Inference of Locus-Specific Population Mixtures From Linked Genome-Wide Allele Frequencies

Carlos S. Reyna-Blanco[a,b,c], Madleina Caduff[a,b,c], Marco Galimberti[1], Christoph Leuenberger[e], Daniel Wegmann[a,b,*]

[a] *These authors contributed equally*
[b] *Department of Biology, Université de Fribourg, Chemin du Musée 10, CH-1700 Fribourg, Switzerland*
[c] *Swiss Institute of Bioinformatics, 1700 Fribourg, Switzerland*
[d] *Yale School of Medicine, New Haven, CT*
[e] *Department of Mathematics, Université de Fribourg, Chemin du Musée 23, CH-1700 Fribourg, Switzerland*

Summary

Admixture between between populations and species is common in nature. Since the influx of new genetic material might be either facilitated or hindered by selection, variation in mixture proportions along the genome is expected in organisms undergoing recombination. Various graph-based models have been developed to better understand these evolutionary dynamics of population splits and mixtures. However, current models assume a single mixture rates for the entire genome and do not explicitly account for linkage. Here, we introduce `TreeSwirl`, a novel method for inferring branch lengths and locus-specific mixture proportions by using genome-wide allele frequency data, assuming that the admixture graph is known or has been inferred. `TreeSwirl` builds upon `TreeMix` that uses Gaussian processes to estimate the presence of gene flow between diverged populations. However, in contrast to `TreeMix`, our model infers locus-specific mixture proportions employing a Hidden Markov Model that accounts for linkage. Through simulated data, we demonstrate that `TreeSwirl` can accurately estimate locus-specific mixture proportions and handle complex demographic scenarios. It also outperforms related D- and F-statistics in terms of accuracy and sensitivity to detect introgressed loci.

Keywords: Gene flow, Admixture, Introgression rate, Gaussian process, Linkage, Hidden Markov Model

1. Introduction

Gene flow, the exchange of genetic material between populations or different species (Slatkin, 1985a), can occur through various mechanisms, such as migration, admixture, hybridization, cross-fertilization, or even by the dispersal of diaspores and pollinators (Barton and Hewitt, 1985; Ellstrand et al., 2003; Tung and Barreiro, 2017; Burgarella et al., 2019). This exchange may play a significant role in the maintenance of genetic variation, but also in the adaptation to multiple ecological niches (Anderson, 1949; Slatkin, 1985b, 1987; Rieseberg and Wendel, 1993; Barton, 2001). At sufficient levels, gene flow can lead to homogenization

*Corresponding author
Email address: `daniel.wegmann@unifr.ch` (Daniel Wegmann)

of populations, particularly in the face of opposing genetic drift (Ellstrand, 2014). Gene flow might also increase genetic variation at a much higher rate than mutation (Grant and Grant, 1994) and impact the process of speciation by becoming a primary source of genetic diversity and adaptive novelty for a population (Ellstrand et al., 2003; Abbott et al., 2013). Several genetic analyses have shown that gene flow, both ancient and present, is a common phenomenon in nature (Grant and Grant, 1992; Mallet, 2005; Patterson et al., 2006; Tung and Barreiro, 2017), and a bifurcating tree, representing population or species historical relationships, fails to account for it (Kulathinal et al., 2009; Reich et al., 2009; Sousa et al., 2009; Green et al., 2010; Durand et al., 2011; Reich et al., 2012). This led to the development of methods that use allele-frequency data and graph-based models to infer population splits and test for the presence of gene flow between divergent populations or species (Pickrell and Pritchard, 2012; Patterson et al., 2012; Yang et al., 2012; Eaton and Ree, 2013; Lipson et al., 2013, 2014; Martin et al., 2013; Kozak et al., 2021), which, for instance, confidently settled the long-standing question whether gene flow occurred between modern humans and archaic hominins. However, these methods assume a genome-wide gene flow rate per migration edge, which is unrealistic in the presence of selection. In theory, the effective gene flow may vary significantly along the genome because of selection and genetic drift (Yamamichi and Innan, 2012), making it essential to quantify these variations to better understand the dynamics that lead to introgression (Racimo et al., 2015, 2017; Suarez-Gonzalez et al., 2018; Sankararaman, 2020).

