic mis-specification, we expect that the domain adaptation framework would also be effective in addressing many other forms of simulation mis-specification, involving factors such as mutation or recombination rates, or the presence of gene conversion. Another interesting application may be to use domain adaptation to accommodate admixed populations. Each ancestry component could be modeled as a distinct target domain using a multi-target domain adaptation technique (Isobe et al., 2021; Nguyen-Meidine et al., 2021; Roy et al., 2021). It is also worth noting that our experiments considered only one, rather simple, strategy for domain adaptation. Since the GRL was proposed, several other architectures for deep domain adaptation have achieved even better empirical performance on computer vision tasks (see: Papers with

Code).

Our domain-adaptation approach leaves simulations unchanged and attempts to "unlearn" their mis-specification, in contrast to other strategies that aim to improve the simulations themselves. For example, the original SIA model was trained with inferred genealogies from the simulated sequences, rather than the true genealogies used to generate the data, to mitigate the effect of genealogy inference error (Hejase et al., 2022). An alternative approach is to use a GAN to train a simulator that accurately mimics the real data (Z. Wang et al., 2021). These methods can require costly preprocessing steps, but they have the advantage of explicitly addressing the simulation mis-specification in an interpretable manner.

It is perhaps worth distinguishing mis-specification along the axis of inference – that is, of target parameters such as the selection coefficient – from mis-specification of other "nuisance" parameters (such as demographic parameters), or similarly, other unmodeled aspects of the data-generating process (such as background selection). From our observations, domain adaptation appears to be effective at addressing mis-specification of nuisance parameters or processes, at least if it is not too severe. Mis-specification of the target parameters, however, is clearly a more challenging problem. For example, it seems unlikely that domain adaptation will ever be able to "extrapolate" beyond the range of the training examples (as it fails to do in Fig. S5 in Appendix C). Hence, it is essential in practical applications to simulate the parameter of interest from an adequately large range. Notably, Burger et al., 2022 recently developed a method that addresses mis-specification in the distribution (but not the range) of a target parameter. Their method improves inference of the scaled mutation rate when regions of the parameter space are under-sampled in the training simulations by adaptively reweighing the training data, effectively improving interpolation (but not extrapolation) from the training distribution. We view these interrelated questions of how to accommodate mis-specification of both nuisance and target parameters as promising areas for future work.

Mis-specification is not only a problem in the simulation-based supervised machine learning setting explored in this work (*simulation* mis-specification), but also arises in many unsupervised methods (such as maximum-likelihood or Bayesian probabilistic models). In these cases, mis-specification typically results from simplified or incorrect assumptions built into a probabilistic model (*model* mis-specification, reviewed in detail by Johri et al., 2022). Such model mis-specification can be difficult and time-consuming to identify and address, usually calling for careful experimental design and model comparison (Johri et al., 2022). In some ways, the simulation mis-specification problem is more straightforward to address through fully empirical, data-driven solutions such

as domain adaptation. It remains to be seen whether these empirical techniques can be used to improve probabilistic-model-based inference methods. Overall, there is rich potential for new work to address a wide variety of misspecification challenges in population genetics, leading to improved accuracy and robustness in inference.

4.5 Methods

4.5.1 Methodological summary of unsupervised domain adaptation

To build domain-adaptive versions of SIA and ReLERNN, we opted for the neural network architecture proposed by Ganin and Lempitsky, 2014, which involved attaching a domain classifier branch via a GRL to a layer of the original neural network where a latent representation of the data is presumably obtained. For example, in a CNN, the attachment point is usually immediately after the convolutional and pooling layers, which are primarily responsible for feature extraction. One possible heuristic for picking the attachment point is to look for a "bottleneck layer" in the original network corresponding to the lowest-dimensional representation of the input. The GRL-containing networks consist of three components–a label predictor branch, a domain classifier branch and a feature extractor common to both branches (Fig. 4.2A&B). During the feedforward step, when data is fed to the neural network to obtain prediction outputs in both branches, the GRL is inactive; it simply passes along any input to the next layer. However, during backpropagation, when the gradient of the loss function with respect to the weights of the network is calculated iteratively backward from the output layer, the GRL inverts the sign of any incoming gradient before passing it back to the previous layer. This operation has the effect of driving the feature extractor away from distinguishing the source and target domains, and consequently encourages it to extract "domain-invariant" features of the data. This effect is manifested during training as the domain-classifier loss being *maximized*. We implemented the GRLs in TensorFlow (v2.4.1) using the `tf.custom_gradient` decorator. On top of each custom GRL, the rest of the model was built using the `tf.keras` functional API (see the GitHub repository for details).

All models were trained with the Adam optimizer using a batch size of 64. For the domain-adaptive models, training consisted of both (1) feeding labeled data from the source domain through the label predictor and obtaining a label prediction loss (cross entropy for classification task, mean squared error for regression task); and (2) feeding a mixture of unlabeled data from

both the source and target domains through the domain classifier, obtaining a domain classification loss (cross entropy) (Fig. 4.2C). In each minibatch, back-propagation from these two steps occurred simultaneously (i.e. the weights of the feature extractor were updated according to the combination of gradient from the label predictor and reversed gradient from the domain classifier). Note that the same source-domain data (but shuffled differently) were used for both steps. Training was accomplished using a custom data generator implemented with `tf.keras.utils.Sequence`. In this study, we simply assigned equal weights to the label-prediction and domain-classification loss functions (following Ganin and Lempitsky, 2014). Nonetheless, the relative weights of the two branches can be tuned via a hyper-parameter λ, with potential implications for performance. Intuitively, the domain classifier should be penalized more when the simulations are more mis-specified. One potential strategy is to leverage the losses and gradients of the domain classifier to guide the choice of λ. Each training epoch took around 300 s for the domain-adaptive SIA model and around 800 s for the domain-adaptive ReLERNN model on a single NVIDIA Tesla V100 GPU. With early-stopping, the models in this study were trained on average for tens of epochs. The runtimes for domain-adaptive SIA and ReLERNN models were therefore on par with their standard versions (on the order of hours) (Adrion, Galloway, et al., 2020; Hejase et al., 2022).

4.5.2 Setup of benchmarking experiments

We designed four benchmarking scenarios to contextualize the performance of the domain-adaptive models (Fig. 4.1C). **i)** In the *simulation benchmark* (*source-matched*) case, we tested the original model trained with source domain data on held-out samples in the source domain. This is how model benchmarks are usually run, with the test data following the same distribution as the training data. Note that for the SIA model, the source domain consists of true genealogies and therefore both training and testing were performed with true trees. **ii)** In the *hypothetical true model* (*target-matched*) case, the original model was trained and tested with labeled target domain data. Here, both training and testing were performed with inferred genealogies for the SIA model. This is a hypothetical case because it is unlikely in the evolution setting to have large quantities of labeled data from the target domain for training (i.e. real population data with known ground truth of evolutionary parameters). This case represents the performance ceiling of a standard machine learning model trained in-domain. **iii)** The *standard model application* recapitulated the usual workflow of supervised machine learning methods, where the model trained with source domain simulations was applied directly to "real" data in the target domain. This was the baseline case to

which we compared the domain-adaptive model. **iv)** *Domain-adaptive application* of supervised machine learning models is the novel approach introduced in this study (see above and Fig. 4.1A).

4.5.3 Background selection experiment with SIA

To assess the robustness of dadaSIA to background selection, we simulated labeled examples (250,000 neutral and 250,000 sweep) in the source domain under demographic equilibrium with $N_e = 10,000$ and $\mu = \rho = 1.25 \times 10^{-8}$/bp/gen. The sweep simulations consisted of 100kb chromosomal segments with a hard sweep at the central nucleotide having selection coefficient $s \in [0.002, 0.01]$. Simulations were performed in SLiM 3 (Haller et al., 2019; Haller and Messer, 2019) followed by recapitation with msprime (Baumdicker et al., 2022), and we kept the true genealogies as source domain data. The unlabeled data in the target domain (with the exception of held-out test dataset with labels retained) were simulated in a similar fashion, albeit with a 10kb segment ("gene") under purifying selection at the center of each 100kb chromosomal segment. All mutations in the central 10kb segment that arose during the forward stage of the simulations (in SLiM), other than the beneficial mutation in sweep simulations, followed a DFE parameterized by a gamma distribution with a mean $\bar{s} = -0.03$, a shape parameter $\alpha = 0.2$ and had dominance coefficient $h = 0.25$ (Boyko et al., 2008). We retained only the sequence data from the target domain simulations and inferred genealogies using Relate (Speidel et al., 2019). The datasets were partitioned following a 90%:2%:8% train-validation-test split.

4.5.4 Demography mis-specification experiment with SIA

In a second set of simulations, we gauged whether domain adaptation also protects SIA against demographic mis-specification. In this case, instead of specifying the degree of mis-specification a priori, we designed an end-to-end workflow that recapitulated how demographic mis-specification arises in a realistic population genetic analysis (Fig. S1A in Appendix C). First, we simulated "real" data (in the target domain) using an assumed demography (Fig. S1A in Appendix C, loosely based on the three-population model in Campagna et al., 2022). Similar to what one would do with actual sequence data, we then used the "real" samples to infer a demography with G-PhoCS (Gronau et al., 2011), pretending that the true demography and genealogies were unknown. The G-PhoCS model assumed constant population sizes between split events and a single pulse migration from population C to B, and therefore was under-parameterized. As shown in Fig. S1A in Appendix C, the inferred

demography was consequently somewhat mis-specified. In addition to errors in population sizes, the split between B and C was inferred to be much more recent compared to the true demographic model. This mis-specified demographic model was then used to simulate labeled training data (in the source domain) for SIA.

With the goal of using SIA to infer selection in population B, we simulated a soft sweep site at the center of a 100kb chromosomal segment with selection coefficient $s \in [0.003, 0.02]$ and initial sweep frequency $f_{\text{init}} \in [0.01, 0.1]$, under positive selection only in population B. To improve computational efficiency, simulations were performed with a hybrid approach where the neutral demographic processes were simulated first with msprime (Baumdicker et al., 2022), followed by positive selection simulated with SLiM 3 (Haller et al., 2019; Haller and Messer, 2019). We produced 200,000 balanced (between neutral and sweep) simulations of "real" data, 10,000 of which were randomly held out as ground-truth test data for benchmarking with their labels preserved (Fig. S1A in Appendix C). The rest remained unlabeled. This corresponded to a train-validation-test split of 93%:2%:5%. We preserved only the sequences and used Relate (Speidel et al., 2019) to infer the ARG of population B from the "real" data. SIA works with a single population and thus the central genealogies containing only samples from population B were encoded as input to the model. For demographic inference, we randomly downsampled 10,000 5kb loci and analyzed them with G-PhoCS, keeping 4 (diploid) individuals from population A and 16 (diploid) individuals each from populations B and C. We took the median of 90,000 MCMC samples (after 10,000 burn-in iterations) as the inferred demography (shown in Fig. S1A in Appendix C). The control file used to run G-PhoCS is available in the GitHub repository. We then simulated true genealogies of population B using the inferred demography, yielding 200,000 balanced samples with neutral/sweep and selection coefficient labels. All SIA models in this study used 64 diploid samples (128 taxa).

4.5.5 Running SIA under varying degrees of simulation mis-specification

To probe the limit of domain adaptation in mitigating simulation mis-specifica -tion, we performed a series of experiments that gradually increased the severity of mis-specification. In all cases, the source domain consisted of 400,000 balanced samples of *true* genealogies simulated under a constant N_e of 10,000. The target domain had a matching size of 400,000 balanced samples of *inferred* genealogies. We used $\mu = \rho = 1.25 \times 10^{-8}$/bp/gen unless otherwise specified. The datasets were partitioned following an 87.5%:2.5%:10% train-validation-

test split. In the "tree inference only" case, the target domain consisted of inferred genealogies simulated under a constant N_e of 10,000 with no demographic mis-specification. In addition, we tested four cases with N_e = 8,000, 5,000, 2,000 or 500 bottlenecks between 1,000 and 2,000 generations before the present, respectively (Fig. S4 in Appendix C). Finally, we tested an "extreme" case with the $N_e = 500$ bottleneck, a mis-specified $\rho = 1 \times 10^{-7}$, as well as background selection in the central 10kb region following a DFE parameterized by a gamma distribution with a mean $\bar{s} = -0.03$, a shape parameter $\alpha = 0.2$ and a dominance coefficient $h = 0.25$.

4.5.6 Updates to genealogical features and deep learning architecture for the SIA model

For this study, we adopted a richer encoding of genealogies than the one used previously for SIA. Instead of simply counting the lineages remaining in the genealogy at discrete time points (Hejase et al., 2022), we fully encoded the topology and branch lengths of the tree using the scheme introduced by (J. Kim et al., 2020). Under this scheme, a genealogy with n taxa is uniquely encoded by an $(n-1) \times (n-1)$ lower-triangular matrix $\mathbf{F}$ and a weight matrix $\mathbf{W}$ of the same shape. Each cell (i, j) of $\mathbf{F}$ records the lineage count between coalescent times t_{n-j} and t_{n-1-i}, whereas each cell (i, j) of $\mathbf{W}$ records the corresponding interval between coalescent times, $t_{n-j} - t_{n-1-i}$ (see Fig. S1B in Appendix C and J. Kim et al., 2020 for details). In addition, we used a third matrix $\mathbf{R}$ to identify the subtree carrying the derived alleles at the site of interest, following the same logic as $\mathbf{F}$ (see Fig. S1B in Appendix C for an example). The $\mathbf{F}$, $\mathbf{W}$ and $\mathbf{R}$ matrices have the same shape and therefore can easily be stacked as input to a convolutional layer with three channels (Fig. 4.2A, 128 taxa yield a $127 \times 127 \times 3$ input tensor).

Unlike the previous reductive encoding of lineage counts, the new scheme is bijective (J. Kim et al., 2020) and therefore contains the entirety of information in the genealogy. To utilize the improved input feature consisting of stacks of matrices, we modified the neural network architecture of SIA and used convolutional layers (Fig. 4.2A). The new feature encoding and CNN architecture resulted in modest gain in performance compared to the original encoding and RNN architecture (Fig. S1C in Appendix C). In this study, both the standard and domain-adaptive SIA models use convolutional layers with the improved feature encoding. The original SIA codebase has been updated to take advantage of the new feature encoding and model architecture as well.

4.5.7 Simulation study of recombination rate inference with ReLERNN

We conducted two sets of simulation experiments to test the same two types of mis-specification as previously described for SIA. Each simulation consisted of 32 haploid samples of 300kb genomic segment with uniformly sampled mutation rate $\mu \sim \mathcal{U}[1.875 \times 10^{-8}, 3.125 \times 10^{-8}]$ and recombination rate $\rho \sim \mathcal{U}[0, 6.25 \times 10^{-8}]$. To test the effect of background selection, the labeled source domain data (with true values of ρ) were simulated under demographic equilibrium with $N_e = 10,000$, whereas the unlabeled target domain data were simulated under the same demography, but with the central 100kb region under purifying selection, as with SIA. To test the effect of demographic mis-specification, we conducted simulations similar to those of Adrion, Galloway, et al., 2020 where labeled source domain data were generated under demographic equilibrium (with $N_e = 6,000$, calculated approximately by $\frac{\hat{\theta}_W}{4\mu}$ where $\hat{\theta}_W$ was estimated from the target domain data) and unlabeled target domain data were generated under a European demography (Tennessen et al., 2012). For each domain, 500,000 simulations were generated with SLiM 3 (background selection experiment) or msprime (demography experiment), and partitioned following an 88%:2%:10% train-validation-test composition. We modified the ReLERNN model to be domain-adaptive (Fig. 4.2B) and used the simulated data to benchmark its performance against the original version of the model.

4.5.8 Application of domain-adaptive SIA model to 1000 Genomes CEU population

Labeled training data (source domain) for SIA were simulated with discoal (Kern and Schrider, 2016) under the European demographic model from Tennessen et al., 2012. Following Hejase et al., 2022, we simulated 500,000 100-kb regions of 198 haploid sequences. The per-base per-generation mutation rate (μ) and recombination rate (ρ) of each simulation were sampled uniformly from the interval $[1.25 \times 10^{-8}, 2.5 \times 10^{-8}]$; the segregating frequency of the beneficial allele (f) was sampled uniformly from $[0.05, 0.95]$; the selection coefficient (s) was sampled from an equal mixture of a uniform and a log-uniform distribution with the support $[1 \times 10^{-4}, 2 \times 10^{-2}]$. An additional 500,000 neutral regions were simulated to train the classification model, under the identical setup sans the positively selected site.

We curated target domain data from the 1000 Genomes CEU population to train the dadaSIA model. The genome was first divided into 2Mb windows

1,111 of which passed three data-quality filters: **1)** contained at least 5,000 variants, **2)** at least 80% of these variants had ancestral allele information, and **3)** at least 60% of nucleotide sites in the window passed *both* the 1000 Genomes strict accessibility mask (Auton et al., 2015) and the deCODE recombination hotspot mask (standardized recombination rate > 10; Kong et al., 2010). In each of these 1,111 windows, we randomly sampled 1,000 variants and extracted genealogical features at those variants from Relate-inferred ARGs (Speidel et al., 2019), yielding around 1 million samples that constituted the unlabeled target domain data. Finally, domain-adaptive SIA models for classifying sweeps and inferring selection coefficients were trained as described previously and applied to a collection of loci of interest (Table 4.1).

Chapter 5

Conclusions and Perspectives

5.1 Summary

There has been astonishing progress in the adoption of AI/ML for population genetics research since we began the works presented in this thesis. This field of research has emerged from prototypical models tailored to relatively bespoke tasks (Sheehan and Song, 2016), gone through speculation and excitement about the promises and pitfalls of a data-driven ML approach to evolutionary modeling (Schrider and Kern, 2018), and eventually accumulated a robust body of literature that spans a range of technical and methodological aspects of ML tailored to diverse empirical problems in population genetics (Huang et al., 2023; Korfmann et al., 2023). My thesis work focuses on utilizing ML to fulfill the potential of making accurate inference with complex genealogical information in the ARG (Chapters 2 & 3) and addressing the fundamental limitation of mis-specified training data for supervised ML models (Chapter 4). This thesis makes a significant contribution to simultaneous efforts in the field that strive to make AI/ML a powerful and accessible inferential framework for profound evolutionary discoveries.

There are rich opportunities to move forward with this line of work. For example, model interpretability remains a crucial subject for further research in the evolutionary applications of deep learning. Population geneticists are interested in not only predictions, but also mechanistic understanding of evolution. Therefore, despite the superb predictive performance of deep learning models, their "black-box" nature presents an obstacle to uncovering the evolutionary machinery driving the genetic diversity observed in populations. Much effort has already been devoted to addressing this issue by incorporating the latest techniques from explainable AI research. This growing body of work is highlighted in Novakovsky et al., 2023. Below, we conclude the thesis by

elaborating on another promising new area that pushes towards a deeper understanding of evolution through ML. Generative AI has achieve remarkable success in a variety of domains, notably NLP and CV, but is still in its infancy in population genetics.

5.2 Evolutionary modeling in the era of generative AI

Generative models capture the underlying probability distribution of observed data and consequently are capable of creating novel data points beyond the observed data by sampling from the captured probability distribution. Many traditional population genetic models are generative models, parameterized under theories of evolution. Although fully interpretable, these models often lack either the scalability to handle large amount of modern genomic data or the versatility to accommodate complex evolutionary processes, as discussed previously (section 1.3). Deep generative models provide an alternative of using neural networks to automatically learn the probability distribution from the training data in a domain-agnostic fashion.

Several early deep generative models have already been applied to population genetic tasks. Restricted Boltzmann machines (RBMs) are energy-based models that map the probability of data to an energy function and have been used to generate artificial genomes mimicking the properties of real ones (Yelmen et al., 2021; Yelmen et al., 2023). A variational autoencoder (VAE) consists of an encoder that maps the input into a latent space defined by a variational distribution and a decoder that can produce different samples from the distribution. VAEs are used to infer population structure and ancestry proportions from large genomic datasets such as the UK Biobank (Meisner and Albrechtsen, 2022). Generative adversarial network (GAN) is another widely popular model architecture and has been applied to infer demography (Z. Wang et al., 2021), selection (Riley et al., 2023), recombination (Gower et al., 2023) as well as to generate synthetic genomes (Yelmen et al., 2021; Yelmen et al., 2023). A GAN contains a generator and a discriminator trained in an adversarial manner, where the discriminator aims to correctly distinguish real data from synthetic data produced by the generator and the generator aims to fool the discriminator by creating realistic synthetic data. These early architectures suffer from various practical issues that limit their applications and have gradually been overshadowed by a new generation of deep generative models (Huang et al., 2023).

The latest and greatest deep generative models are diffusion models (Sohl-

Dickstein et al., 2015) and transformers (Vaswani et al., 2017). Diffusion models typically contain a forward diffusion process where noise is injected at each step and a reverse denoising process where neural networks recover the input by attempting to remove the noise. Diffusion models are most notably used for text-to-image generation and yield impressive results. Transformers pioneered the self-attention mechanism and have achieved unparalleled performance for many NLP tasks. In particular, transformers power large language models (LLMs) and help establish a new paradigm where large foundation models such as LLMs are pre-trained in an unsupervised or self-supervised manner and subsequently fine-tuned to use cases across a wide range of domains.

In light of the striking success of generative AI models across many fields, it is timely for applications of AI/ML in population genetics to move beyond optimization of simple prediction tasks and towards fully generative evolutionary models. Here we introduce two specific avenues of future work among myriad other possibilities in this rising field of research.

As mentioned previously, a great deal of efforts have been devoted to learning evolutionary parameters from large-scale population sequencing data using GANs. The fundamental challenge to this approach is that the generator component of the GAN is usually a population genetic simulator and therefore non-differentiable. A prototype of this model called "pg-gan" has been trained with a gradient-free method – simulated annealing, which can be computationally prohibitive for high-dimensional search spaces, hence limiting the complexity of the evolutionary model (Z. Wang et al., 2021). This problem is reminiscent of non-differentiable criteria encountered in training LLMs such as human feedback, which have found effective solutions through either reinforcement learning (Christiano et al., 2017; Hui et al., 2021) or zeroth-order optimizers using gradient approximation techniques such as ZO-SGD (Spall, 1992) and MeZO (Malladi et al., 2023). There is great potential in developing GANs tailored to complex evolutionary generators where the simulators are trained with either reinforcement learning or zeroth-order optimization algorithms. This new approach has the prospect of accommodating a large set of evolutionary parameters while maintaining computational efficiency.

Pushing the idea of generative models for evolutionary analyses further, an ultimate vision is to make large foundation models of evolution a reality. The emergence of generative AI models has made a profound impact on biomedical research. For example, LLMs of protein and DNA have already shown outstanding performance in a variety of problems in molecular biology such as protein structure (Lin et al., 2023) or variant effect prediction (Benegas et al., 2023; Cheng et al., 2023). Similarly, foundation models of evolution based

on genealogical embeddings of the ARG have the potential to revolutionize population genetic research. In order to build such models, an auto-regressive training procedure for genealogies needs to be developed. One possibility is to borrow the idea of "threading" from ARGweaver (Rasmussen et al., 2014; Hubisz et al., 2020, see section 1.1) where one left-out sample or subtree is "re-threaded" into the genealogy. We can similarly train a neural network by "masking" a sample or a subtree and optimizing it to complete the genealogy, akin to masked language modeling for LLMs. This will pave the way to foundation models pre-trained in a self-supervised manner with an incredibly wide range of simulations of many evolutionary processes. Such models have the potential to "understand" the grammar and logic of how evolutionary histories manifest in different topologies of the ARG, much like the way large language models "understand" natural languages. Since access to computational resources is a limiting factor for many academic researchers, the need to train highly parameterized models from scratch remains a hurdle for the adoption of deep learning in population genetics. Foundation models of evolution will create a new paradigm where empiricists can fine-tune high-performing pre-trained models even with limited amount of compute or labeled data instead of creating less powerful models from scratch for each new application to a different population or organism.

Generative AI models of evolution will empower rapid scientific discovery that keeps pace with the ever-growing scale of genomic datasets and move the field beyond solving isolated inference problems into a holistic view of evolution.

Bibliography

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R. A., Durvasula, A., Gronau, I., Kim, B. Y., McKenzie, P., Messer, P. W., Noskova, E., Ortega-Del Vecchyo, D., ... Kern, A. D. (2020). A community-maintained standard library of population genetic models (G. Coop, P. J. Wittkopp, J. Novembre, A. Sethuraman, & S. Mathieson, Eds.). *eLife*, *9*, e54967. https://doi.org/10.7554/eLife.54967

Adrion, J. R., Galloway, J. G., & Kern, A. D. (2020). Predicting the Landscape of Recombination Using Deep Learning. *Molecular Biology and Evolution*, *37*(6), 1790–1808. https://doi.org/10.1093/molbev/msaa038

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., et al. (2016). 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell*, *166*(2), 481–491. https://doi.org/10.1016/j.cell.2016.05.063

Arenas, M. (2013). The importance and application of the ancestral recombination graph. *Frontiers in Genetics*, *4*. https://doi.org/10.3389/fgene.2013.00206

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., ... National Eye Institute, N. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Bandelt, H.-J., Forster, P., & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, *16*(1), 37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., ... Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3), iyab229. https://doi.org/10.1093/genetics/iyab229

Benegas, G., Batra, S. S., & Song, Y. S. (2023). Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, *120*(44), e2311219120. https://doi.org/10.1073/pnas.2311219120

Berg, J. J., & Coop, G. (2014). A population genetic signal of polygenic adaptation. *PLoS genetics*, *10*(8), e1004412. https://doi.org/10.1371/journal.pgen.1004412

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., & Hirschhorn, J. N. (2004). Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *The American Journal of Human Genetics*, *74*(6), 1111–1120. https://doi.org/10.1086/421051

Blischak, P. D., Barker, M. S., & Gutenkunst, R. N. (2021). Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks. *Molecular Ecology Resources*, *21*(8), 2676–2688. https://doi.org/10.1111/1755-0998.13355

Blount, Z. D., Lenski, R. E., & Losos, J. B. (2018). Contingency and determinism in evolution: Replaying life's tape. *Science*, *362*(6415), eaam5979. https://doi.org/10.1126/science.aam5979

Boitard, S., Schlotterer, C., & Futschik, A. (2009). Detecting selective sweeps: A new approach based on hidden markov models. *Genetics*, *181*(4), 1567–1578. https://doi.org/10.1534/genetics.108.100032

Bourgeois, Y. X., Delahaie, B., Gautier, M., Lhuillier, E., Malé, P.-J. G., Bertrand, J. A., Cornuault, J., Wakamatsu, K., Bouchez, O., Mould, C., et al. (2017). A novel locus on chromosome 1 underlies the evolution of a melanic plumage polymorphism in a wild songbird. *Royal Society open science*, *4*(2), 160805. https://doi.org/10.1098/rsos.160805

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R., Clark, A. G., & Bustamante, C. D. (2008). Assessing the Evolutionary Impact of Amino Acid Mu-

tations in the Human Genome. *PLOS Genetics*, *4*(5), e1000083. https://doi.org/10.1371/journal.pgen.1000083

Brandt, D. Y., Wei, X., Deng, Y., Vaughn, A. H., & Nielsen, R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, *221*(1), iyac044. https://doi.org/10.1093/genetics/iyac044

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Burger, K. E., Pfaffelhuber, P., & Baumdicker, F. (2022). Neural networks for self-adjusting mutation rate estimation when the recombination rate is unknown. *PLOS Computational Biology*, *18*(8), e1010407. https://doi.org/10.1371/journal.pcbi.1010407

Caldas, I. V., Clark, A. G., & Messer, P. W. (2022). Inference of selective sweep parameters through supervised learning. https://doi.org/10.1101/2022.07.19.500702

Campagna, L., Gronau, I., Silveira, L. F., Siepel, A., & Lovette, I. J. (2015). Distinguishing noise from signal in patterns of genomic divergence in a highly polymorphic avian radiation. *Molecular Ecology*, *24*(16), 4238–4251. https://doi.org/10.1111/mec.13314

Campagna, L., Mo, Z., Siepel, A., & Uy, J. A. C. (2022). Selective sweeps on different pigmentation genes mediate convergent evolution of island melanism in two incipient bird species. *PLOS Genetics*, *18*(11), e1010474. https://doi.org/10.1371/journal.pgen.1010474

Campagna, L., Repenning, M., Silveira, L. F., Fontana, C. S., Tubaro, P. L., & Lovette, I. J. (2017). Repeated divergent selection on pigmentation genes in a rapid finch radiation. *Science Advances*, *3*(5), e1602404. https://doi.org/10.1126/sciadv.1602404

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S., & Yandell, M. (2008). Maker: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research*, *18*(1), 188–196. https://doi.org/10.1101/gr.6743907

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, *381*(6664), eadg7492. https://doi.org/10.1126/science.adg7492

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In

I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf

Cochran, K., Srivastava, D., Shrikumar, A., Balsubramani, A., Hardison, R. C., Kundaje, A., & Mahony, S. (2022). Domain-adaptive neural networks improve cross-species prediction of transcription factor binding. *Genome Research*, *32*(3), 512–523. https://doi.org/10.1101/gr.275394.121