Introgression is a lasting consequence of gene flow that leads to the assimilation of variants into the local gene pool through repeated back-crossing, resulting in their permanent inclusion (Anderson and Hubricht, 1938). When introgressed loci increase the fitness of the recipient population, this is known as "adaptive introgression". Unlike neutral introgression, which can be lost over time due to drift, adaptive introgression is sustained by selection and can eventually lead to fixation (Zhang et al., 2021). The classic way to identify introgressed loci is by using population genetic summary statistics. Patterson's D, for example, has been estimated in sliding windows along the genome to identify introgressed loci (Dasmahapatra et al., 2012; Kronforst et al., 2013; Smith and Kronforst, 2013; Rheindt et al., 2014; Fontaine et al., 2015). Since it was originally intended for genome-wide analysis (Martin et al., 2015), more suitable related statistics have been used for analyzing specific short genomic regions, such as f_d , f_{dM}, and d_f (Martin et al., 2015; Malinsky et al., 2015; Pfeifer and Kapan, 2019; Malinsky et al., 2021). There are other statistics, for instance, S* and its variants that use linkage disequilibrium information to detect long introgressed haplotypes (Plagnol and Wall, 2006; Wall et al., 2009; Vernot and Akey, 2014; Vernot et al., 2016; Browning et al., 2018) or ArchIE that combines diverse summary statistics to detect introgressed haplotypes without a reference (Durvasula and Sankararaman, 2019, 2020). However, outlier scans based on such statistics are likely to ignore valuable information present in the full data, do not model linkage explicitly or require an arbitrary choice of large window-size and outliers identification. To overcome these constraints, probabilistic frameworks such as Hidden Markov Models (HMMs) (Rabiner and Juang, 1986; Prüfer et al., 2014; Seguin-Orlando et al., 2014; Skov et al., 2018; Steinrücken et al., 2018), and Conditional Random Fields (CRF) (Sankararaman et al., 2014) have been applied to infer the ancestry state of each site. These methods are extensions of models that infer local ancestry from genotyping data (Tang et al., 2006; Price et al., 2009; Wegmann et al., 2011; Lawson et al., 2012; Maples et al., 2013) and while explicitly accounting for demographic history and linkage, they rely

on phased and training sequence data, unadmixed or archaic reference, and detailed demographic models. As a consequence, such approaches are not easily applicable to non-model species for which more limited data and knowledge is available.

To complement these methods, we here propose a model that makes use of Gaussian processes to infer locus-specific mixture proportions. Gaussian processes have a rather long history to model allele frequency differences between populations (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1981), but have recently seen a surge in applications due to the development of the popular tool `TreeMix` (Pickrell and Pritchard, 2012). Our method, `TreeSwirl`, explicitly takes an admixture graph (e.g. inferred by TreeMix) and genome-wide allele frequencies to infer locus-specific mixture proportions. To account for linkage, we make use of a Hidden Markov Model (HMM), wherein the hidden states are represented by the proportion of the mixture at a particular site and the observed data is represented by the sampled allele frequencies. To evaluate the performance of our method against other tools, we simulated data using various demographic models. We estimated the mixture proportions with `TreeSwirl` and computed related D- and F-statistics using D-suite Dinvestigate (Malinsky et al., 2021). Our findings revealed that `TreeSwirl` surpasses the summary statistics estimates in detecting the simulated signal of introgression under different scenarios, although at an additional computational cost. Furthermore, by appling `TreeSwirl` to real data cases, we succesfully identified candiate genomic regions where migration rates fluctuate and may be subject to selection.

2. Materials and Methods

2.1. The Model

Consider a set of populations $m = 1, 2, ..., M$ that are linked by a graph $\mathcal{G}$ which represents their population history in terms of population splits and migration events. Consider as well a series of diploid, bi-allelic loci $l = 1, 2, ..., L$, where the total number of loci L might constitute, for instance, consecutive SNPs along the genome. At each locus l, a total number of $\boldsymbol{N}_l = (N_{l1}, \ldots, N_{lM})$ alleles have been observed across the M populations, of which $\boldsymbol{n}_l = (n_{l1}, \ldots, n_{lM})$ were derived and the remaining ancestral (or otherwise polarized). To model sampled allele counts $\boldsymbol{n}_l|\boldsymbol{N}_l$ we distinguish two processes: the first models the distribution of the vector of the actual but unknown population frequencies $\boldsymbol{y}_l = (y_{l1}, \ldots, y_{lM})'$ given the graph $\mathcal{G}$, and the second the distribution of the sampled allele counts $\boldsymbol{n}_l|\boldsymbol{N}_l$ given $\boldsymbol{y}_l$ (Fig 1A).

2.1.1. Evolution along the graph $\mathcal{G}$

We assume, as in (Pickrell and Pritchard, 2012), that the change in allele frequencies from the root to the tips of $\mathcal{G}$ is modeled as a Brownian motion (BM) process. For each locus l, the BM process starts at the root of $\mathcal{G}$ at a value of allele frequency which we denote by ν_l. It proceeds along the branches of $\mathcal{G}$ and finally gives rise to the above-mentioned random vector $\boldsymbol{y}_l$ at the leaves of $\mathcal{G}$. The probability of $\boldsymbol{y}_l$ is given by the multivariate normal density

$$\pi(\boldsymbol{y}_l|\nu_l, \mathcal{G}) = \mathcal{N}\left(\boldsymbol{\nu}_l, \boldsymbol{V}(\nu_l)\right),$$

where $\boldsymbol{\nu}_l = (\nu_l, \ldots, \nu_l)'$ is the mean vector and $\boldsymbol{V}(\nu_l)$ is the variance-covariance matrix corresponding to the BM on $\mathcal{G}$. For the construction of $\boldsymbol{V}(\nu_l)$, which depends on the topology

of $\mathcal{G}$ and and the migration rates, we follow (Pickrell and Pritchard, 2012). We set

$$\boldsymbol{V}(\nu_l) = \nu_l(1-\nu_l)\boldsymbol{W}_l, \tag{1}$$

where $\boldsymbol{W}_l$ only depends on the tree topology, the branch lengths and the migration rates.

However, it was long recognized that BM with constant variance is not adequately describing allele frequency changes, especially close to boundaries and various transformations to alleviate the problem have been proposed (Felsenstein, 1981). Here we will consider the transformation

$$\mu_l = \arcsin(2\nu_l - 1) \tag{2}$$

from the interval $[0,1]$ onto $[-\pi/2, \pi/2]$. This has the advantage that all factors of $\nu_l(1-\nu_l)$ in front of the variance matrices will be canceled. We thus replace (eq. 1) by

$$\boldsymbol{W}_l = \left(\frac{\mathrm{d}\mu_l}{\mathrm{d}\nu_l}\right)^2 \boldsymbol{V}(\nu_l). \tag{3}$$

Let $\boldsymbol{x}_l = (x_{l1}, \ldots, x_{lM}), x_{lm} = \arcsin(2y_{lm} - 1)$ denote the transformed population allele frequencies. The distribution of $\boldsymbol{x}_l$ thus follows the multivariate normal density

$$\pi(\boldsymbol{x}_l|\mu_l, \mathcal{G}) = \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{W}_l) \tag{4}$$

with $\boldsymbol{\mu}_l = (\mu_l, \ldots, \mu_l) = \mu_l\mathbf{1}$.

The matrix $\boldsymbol{W}_l$ is constructed as follows. Let $\mathcal{T}$ be a rooted population tree with K oriented branches $k = 2, \ldots, K$ of length c_k; the orientation of the branches points in direction of the leaves. We assume that the tree also contains I oriented migration edges τ_i, $i = 1, \ldots, I$, to which we assign no branch length. The migration edges should be placed such that there are no cycles in the tree. We now consider paths leading from the root of the tree to a leaf taking some of the migration edges (open edges) and leaving others out (closed edges). More precisely, let

$$\boldsymbol{b} = (b_1, \ldots, b_I)$$

be a binary vector indicating a certain configuration of open and closed migration edges: a bit $b_i = 1$ indicates that the migration edge τ_i is open and $b_i = 0$ that the migration edge τ_i is closed (Fig 1B). We denote by w_{li} the migration rate, i.e. the probability of edge τ_i to be open, and thus we assign to the configuration $\boldsymbol{b}$ the probability

$$w_l(\boldsymbol{b}) = \prod_{i=1}^{I} w_{li}^{b_i}(1-w_{li})^{1-b_i}. \tag{5}$$