Conte, G. L., Arnegard, M. E., Peichel, C. L., & Schluter, D. (2012). The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1749), 5039–5047. https://doi.org/10.1098/rspb.2012.2146

Cooper, E. A., & Uy, J. A. C. (2017). Genomic evidence for convergent evolution of a key trait underlying divergence in island birds. *Molecular Ecology*, *26*(14), 3760–3774. https://doi.org/10.1111/mec.14116

Csurka, G. (2017). A Comprehensive Survey on Domain Adaptation for Visual Applications. In G. Csurka (Ed.), *Domain Adaptation in Computer Vision Applications* (pp. 1–35). Springer International Publishing. https://doi.org/10.1007/978-3-319-58347-1_1

Currat, M., Trabuchet, G., Rees, D., Perrin, P., Harding, R. M., Clegg, J. B., Langaney, A., & Excoffier, L. (2002). Molecular Analysis of the β-Globin Gene Cluster in the Niokholo Mandenka Population Reveals a Recent Origin of the βS Senegal Mutation. *The American Journal of Human Genetics*, *70*(1), 207–223. https://doi.org/10.1086/338304

Dai, W., Yang, Q., Xue, G.-R., & Yu, Y. (2007). Boosting for transfer learning. *Proceedings of the 24th international conference on Machine learning*, 193–200. https://doi.org/10.1145/1273496.1273521

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., et al. (2011). The variant call format and vcftools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Daumé III, H. (2009). Frustratingly Easy Domain Adaptation. https://doi.org/10.48550/arXiv.0907.1815

Ducrest, A.-L., Keller, L., & Roulin, A. (2008). Pleiotropy in the melanocortin system, coloration and behavioural syndromes. *Trends in ecology & evolution*, *23*(9), 502–510. https://doi.org/10.1016/j.tree.2008.06.001

Edge, M. D., & Coop, G. (2019). Reconstructing the history of polygenic scores using coalescent trees. *Genetics*, *211*(1), 235–262. https://doi.org/10.1534/genetics.118.301687

Eriksson, N., Macpherson, J. M., Tung, J. Y., Hon, L. S., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe'er, I., & Mountain, J. (2010). Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits. *PLOS Genetics*, *6*(6), e1000993. https://doi.org/10.1371/journal.pgen.1000993

Fay, J. C., & Wu, C.-I. (2000). Hitchhiking Under Positive Darwinian Selection. *Genetics*, *155*(3), 1405–1413. Retrieved June 25, 2019, from https://www.genetics.org/content/155/3/1405

Felsenstein, J., & Churchill, G. A. (1996). A hidden markov model approach to variation among sites in rate of evolution. *Molecular biology and evolution*, *13*(1), 93–104. https://doi.org/10.1093/oxfordjournals.molbev.a025575

Fernando, B., Habrard, A., Sebban, M., & Tuytelaars, T. (2013). Unsupervised Visual Domain Adaptation Using Subspace Alignment. *2013 IEEE International Conference on Computer Vision*, 2960–2967. https://doi.org/10.1109/ICCV.2013.368

Flagel, L., Brandvain, Y., & Schrider, D. R. (2019). The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference (Y. Kim, Ed.). *Molecular Biology and Evolution*, *36*(2), 220–238. https://doi.org/10.1093/molbev/msy224

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). Repeatmodeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R. B., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A.-M., Ness, A. R., Ebrahim, S., ... McCarthy, M. I. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science (New York, N.Y.)*, *316*(5826), 889–894. https://doi.org/10.1126/science.1141634

Fu, W., & Akey, J. M. (2013). Selection and Adaptation in the Human Genome. *Annual Review of Genomics and Human Genetics*, *14*(1), 467–489. https://doi.org/10.1146/annurev-genom-091212-153509

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., & Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolu-

tion. *PLoS genetics*, *7*(11), e1002355. https://doi.org/10.1371/journal.pgen.1002355

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *International Conference on Machine Learning*, 1050–1059. Retrieved June 10, 2021, from http://proceedings.mlr.press/v48/gal16.html

Ganin, Y., & Lempitsky, V. (2014). Unsupervised Domain Adaptation by Backpropagation. https://doi.org/10.48550/arXiv.1409.7495

Garcia-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., & Conesa, A. (2012). Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics*, *28*(20), 2678–2679. https://doi.org/10.1093/bioinformatics/bts503

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLoS Genetics*, *11*(2), e1005004. https://doi.org/10.1371/journal.pgen.1005004

Ghifary, M., Kleijn, W. B., Zhang, M., Balduzzi, D., & Li, W. (2016). Deep Reconstruction-Classification Networks for Unsupervised Domain Adaptation. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 597–613). Springer International Publishing. https://doi.org/10.1007/978-3-319-46493-0_36

Gould, S. J. (1989). *Wonderful life: The burgess shale and the nature of history*. WW Norton & Company.

Gower, G., Iáñez Picazo, P., Lindgren, F., & Racimo, F. (2023). Inference of population genetics parameters using discriminator neural networks: An adversarial monte carlo approach. *bioRxiv*, 2023–04. https://doi.org/10.1101/2023.04.27.538386

Gower, G., Picazo, P. I., Fumagalli, M., & Racimo, F. (2021). Detecting adaptive introgression in human evolution using convolutional neural networks. *Elife*, *10*, e64669. https://doi.org/10.7554/eLife.64669

Grabherr, M. G., Russell, P., Meyer, M., Mauceli, E., Alföldi, J., Di Palma, F., & Lindblad-Toh, K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*, *26*(9), 1145–1151. https://doi.org/10.1093/bioinformatics/btq102

Grant, P. R., & Grant, B. R. (2002). Unpredictable evolution in a 30-year study of darwin's finches. *Science*, *296*(5568), 707–711. https://doi.org/10.1126/science.1070315

Griffiths, R. C., & Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology:*

A Journal of Computational Molecular Cell Biology, *3*(4), 479–502. https://doi.org/10.1089/cmb.1996.3.479

Griffiths, R. C., Marjoram, P., Donnelly, P., & Tavaré, S. (1997). Progress in population genetics and human evolution. *Progress in Population Genetics and Human Evolution*, *87*, 257.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, *43*(10), 1031–1034. https://doi.org/10.1038/ng.937

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). Quast: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic acids research*, *41*(13), e129–e129. https://doi.org/10.1093/nar/gkt371

Hahn, M. W. (2018). *Molecular population genetics*. Oxford University Press.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., & Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, *19*(2), 552–566. https://doi.org/10.1111/1755-0998.12968

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, *36*(3), 632–637. https://doi.org/10.1093/molbev/msy228

Han, J., Kraft, P., Nan, H., Guo, Q., Chen, C., Qureshi, A., Hankinson, S. E., Hu, F. B., Duffy, D. L., Zhao, Z. Z., Martin, N. G., Montgomery, G. W., Hayward, N. K., Thomas, G., Hoover, R. N., Chanock, S., & Hunter, D. J. (2008). A Genome-Wide Association Study Identifies Novel Alleles Associated with Hair Color and Skin Pigmentation. *PLOS Genetics*, *4*(5), e1000074. https://doi.org/10.1371/journal.pgen.1000074

Harding, R. M., Healy, E., Ray, A. J., Ellis, N. S., Flanagan, N., Todd, C., Dixon, C., Sajantila, A., Jackson, I. J., Birch-Machin, M. A., & Rees, J. L. (2000). Evidence for Variable Selective Pressures at MC1R. *The American Journal of Human Genetics*, *66*(4), 1351–1361. https://doi.org/10.1086/302863

Hayman, E., Ignatieva, A., & Hein, J. (2023). Recoverability of ancestral recombination graph topologies. *Theoretical Population Biology*, *154*, 27–39. https://doi.org/10.1016/j.tpb.2023.07.004

Hebert, P. D., Ratnasingham, S., & De Waard, J. R. (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related

species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(suppl_1), S96–S99. https://doi.org/10.1098/rsbl.2003.0025

Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, *36*(4), 396–405. https://doi.org/10.1007/BF00182187

Hejase, H. A., Dukler, N., & Siepel, A. (2020). From Summary Statistics to Gene Trees: Methods for Inferring Positive Selection. *Trends in Genetics*, *36*(4), 243–258. https://doi.org/10.1016/j.tig.2019.12.008

Hejase, H. A., Mo, Z., Campagna, L., & Siepel, A. (2022). A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph. *Molecular Biology and Evolution*, *39*(1), msab332. https://doi.org/10.1093/molbev/msab332

Hejase, H. A., Salman-Minkov, A., Campagna, L., Hubisz, M. J., Lovette, I. J., Gronau, I., & Siepel, A. (2020). Genomic islands of differentiation in a rapid avian radiation have been driven by recent selective sweeps. *Proceedings of the National Academy of Sciences*, *117*(48), 30554–30565. https://doi.org/10.1073/pnas.2015987117

Henry, R. J., & Nevo, E. (2014). Exploring natural selection to guide breeding for agriculture. *Plant biotechnology journal*, *12*(6), 655–662. https://doi.org/10.1111/pbi.12215

Hiragaki, T., Inoue-Murayama, M., Miwa, M., Fujiwara, A., Mizutani, M., Minvielle, F., & Ito, S. (2008). Recessive black is allelic to the yellow plumage locus in japanese quail and associated with a frameshift deletion in the asip gene. *Genetics*, *178*(2), 771–775. https://doi.org/10.1534/genetics.107.077040

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoekstra, H. E., Drumm, K. E., & Nachman, M. W. (2004). Ecological genetics of adaptive color polymorphism in pocket mice: Geographic variation in selected and neutral genes. *Evolution*, *58*(6), 1329–1341. https://doi.org/10.1111/j.0014-3820.2004.tb01711.x

Huang, X., Rymbekova, A., Dolgova, O., Lao, O., & Kuhlwilm, M. (2023). Harnessing deep learning for population genetic inference. *Nature Reviews Genetics*, 1–18. https://doi.org/10.1038/s41576-023-00636-3

Hubisz, M. J., & Siepel, A. (2020). Inference of ancestral recombination graphs using argweaver. *Statistical population genomics*, 231–266. https://doi.org/10.1007/978-1-0716-0199-0_10

Hubisz, M. J., Williams, A. L., & Siepel, A. (2020). Mapping gene flow between ancient hominins through demography-aware inference of the ancestral

recombination graph. *PLOS Genetics*, *16*(8), e1008895. https://doi.org/10.1371/journal.pgen.1008895

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, *7*, 1–44. Retrieved August 20, 2019, from https://www.cabdirect.org/cabdirect/abstract/19910191040

Hui, Z., Li, J., Wang, X., & Gao, X. (2021). Learning the non-differentiable optimization for blind super-resolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2093–2102.

Irwin, D. E. (2018). Sex chromosomes and speciation in birds and other zw systems. *Molecular Ecology*, *27*(19), 3831–3851. https://doi.org/10.1111/mec.14537

Isobe, T., Jia, X., Chen, S., He, J., Shi, Y., Liu, J., Lu, H., & Wang, S. (2021). Multi-Target Domain Adaptation With Collaborative Consistency Learning, 8187–8196. Retrieved October 27, 2022, from https://openaccess.thecvf.com/content/CVPR2021/html/Isobe_Multi-Target_Domain_Adaptation_With_Collaborative_Consistency_Learning_CVPR_2021_paper.html

Jacquin, L., Lenouvel, P., Haussy, C., Ducatez, S., & Gasparini, J. (2011). Melanin-based coloration is related to parasite intensity and cellular immune response in an urban free living bird: The feral pigeon columba livia. *Journal of Avian Biology*, *42*(1), 11–15. https://doi.org/10.1111/j.1600-048X.2010.05120.x

Jarvis, J. P., Scheinfeldt, L. B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J.-M., Beggs, W., Hoffman, G., Mezey, J., & Tishkoff, S. A. (2012). Patterns of Ancestry, Signatures of Natural Selection, and Genetic Association with Stature in Western African Pygmies. *PLOS Genetics*, *8*(4), e1002641. https://doi.org/10.1371/journal.pgen.1002641

Johri, P., Aquadro, C. F., Beaumont, M., Charlesworth, B., Excoffier, L., Eyre-Walker, A., Keightley, P. D., Lynch, M., McVean, G., Payseur, B. A., Pfeifer, S. P., Stephan, W., & Jensen, J. D. (2022). Recommendations for improving statistical inference in population genomics. *PLOS Biology*, *20*(5), e3001669. https://doi.org/10.1371/journal.pbio.3001669

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quan-

tified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Karlsson, E. K., Kwiatkowski, D. P., & Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nature Reviews Genetics*, *15*(6), 379–393. https://doi.org/10.1038/nrg3734

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, *12*(5), e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Kelleher, J., Thornton, K. R., Ashander, J., & Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLoS computational biology*, *14*(11), e1006581. https://doi.org/10.1371/journal.pcbi.1006581

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., & McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, *51*(9), 1330–1338. https://doi.org/10.1038/s41588-019-0483-y

Kelley, J. L., & Swanson, W. J. (2008). Positive Selection in the Human Genome: From Genome Scans to Biological Significance. *Annual Review of Genomics and Human Genetics*, *9*(1), 143–160. https://doi.org/10.1146/annurev.genom.9.081307.164411

Kelly, J. K. (1997). A Test of Neutrality Based on Interlocus Associations. *Genetics*, *146*(3), 1197–1206. Retrieved June 24, 2019, from https://www.genetics.org/content/146/3/1197

Kenny, E. E., Timpson, N. J., Sikora, M., Yee, M.-C., Moreno-Estrada, A., Eng, C., Huntsman, S., Burchard, E. G., Stoneking, M., Bustamante, C. D., & Myles, S. (2012). Melanesian blond hair is caused by an amino acid change in TYRP1. *Science (New York, N.Y.)*, *336*(6081), 554. https://doi.org/10.1126/science.1217849

Kern, A. D., & Haussler, D. (2010). A population genetic hidden markov model for detecting genomic regions under selection. *Molecular biology and evolution*, *27*(7), 1673–1685. https://doi.org/10.1093/molbev/msq053

Kern, A. D., & Schrider, D. R. (2016). Discoal: Flexible coalescent simulations with selection. *Bioinformatics*, *32*(24), 3839–3841. https://doi.org/10.1093/bioinformatics/btw556

Kern, A. D., & Schrider, D. R. (2018). diploS/HIC: An Updated Approach to Classifying Selective Sweeps. *G3: Genes, Genomes, Genetics*, *8*(6), 1959–1970. https://doi.org/10.1534/g3.118.200262

Khomutov, E., Arzymatov, K., & Shchur, V. (2021). Deep learning based methods for estimating distribution of coalescence rates from genome-wide data. *Journal of Physics: Conference Series*, *1740*(1), 012031. https://doi.org/10.1088/1742-6596/1740/1/012031

Kijas, J., Wales, R., Törnsten, A., Chardon, P., Moller, M., & Andersson, L. (1998). Melanocortin receptor 1 (mc1r) mutations and coat color in pigs. *Genetics*, *150*(3), 1177–1185. https://doi.org/10.1093/genetics/150.3.1177

Kim, J., Rosenberg, N. A., & Palacios, J. A. (2020). Distance metrics for ranked evolutionary trees. *Proceedings of the National Academy of Sciences*, *117*(46), 28876–28886. https://doi.org/10.1073/pnas.1922851117

Kim, Y., & Nielsen, R. (2004). Linkage Disequilibrium as a Signature of Selective Sweeps. *Genetics*, *167*(3), 1513–1524. https://doi.org/10.1534/genetics.103.025387

Kingman, J. F. C. (1982a). The coalescent. *Stochastic processes and their applications*, *13*(3), 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Kingman, J. F. C. (1982b). On the genealogy of large populations. *Journal of applied probability*, *19*(A), 27–43. https://doi.org/10.2307/3213548

Kingsley, E. P., Manceau, M., Wiley, C. D., & Hoekstra, H. E. (2009). Melanism in peromyscus is caused by independent mutations in agouti. *PloS one*, *4*(7), e6435. https://doi.org/10.1371/journal.pone.0006435

Kokko, H., Brooks, R., McNamara, J. M., & Houston, A. I. (2002). The sexual selection continuum. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *269*(1498), 1331–1340. https://doi.org/10.1098/rspb.2002.2020

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G. B., Jonasdottir, A., Gylfason, A., Kristinsson, K. T., Gudjonsson, S. A., Frigge, M. L., Helgason, A., Thorsteinsdottir, U., & Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, *467*(7319), 1099–1103. https://doi.org/10.1038/nature09525

Korfmann, K., Gaggiotti, O. E., & Fumagalli, M. (2023). Deep Learning in Population Genetics. *Genome Biology and Evolution*, *15*(2), evad008. https://doi.org/10.1093/gbe/evad008

Kuhner, M. K. (2006). LAMARC 2.0: Maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, *22*(6), 768–770. https://doi.org/10.1093/bioinformatics/btk051

Lachance, J., Vernot, B., Elbers, C. C., Ferwerda, B., Froment, A., Bodo, J.-M., Lema, G., Fu, W., Nyambo, T. B., Rebbeck, T. R., Zhang, K., Akey, J. M., & Tishkoff, S. A. (2012). Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. *Cell*, *150*(3), 457–469. https://doi.org/10.1016/j.cell.2012.07.009

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lauterbur, M. E., Cavassim, M. I. A., Gladstein, A. L., Gower, G., Pope, N. S., Tsambos, G., Adrion, J., Belsare, S., Biddanda, A., Caudill, V., Cury, J., Echevarria, I., Haller, B. C., Hasan, A. R., Huang, X., Iasi, L. N. M., Noskova, E., Obšteter, J., Pavinato, V. A. C., … Gronau, I. (2022). Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. https://doi.org/10.1101/2022.10.29.514266

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Leigh, J. W., & Bryant, D. (2015). Popart: Full-feature software for haplotype network construction. *Methods in ecology and evolution*, *6*(9), 1110–1116. https://doi.org/10.1111/2041-210X.12410

Lewanski, A. L., Grundler, M. C., & Bradburd, G. S. (2023). The era of the arg: An empiricist's guide to ancestral recombination graphs. https://doi.org/10.48550/arXiv.2310.12070

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, 1. G. P. D. P. (2009). The sequence alignment/map format and samtools. *bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, *165*(4), 2213–2233. https://doi.org/10.1093/genetics/165.4.2213

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Liu, F., Wollstein, A., Hysi, P. G., Ankra-Badu, G. A., Spector, T. D., Park, D., Zhu, G., Larsson, M., Duffy, D. L., Montgomery, G. W., Mackey, D. A., Walsh, S., Lao, O., Hofman, A., Rivadeneira, F., Vingerling, J. R., Uitterlinden, A. G., Martin, N. G., Hammond, C. J., & Kayser, M. (2010). Digital Quantification of Human Eye Color Highlights Genetic Association of Three New Loci. *PLOS Genetics*, *6*(5), e1000934. https://doi.org/10.1371/journal.pgen.1000934

Liu, M.-Y., & Tuzel, O. (2016). Coupled Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, *29*. Retrieved October 28, 2022, from https://papers.nips.cc/paper/2016/hash/502e4a16930e414107ee22b6198c578f-Abstract.html

Lu, D., Vage, D. I., & Cone, R. D. (1998). A ligand-mimetic model for constitutive activation of the melanocortin-1 receptor. *Molecular Endocrinology*, *12*(4), 592–604. https://doi.org/10.1210/mend.12.4.0091

Lyssenko, V., Lupi, R., Marchetti, P., Guerra, S. D., Orho-Melander, M., Almgren, P., Sjögren, M., Ling, C., Eriksson, K.-F., Lethagen, Å.-L., Mancarella, R., Berglund, G., Tuomi, T., Nilsson, P., Prato, S. D., & Groop, L. (2007). Mechanisms by which common variants in the *TCF7L2* gene increase risk of type 2 diabetes. *The Journal of Clinical Investigation*, *117*(8), 2155–2163. https://doi.org/10.1172/JCI30706

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150.

Mahler, D. L., Ingram, T., Revell, L. J., & Losos, J. B. (2013). Exceptional convergence on the macroevolutionary landscape in island lizard radiations. *Science*, *341*(6143), 292–295. https://doi.org/10.1126/science.1232392

Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J. D., Chen, D., & Arora, S. (2023). Fine-tuning language models with just forward passes. *arXiv preprint arXiv:2305.17333*. https://doi.org/10.48550/arXiv.2305.17333

Mallick, S., Micco, A., Mah, M., Ringbauer, H., Lazaridis, I., Olalde, I., Patterson, N., & Reich, D. (2023). The allen ancient dna resource (aadr): A curated compendium of ancient human genomes. *bioRxiv*. https://doi.org/10.1101/2023.04.06.535797

Marcondes, R. S., Nations, J. A., Seeholzer, G. F., & Brumfield, R. T. (2021). Rethinking gloger's rule: Climate, light environments, and color in a large family of tropical birds (furnariidae). *The American Naturalist*, *197*(5), 592–606. https://doi.org/10.1086/713386

Marcus, J. H., & Novembre, J. (2017). Visualizing the geography of genetic variants. *Bioinformatics*, *33*(4), 594–595. https://doi.org/10.1093/bioinformatics/btw643

Marjoram, P., & Tavaré, S. (2006). Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics*, *7*(10), 759–770. https://doi.org/10.1038/nrg1961

Mathieson, I. (2020). *Estimating time-varying selection coefficients from time series data of allele frequencies* (tech. rep.). https://doi.org/10.1101/2020.11.17.387761

Mathieson, I., & Scally, A. (2020). What is ancestry? *PLOS Genetics*, *16*(3), e1008624. https://doi.org/10.1371/journal.pgen.1008624

Mathieson, S., & Mathieson, I. (2018). FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology and Evolution*, *35*(12), 2957–2970. https://doi.org/10.1093/molbev/msy180

Mayr, E. (1999). *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press.

Mayr, E., & Diamond, J. M. (2001). *The birds of northern melanesia: Speciation, ecology & biogeography*. Oxford University Press.

Meisner, J., & Albrechtsen, A. (2022). Haplotype and population structure inference using neural networks in whole-genome sequencing data. *Genome Research*, *32*(8), 1542–1552. https://doi.org/10.1101/gr.275994.121

Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, *28*(11), 659–669. https://doi.org/10.1016/j.tree.2013.08.003

Minichiello, M. J., & Durbin, R. (2006). Mapping Trait Loci by Use of Inferred Ancestral Recombination Graphs. *The American Journal of Human Genetics*, *79*(5), 910–922. https://doi.org/10.1086/508901

Montserrat, D. M., Bustamante, C., & Ioannidis, A. (2020). Lai-net: Local-ancestry inference with neural networks. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1314–1318. https://doi.org/10.1109/ICASSP40776.2020.9053662

Mundy, N. I. (2005). A window on the genetics of evolution: Mc1r and plumage colouration in birds. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1573), 1633–1640. https://doi.org/10.1098/rspb.2005.3107

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, *156*(1), 297–304. Retrieved June 10, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461236/

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(10), 5269–

5273. Retrieved June 24, 2019, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC413122/

Nguyen-Meidine, L. T., Belal, A., Kiran, M., Dolz, J., Blais-Morin, L.-A., & Granger, E. (2021). Unsupervised Multi-Target Domain Adaptation Through Knowledge Distillation, 1339–1347. Retrieved October 27, 2022, from https://openaccess.thecvf.com/content/WACV2021/html/Le_Thanh_Nguyen-Meidine_Unsupervised_Multi-Target_Domain_Adaptation_Through_Knowledge_Distillation_WACV_2021_paper.html

Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., & Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, *24*(2), 125–137. https://doi.org/10.1038/s41576-022-00532-2

O'Fallon, B. D. (2013). ACG: Rapid inference of population history from recombining nucleotide sequences. *BMC Bioinformatics*, *14*(1), 40. https://doi.org/10.1186/1471-2105-14-40

Ohashi, J., Naka, I., Patarapotikul, J., Hananantachai, H., Brittenham, G., Looareesuwan, S., Clark, A. G., & Tokunaga, K. (2004). Extended Linkage Disequilibrium Surrounding the Hemoglobin E Variant Due to Malarial Selection. *The American Journal of Human Genetics*, *74*(6), 1198–1208. https://doi.org/10.1086/421330

Ohashi, J., Naka, I., & Tsuchiya, N. (2011). The Impact of Natural Selection on an ABCC11 SNP Determining Earwax Type. *Molecular Biology and Evolution*, *28*(1), 849–857. https://doi.org/10.1093/molbev/msq264

Oksanen, J. (2022). *Vegan: Community ecology package*. https://CRAN.R-project.org/package=vegan

Oriol Sabat, B., Mas Montserrat, D., Giro-i-Nieto, X., & Ioannidis, A. G. (2022). Salai-net: Species-agnostic local ancestry inference network. *Bioinformatics*, *38*(Supplement_2), ii27–ii33. https://doi.org/10.1093/bioinformatics/btac464

Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2011). Domain Adaptation via Transfer Component Analysis. *IEEE Transactions on Neural Networks*, *22*(2), 199–210. https://doi.org/10.1109/TNN.2010.2091281

Pavlidis, P., Jensen, J. D., & Stephan, W. (2010). Searching for footprints of positive selection in whole-genome snp data from nonequilibrium populations. *Genetics*, *185*(3), 907–922. https://doi.org/10.1534/genetics.110.116459

Peter, B. M., Huerta-Sanchez, E., & Nielsen, R. (2012). Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLOS Genetics*, *8*(10), e1003011. https://doi.org/10.1371/journal.pgen.1003011

Plummer, M., Best, N., Cowles, K., Vines, K., et al. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R news*, *6*(1), 7–11.

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Müller, I., Baglione, V., Unneberg, P., Wikelski, M., Grabherr, M. G., et al. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, *344*(6190), 1410–1414. https://doi.org/10.1126/science.1253226

Prangle, D. (2018). Summary statistics. In *Handbook of approximate bayesian computation* (pp. 125–152). Chapman; Hall/CRC. https://doi.org/10.48550/arXiv.1512.05633

Price, T. (2007). *Speciation in birds*. Roberts & Co. Publishers.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Racimo, F., Berg, J. J., & Pickrell, J. K. (2018). Detecting polygenic adaptation in admixture graphs. *Genetics*, *208*(4), 1565–1584. https://doi.org/10.1534/genetics.117.300489

Ralph, P., Thornton, K., & Kelleher, J. (2020). Efficiently summarizing relationships in large samples: A general duality between statistics of genealogies and genomes. *Genetics*, *215*(3), 779–797. https://doi.org/10.1534/genetics.120.303253

Rasmussen, M. D., Hubisz, M. J., Gronau, I., & Siepel, A. (2014). Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genetics*, *10*(5), e1004342. https://doi.org/10.1371/journal.pgen.1004342

Ray, D. D., Flagel, L., & Schrider, D. R. (2023). Introunet: Identifying introgressed alleles via semantic segmentation. *bioRxiv*. https://doi.org/10.1101/2023.02.07.527435

Revell, L. J. (2012). Phytools: An r package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution*, (2), 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Riley, R., Mathieson, I., & Mathieson, S. (2023). Interpreting generative adversarial networks to infer natural selection from genetic data. *bioRxiv*, 2023–03. https://doi.org/10.1101/2023.03.07.531546

Roff, D. A. (1994). The evolution of flightlessness: Is history important? *Evolutionary Ecology*, *8*, 639–657. https://doi.org/10.1007/BF01237847

Roy, S., Krivosheev, E., Zhong, Z., Sebe, N., & Ricci, E. (2021). Curriculum Graph Co-Teaching for Multi-Target Domain Adaptation, 5351–5360. Retrieved October 27, 2022, from https://openaccess.thecvf.com/content/CVPR2021/html/Roy_Curriculum_Graph_Co-Teaching_for_Multi-Target_Domain_Adaptation_CVPR_2021_paper.html

Rozantsev, A., Salzmann, M., & Fua, P. (2019). Beyond Sharing Weights for Deep Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(4), 801–814. https://doi.org/10.1109/TPAMI.2018.2814042

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., & Lander, E. S. (2006). Positive Natural Selection in the Human Lineage. *Science*, *312*(5780), 1614–1620. https://doi.org/10.1126/science.1124309

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., & The International HapMap Consortium. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250

Sanchez, T., Cury, J., Charpiat, G., & Jay, F. (2021). Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation. *Molecular Ecology Resources*, *21*(8), 2645–2660. https://doi.org/10.1111/1755-0998.13224

Scally, A., & Durbin, R. (2012). Revising the human mutation rate: Implications for understanding human evolution. *Nature Reviews Genetics*, *13*(10), 745–753. https://doi.org/10.1038/nrg3295

Schrider, D. R., & Kern, A. D. (2016). S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLOS Genetics*, *12*(3), e1005928. https://doi.org/10.1371/journal.pgen.1005928

Schrider, D. R., & Kern, A. D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*, *34*(4), 301–312. https://doi.org/10.1016/j.tig.2017.12.005

Schrider, D. R., Mendes, F. K., Hahn, M. W., & Kern, A. D. (2015). Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics*, *200*(1), 267–284. https://doi.org/10.1534/genetics.115.174912

Schubert, M., Lindgreen, S., & Orlando, L. (2016). Adapterremoval v2: Rapid adapter trimming, identification, and read merging. *BMC research notes*, *9*(1), 1–7. https://doi.org/10.1186/s13104-016-1900-2

Semenov, G. A., Linck, E., Enbody, E. D., Harris, R. B., Khaydarov, D. R., Alström, P., Andersson, L., & Taylor, S. A. (2021). Asymmetric introgression reveals the genetic architecture of a plumage trait. *Nature Communications*, *12*(1), 1019. https://doi.org/10.1038/s41467-021-21340-y

Sheehan, S., & Song, Y. S. (2016). Deep Learning for Population Genetic Inference. *PLOS Computational Biology*, *12*(3), e1004845. https://doi.org/10.1371/journal.pcbi.1004845

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*(2), 227–244. https://doi.org/10.1016/S0378-3758(00)00115-4

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). Busco: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smeds, L., Qvarnström, A., & Ellegren, H. (2016). Direct estimate of the rate of germline mutation in a bird. *Genome research*, *26*(9), 1211–1218. https://doi.org/10.1101/gr.204669.116

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *International conference on machine learning*, 2256–2265.

Song, Y. S., & Hein, J. (2005). Constructing Minimal Ancestral Recombination Graphs. *Journal of Computational Biology*, *12*(2), 147–169. https://doi.org/10.1089/cmb.2005.12.147

Spall, J. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, *37*(3), 332–341. https://doi.org/10.1109/9.119632

Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, *51*(9), 1321–1329. https://doi.org/10.1038/s41588-019-0484-x

Spellicy, C. J., Harding, M. J., Hamon, S. C., Mahoney, J. J., Reyes, J. A., Kosten, T. R., Newton, T. F., Garza, R. D. L., & Nielsen, D. A. (2014). A variant in ANKK1 modulates acute subjective effects of cocaine: A

preliminary study. *Genes, Brain and Behavior*, *13*(6), 559–564. https://doi.org/10.1111/gbb.12121

Stamatakis, A. (2014). Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stern, A. J., Speidel, L., Zaitlen, N. A., & Nielsen, R. (2021). Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *The American Journal of Human Genetics*, *108*(2), 219–239. https://doi.org/10.1016/j.ajhg.2020.12.005

Stern, A. J., Wilton, P. R., & Nielsen, R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data (R. D. Hernandez, Ed.). *PLOS Genetics*, *15*(9), e1008384. https://doi.org/10.1371/journal.pgen.1008384

Sturm, R. A., Duffy, D. L., Zhao, Z. Z., Leite, F. P. N., Stark, M. S., Hayward, N. K., Martin, N. G., & Montgomery, G. W. (2008). A Single SNP in an Evolutionary Conserved Region within Intron 86 of the HERC2 Gene Determines Human Blue-Brown Eye Color. *The American Journal of Human Genetics*, *82*(2), 424–431. https://doi.org/10.1016/j.ajhg.2007.11.005

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., Jakobsdottir, M., Steinberg, S., Pálsson, S., Jonasson, F., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Aben, K. K., ... Stefansson, K. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genetics*, *39*(12), 1443–1452. https://doi.org/10.1038/ng.2007.13

Sun, B., Feng, J., & Saenko, K. (2016). Return of frustratingly easy domain adaptation. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2058–2065.

Tajima, F. (1983). Evolutionary relationship of dna sequences in finite populations. *Genetics*, *105*(2), 437–460. https://doi.org/10.1093/genetics/105.2.437

Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, *123*(3), 585–595.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., ... ON BEHALF OF THE NHLBI EXOME SEQUENCING PROJECT. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*, *337*(6090), 64–69. https://doi.org/10.1126/science.1219240

Theron, E., Hawkins, K., Bermingham, E., Ricklefs, R. E., & Mundy, N. I. (2001). The molecular basis of an avian plumage polymorphism in the wild: A melanocortin-1-receptor point mutation is perfectly associated with the melanic plumage morph of the bananaquit, coereba flaveola. *Current Biology*, *11*(8), 550–557. https://doi.org/10.1016/S0960-9822(01)00158-0

Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, L., Mathieson, S., & Fumagalli, M. (2019). ImaGene: A convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics*, *20*(S9), 337. https://doi.org/10.1186/s12859-019-2927-x

Turbek, S. P., Browne, M., Giacomo, A. S. D., Kopuchian, C., Hochachka, W. M., Estalles, C., Lijtmaer, D. A., Tubaro, P. L., Silveira, L. F., Lovette, I. J., Safran, R. J., Taylor, S. A., & Campagna, L. (2021). Rapid speciation via the evolution of pre-mating isolation in the Iberá Seedeater. *Science*, *371*(6536). https://doi.org/10.1126/science.abc0256

Turner, S. D. (2018). Qqman: An r package for visualizing gwas results using q-q and manhattan plots. *Journal of Open Source Software*, *3*(25), 731. https://doi.org/10.21105/joss.00731

Uy, J. A. C., Cooper, E. A., Cutie, S., Concannon, M. R., Poelstra, J. W., Moyle, R. G., & Filardi, C. E. (2016). Mutations in different pigmentation genes are associated with parallel melanism in island flycatchers. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1834), 20160731. https://doi.org/10.1098/rspb.2016.0731

Uy, J. A. C., Irwin, D. E., & Webster, M. S. (2018). Behavioral isolation and incipient speciation in birds. *Annual Review of Ecology, Evolution, and Systematics*, *49*, 1–24. https://doi.org/10.1146/annurev-ecolsys-110617-062646

Uy, J. A. C., Moyle, R. G., & Filardi, C. E. (2009). Plumage and song differences mediate species recognition between incipient flycatcher species

of the solomon islands. *Evolution*, *63*(1), 153–164. https://doi.org/10.1111/j.1558-5646.2008.00530.x

Uy, J. A. C., Moyle, R. G., Filardi, C. E., & Cheviron, Z. A. (2009). Difference in plumage color used in species recognition between incipient species is linked to a single amino acid substitution in the melanocortin-1 receptor. *The American Naturalist*, *174*(2), 244–254. https://doi.org/10.1086/600084

Uy, J. A. C., & Safran, R. J. (2013). Variation in the temporal and spatial use of signals and its implications for multimodal communication. *Behavioral Ecology and Sociobiology*, *67*, 1499–1511. https://doi.org/10.1007/s00265-013-1492-y

Uy, J. A. C., & Vargas-Castro, L. E. (2015). Island size predicts the frequency of melanic birds in the color-polymorphic flycatcher monarcha castaneiventris of the solomon islands. *The Auk: Ornithological Advances*, *132*(4), 787–794. https://doi.org/10.1642/AUK-14-284.1

Våge, D. I., Klungland, H., Lu, D., & Cone, R. D. (1999). Molecular and pharmacological characterization of dominant black coat color in sheep. *Mammalian Genome*, *10*(1), 39–43. https://doi.org/10.1007/s003359900939

Van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the cloud: Using docker, gatk, and wdl in terra*. O'Reilly Media.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual review of genetics*, *47*, 97–120. https://doi.org/10.1146/annurev-genet-111212-133526

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome (L. Hurst, Ed.). *PLoS Biology*, *4*(3), e72. https://doi.org/10.1371/journal.pbio.0040072

Wakeley, J. (2005). *Coalescent theory, an introduction*. Roberts; Company.

Walsh, J., Campagna, L., Feeney, W. E., King, J., & Webster, M. S. (2021). Patterns of genetic divergence and demographic history shed light on island-mainland population dynamics and melanic plumage evolution in the white-winged fairywren. *Evolution*, *75*(6), 1348–1360. https://doi.org/10.1111/evo.14185

Wang, L., Zhang, K., & Zhang, L. (2001). Perfect phylogenetic networks with recombination. *Proceedings of the 2001 ACM symposium on Applied computing*, 46–50. https://doi.org/10.1145/372202.372271

Wang, Z., Wang, J., Kourakos, M., Hoang, N., Lee, H. H., Mathieson, I., & Mathieson, S. (2021). Automatic inference of demographic parameters using generative adversarial networks. *Molecular Ecology Resources*, *21*(8), 2689–2705. https://doi.org/10.1111/1755-0998.13386

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*(2), 256–276. https://doi.org/10.1016/0040-5809(75)90020-9

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., Hollfelder, N., Potekhina, I. D., Schier, W., Thomas, M. G., & Burger, J. (2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proceedings of the National Academy of Sciences*, *111*(13), 4832–4837. https://doi.org/10.1073/pnas.1316513111

Wilson, G., & Cook, D. J. (2020). A Survey of Unsupervised Deep Domain Adaptation. *ACM Transactions on Intelligent Systems and Technology*, *11*(5), 51:1–51:46. https://doi.org/10.1145/3400066

Wiuf, C., & Hein, J. (1999). Recombination as a Point Process along Sequences. *Theoretical Population Biology*, *55*(3), 248–259. https://doi.org/10.1006/tpbi.1998.1403

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., & McVean, G. (2022). A unified genealogy of modern and ancient genomes. *Science*, *375*(6583), eabi8264. https://doi.org/10.1126/science.abi8264

Wong, Y., Ignatieva, A., Koskela, J., Gorjanc, G., Wohns, A. W., & Kelleher, J. (2023). A general and efficient representation of ancestral recombination graphs. *bioRxiv*, 2023–11. https://doi.org/10.1101/2023.11.03.565466

Wright, N. A., Steadman, D. W., & Witt, C. C. (2016). Predictable evolution toward flightlessness in volant island birds. *Proceedings of the National Academy of Sciences*, *113*(17), 4765–4770. https://doi.org/10.1073/pnas.1522931113

Yelmen, B., Decelle, A., Boulos, L. L., Szatkownik, A., Furtlehner, C., Charpiat, G., & Jay, F. (2023). Deep convolutional and conditional neural networks for large-scale genomic data generation. *bioRxiv*, 2023–03. https://doi.org/10.1371/journal.pcbi.1011584

Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., Furtlehner, C., Pagani, L., & Jay, F. (2021). Creating artificial human genomes using generative neural networks. *PLoS genetics*, *17*(2), e1009303. https://doi.org/10.1371/journal.pgen.1009303

Yoshiura, K.-i., Kinoshita, A., Ishida, T., Ninokata, A., Ishikawa, T., Kaname, T., Bannai, M., Tokunaga, K., Sonoda, S., Komaki, R., Ihara, M., Saenko, V. A., Alipov, G. K., Sekine, I., Komatsu, K., Takahashi, H., Nakashima, M., Sosonkina, N., Mapendano, C. K., ... Niikawa, N. (2006). A SNP in the ABCC11 gene is the determinant of human earwax type. *Nature Genetics*, *38*(3), 324–330. https://doi.org/10.1038/ng1733

Zhang, Y., Zhu, Q., Shao, Y., Jiang, Y., Ouyang, Y., Zhang, L., & Zhang, W. (2023). Inferring historical introgression with deep learning. *Systematic Biology*, syad033. https://doi.org/10.1093/sysbio/syad033

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of snp data. *Bioinformatics*, *28*(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, *11*(4), 407–409. https://doi.org/10.1038/nmeth.2848

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The masurca genome assembler. *Bioinformatics*, *29*(21), 2669–2677. https://doi.org/10.1093/bioinformatics/btt476

Appendix A

Supplementary material for Chapter 2

The simulation scripts and code for building and training the SIA model are publicly available on GitHub. Supplementary figures are included in this appendix.

Supplemental Figures

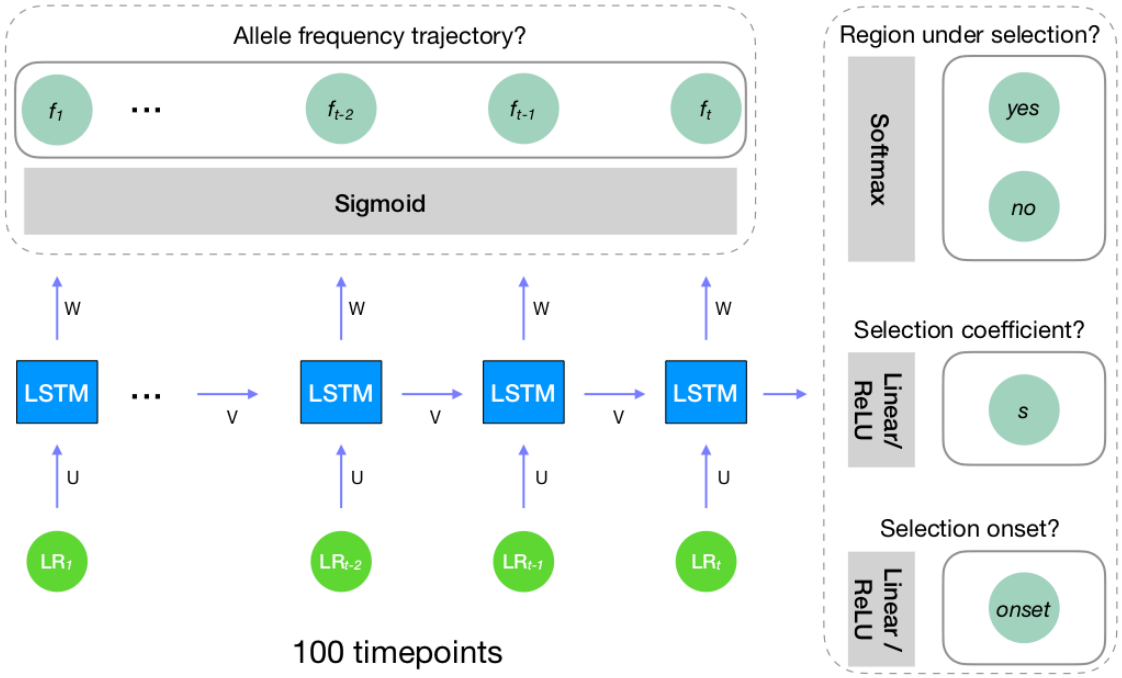


Figure S1: Overview of the deep learning architecture. A form of Recurrent Neural Networks called Long-Short Term Memory (LSTM) was used for sweep prediction. LSTMs are designed to handle the temporal nature of our feature set and account for long term dependencies. Our model has 100 timepoints with the final target output differing in terms of the task at hand (i.e. classification or regression task). For the classification task, the final target output is a label for a binary classification problem predicting whether a region is under selection or neutrality. For the regression task, the final target output is a continuous value, representing the selection coefficient or the selection onset. We also took a many-to-many approach to model the allele frequency trajectory for the site under selection.

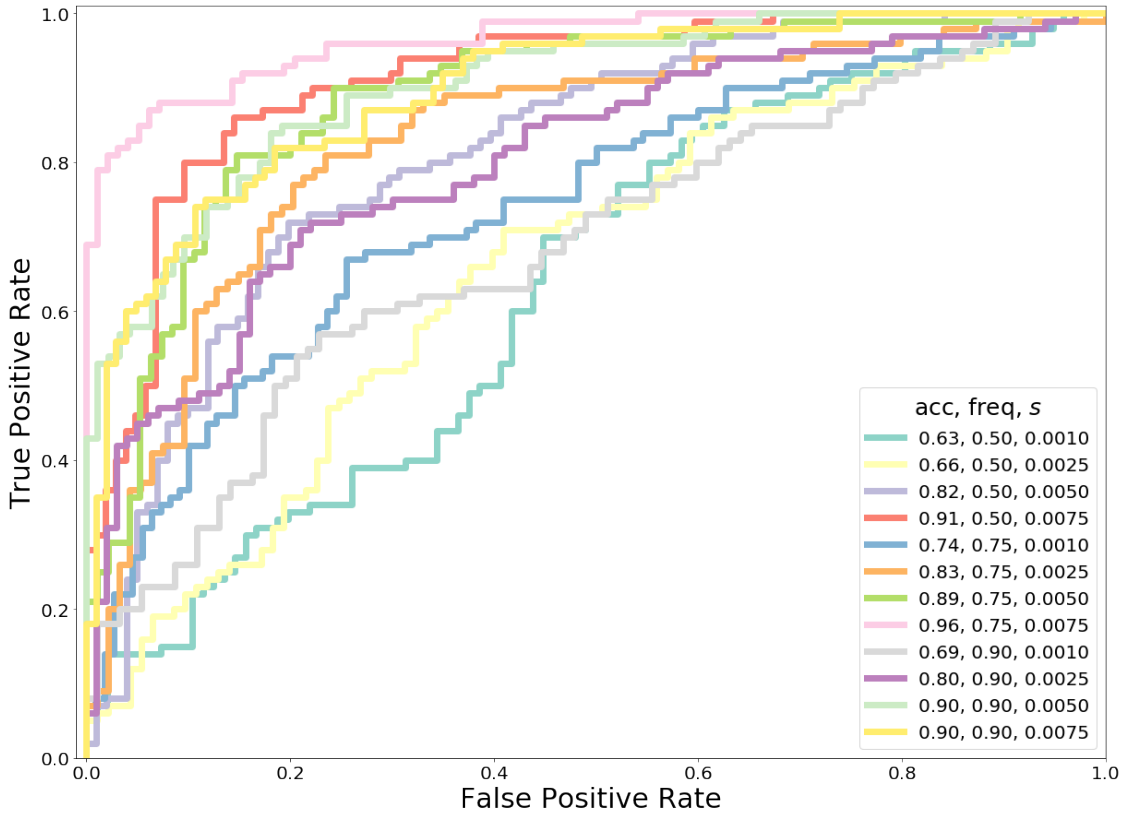


Figure S2: The impact of selection on the performance of SIA on the *S. hypoxantha* population. Data was simulated under a variety of selection regimes and segregating frequencies for the beneficial mutation under selection (shown in the legend under freq and *s*). The prediction task involves two classes: neutral versus soft sweep. SIA was tested on a set of 200 regions per ROC curve (100 per class), and the receiver operating characteristic (ROC) curve records the true positive rate (TPR) as a function of the false positive rate (FPR). The curve associated with the binary prediction task (neutral vs. soft sweep) is obtained by varying the prediction threshold from 0 to 1 and recording for each threshold the number of regions correctly assigned (TPs) and misassigned (FPs) (with prediction probability above the threshold). The performance of SIA was evaluated based on the area under its ROC curve, or AUROC (shown in the legend under acc). We report SIA's AUROC as an average across 200 replicate datasets for each ROC curve.

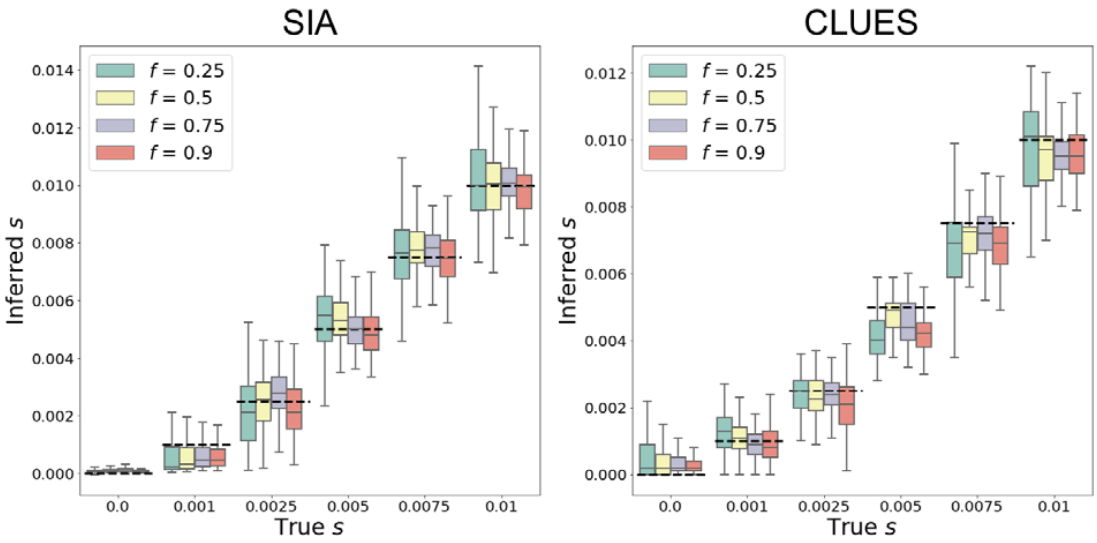


Figure S3: Predictions of selection coefficient on simulations using true gene trees. Results are binned by segregating frequency for each selection regime. Each model condition (i.e. box plot) represents a set of 100 replicates. Figure layout and description are otherwise similar to **Figure 3A**.

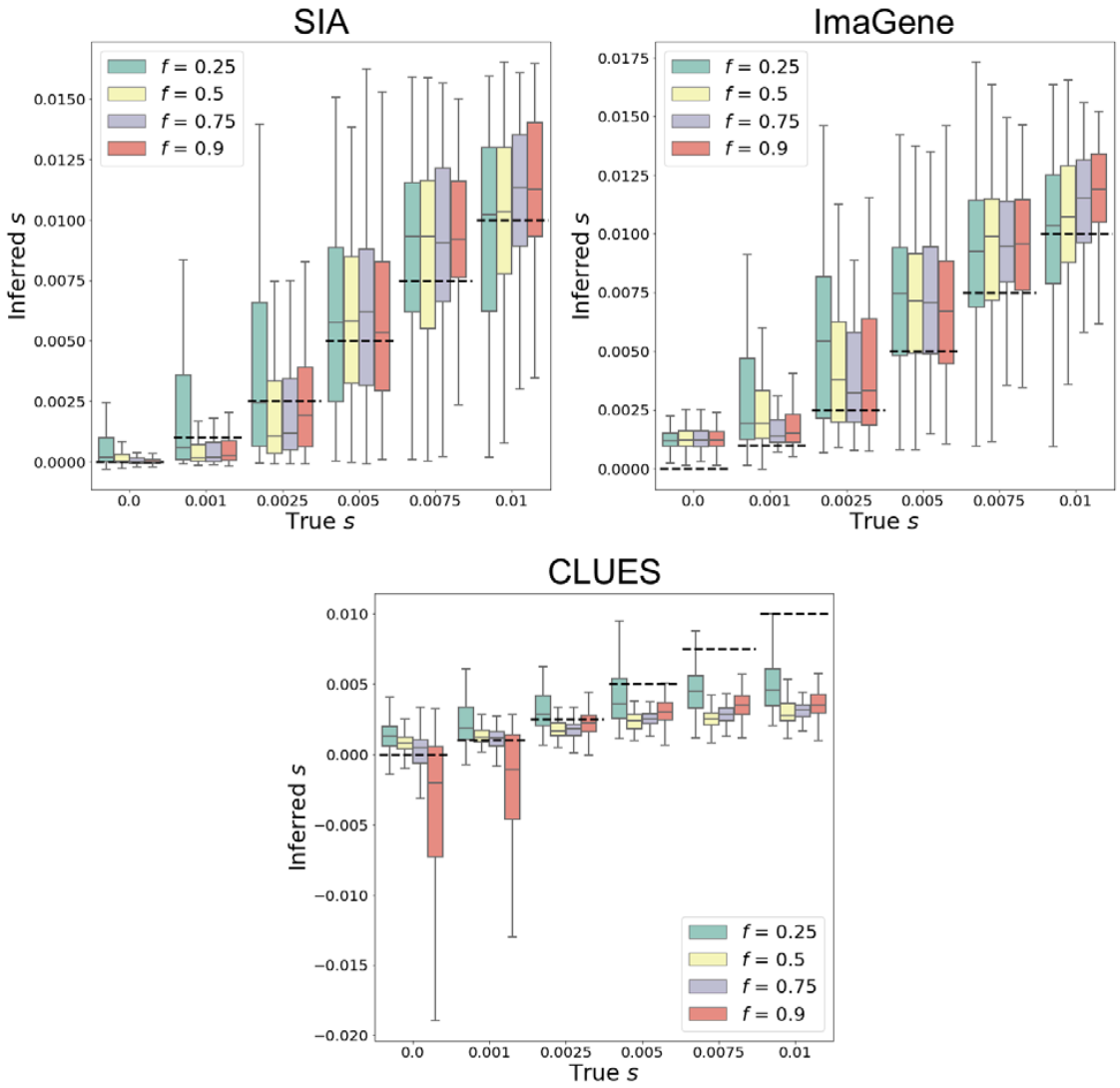


Figure S4: Predictions of selection coefficient on simulations using SIA, ImaGene, and CLUES. Results are binned by segregating frequency for each selection regime. Each model condition (i.e. box plot) represents a set of 100 replicates. The simulations are based on the CEU demographic model where inferred genealogies were used as input to SIA and CLUES. Figure layout and description are otherwise similar to **Figure 3A**.

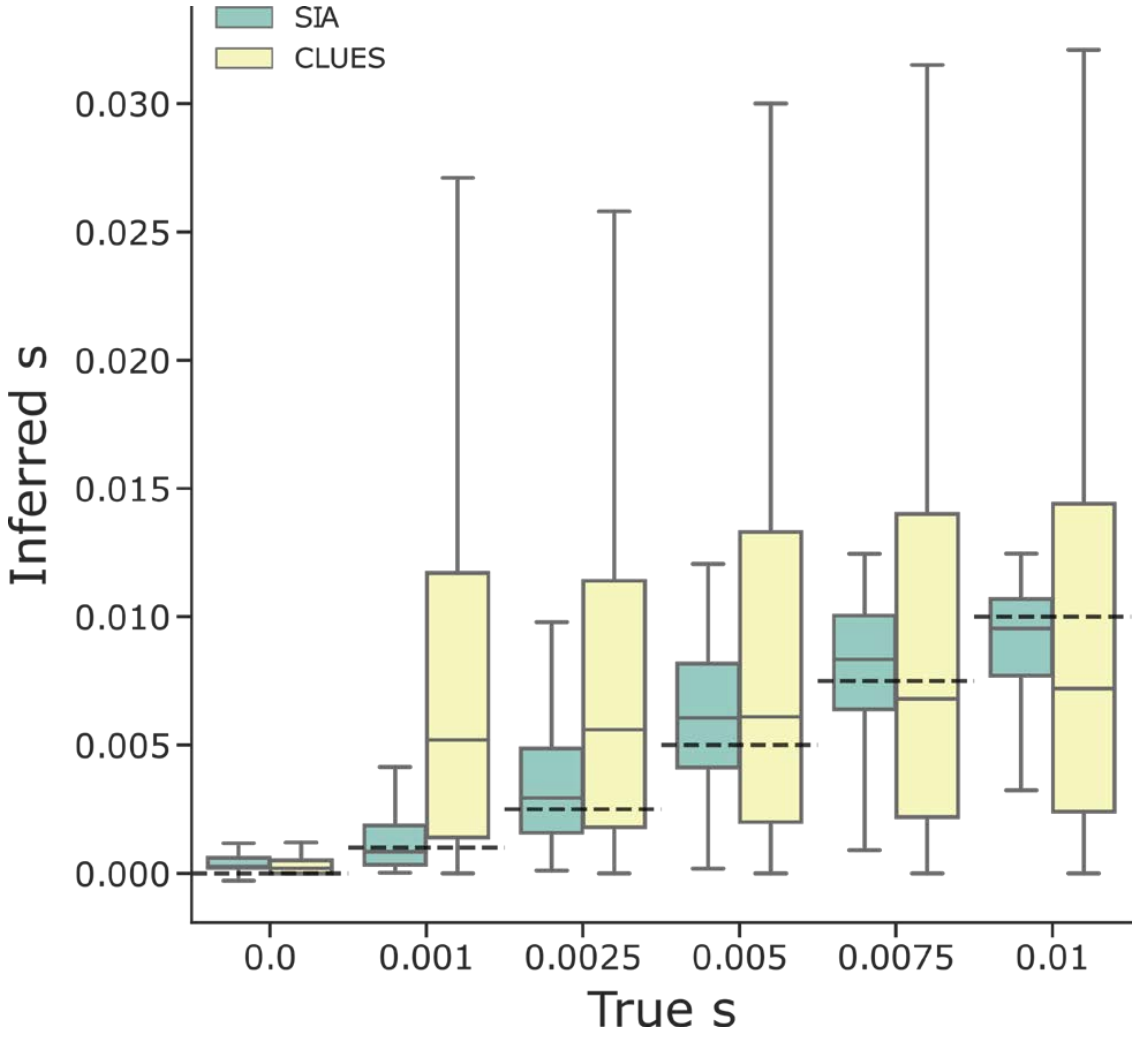


Figure S5: Predictions of selection coefficient on simulated regions using SIA based on inferred genealogies and CLUES with ARGweaver-sampled genealogies. Both methods were evaluated on a test set of 10,000 neutral and 10,000 sweep simulations of 32 haplotypes under a constant-sized demography with N_e=10,000. Figure layout and description are otherwise similar to **Figure 3A.**

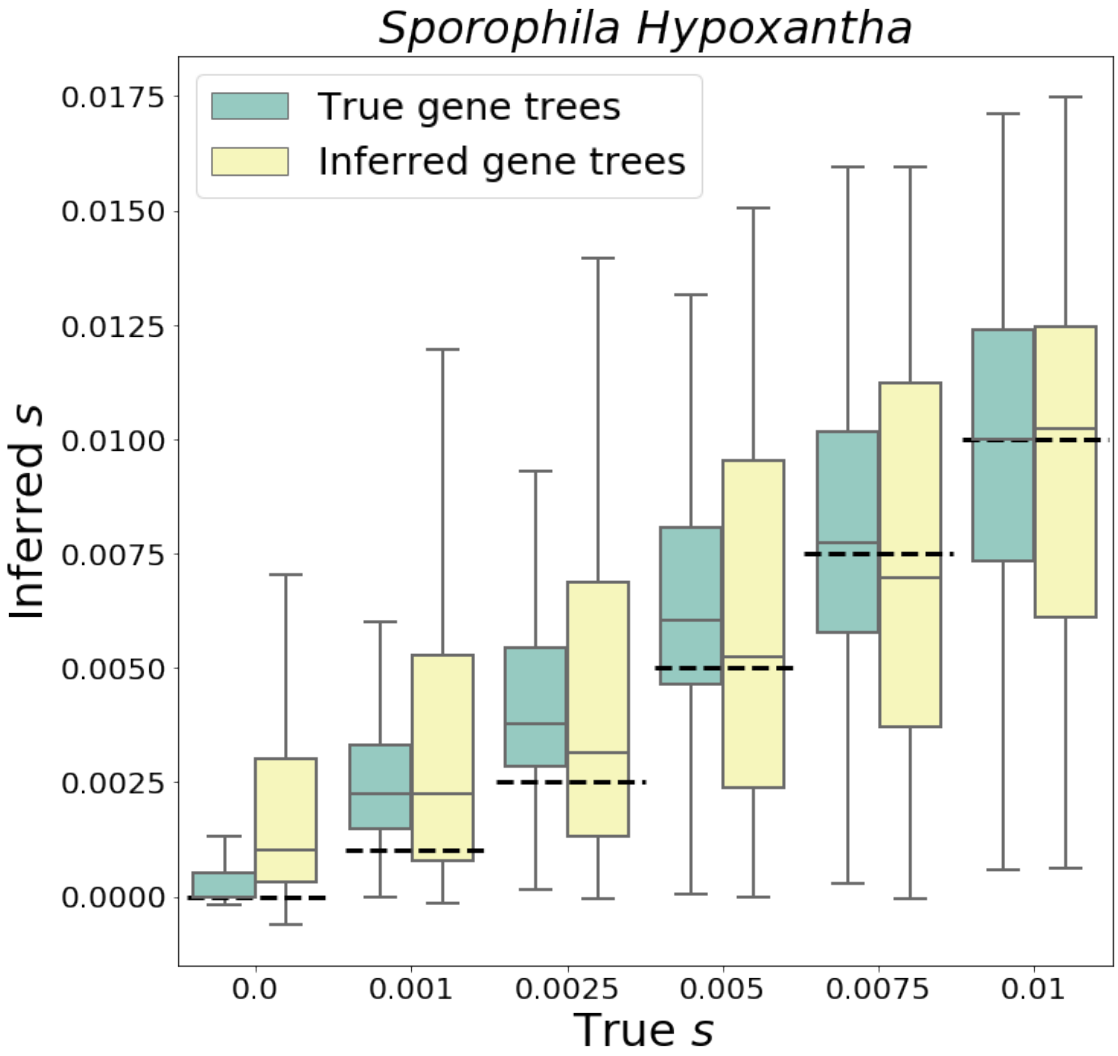


Figure S6: Predictions of selection coefficient on southern capuchino simulations using SIA. The distribution of inferred selection coefficients for SIA on *S. hypoxantha* and each model condition is reported using a box plot. The simulations are based on the capuchinos demographic model where true or inferred genealogies were used as input to SIA. Each model condition (i.e. box plot) represents a set of 400 replicates. Figure layout and description are otherwise similar to **Figure 3A**.