Now, for a given configuration $\boldsymbol{b}$, pick a population (leaf) m and a branch k. There is at most one path leading from the root to the population m and taking exactly the open migration edges according to $\boldsymbol{b}$. If, moreover, this path contains the branch k, we set the indicator function $I_{mk}(\boldsymbol{b})$ equal to 1. Otherwise we set $I_{mk}(\boldsymbol{b}) = 0$.

Using this notation, we can now define the $M \times M$-matrices $\boldsymbol{J}_{lk}$ for each branch k element-wise by

$$[\boldsymbol{J}_{lk}]_{mn} = \sum_{\boldsymbol{b}} w_l(\boldsymbol{b}) I_{mk}(\boldsymbol{b}) \sum_{\boldsymbol{b}'} w_l(\boldsymbol{b}') I_{nk}(\boldsymbol{b}'), \tag{6}$$

where each sum runs over all the 2^I possible configurations of $\boldsymbol{b}$ and $\boldsymbol{b}'$, respectively. Each matrix $\boldsymbol{J}_{lk}$ thus reflects the probabilities that branch k was common for any pair of leaves.

The matrix $\boldsymbol{W}_l$, after all, is given by

$$\boldsymbol{W}_l(\boldsymbol{w}) = \sum_{k=1}^{K} c_k \boldsymbol{J}_{lk}. \tag{7}$$

This construction of the variance matrix $\boldsymbol{W}_l(\boldsymbol{w}_l)$ is a generalized reformulation of an argument given in (Pickrell and Pritchard, 2012).

To unclutter the notation, we will use $\boldsymbol{W}_l = \boldsymbol{W}_l(\boldsymbol{w}_l)$ in the rest of this article and thus not indicate its dependence on the migration rates $\boldsymbol{w}_l = (w_{l1}, \ldots, w_{lI})$.

2.1.2. Sampling

We assume that the observed allele counts $\boldsymbol{n}_{lm}$ at locus l and population m follow a binomial distribution with parameters N_{lm} and y_{lm}, where y_{lm} is the true allele frequency in population m. By independence of the samples, we have

$$\pi(\boldsymbol{n}_l|\boldsymbol{y}_l) = \prod_{m=1}^{M} \mathrm{Bin}(n_{lm}|N_{lm}, y_{lm}). \tag{8}$$

If the sample sizes are sufficiently large, we can approximate this distribution by a multivariate density. Let $\mathbf{f}_l = (f_{l1}, \ldots, f_{lM})$ with $f_{lm} = n_{lm}/N_{lm}$ denote the observed allele frequencies at locus l, which are approximately normally distributed with with mean $\boldsymbol{y}_l$ and a diagonal variance-covariance matrix:

$$\mathrm{diag}\left[\frac{y_{l1}(1-y_{l1})}{N_{l1}}, \ldots, \frac{y_{lM}(1-y_{lM})}{N_{lM}}\right]. \tag{9}$$

The transformed observed allele frequencies $\mathbf{d}_l = (d_{l1}, \ldots, d_{lM})$ with $d_{lm} = \arcsin(2f_{lm} - 1)$, are then approximated by a the multivariate density

$$\pi(\mathbf{d}_l|\boldsymbol{x}_l) \approx \mathcal{N}(\boldsymbol{x}_l, \boldsymbol{\Sigma}_l) \tag{10}$$

with

$$\boldsymbol{\Sigma}_l = \mathrm{diag}\left[\frac{1}{N_{l1}}, \ldots, \frac{1}{N_{lM}}\right]$$

because the factors $y_{l1}(1 - y_{l1})$ are transformed away from the variance-covariance matrix (eq. 9) similar to (eq. 3).