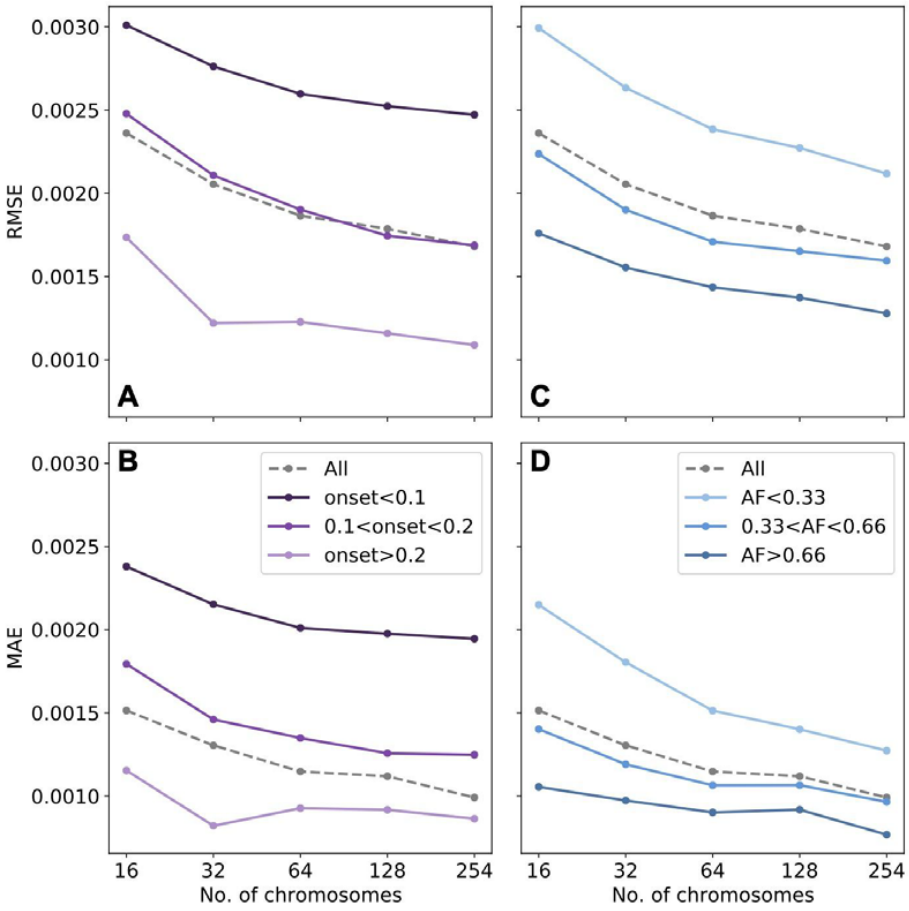


Figure S7: Performance of SIA in selection-coefficient inference with different sample sizes. A separate SIA model was trained using 100,000 neutral and 100,000 sweep simulations (95%-5% train-validation split) under a constant N_e = 10,000 for 16, 32, 64, 128 and 254 haploid genomes. The performance of each model in selection-coefficient inference was evaluated on a test set of 10,000 neutral and 10,000 sweep simulations using root mean square error (RMSE, top) and mean absolute error (MAE, bottom), stratified by time of emergence (in coalescent unit of $4N_e$) of the *de novo* beneficial allele (left), and by the current derived allele frequency (right). Grey dots indicate the overall performance on the entire test set.

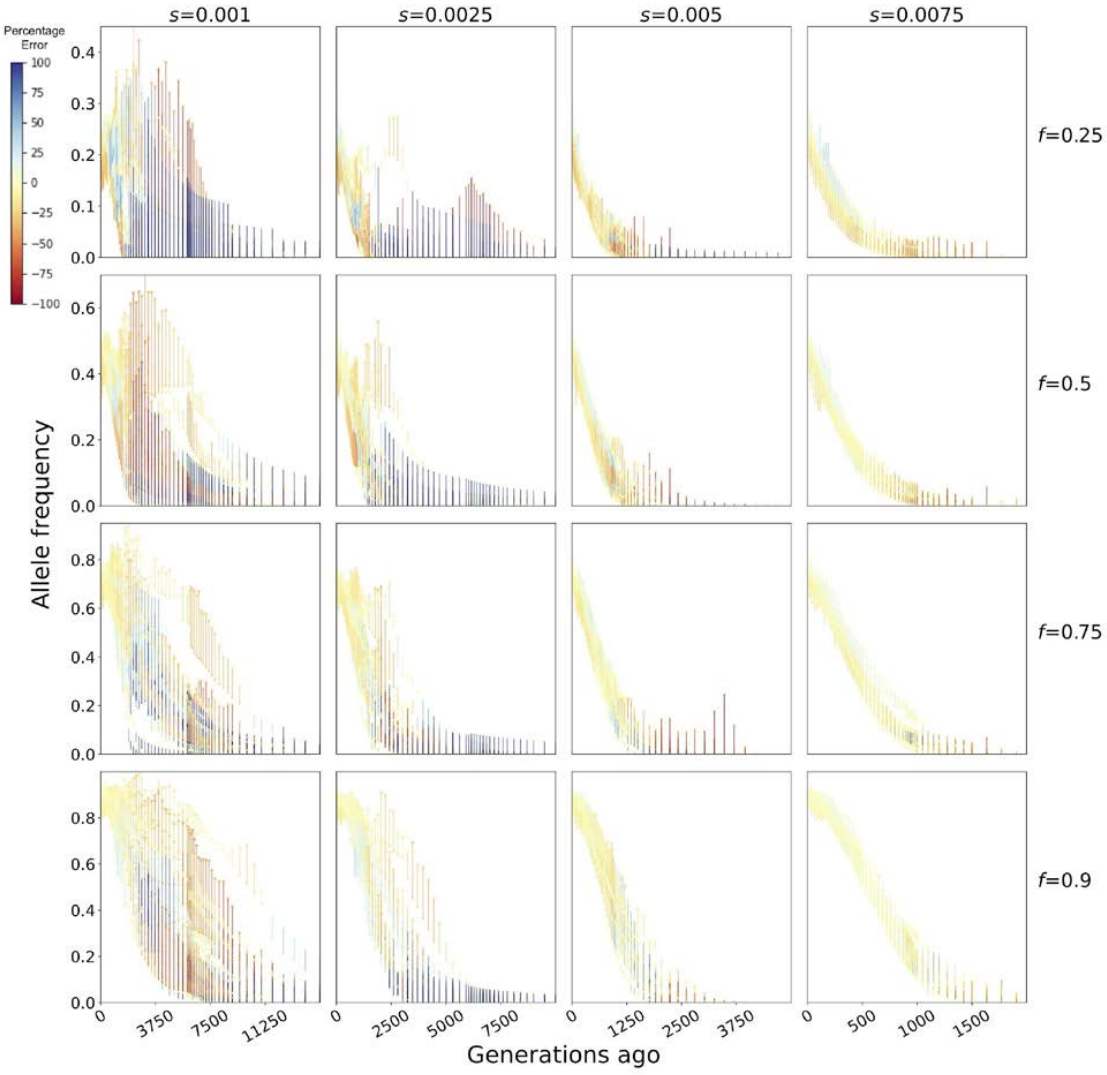


Figure S8: Allele frequency (AF) trajectories inferred with SIA using true genealogies of simulations under the CEU demography. Each panel shows 20 randomly selected examples of AF trajectories for a particular combination of selection coefficient and current AF. For each example, the true and inferred AF at each time point are connected by a vertical line with color scaled to the percentage error.

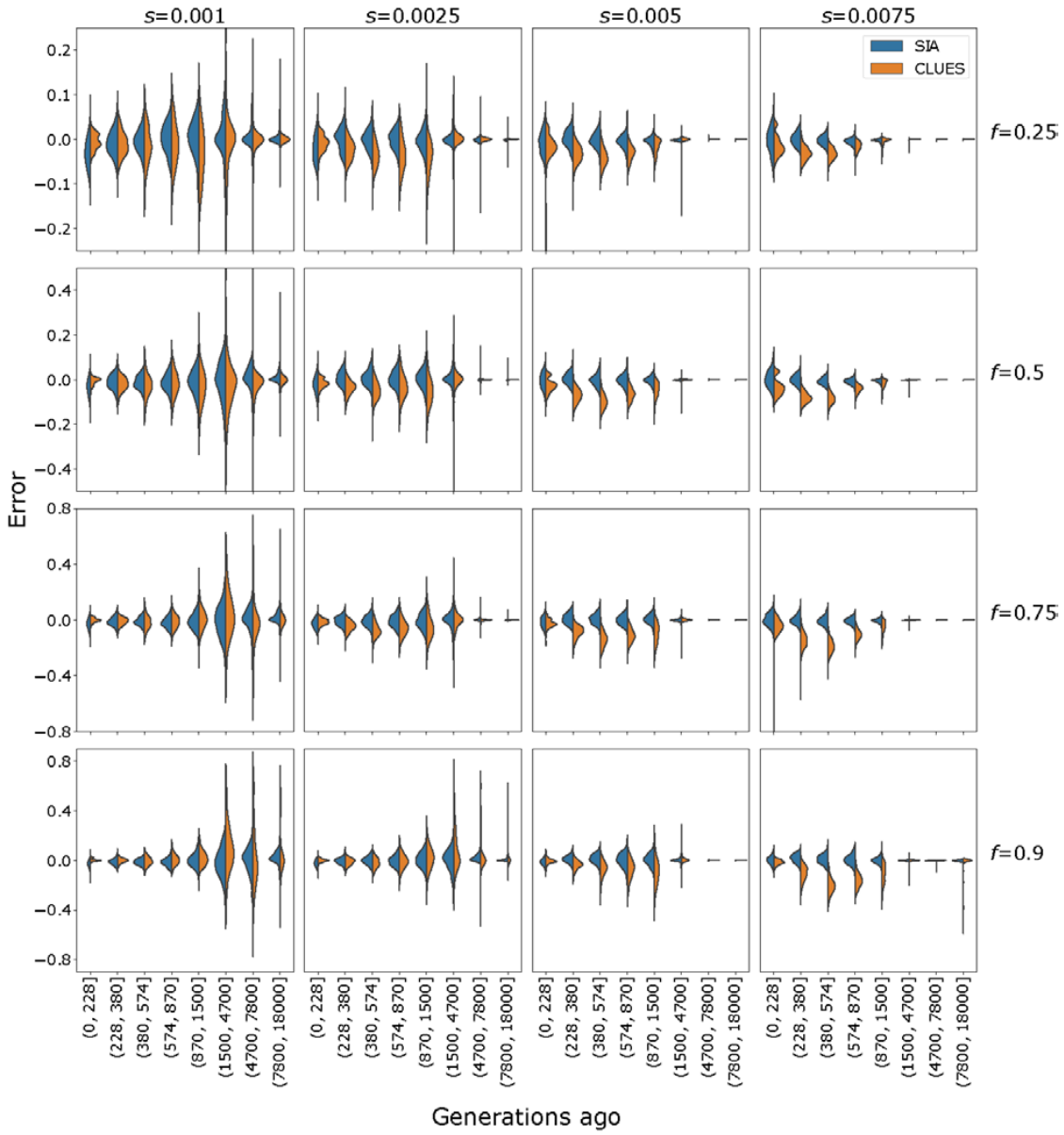


Figure S9: Distribution of error in allele frequency trajectory inference with true genealogies of simulations under the CEU demography. The performance of SIA and CLUES for AF trajectory inference was evaluated with the same set of 100 simulations under each combination of selection coefficient and current AF. Violin plots in each panel show the distribution of absolute error in AF estimation in time point bins indicated on the *x*-axis. Note that the *y*-axes limits for each row of panels with current AF equals *f* were set to be (-*f*, +*f*).

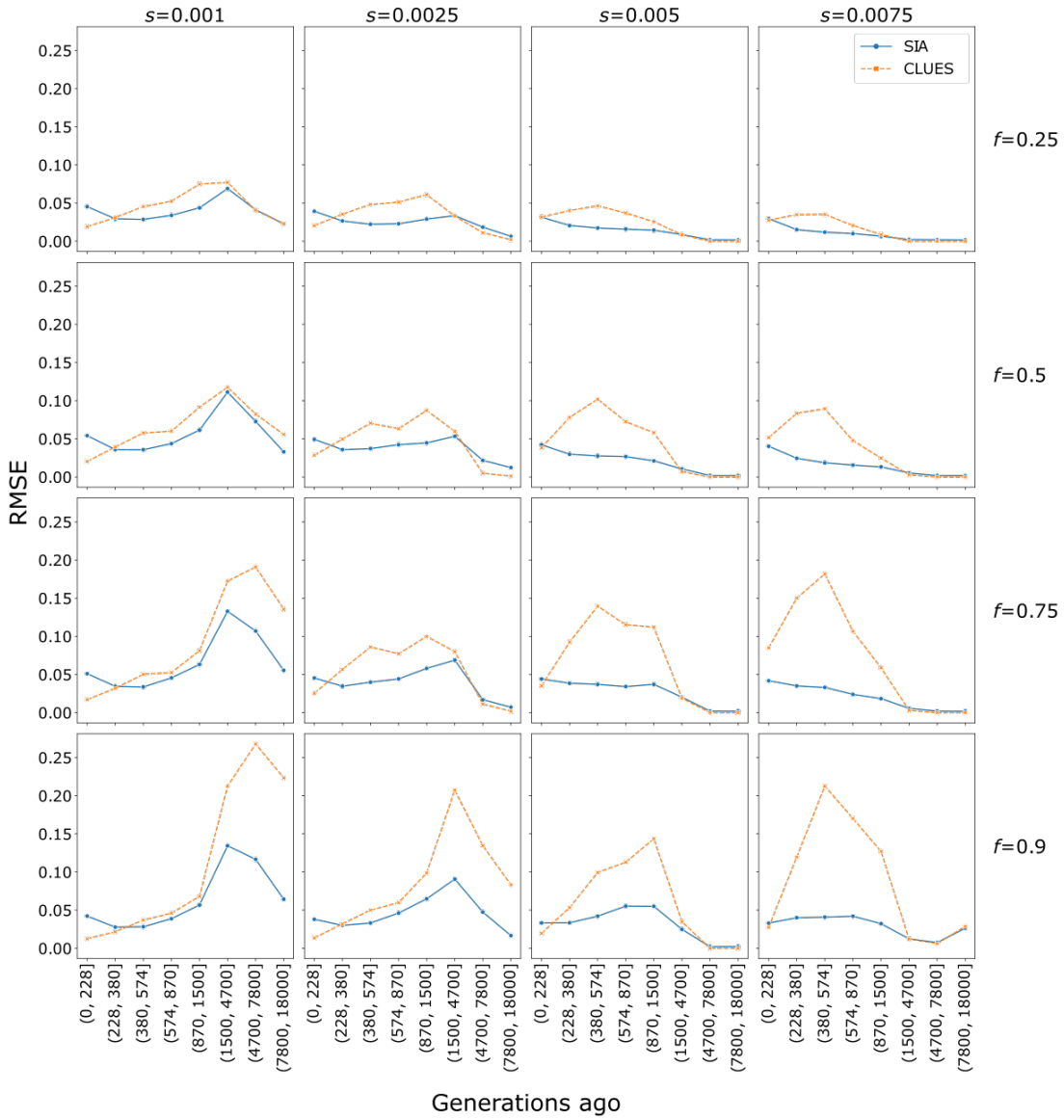
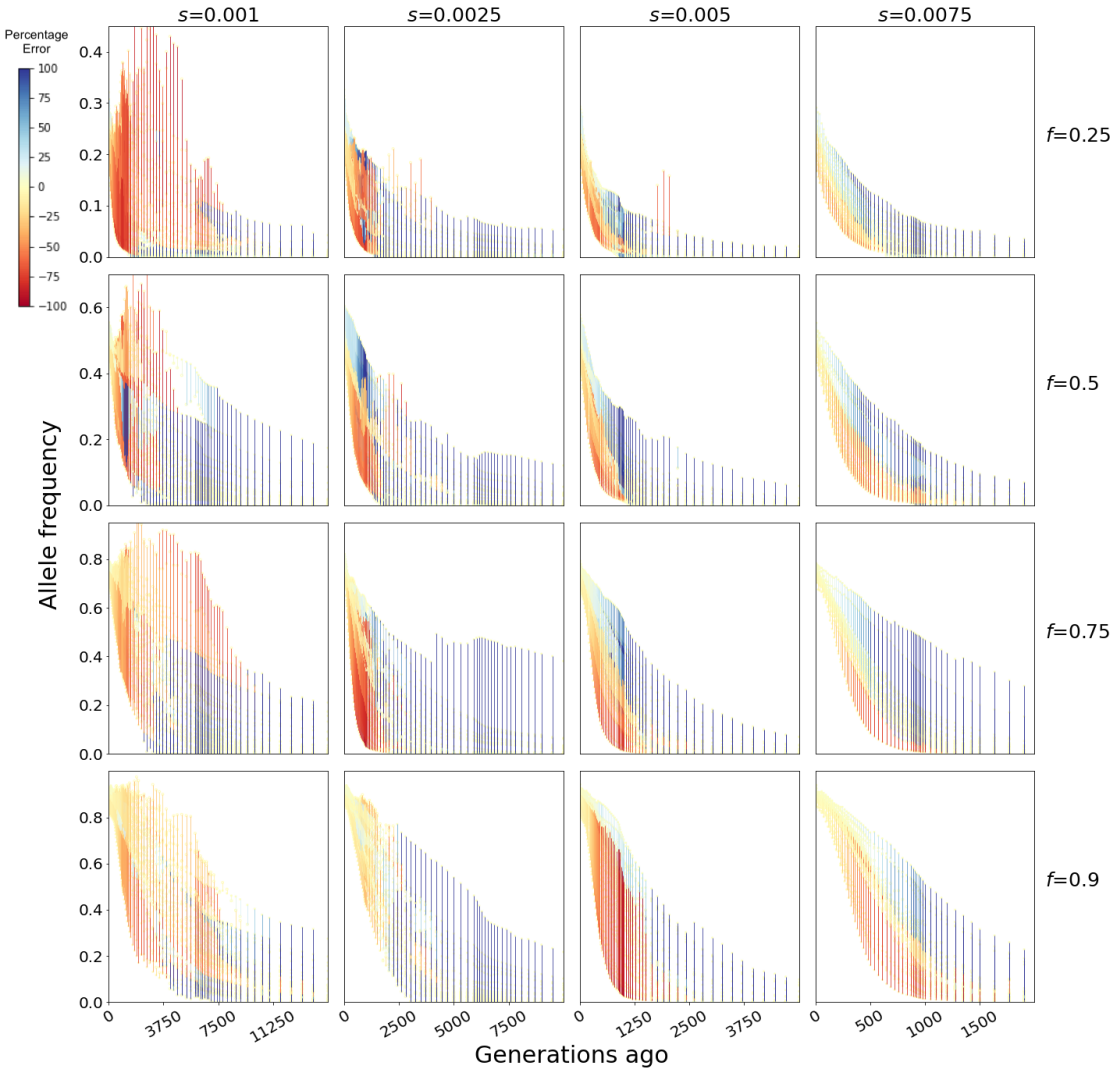


Figure S10: Root-mean-square error (RMSE) in allele frequency trajectory inference with true genealogies of simulations under the CEU demography. The error distributions visualized in **Figure S9** are summarized here as RMSE.



64 **Figure S11: Examples of allele frequency (AF) trajectories inferred with SIA using**
65 **genealogies inferred from data simulated under the CEU demography.** Figure layout and
66 description are identical to that of **Figure S8**.

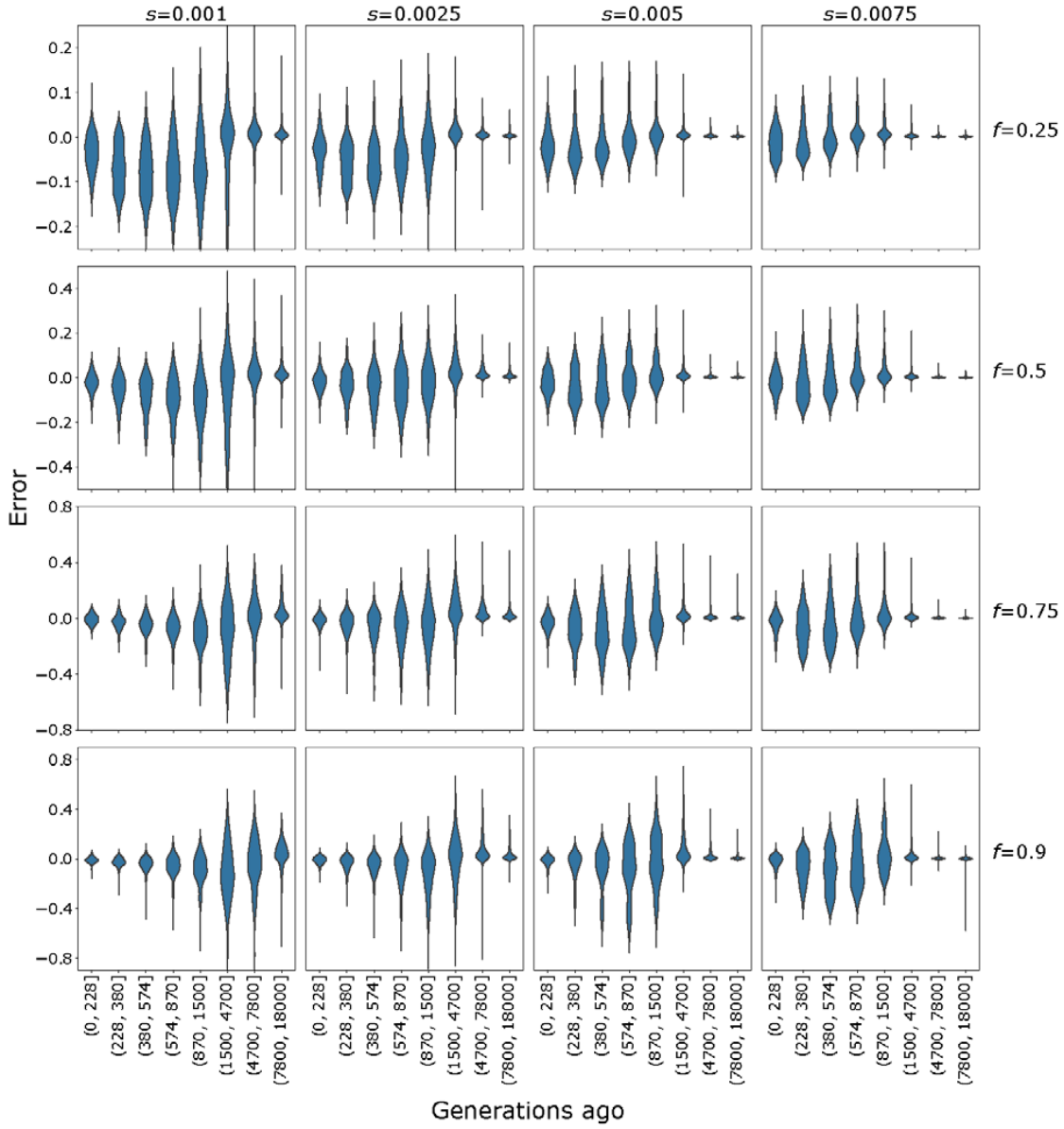


Figure S12: Distribution of error in allele frequency trajectory inferred with SIA using genealogies inferred from data simulated under the CEU demography. Figure layout and description are identical to that of **Figure S9**.