2.1.3. Full likelihood for one locus

Given the ancestral frequency μ_l, we obtain the likelihood by combining (eq. 4) and (eq. 10) and integrating out:

$$\pi(\mathbf{d}_l|\mu_l, \mathcal{G}) = \int \pi(\mathbf{d}_l|\boldsymbol{x}_l)\pi(\boldsymbol{x}_l|\mu_l, \mathcal{G})\mathrm{d}\boldsymbol{x}_l. \tag{11}$$

Using well-known formulae for linear systems (see Thm. 4.4.1 in (Murphy, 2012)) we obtain for the likelihood (eq. 11) the following approximation:

$$\pi(\mathbf{d}_l|\mu_l, \mathcal{G}) \approx \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l + \boldsymbol{W}). \tag{12}$$

We now set a normal prior on μ_l, namely we assume that

$$\pi(\mu_l) = \mathcal{N}(\mu, \sigma^2).$$

Again from Thm. 4.4.1 in (Murphy, 2012) we conclude that

$$\pi(\mathbf{d}_l|\mu_l, \sigma^2, \mathcal{G}) = \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{S}_l) \tag{13}$$

with

$$\boldsymbol{\mu}_l = \mu_l \mathbf{1}, \quad \boldsymbol{S}_l = \boldsymbol{\Sigma}_l + \boldsymbol{W}_l + \sigma^2 \mathbf{1}\mathbf{1}'. \tag{14}$$

Explicitly

$$\pi(\mathbf{d}_l|\mu_l, \sigma^2, \mathcal{G}) \approx \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{S}_l|}} \exp\left[-\frac{(\mathbf{d}_l - \mu_l \mathbf{1})' \boldsymbol{S}_l^{-1} (\mathbf{d}_l - \mu_l \mathbf{1})}{2}\right]. \tag{15}$$

2.2. Hidden Markov Model

We develop a Hidden Markov Model (HMM) for multiple loci $l = 1, \ldots, L$ with varying migration rates for each of the I migration edges of graph $\mathcal{G}$. We assume that the locus and specific migration rates w_{li} take values out of a small set of discrete numbers between 0 and 1:

$$w_{li} \in \{w_{i1}, w_{i2}, \ldots, w_{iJ_i}\}.$$

We thus have $J_1 \cdot J_2 \cdot \ldots \cdot J_I$ possible combinations and these combinations will constitute the hidden states of our Markov model. We denote the hidden state at locus l by z_l. Each state z_l corresponds to a multiindex

$$j = (j_1, j_2, \ldots, j_I)$$

that defines the migration values $(w_{1j_1}, \ldots, w_{Ij_I})$ of the migration edges. Thus, knowing the state z_l is tantamount to knowing the combination of migration rates at the given site which in turn determines the matrix $\boldsymbol{W}$ in eq. (eq. 7) via (eq. 5) and (eq. 6).

To account for linkage between loci, we assume that the locus-specific transition matrix $\mathbb{P}(z_l = j'|z_{l-1} = j)$ is based on physical or genetic distances δ_l between loci. We assume independence of the transition probabilities of the different migration edges:

$$\mathbb{P}(z_l = j'|z_{l-1} = j) = \mathbb{P}_l(j, j') = \prod_{i=1}^{I} \boldsymbol{P}_{li}(j_i, j_i').$$

Each one of the factors in this product is an element of a a ladder-type Markov matrix $\boldsymbol{P}_{li}$ which is defined via a transition rate matrix $\kappa_i \boldsymbol{\Lambda}_i$:

$$\boldsymbol{P}_{li} = e^{\delta_l \kappa_i \boldsymbol{\Lambda}_i}. \tag{16}$$

Here, κ_i is a positive scaling parameter pertaining to migration edge i, the distances δ_l are known constants corresponding to the linking distances.