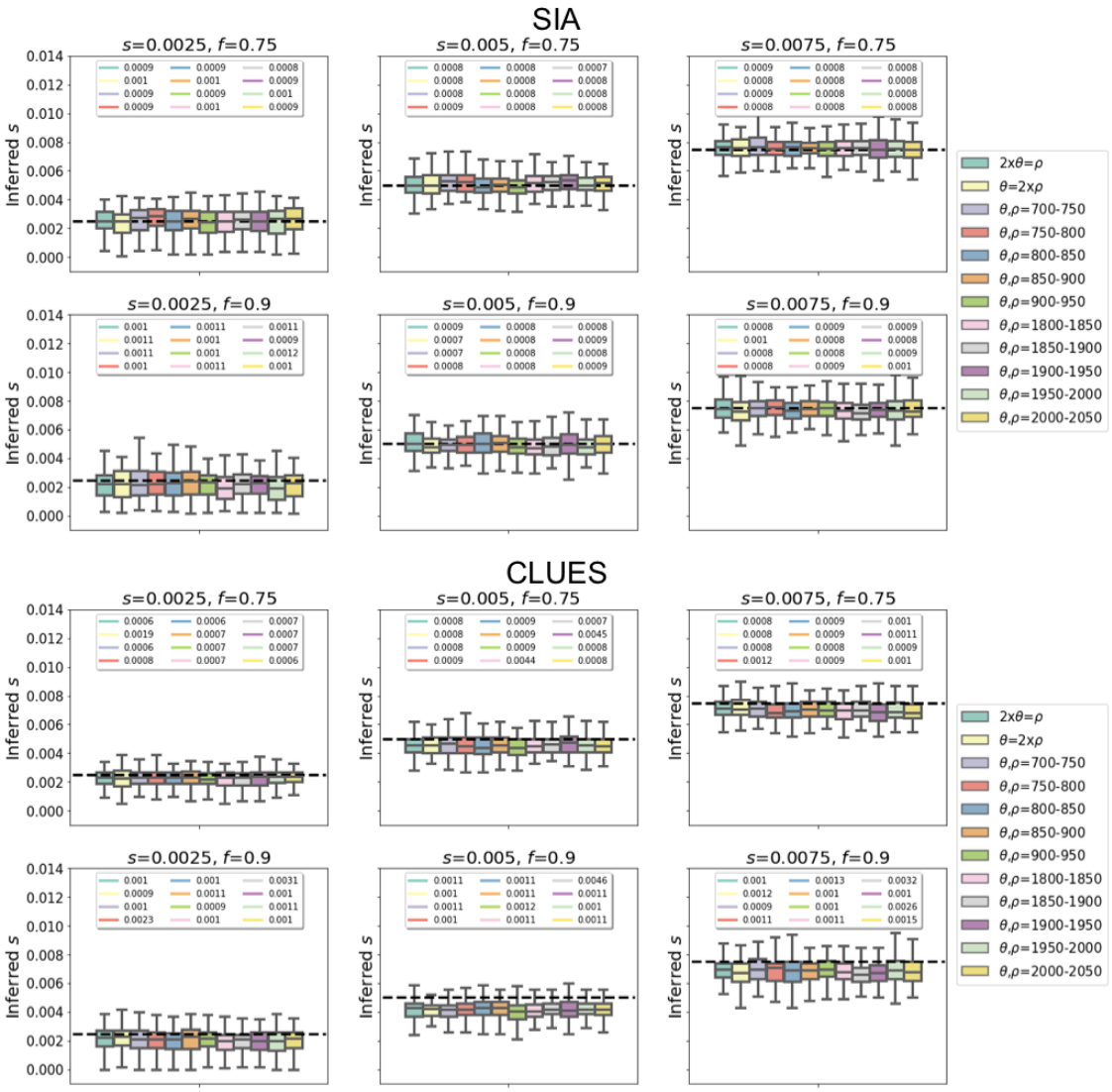


Figure S13: Performance of SIA and CLUES models on selection coefficient inference, tested on true genealogies simulated under variable combinations of population-scaled mutation rate ($\theta=4N_e\mu L$) and population-scaled recombination rate ($\rho=4N_erL$). Each panel shows the model predictions for simulations of a particular selection coefficient (s) and current derived allele frequency (f). Each box represents a group of 100 simulations under θ and ρ either specified by a fixed ratio, or sampled independently and uniformly within a particular range, as indicated in the legend. The dashed line marks the target value of s. The root mean squared error (RMSE) of the model predictions for each group of simulations is indicated at the top of the panel. For reference, the SIA model tested here was trained with true genealogies simulated under combinations of θ and ρ sampled independently and uniformly from a range of [940, 1880].

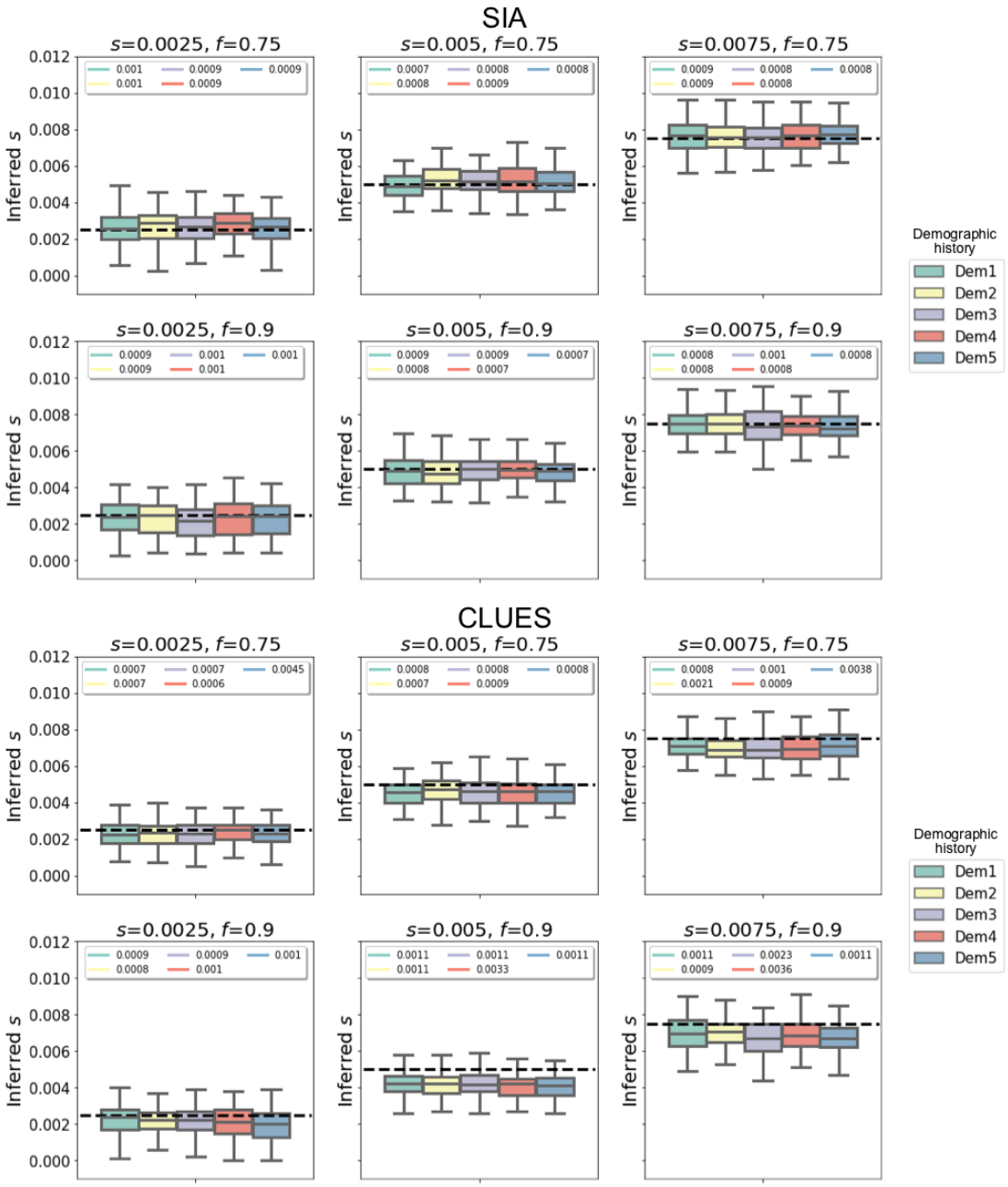


Figure S14: Performance of SIA and CLUES models on selection coefficient inference, tested on true genealogies simulated under five alternative demographies. Each demography is obtained from the CEU demography by modifying the population size at one time point during the recent population expansion phase (see **Figure S19** for more details). Each box represents a group of 100 simulations under the demography as indicated in the legend. Figure layout and description are otherwise similar to **Figure S13**.

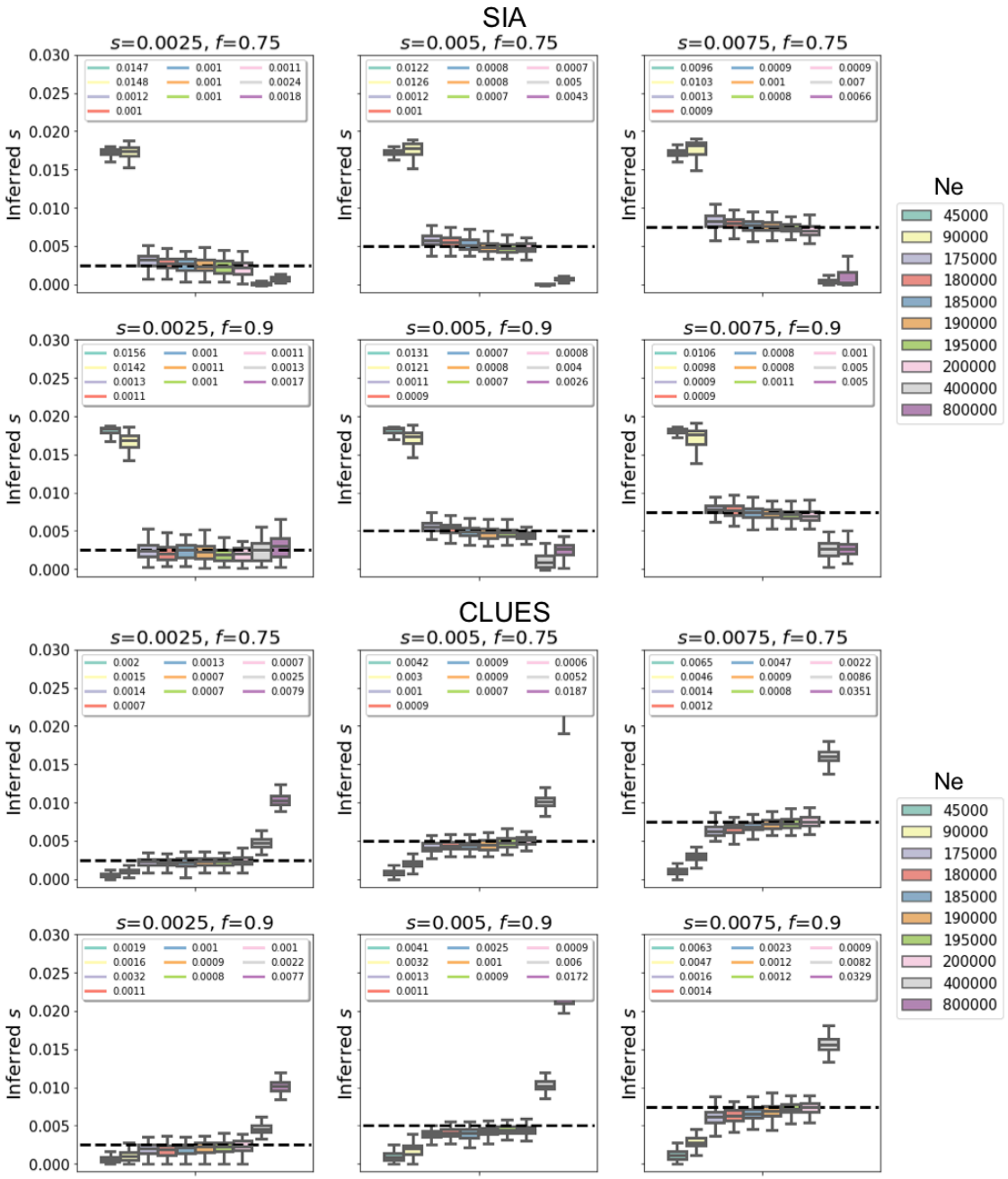


Figure S15: Performance of SIA and CLUES models on selection coefficient inference, tested on true genealogies simulated under the CEU demography scaled to different present-day N_e. Each group of 100 simulations of specific *s* and *f* under a particular present-day N_e (i.e. a box) were performed with a globally scaled CEU demography such that the resulting demography has a present-day N_e indicated by the legend (i.e. relative population size changes are preserved). For reference, the SIA model was trained with true genealogies

simulated under a present-day N_e of 188,088 (standard CEU). Figure layout and description are otherwise similar to **Figure S13**.

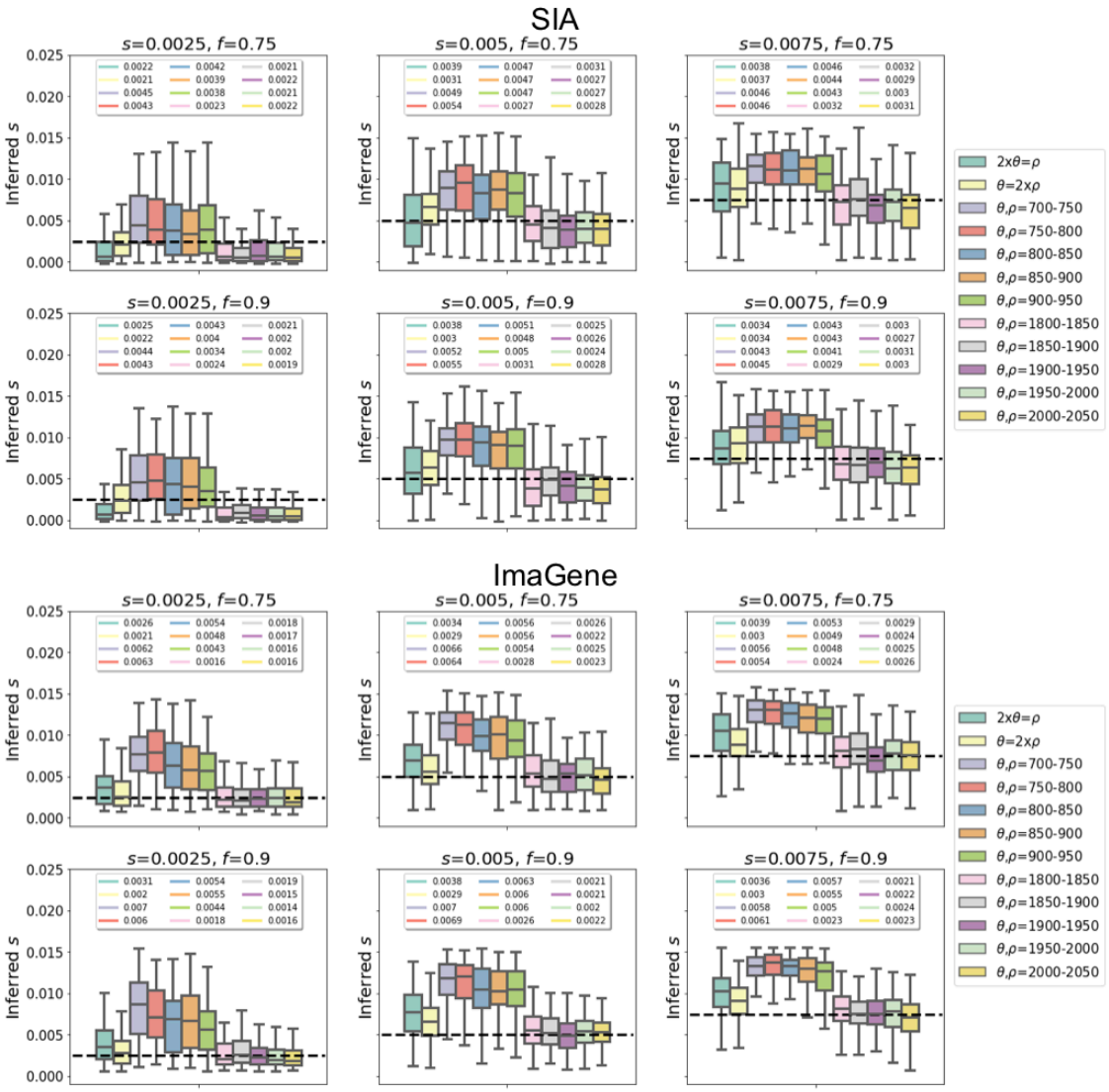


Figure S16: Performance of SIA and ImaGene models on selection coefficient inference, tested on genealogies inferred by Relate from simulations under variable combinations of population-scaled mutation rate *θ* and population-scaled recombination rate *ρ*. Note that the training data for both SIA and ImaGene are generated from identical sets of simulations, which were performed under combinations of *θ* and *ρ* sampled independently and uniformly from a range of [940, 1880]. Figure layout and description are otherwise similar to **Figure S13**.

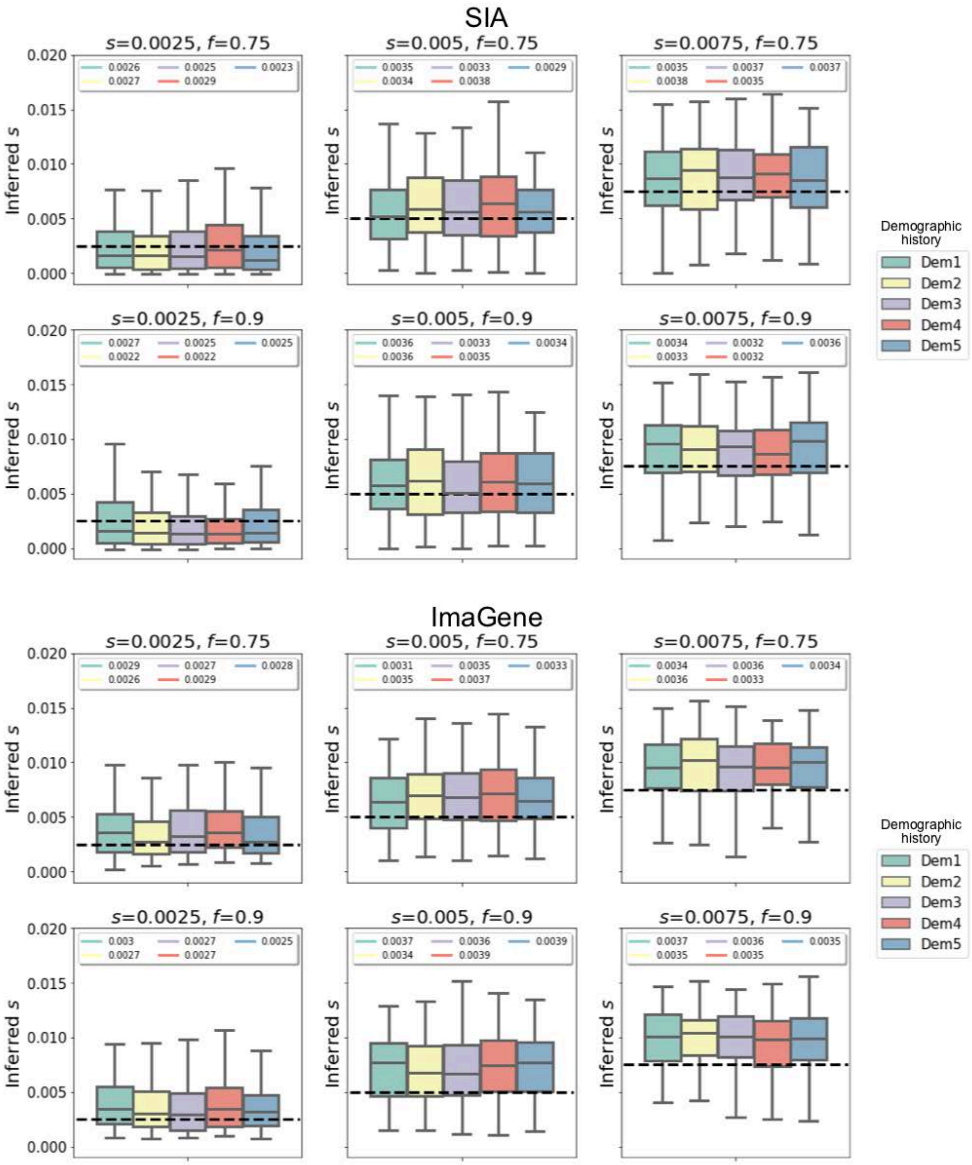


Figure S17: Performance of SIA and ImaGene models on selection coefficient inference, tested on genealogies inferred by Relate from simulations under five alternative demographies. Figure layout and description are otherwise similar to **Figure S14**. See **Figure S19** for details of the demography.

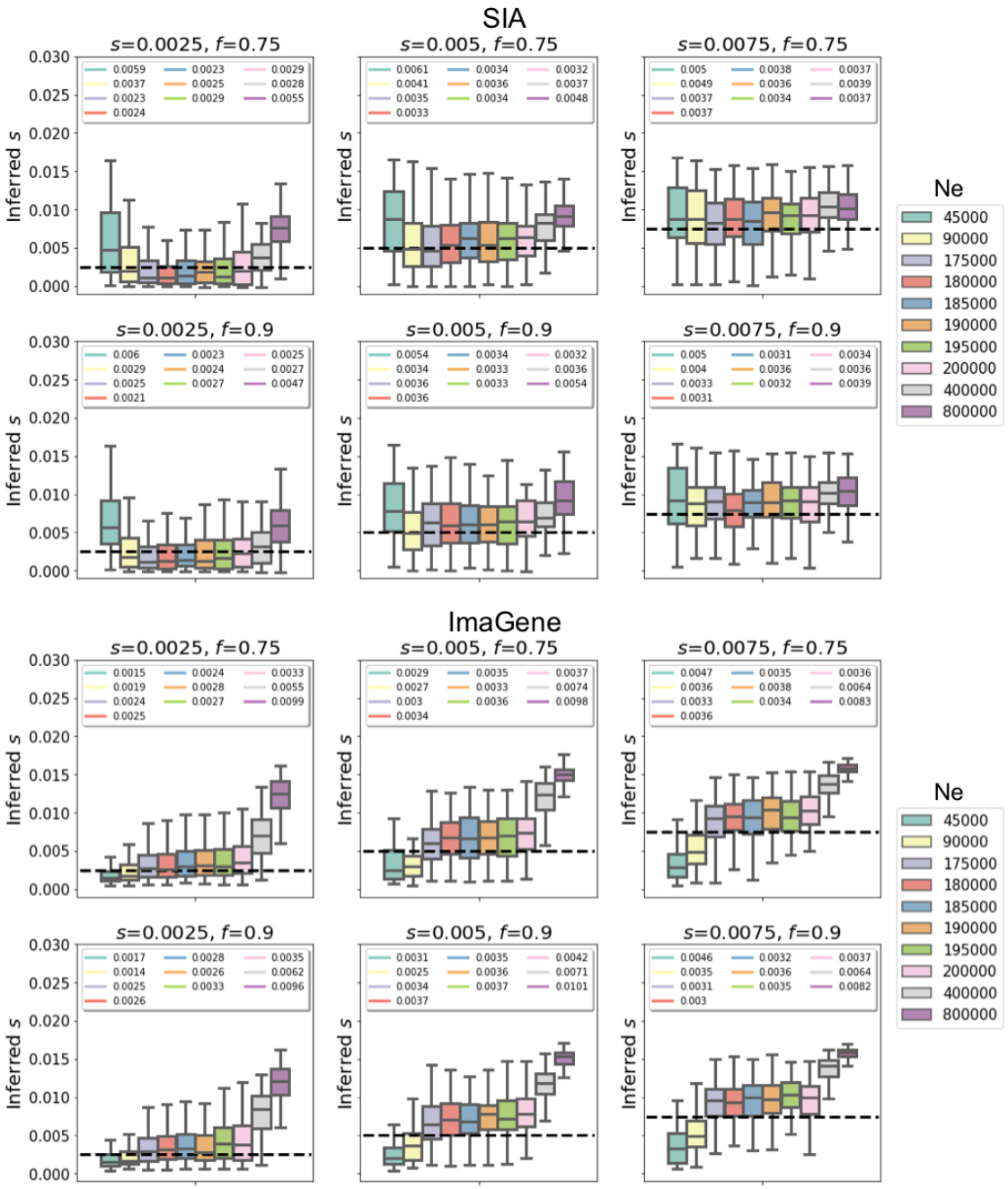


Figure S18: Performance of SIA and ImaGene models on selection coefficient inference, tested on genealogies inferred by Relate from simulations under the CEU demography scaled to different present-day N_e. Note that the training data for both SIA and ImaGene are generated from identical sets of simulations, which were performed under a present-day N_e of 188,088 (standard CEU). Figure layout and description are otherwise similar to **Figure S15.**

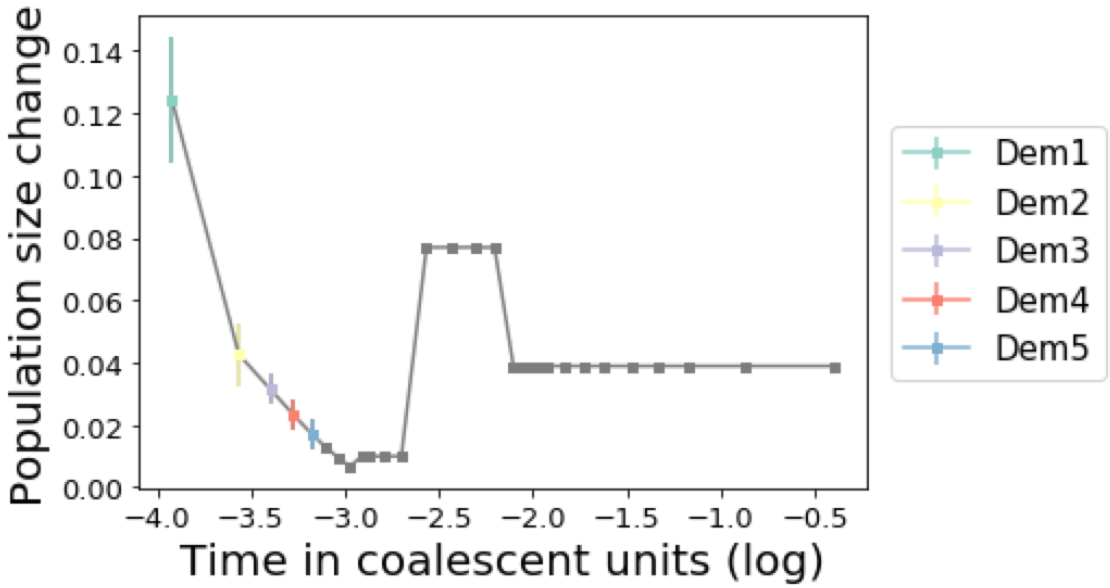


Figure S19: Illustration of alternative demographies used to simulate test data plotted in Figure S14 and S17. Squares indicate population size changes of the Tennessen *et al.* CEU model. For a simulation under a particular alternative demography, population size at the time point with matching color to the legend was modified by randomly sampling from a range centered on the original value ([-0.02, +0.02] for Dem1, [-0.01, +0.01] for Dem2, and [-0.005, +0.005] for Dem3-Dem5, as indicated by the vertical bar in the plot). Population sizes at all other time points were kept identical to the CEU demography.

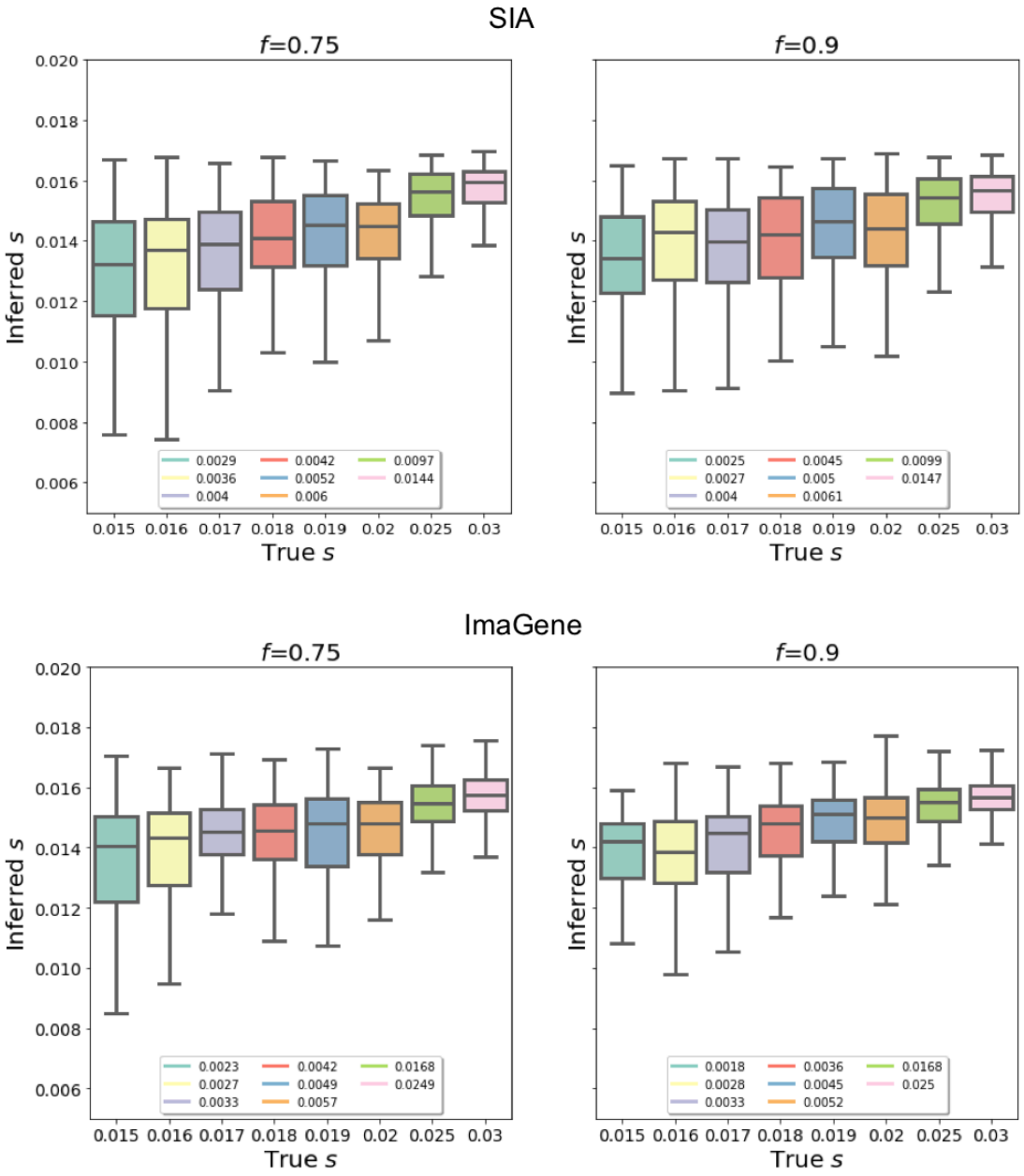


Figure S20: Performance of SIA and ImaGene models on selection coefficient inference, tested on genealogies inferred by Relate from simulations under selection coefficients (*s*) beyond the range used for simulating training data. Note that the training data for both SIA and ImaGene are generated from identical sets of simulations. Selection coefficients of sweep simulations constituting the training set were sampled uniformly from a range of [0.001, 0.02]. Figure layout and description are otherwise similar to **Figure S13**.