Further, the $J_i \times J_i$-matrices $\mathbf{\Lambda}_i$ reflect a transition model similar to that of (Galimberti et al., 2020), which is governed by an attractor state $a_i \in \{w_{i1}, \ldots, w_{iJ_i}\}$ reflecting the background migration rate and two parameters ϕ_i and ζ_i describing the number of loci deviating from the attractor state and the degree of that deviation, respectively (see Galimberti et al., 2020, for an illustration). Specifically, we have

$$\mathbf{\Lambda}_i = \begin{pmatrix} -1 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ \zeta_i & -1-\zeta_i & 1 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & \zeta_i & -1-\zeta_i & 1 & \ldots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & -1-\zeta_i & \zeta_i & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -1-\zeta_i & \zeta_i \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & 1 & -1 \end{pmatrix}$$

with the attractor row given by

$$\begin{pmatrix} 0 & \ldots & 0 & \phi_i\zeta_i & -2\phi_i\zeta_i & \phi_i\zeta_i & 0 & \ldots & 0 \end{pmatrix}. \tag{17}$$

See supplementary text for some examples.

Note that the κ_i, ϕ_i and ζ_i all must be strictly positive. However, we limit ϕ_i and ζ_i to the range (0,1] to ensure that the stationary probability of the attractor state a_i is higher than for any other state.

We can also easily define a transition rate matrix that does not depend on an attractor state a_i and the parameters, ϕ_i and ζ_i. This can be done as follows:

$$\mathbf{\Lambda}_i = \begin{pmatrix} -1 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -1 \end{pmatrix}$$

Note that this simplifies the inference of transition probabilities, as they now depend solely on the scaling factor κ_i rather than four parameters (κ_i, a_i, ϕ_i, and ζ_i). As a result, instead of using a Nelder-Mead (Nelder and Mead, 1965) optimization to maximize the Q-function of the transition probabilities, it is now feasible to numerically solve it with a linear search. This approach could be more realistic for certain cases where there is no clear background migration rate.

Finally, the emission probabilities are generated via the marginal likelihood (eq. 15):

$$\mathbb{P}(\mathbf{d}_l | z_l = j) = \pi(\mathbf{d}_l | \mu_l, \sigma^2, \mathcal{G}_j), \tag{18}$$

where $\mathcal{G}_j$ denotes the population graph with migration rates according to the state $z_l = j$ and μ_l is the root state at site l.

2.3. Inference

We developed an empirical Bayes inference scheme for the hidden states under the assumption that the topology of the admixture graph is either known or was previously obtained. Specifically, we first infer both the emission and transition probabilities using the Baum-Welch algorithm (Baum et al., 1970) and then posterior state probabilities under the inferred parameters. As detailed in the Supplementary Information (see section "Baum-Welch"), the Baum-Welch algorithm requires numerical optimization in each iteration. While the parameter of the root prior μ can be optimized analytically, we resort to Newton-Raphson optimization (Nocedal and Wright, 2006; Lange, 2010) for the root prior σ_2 and for parameters of the population graph (i.e. the branch lengths $c_i, \ldots, c_K$) and to Nelder-Mead optimization (Nelder and Mead, 1965) for the parameters regarding the transition matrices with attractors (i.e. the κ_i, ϕ_i, ζ_i and a_i) or a linear search for transition matrices with no attractors (i.e. the κ_i).

The Baum-Welch algorithm may be sensitive to initial conditions. We obtain initial estimated of all parameter values as follows (see Supplementary Information for more details):

1. We use the observed variance-covariance matrix of the transformed observed frequencies as an initial guess of the variance covariance matrix $\boldsymbol{W}$.
2. To account for variation in $\boldsymbol{W}$ among loci, we refine this initial estimates using a Gaussian Mixture Model (GMM) under which the transformed observed frequencies are modeled by one of $q = 1, \ldots, Q$ multi-variate Gaussian distributions with variance-covariances matrices $\boldsymbol{W}_q$ but shared root priors μ and σ_2. This model assumes no constraints regarding the structure of the $\boldsymbol{W}_q$ and can be optimized with an Expectation-Maximization (EM) algorithm with analytic updates.
3. We next use a Nelder-Mead algorithm to coerce the inferred variance-covariance matrices $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_Q$ onto the population graph. Specifically, we seek to find the set of branch lengths $\boldsymbol{c}_1 ... \boldsymbol{c}_k$ and partition-specific migration rate $\boldsymbol{w}_q$ that best explain the previously learned variance-covariance matrices using the weighted Residuals Sum of Squares.
4. To initialize the transition parameters, we first determine the posterior mean state p_{il} for each each migration edge i and locus l under uniform priors and the above learned branch lengths and root prior. We then infer the transition parameters κ_i, ϕ_i, ζ_i and a_i using a simplified HMM that models the p_{il} using beta distributions.