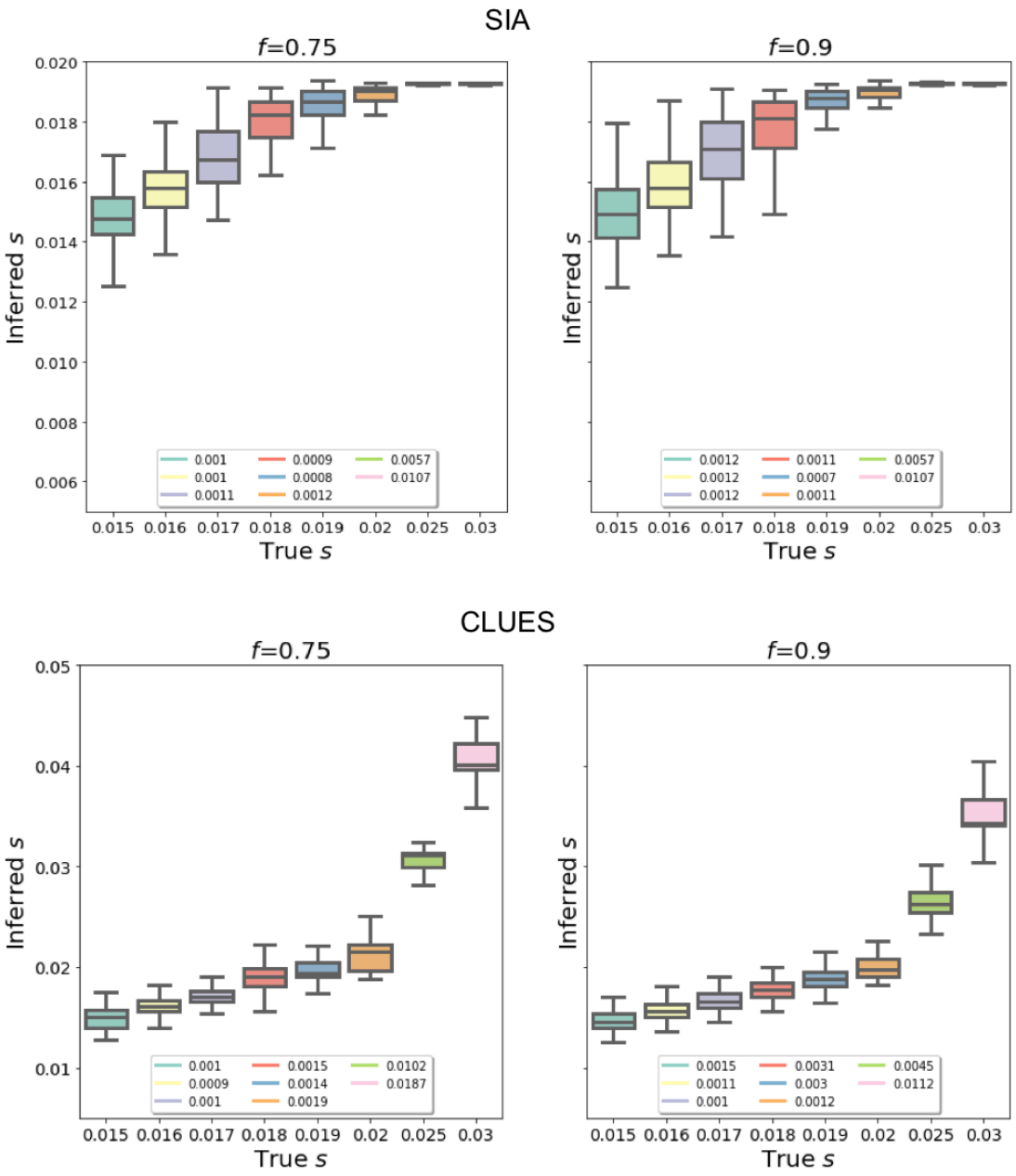


Figure S21: Performance of SIA and CLUES models on selection coefficient inference, tested on true genealogies simulated under selection coefficients (*s*) beyond the range used for simulating SIA training data. For reference, selection coefficients of sweep simulations constituting the SIA training set were sampled uniformly from a range of [0.001, 0.02]. Figure layout and description are otherwise similar to **Figure S20**.

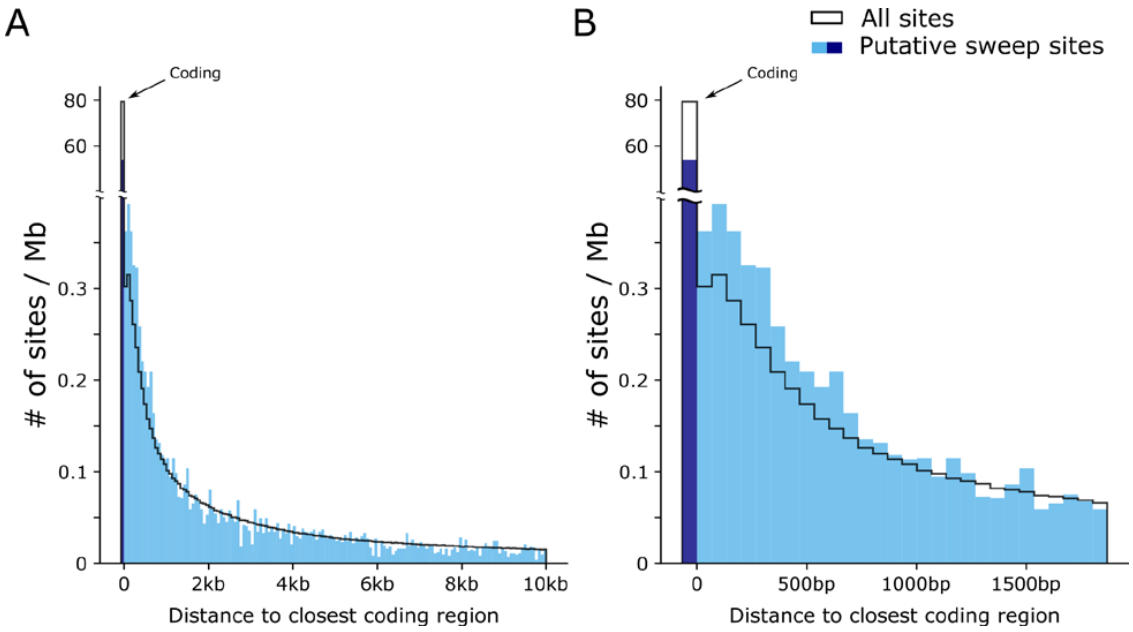


Figure S22: Distribution of putative soft sweep sites in *S. hypoxantha* with respect to the nearest coding regions. Panels (A) and (B) show the distribution of 15,551 sites (blue) across 333 scaffolds at different scales. For reference, the expected distribution of sites randomly drawn from all polymorphic sites is shown in black. Note that the sites that fall in coding regions are plotted in a separate bin.

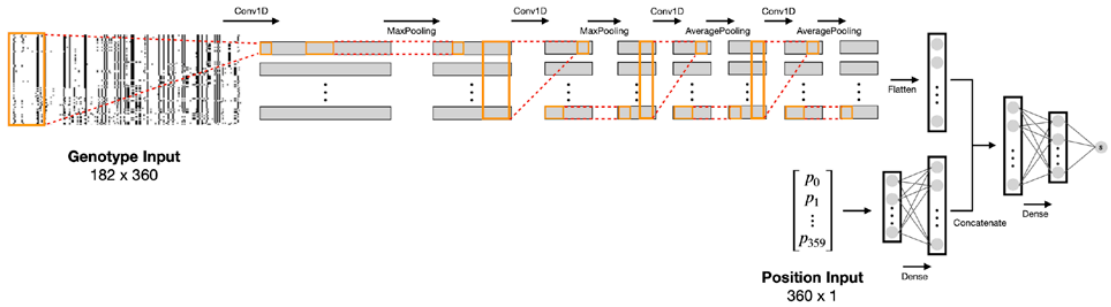


Figure S23: Architecture of genotyped-based CNN for selection inference. The model shares the exact same architecture as presented by Flagel et al. (Flagel et al. 2019) with one modification — the original softmax output layer for classification was replaced by a linear output layer for selection coefficient inference.

Appendix B

Supplementary material for Chapter 3

The computer code for this project has been deposited in GitHub repos, bird_capuchino_analysis and arg-selection. Genomic data have been archived in GenBank (BioProject ID PRJNA835722). Supplementary tables and figures are included in this appendix.

Table S1: Genes within the association peaks.

Comparison	Contig	Chromosome	Number of annotations	Genes
SASC vs. Makira	400	11	15	*TANGO6, GAS8, DBNDD1, DEF8, TUBB3, MC1R, TCF25, SPIRE2, FANCA, ZNF276, VPS9D1, SPATA2L, CDK10, LOC100232461, LOC115496849*
Ugi vs. Makira	1042	Z	1	*LOC115491070*
Ugi vs. Makira	224	Z	15	*SLC44A1, SLC27A6, ISOC1, ADAMTS19, MINAR2, CHSY3, HINT1, LYRM7, CDC42SE2, SPTLC1, ROR2, NFIL3, DIRAS2, GADD45G, LOC116806816*
Ugi vs. Makira	5	1	23	*EFHC2, FUNDC1, DIPK2B, MIR221, ICOSLG, GATD3A, PWP2, TRAPPC10, AGPAT3, PDXK, RRP1B, HSF2BP, SIK1, CRYAA, U2AF1, LOC116808847, LOC115496971, LOC115496975, LOC115496977, LOC115497347, LOC115496981, LOC115496994, LOC115497018*
Ugi vs. Makira	866	20	10	*TP53INP2, NCOA6, GGT7, RAB5IF, MYL9, TGIF2, DLGAP4, EPB41L1, SRSF6, L3MBTL1*
Ugi vs. Makira	947	20	14	*RALY, EIF2S2, ASIP, AHCY, ITCH, DYNLRB1, FSIP2, CEP250, GDF5, FAM83C, EIF6, MMP24, LOC116806597, LOC116809171*
Ugi vs. Makira	62	Z	19	*PLPP1, MSMP, GALT, SIGMAR1, DCTN3, ENHO, FAM219A, MYORG, NUDT2, UBAP1, DCAF12, NOL6, AQP3, KIAA1328, MAPK6, LOC115491277, LOC105760850, LOC115491042, LOC100226213*
Ugi vs. Makira	318	6	1	*LOC116808540*

Table S2: Details for the samples used in this study.

Subspecies	Island	Collection date	Locality	Latitude	Longitude	Ventral Coloration	Sample ID	Sex*
megarhynchus	Makira	5/16/09	Kira Kira	-10.5	161.9	Chestnut	MA132	M
megarhynchus	Makira	5/16/09	Kira Kira	-10.5	161.9	Chestnut	MA133	M
megarhynchus	Makira	5/22/09	Kira Kira	-10.5	161.9	Intermediate	MA166	M
megarhynchus	Makira	6/21/09	Star Harbour	-10.8	162.2	Chestnut	MA175	M
megarhynchus	Makira	6/22/09	Star Harbour	-10.8	162.2	Chestnut	MA180	M
megarhynchus	Makira	6/22/09	Star Harbour	-10.8	162.2	Intermediate	MA182	F
megarhynchus	Makira	6/23/09	Star Harbour	-10.8	162.2	Chestnut	MA183	M
megarhynchus	Makira	6/23/09	Star Harbour	-10.8	162.2	Chestnut	MA184	M
megarhynchus	Makira	6/24/09	Star Harbour	-10.8	162.2	Intermediate	MA185	F
megarhynchus	Makira	6/24/09	Star Harbour	-10.8	162.2	Chestnut	MA187	M
megarhynchus	Makira	6/15/11	Kira Kira	-10.5	161.9	Chestnut	MA193	F
megarhynchus	Makira	3/15/12	Kira Kira	-10.5	161.9	Chestnut	MA230	F
megarhynchus	Makira	3/15/12	Kira Kira	-10.5	161.9	Chestnut	MA231	F
megarhynchus	Makira	6/30/12	Kira Kira	-10.5	161.9	Chestnut	MA250	F
megarhynchus	Makira	3/11/14	Waimasi	-10.4	161.7	Intermediate	MA434	M
megarhynchus	Makira	3/11/14	Waimasi	-10.4	161.7	Intermediate	MA435	F
megarhynchus	Makira	3/12/14	Waimasi	-10.4	161.7	Chestnut	MA440	M
megarhynchus	Makira	3/12/14	Waimasi	-10.4	161.7	Intermediate	MA441	F
megarhynchus	Makira	6/15/18	Waimasi	-10.4	161.7	Chestnut	MA704	M
megarhynchus	Makira	6/15/18	Waimasi	-10.4	161.7	Melanic	MA705	M
megarhynchus	Makira	7/2/18	Waimasi	-10.4	161.7	Chestnut	MA714	M
megarhynchus	Makira	5/15/09	Kira Kira	-10.5	161.9	Chestnut	MA129**	M
ugiensis	Santa Ana	8/8/06	Gupuna	-10.8	162.5	Melanic	SA082	M
ugiensis	Santa Ana	8/10/06	Gupuna	-10.8	162.5	Intermediate	SA085	M
ugiensis	Santa Ana	8/10/06	Gupuna	-10.8	162.5	Melanic	SA087	F

ugiensis	Santa Ana	8/11/06	Gupuna	-10.8	162.5	Melanic	SA095	M
ugiensis	Santa Ana	6/17/07	Gupuna	-10.8	162.5	Melanic	SA105	M
ugiensis	Santa Ana	6/17/07	Gupuna	-10.8	162.5	Melanic	SA106	M
ugiensis	Santa Ana	6/17/07	Gupuna	-10.8	162.5	Melanic	SA107	M
ugiensis	Santa Ana	5/7/08	Gupuna	-10.8	162.5	Melanic	SA121	M
ugiensis	Santa Ana	5/7/08	Gupuna	-10.8	162.5	Melanic	SA123	M
ugiensis	Santa Ana	5/8/08	Gupuna	-10.8	162.5	Melanic	SA124	M
ugiensis	Santa Ana	5/7/08	Gupuna	-10.8	162.5	Melanic	SA125	F
ugiensis	Santa Ana	3/20/13	Gupuna	-10.8	162.5	Melanic	SA267	M
ugiensis	Santa Catalina	6/22/13	Santa Catalina	-10.9	162.5	Intermediate	SC275	F
ugiensis	Santa Catalina	6/22/13	Santa Catalina	-10.9	162.5	Intermediate	SC277	F
ugiensis	Santa Catalina	6/22/13	Santa Catalina	-10.9	162.5	Intermediate	SC278	M
ugiensis	Santa Catalina	6/22/13	Santa Catalina	-10.9	162.5	Melanic	SC283	M
ugiensis	Santa Catalina	6/24/13	Santa Catalina	-10.9	162.5	Melanic	SC296	M
ugiensis	Santa Catalina	6/24/13	Santa Catalina	-10.9	162.5	Melanic	SC402	M
ugiensis	Santa Catalina	6/24/13	Santa Catalina	-10.9	162.5	Melanic	SC404	M
ugiensis	Ugi	4/26/08	Pawa	-10.3	161.7	Melanic	UG108	M
ugiensis	Ugi	4/27/08	Pawa	-10.3	161.7	Melanic	UG115	M
ugiensis	Ugi	4/27/08	Pawa	-10.3	161.7	Melanic	UG116	F
ugiensis	Ugi	5/18/09	Pawa	-10.3	161.7	Melanic	UG147	M
ugiensis	Ugi	5/18/09	Pawa	-10.3	161.7	Melanic	UG148	M
ugiensis	Ugi	5/19/09	Pawa	-10.3	161.7	Melanic	UG152	M
ugiensis	Ugi	5/19/09	Pawa	-10.3	161.7	Melanic	UG155	M
ugiensis	Ugi	5/19/09	Pawa	-10.3	161.7	Melanic	UG156	M
ugiensis	Ugi	5/20/09	Pawa	-10.3	161.7	Melanic	UG160	F
ugiensis	Ugi	5/20/09	Pawa	-10.3	161.7	Melanic	UG161	M
ugiensis	Ugi	6/8/11	Pawa	-10.3	161.7	Melanic	UG197	M
ugiensis	Ugi	3/13/12	Pawa	-10.3	161.7	Melanic	UG220	M

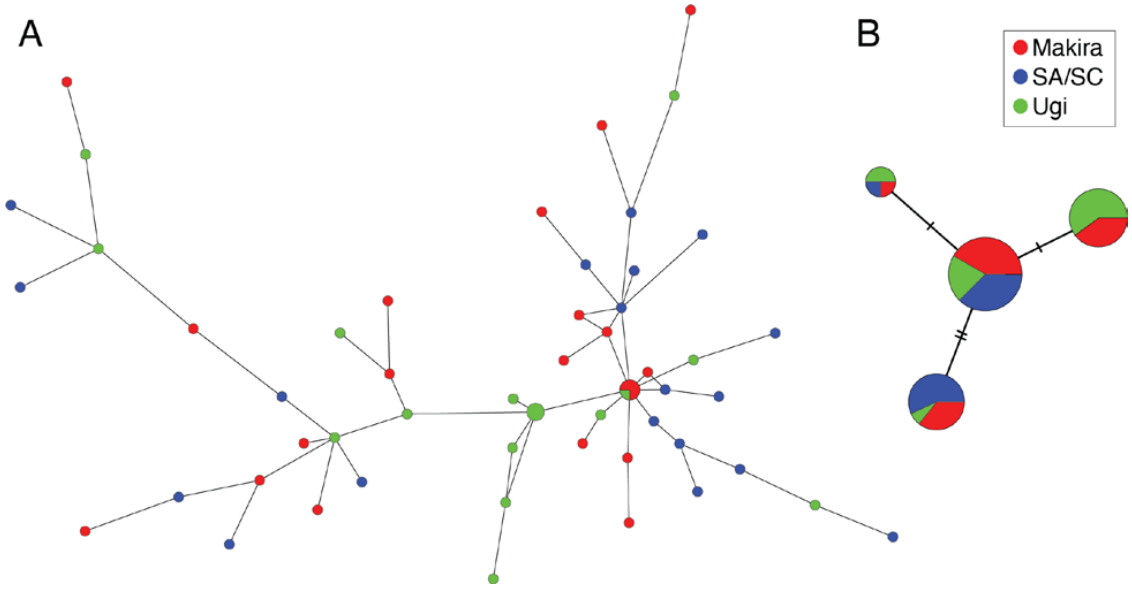
ugiensis	Ugi	7/4/13	Bio	-10.2	161.7	Melanic	UG407	M
ugiensis	Ugi	3/15/14	Pawa	-10.3	161.7	Melanic	UG448	M
ugiensis	Ugi	3/15/14	Pawa	-10.3	161.7	Melanic	UG450	F
ugiensis	Ugi	3/15/14	Pawa	-10.3	161.7	Melanic	UG451	M

*We determined sex by calculating the average depth of coverage across all the positions in each of six different contigs. Three of these contigs were autosomal and three were part of the Z chromosome. We subsequently averaged the depth of coverage for the three Z-linked contigs and divided it by the average from the three autosomal contigs. This process produced values around 0.5 for heterogametic females, and values close to 1 for males. Additionally, for a subset of 39 individuals we also determined sex through PCR as described in reference (1). Both methods produced congruent results.

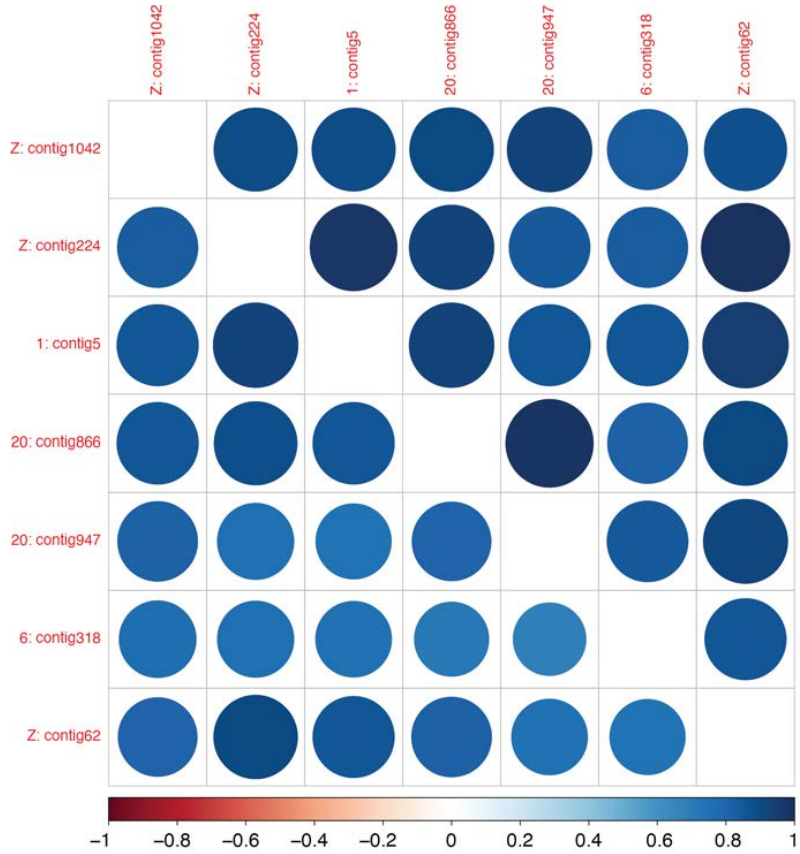
**This sample is labelled as UG129 and not MA129 in the different files from our bioinformatics pipeline (e.g., vcf files), yet was treated correctly as a sample originating from Makira and not from Ugi (as other samples denoted with UG).

References

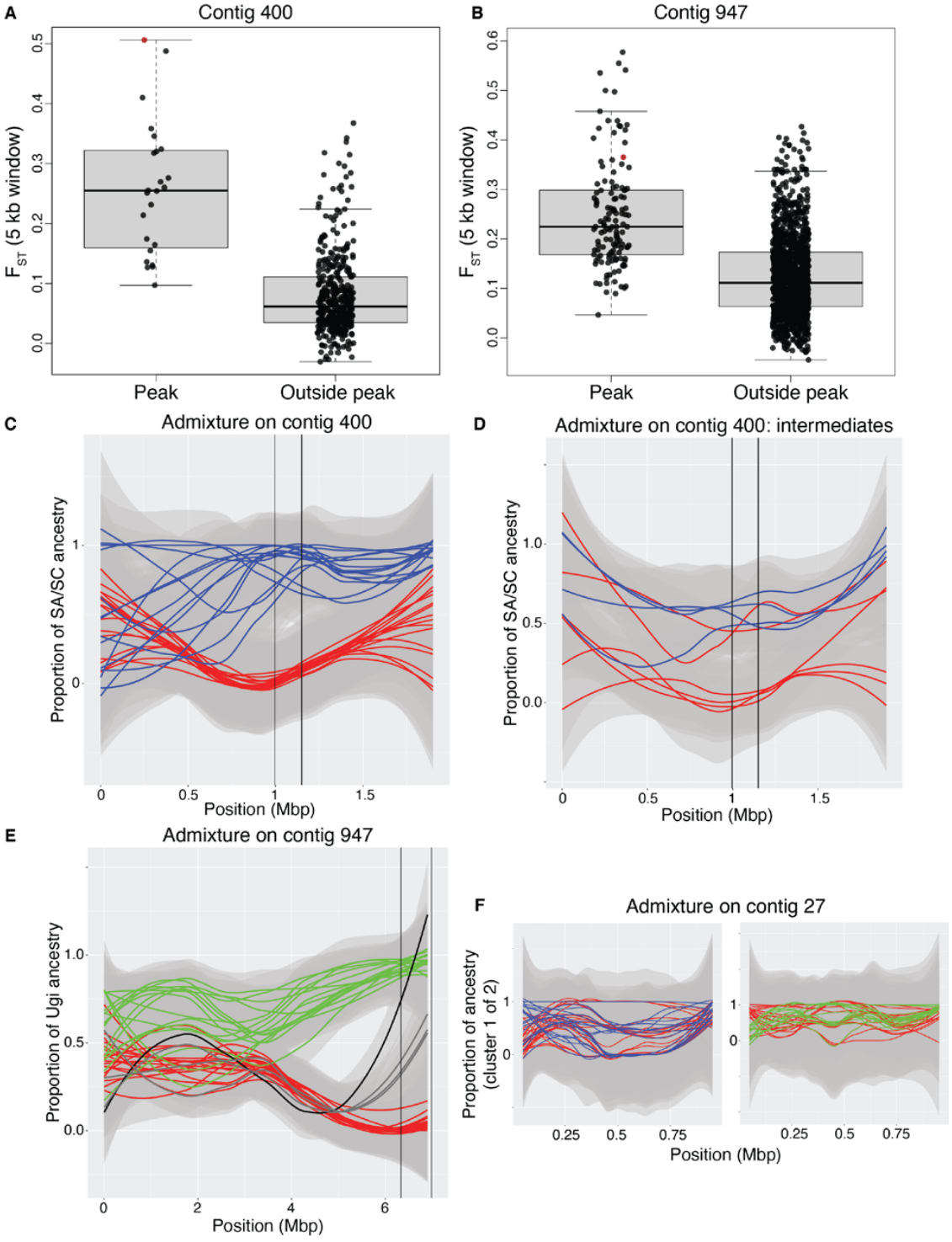
1. N. W. Kahn, J. St. John, T. W. Quinn, Chromosome-specific intron size differences in the avian CHD gene provide an efficient method for sex identification in birds. *The Auk*, 1074–1078 (1998).



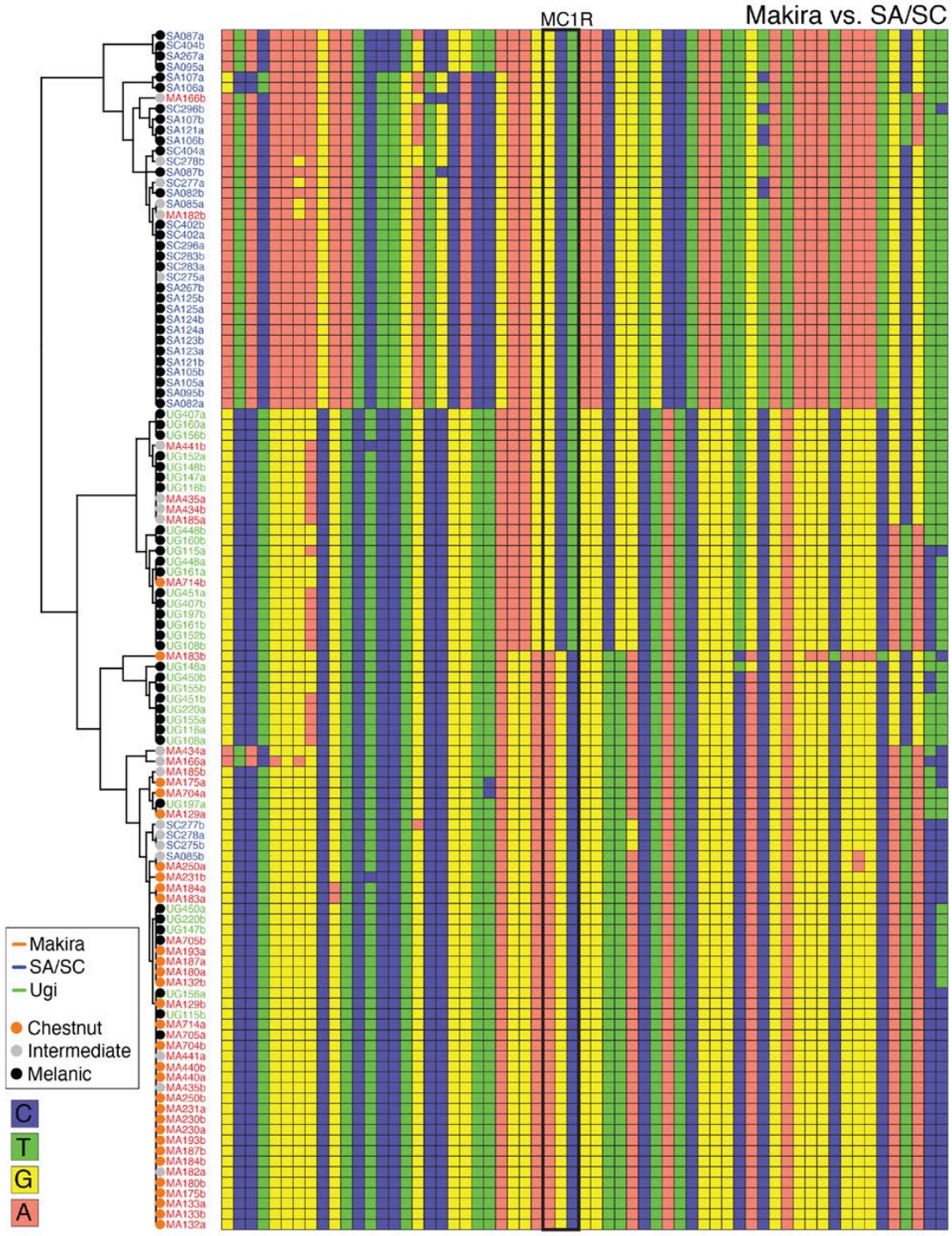
S1 Fig. Mitochondrial minimum spanning networks. **A**. Haplotype network based on a ~17 kbp alignment of the mitochondrial genome. **B**. Haplotype network based on 650 bp of the mitochondrial COI gene, commonly used for species identification. Branch lengths are proportional to the number of nucleotide differences between haplotypes, which are indicated by short lines on each branch (omitted for simplicity in the case of the full mitochondrial network).



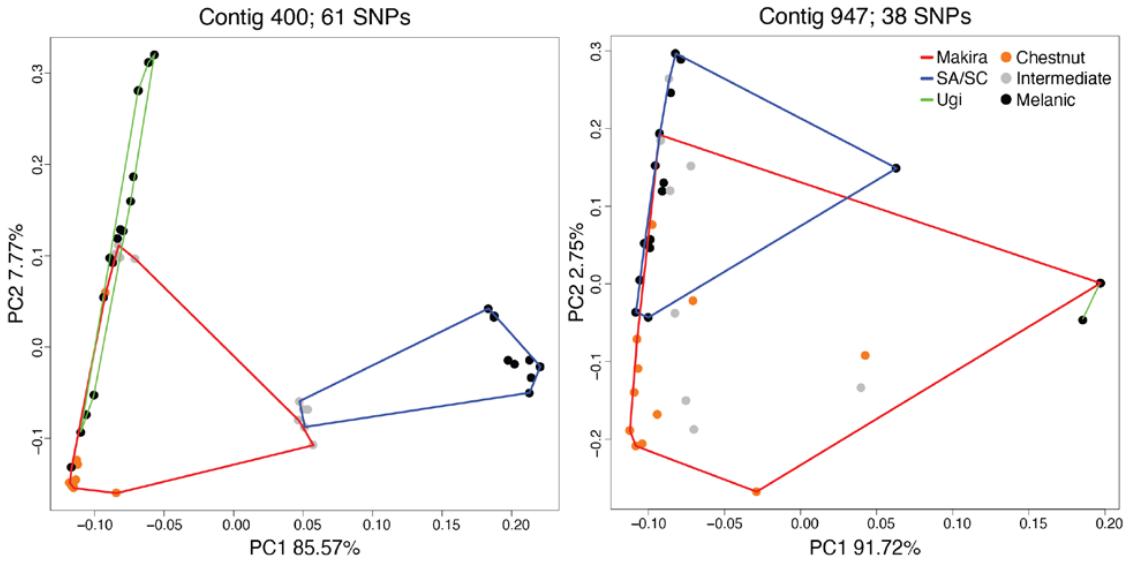
S2 Fig. Linkage disequilibrium among association peaks identified in the GWAS conducted with Makira and Ugi individuals. Average (below the diagonal) and maximum (above the diagonal) R2 values among all the statistical outlier sites in the GWAS from different pairs of association peaks. The chromosome and contig to which each peak belongs is indicated in red, and the size and color of the circles denotes the magnitude of LD.


A
Contig 400
F_{ST} (5 kb window)
Peak
Outside peak
B
Contig 947
F_{ST} (5 kb window)
Peak
Outside peak
C
Admixture on contig 400
Proportion of SA/SC ancestry
Position (Mbp)
D
Admixture on contig 400: intermediates
Proportion of SA/SC ancestry
Position (Mbp)
E
Admixture on contig 947
Proportion of Ugi ancestry
Position (Mbp)
F
Admixture on contig 27
Proportion of ancestry (cluster 1 of 2)
Position (Mbp)

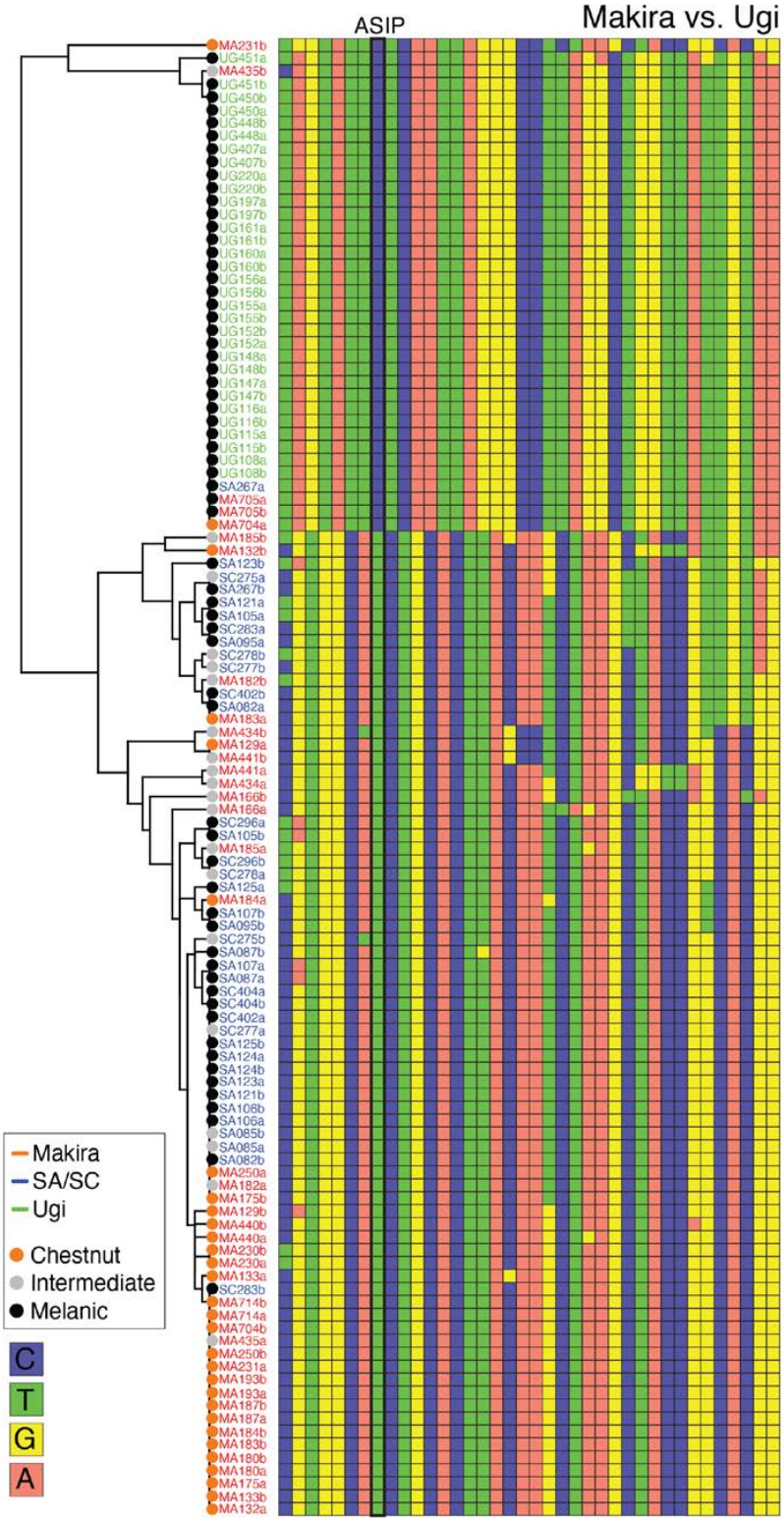
S3 Fig. Genetic differentiation within and outside of association peaks. **A and B**. F_{ST} values calculated for 5 kb windows inside and outside association peaks. The plot for contig 400 compares individuals from SA/SC and Makira, while the plot for contig 947 compares individuals from Ugi and Makira. The red dot denotes the window containing the *MC1R* and *ASIP* genes. **C to F**. Smoothed ancestry values across contigs, with association peaks indicated between vertical lines. Ancestry values were calculated in 100 kb sliding windows in Admixture, and the analysis was restricted to Makira vs. SA/SC for contig 400 and Makira vs. Ugi for contig 947. The plot in (**C**) shows increased separation between individuals from Makira and SA/SC in the peak region on contig 400 and extending approximately 0.5 Mbp in each direction. The plot includes only individuals from Makira and SA/SC that were homozygotes for the *Asp119* or *Asn119 MC1R* mutation, respectively. The plot in (**D**) shows the six individuals from Makira and the four individuals from SA/SC which were heterozygotes for this mutation and had intermediate plumage, and shows overall more admixture than what is seen in (**C**). **E**. Ancestry values across contig 947 showing increased resolution in the peak region and extending approximately 2 Mb downstream. Three individuals from Makira which were heterozygotes for the *Tre*55 *ASIP* mutation are labeled in gray (only one of these individuals had intermediate plumage). The single melanic individual from Makira (MA705), which was homozygous for the derived *Tre*55 mutation, is labelled in black. For all four individuals, Ugi ancestry decreases to levels comparable to other Makira individuals about 1 Mb downstream of the peak. **F**. Ancestry across contig 27 shows little resolution compared to the association peaks on contigs 400 and 947. We note that ancestry values range from 0 to 1, but that the plots extend beyond this range because of the smoothing algorithm and particularly the uncertainty shown by the confidence bands. Values beyond the [0,1] interval are therefore meaningless.


MC1R
Makira vs. SA/SC
Makira
SA/SC
Ugi
Chestnut
Intermediate
Melanic
C
T
G
A

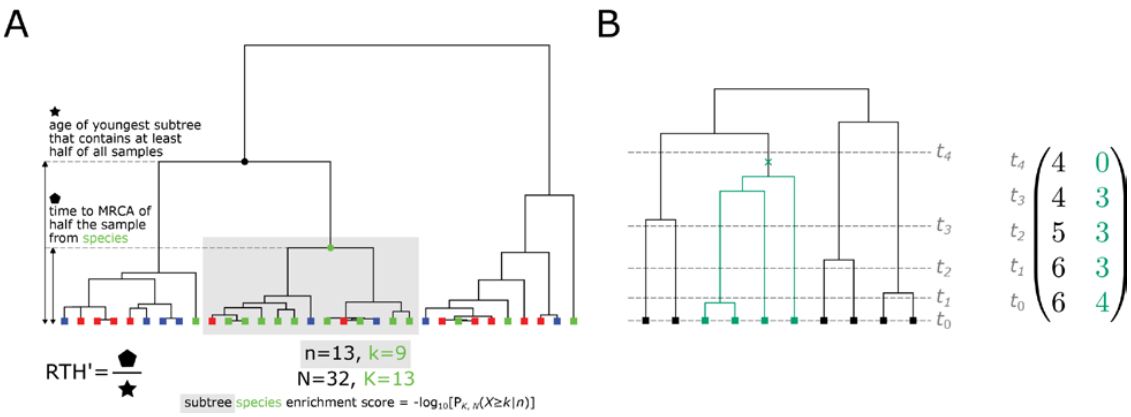
S4 Fig. Clustering of haplotypes obtained from the association peak on contig 400. Phased genotypes for the 61 SNPs located in the association peak on contig 400. Rows represent single chromosomes, therefore individuals are represented twice in the clustering tree on the left. The four nucleotides, the collection locality and the coloration phenotype are color-coded as indicated at left. The three SNPs within the *MC1R* coding region are indicated with a black rectangle. All individuals from SA/SC contained at least one haplotype in the region delimited by the SNPs with significant association scores around *MC1R* that differed from the one present in most individuals from Makira and Ugi. All the melanic individuals from SA/SC possessed two copies of this haplotype, while the four individuals with intermediate coloration possessed one of each, as was the case for two of the six individuals with intermediate coloration from Makira. All melanic individuals from SA/SC carried two copies of the derived *Asn119* mutation, while all but one of the chestnut-bellied individuals from Makira had two copies of *Asp119*. The individuals with intermediate coloration (from either Makira or SA/SC) were heterozygotes for this coding mutation. The exception to this pattern was a single chestnut-bellied individual from Makira (MA714), which was a heterozygote yet was scored in the field under heavy molt and may have been incorrectly classified as having a chestnut belly. Finally, the derived *Asn199* mutation existed primarily on the haplotype background found on SA/SC, but also to a lesser extent on the haplotype background found on Makira and Ugi. We note that the chestnut-bellied MA714 bird carried the *Asn119* mutation on the most common haplotype background observed on Makira. The birds from Ugi, which are all melanic, were sometimes homozygous for either allele or heterozygotes.



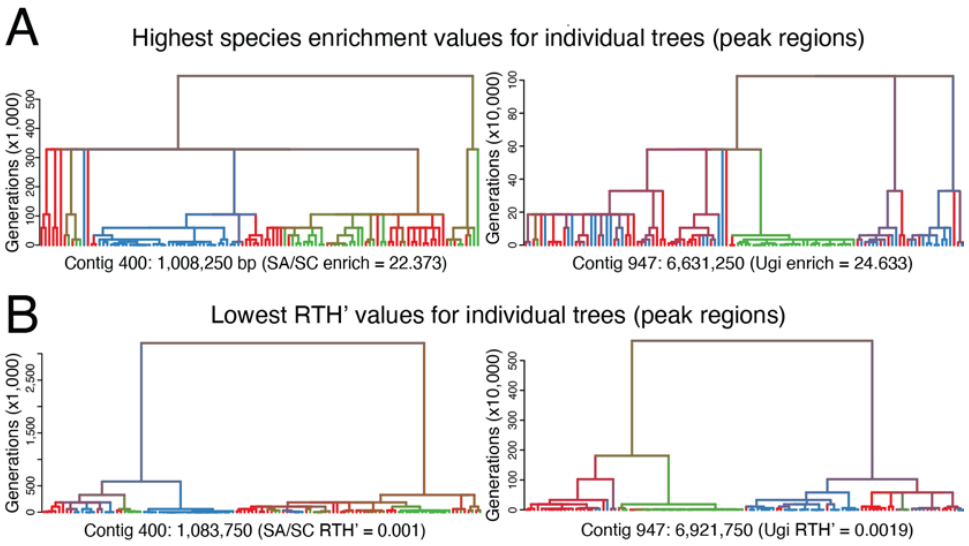
S5 Fig. Principal component analyses derived from the SNPs within the association peaks. PCAs from the variants within the association peaks on contig 400 and contig 947. Coloration phenotype and the island where individuals were sampled are color-coded.


ASIP
Makira vs. Ugi
MA231b
UG451a
MA435b
UG451b
UG450b
UG450a
UG448b
UG448a
UG407a
UG407b
UG220a
UG220b
UG197a
UG197b
UG161a
UG161b
UG160a
UG160b
UG156a
UG156b
UG155a
UG155b
UG152b
UG152a
UG148a
UG148b
UG147a
UG147b
UG116a
UG116b
UG115a
UG115b
UG108a
UG108b
SA267a
MA705a
MA705b
MA704a
MA185b
MA132b
SA123b
SC275a
SA267b
SA121a
SA105a
SC283a
SA065a
SC278b
SC277b
MA182b
SC402b
SA082a
MA183a
MA434b
MA129a
MA441b
MA441a
MA434a
MA166b
MA166a
SC296a
SA105b
MA185a
SC296b
SC278a
SA125a
MA184a
SA107b
SA095b
SC275b
SA087b
SA107a
SA087a
SC404a
SC404b
SC402a
SC277a
SA125b
SA124a
SA124b
SA123a
SA121b
SA106b
SA106a
SA085b
SA085a
SA082b
MA250a
MA182a
MA175b
MA129b
MA440b
MA440a
MA230b
MA230a
MA133a
SC283b
MA714b
MA714a
MA704b
MA435a
MA250b
MA231a
MA193b
MA193a
MA187b
MA187a
MA184b
MA183b
MA180b
MA180a
MA175a
MA133b
MA132a
Makira
SA/SC
Ugi
Chestnut
Intermediate
Melanic
C
T
G
A

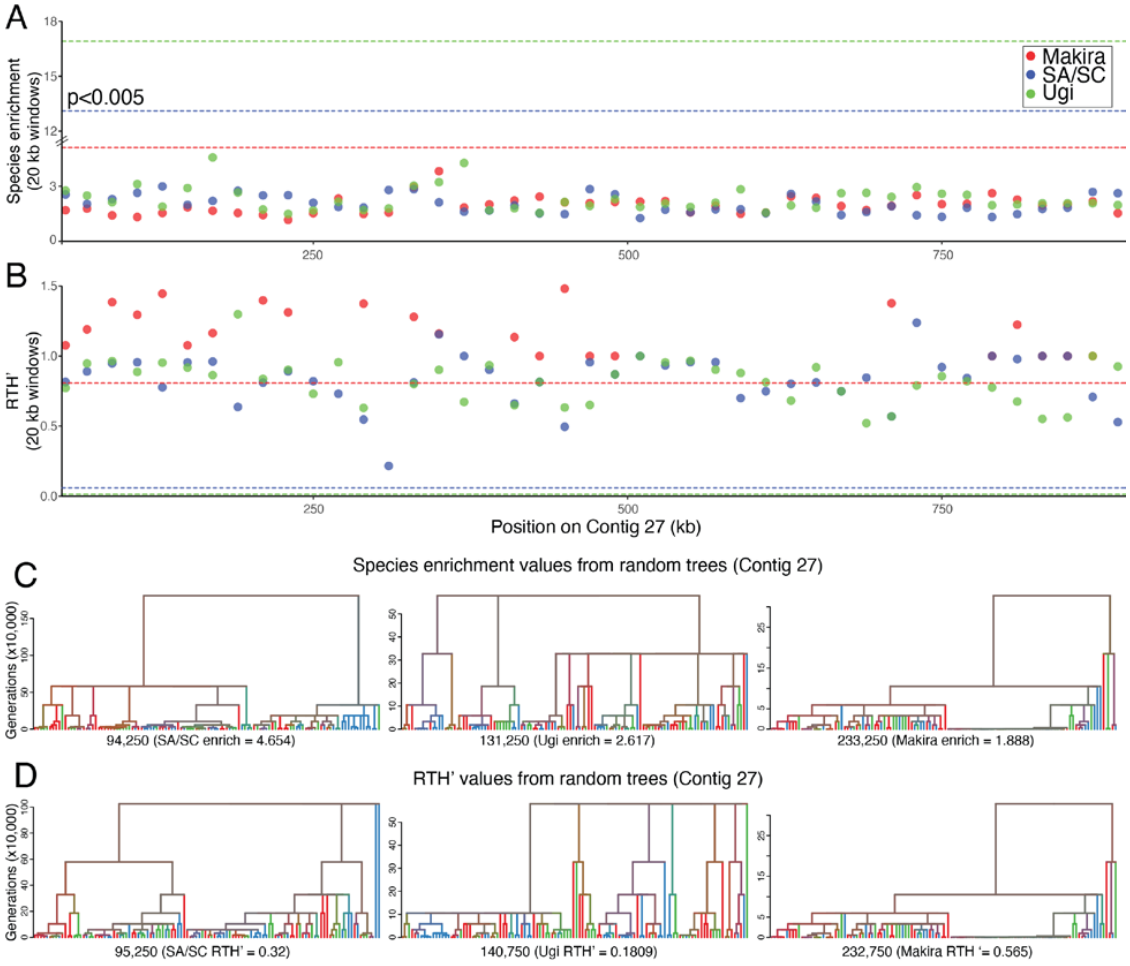
S6 Fig. Clustering of haplotypes obtained from the association peak on contig 947. Details as in **S4 Fig**. Ugi individuals had two copies of a haplotype that was different from the one present in SA/SC and Makira individuals. The only melanic individual from Makira (MA705) also carried two copies of the derived Ugi haplotype. We found a few heterozygote individuals but there was no obvious pattern with respect to their coloration.



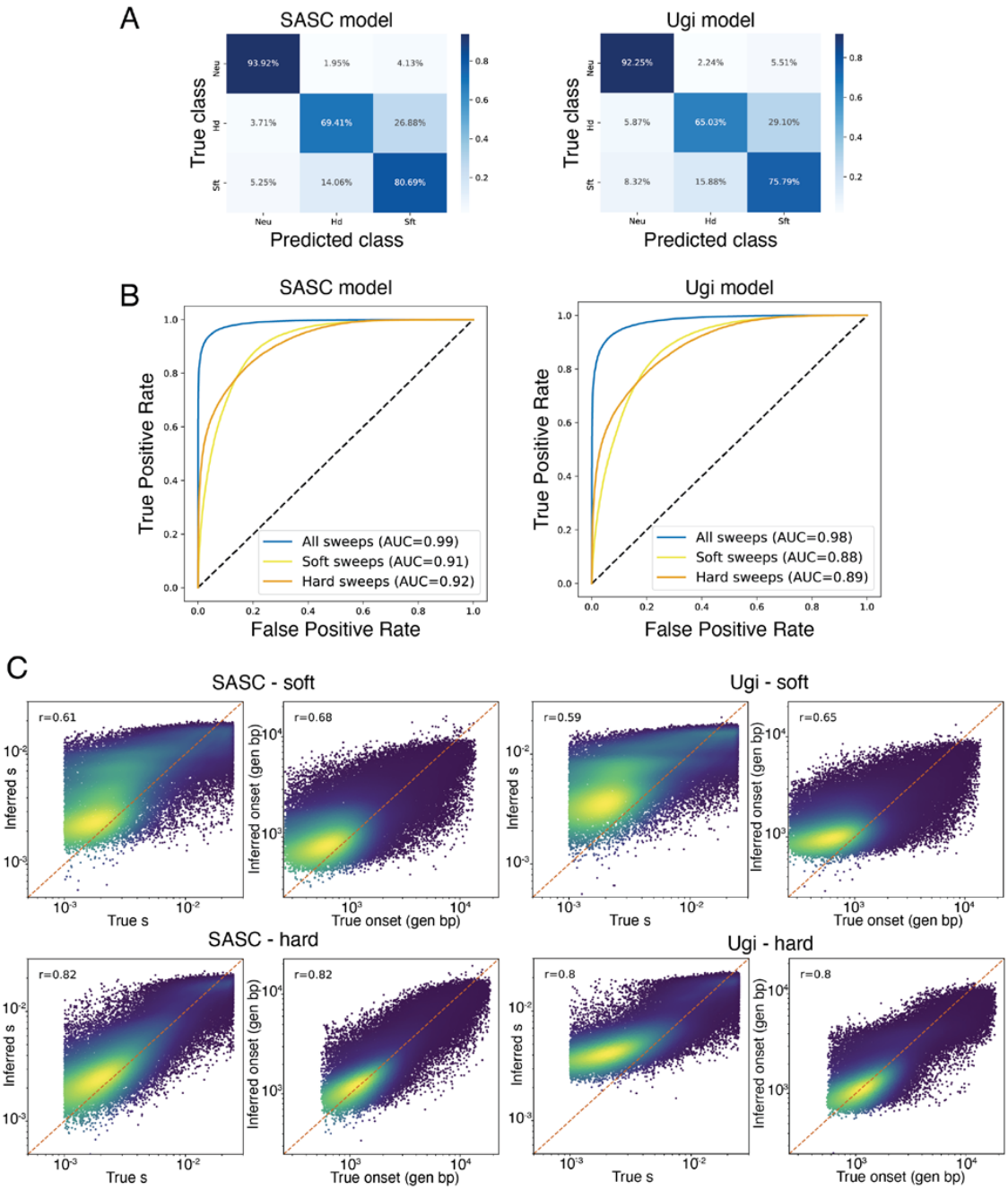
S7 Fig. Statistics and feature encoding of the genealogy. **A.** RTH' tests for the reduction in within-species TMRCA and is defined as the ratio between the TMRCA of half of the samples from *a given species* and the age of the youngest subtree that contains at least half of *all* samples. The species enrichment score tests for species differentiation in local trees and is defined as the maximum score associated with a given species in a subtree of the full genealogy. For a subtree, the species enrichment score is the probability of observing the number of samples of a particular species in that subtree under a hypergeometric distribution. Here the coloring of the leaves indicates hypothetical species and the example illustrates the statistics with respect to the green species. **B.** Genealogical features for the SIA model consist of the number of lineages in the genealogy at a set of discrete time points (t_0, t_1, …). The time points are chosen in an approximately log-uniform manner resulting in finer discretization of more recent time scales. In addition, when encoding the genealogical features at a particular site of interest, we encode separately the counts of ancestral (shown in black in the example) and derived (shown in aquamarine) lineages. The aquamarine cross indicates the branch where the mutation occurred.



S8 Fig. Highest species enrichment scores and lowest RTH' values for peak regions on contigs 400 and 947. Representative trees from the peak regions which show extreme values of species-enrichment (**A**) and RTH' (**B**) for SA/SC and Ugi.

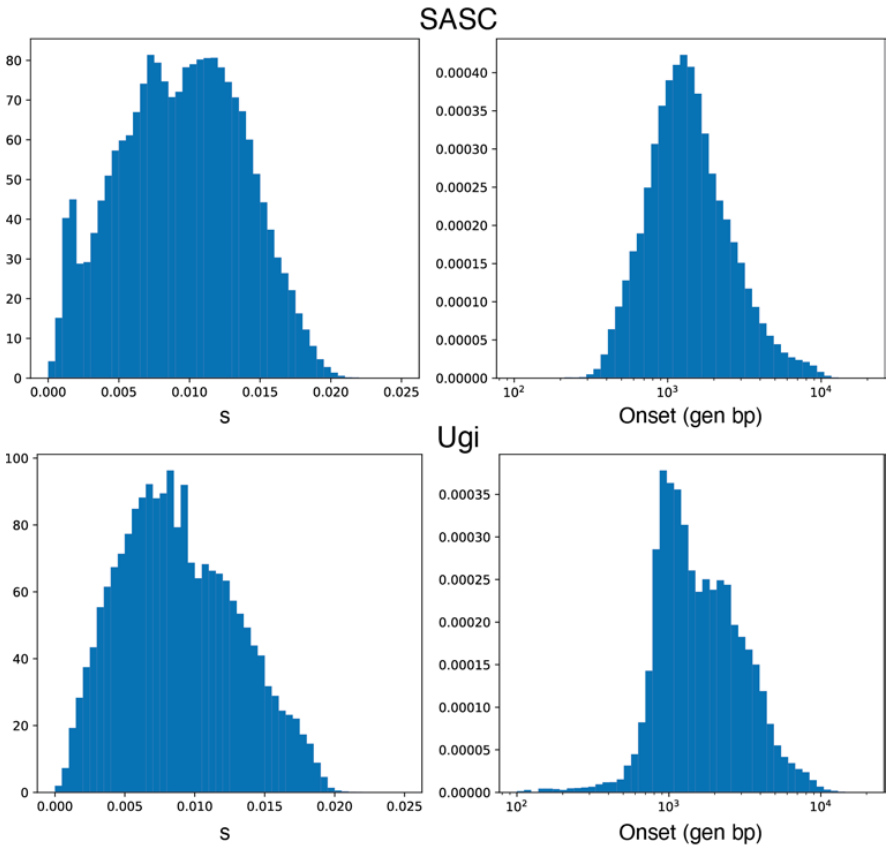


S9 Fig. Species enrichment scores and RTH' values for a control contig. Plots showing species enrichment (**A**) and RTH' (**B**) in 20 kb windows along contig 27. Horizontal lines show species-specific levels of statistical significance (p<0.005). Trees obtained from random positions on contig 27 (each position is shown under the tree) indicating species enrichment scores (**C**) and RTH' (**D**) values for each population.



S10 Fig. Benchmarking of the SIA models. **A**. Confusion matrices generated by applying the SA/SC and Ugi models to simulated data. Instead of applying a specific probability threshold, the predicted class was identified as the one with the highest probability according to the model. Under this maximum likelihood classification scheme, both models perform best distinguishing neutral from selected sites, and moderately when identifying soft or hard sweeps. **B**. One-versus-rest (OvR) receiver

operating characteristic (ROC) curves of the model classification performance on simulated data. The models perform very well on distinguishing sweeps, and less so on precisely identifying hard or soft sweeps. **C**. Comparisons between true and inferred selection coefficients and time of selection onset (expressed in generations before present) for SA/SC and Ugi models trained to detect soft or hard sweeps. Models trained to predict hard sweeps generally perform better than those trained to predict soft sweeps, which tend to overestimate small selection coefficients and recent selection onset times.



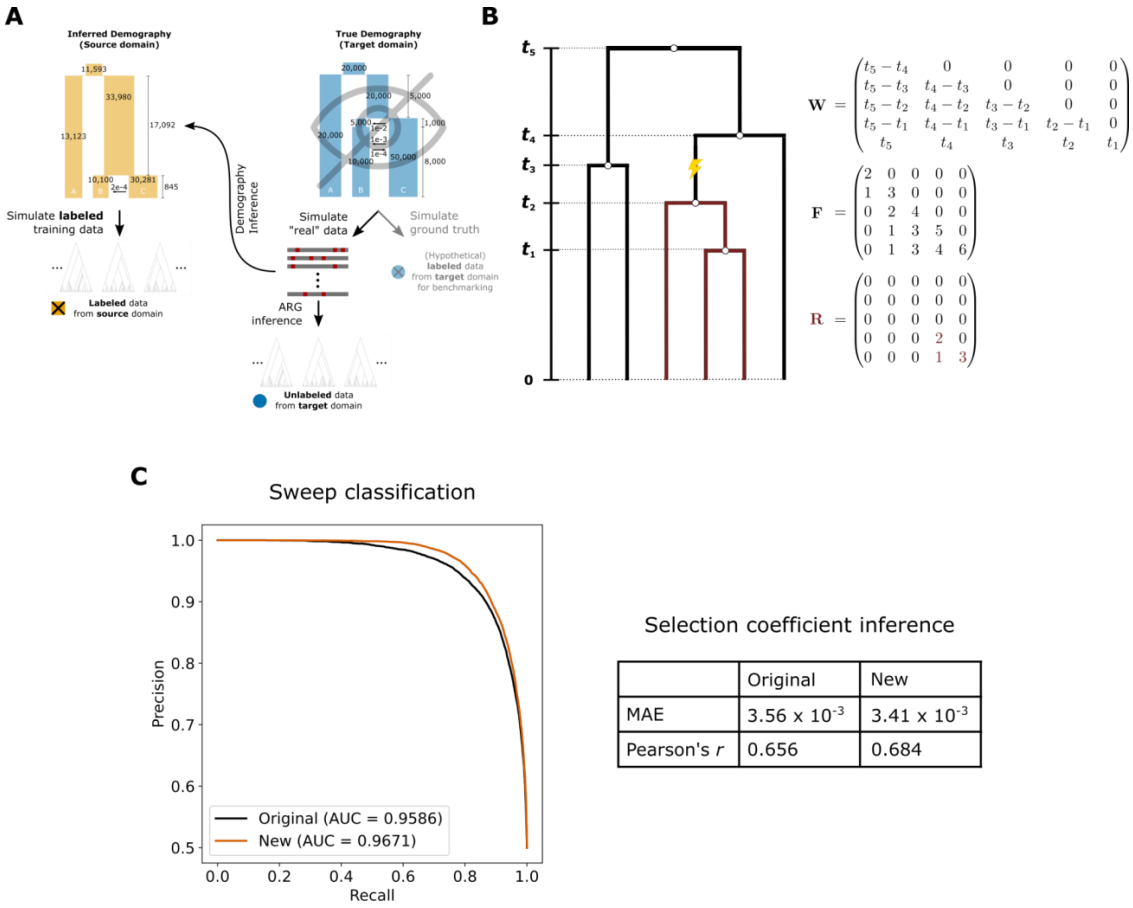
S11 Fig. Genome-wide estimates of selection coefficients and time of selection onset. Predictions are based on all the variants from the 190 scaffolds longer than 100kb.

Appendix C

Supplementary material for Chapter 4

All code used in this study are available at GitHub. The 1000 Genomes data are available online. Supplementary figures are included in this appendix.

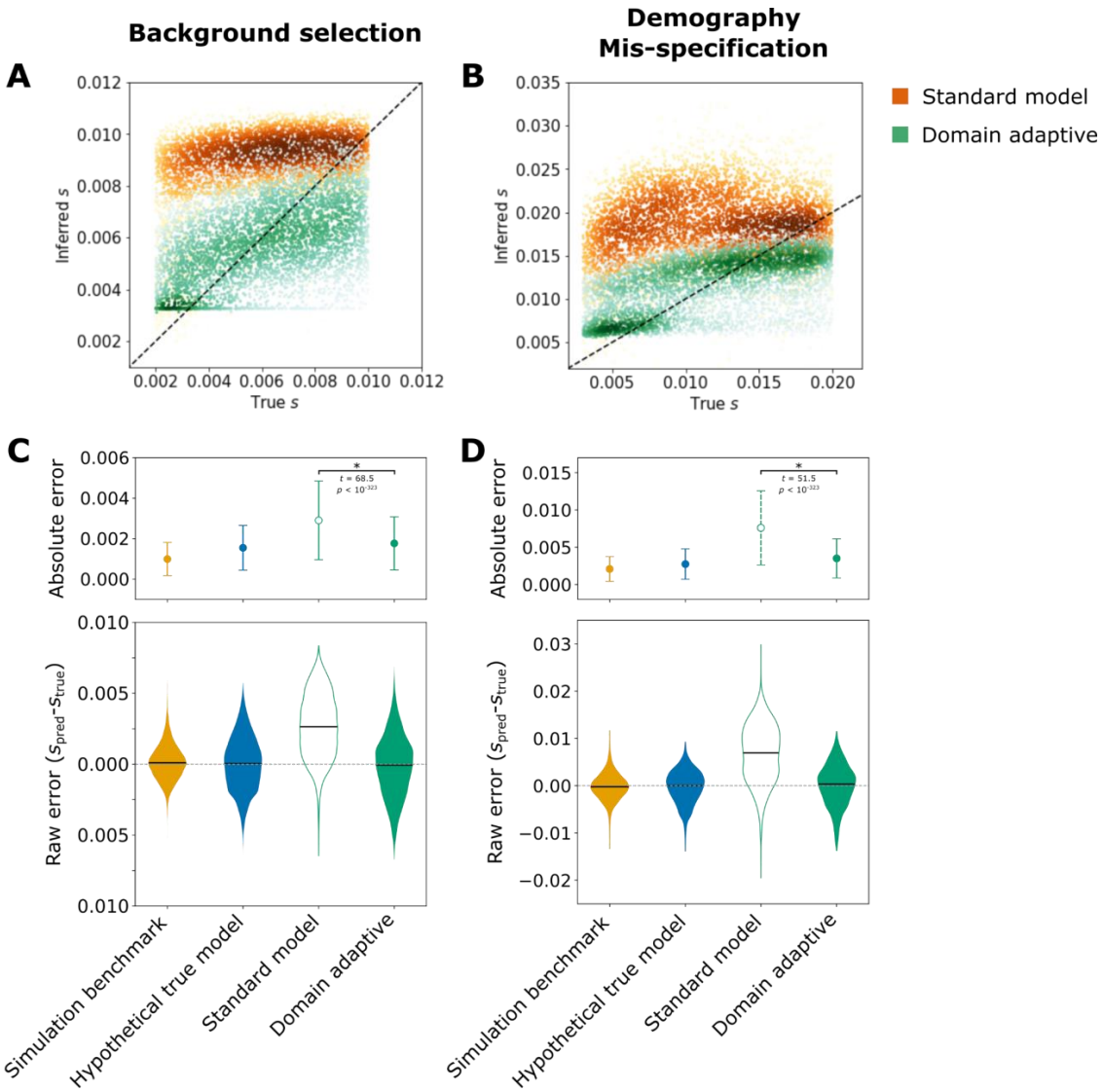
Supplementary Figures



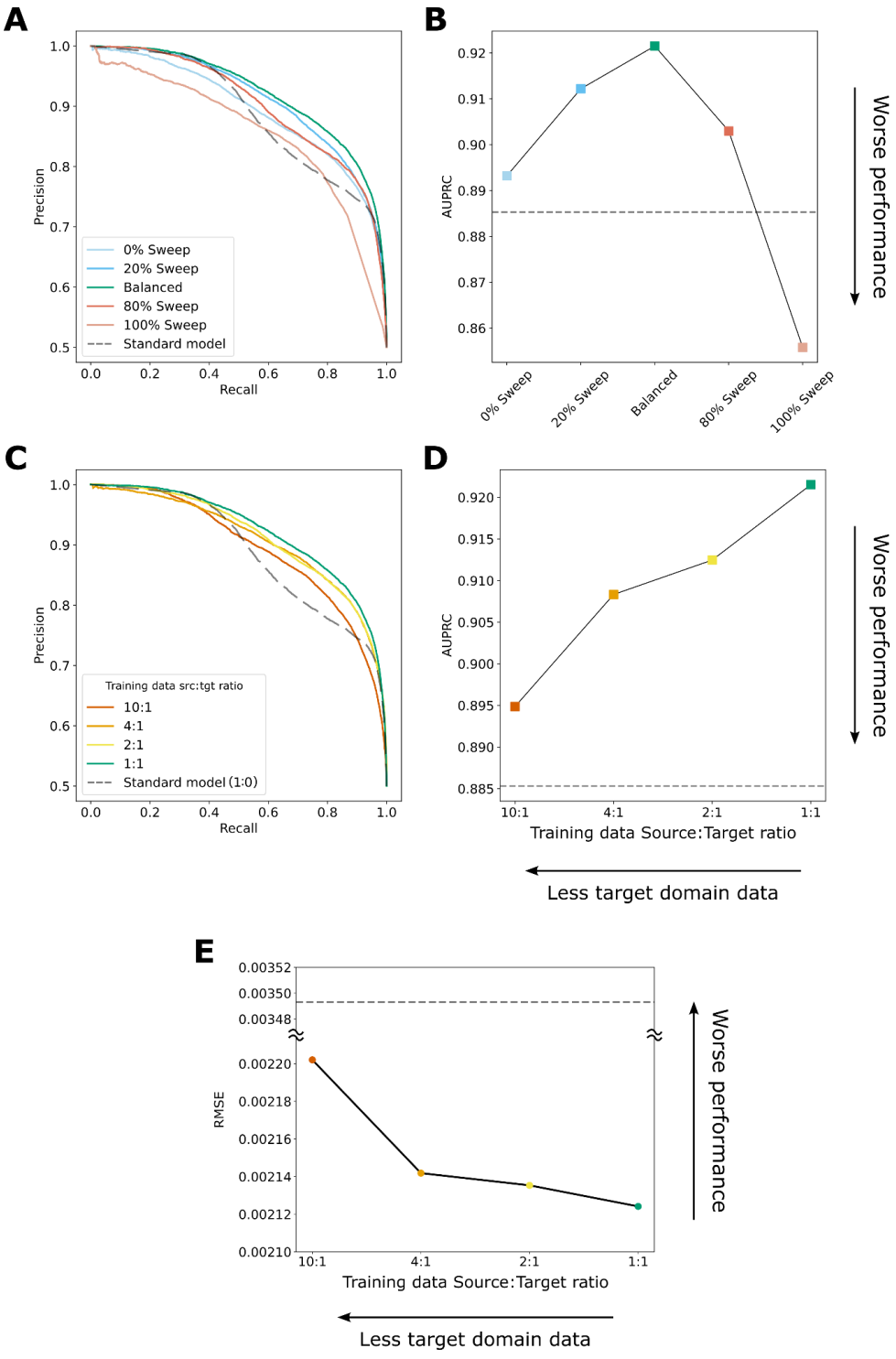
	Original	New
MAE	3.56 x 10^{-3}	3.41 x 10^{-3}
Pearson's *r*	0.656	0.684

Supplementary Figure 1. Domain-adaptive SIA. A) The workflow of a simulation study that aims to benchmark the performance of the domain-adaptive SIA model in a realistic setting of demographic mis-specification. **B)** An improved version of SIA input features that encodes the full genealogy (adapted from [59]). A genealogy with *n* taxa at a polymorphic site is uniquely encoded by three (*n*-1) x (*n*-1) lower triangular matrices. The weight matrix **W** encodes the coalescent intervals where $w_{ij} = t_{n-j} - t_{n-1-i}, \forall\, i \geq j$, and the topology matrix **F** encodes the number of lineages persistent in the coalescent intervals *corresponding* to **W** (i.e. $f_{ij} = \#\ of\ lineages\ between\ t_{n-j} and\ t_{n-1-i}, \forall i \geq j$). The derived lineage matrix **R** encodes only the subtree subtending the branch where the mutation occurred (red lightning symbol), following the same scheme as **F**. Note that the **W** matrix is a redundant encoding of the *n*-1 coalescent times (t_1,t_2, …,t_{n-1}), which contains information roughly equivalent to the original SIA input features [12]. **C)** Comparison of

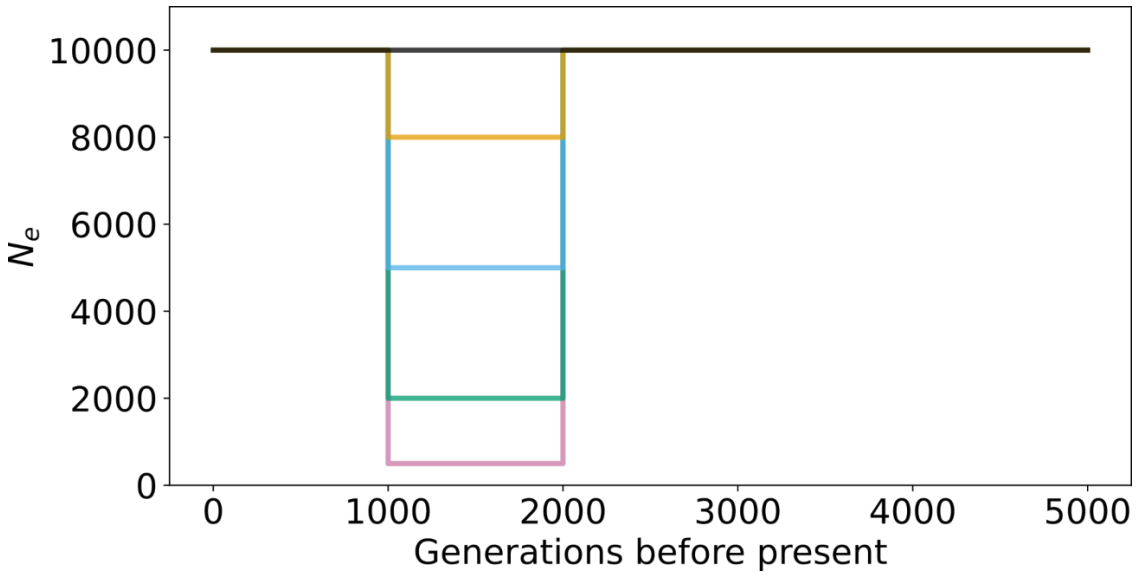
the performance of the new SIA input features in (**B**) to that of the original SIA input features.



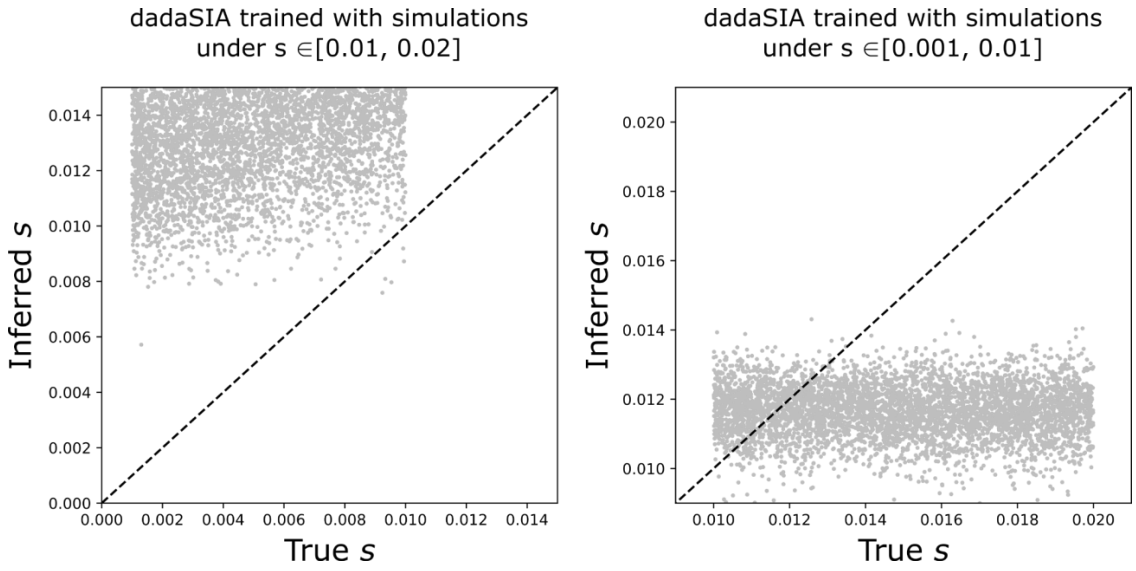
Supplementary Figure 2. Selection coefficient inference performance of SIA models. Raw data used to plot **Figs. 3B** and **3D** are presented in (**A**) and (**B**), respectively. Performance of SIA models in the simulation experiment of failure to account for background selection (**C**) and in the simulation experiment of demographic model mis-specification (**D**) is presented in terms of mean and standard deviation of the absolute error (top) as well as the distribution of raw error (bottom). Statistical significance (*) of the difference between the absolute error of the standard model and that of the domain-adaptive model is evaluated with Welch's *t*-test. See **Fig. 1C** for definition of the model labels.


A
Precision
Recall
0% Sweep
20% Sweep
Balanced
80% Sweep
100% Sweep
Standard model
B
AUPRC
0% Sweep
20% Sweep
Balanced
80% Sweep
100% Sweep
Worse performance
C
Precision
Recall
Training data src:tgt ratio
10:1
4:1
2:1
1:1
Standard model (1:0)
D
AUPRC
10:1
4:1
2:1
1:1
Training data Source:Target ratio
Worse performance
Less target domain data
E
RMSE
10:1
4:1
2:1
1:1
Training data Source:Target ratio
Worse performance
Less target domain data

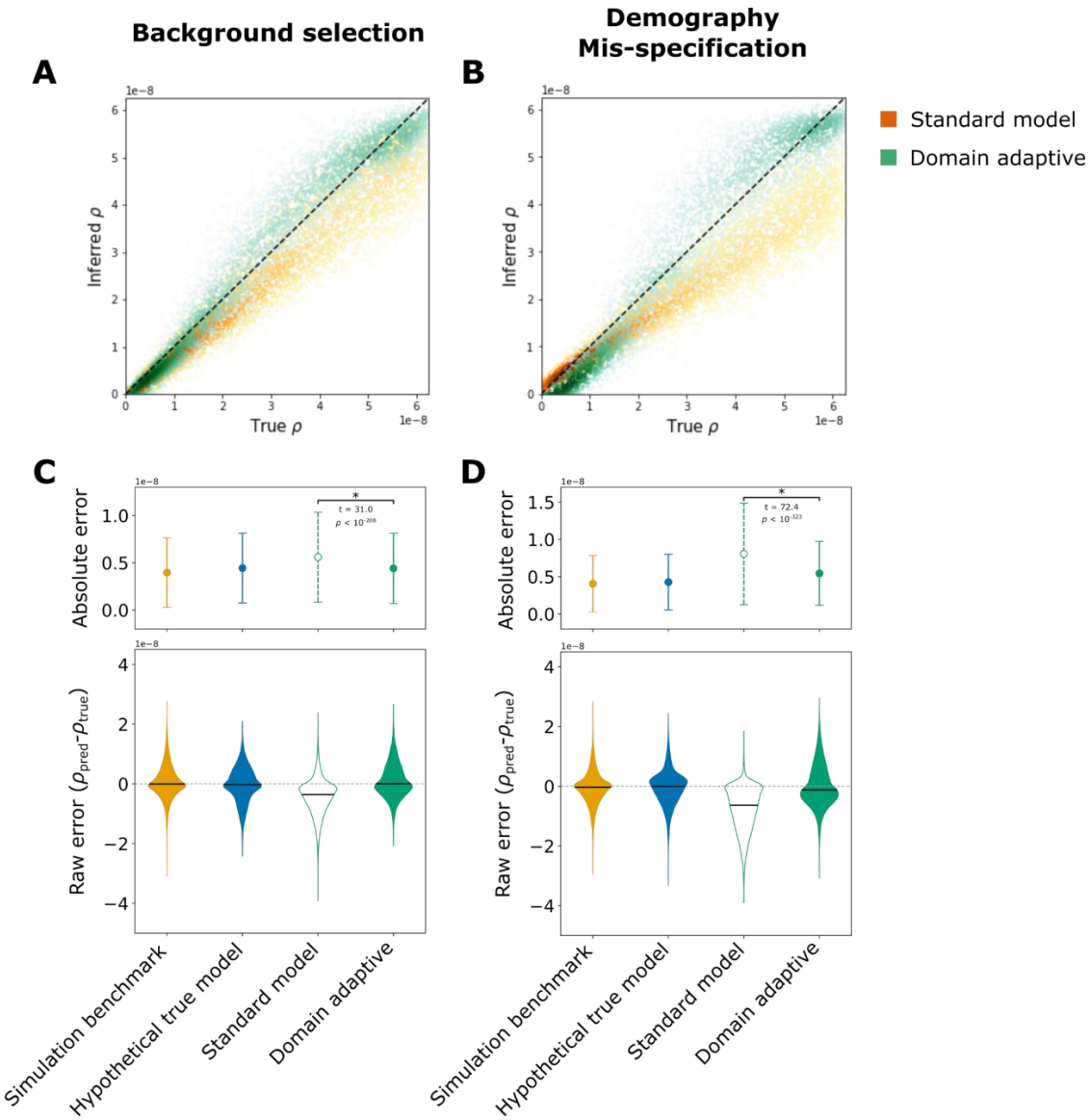
Supplementary Figure 3. Performance of dadaSIA models trained with imbalanced data. The sweep classification performance of dadaSIA models trained with different proportions of sweep vs. neutral examples in the target domain is shown in the form of precision-recall curves (**A**) and the area under precision-recall curve (AUPRC) (**B**). Note that the performance is always evaluated on a balanced test set. The performance of dadaSIA models trained with less target domain data than source domain data is shown in the form of precision-recall curves (**C**) and the values of AUPRC (**D**) for the classification task, and in the form of root mean squared error (RMSE) (**E**) for the selection coefficient inference task. The dashed lines in (**B**), (**D**) and (**E**) indicate performance of the standard model.



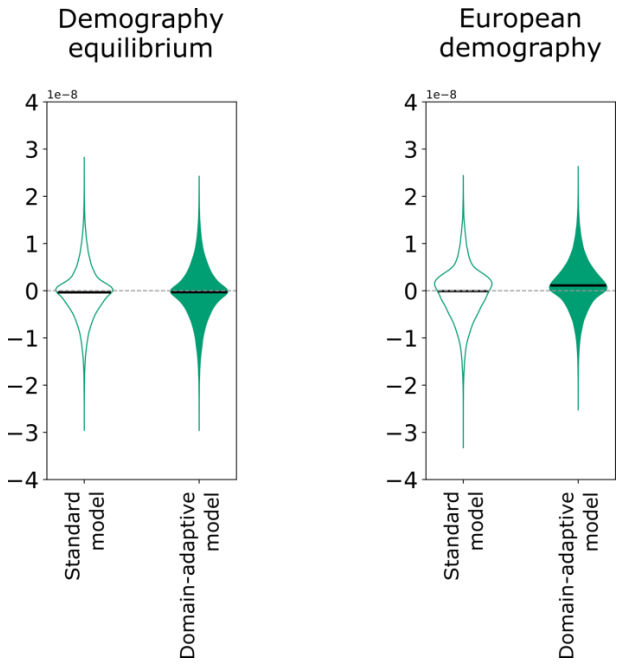
Supplementary Figure 4. Demographic mis-specification in the form of different degrees of bottlenecks tested in Fig. 5 experiments.



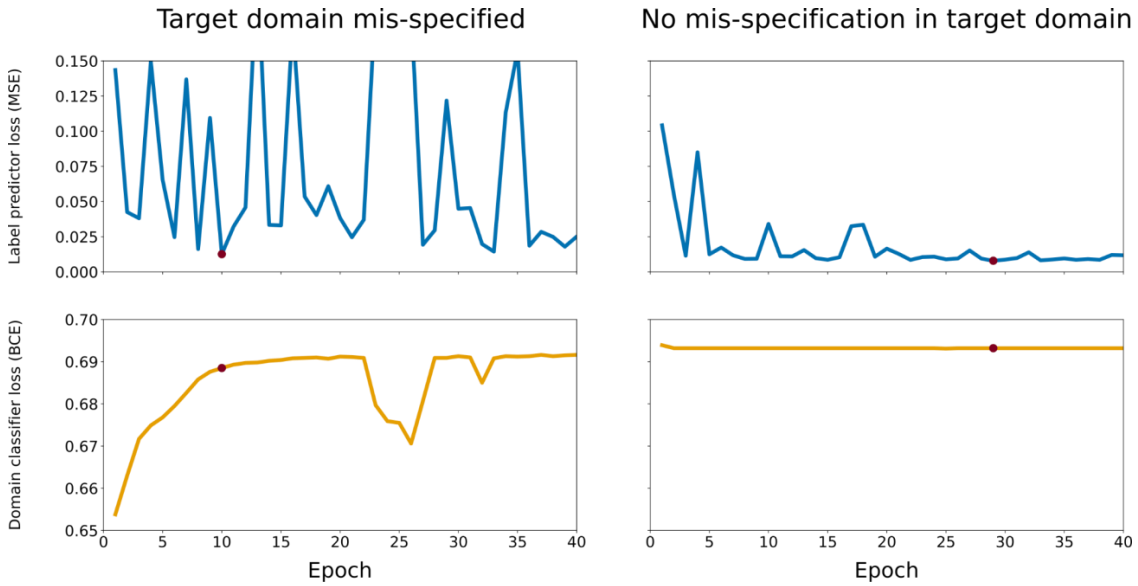
Supplementary Figure 5. **Inference of out-of-range selection coefficients in the target domain using the dadaSIA model.** The dadaSIA model trained with source domain data under $s \in [0.01, 0.02]$ failed to meaningfully infer any value lower than 0.01, even when examples of $s \in [0.001, 0.01]$ were supplied to the model as "unlabeled" target domain data, and vice versa.



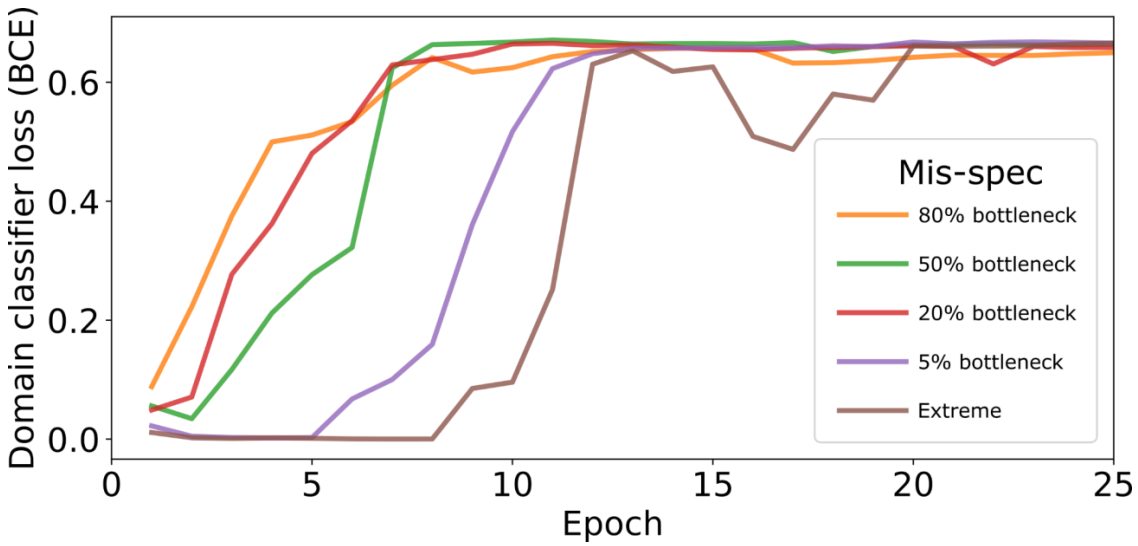
Supplementary Figure 6. **Recombination rate inference performance of ReLERNN models.** Raw data used to plot **Figs. 4A** and **4B** are presented in (**A**) and (**B**), respectively. Performance of ReLERNN models in the simulation experiment of failure to account for background selection (**C**) and in the simulation experiment of demographic model mis-specification (**D**) is presented in terms of mean and standard deviation of the absolute error (top) as well as the distribution of raw error (bottom). Statistical significance (*) of the difference between the absolute error of the standard model and that of the domain-adaptive model is evaluated with Welch's *t*-test. See **Fig. 1C** for definition of the model labels.



Supplementary Figure 7. Distribution of raw error of the ReLERNN models inferring recombination rate without simulation mis-specification. The respective mean absolute error (MAE) of the standard and domain-adaptive models are 4.05 x 10^{-9} and 4.13 x 10^{-9}, under demography equilibrium, and 4.28 x 10^{-9} and 3.93 x 10^{-9}, under a European demography. Note that the domain-adaptive model has a slight upward bias in its estimates in the case of European demography.



Supplementary Figure 8. Validation loss of the label predictor branch (mean squared error) and the domain classifier branch (binary cross entropy) over training epochs. The losses of the domain-adaptive ReLERNN models during training are plotted with and without simulation mis-specification. The red dot marks the early-stopping epoch (i.e. epoch with the lowest validation loss for the label predictor).



Supplementary Figure 9. **Domain classifier loss of dadaSIA models under different degrees of simulation mis-specification.** See **Fig. 5** and **Methods** for details of the types of mis-specification.