Despite this initialization, we noticed that the Baum-Welch algorithm may settle on a non-optimal attractor state a_i too early. After initial convergence of the algorithm we therefore check if some neighboring attractor states may lead to a higher likelihood when allowed a few additional Baum-Welch iterations.

Once maximum likelihood estimates for the branch lengths $c_i, \ldots, c_K$, the transition parameters κ_i, ϕ_i, ζ_i and a_i as well as the root prior μ and σ_2 are obtained, we infer state posterior probabilities $P(z_l|\mathbf{d}, \boldsymbol{\theta})$ given the full data $\mathbf{d}$ and the learned parameters collectively denoted by $\boldsymbol{\theta}$, see Fig 1C. We further determined the posterior mean migration rates as

$$\bar{w}_{il} = \sum_j w_{ij_i} \mathbb{P}(z_l = j|\mathbf{d}, \boldsymbol{\theta}). \tag{19}$$

To identify candidate regions under selection, i.e. exhibiting either excess or dearth introgression compared to the genome-wide average, we summarized these posterior probabilities as

$$\mathbb{P}(z_l > a_i|\mathbf{d}, \boldsymbol{\theta}) = \sum_j \mathcal{I}(j_i > a_i)\mathbb{P}(z_l|\mathbf{d}, \boldsymbol{\theta}),$$
$$\mathbb{P}(z_l < a_i|\mathbf{d}, \boldsymbol{\theta}) = \sum_j \mathcal{I}(j_i < a_i)\mathbb{P}(z_l|\mathbf{d}, \boldsymbol{\theta}),$$

where $\mathcal{I}(\cdot)$ denotes the indicator function. We then determined for each locus l the false discovery rates (FDR) for excess ($q_e(l)$) and dearth ($q_d(l)$) introgression as

$$q_e(l) = 1 - \mathbb{P}(z_l > a_i|\mathbf{d}, \boldsymbol{\theta}),$$
$$q_d(l) = 1 - \mathbb{P}(z_l < a_i|\mathbf{d}, \boldsymbol{\theta}).$$

2.4. Implementation

We implemented the proposed inference scheme as a user-friendly `C++` program `TreeSwirl`, which is available, along with documentation, through a git repository at `https://bitbucket.org/wegmannlab/treeswirl`.

To streamline computations, we employ a straightforward clustering method to reduce the number of sampling size variance matrices Σ_l that need to be considered to either a default or user-specified number, following these steps:

1. Sort the vector of sample sizes according to the frequency of each occurrence.
2. To cluster, identify the pair of vectors with the least occurrences and compute their weighted average.
3. Retain the weighted vector of sample sizes, remove the pair, and update the occurrence count as the sum of the deleted pair counts.
4. Repeat steps 1 through 3 until the desired number of Σ_l is obtained.

Given a limited number u of such matrices and given that we use a finite number of discrete migration rates, there exist also an only finite number of matrices $\boldsymbol{S}_l$ that can be pre-computed in each Baum-Welch iteration to speed up the forward-backward pass through the HMM.

2.5. Simulations

2.5.1. `fastsimcoal2`

To compare `TreeSwirl` to competing methods, we used `fastsimcoal2` (Excoffier et al., 2021) to simulate genomic data under five different demographic scenarios only consisting of population splits and admixture pulses (but no population growth or continuous migration, Figure 2). We maintained a constant effective population size of $N_e = 10,000$ and used a sample size of $N = 100$ for each population in all cases.

To simulate variation in admixture pulses along chromosomes, we composed each chromosome of seven blocks, each containing many independent loci of length 1000 bp, fully-linked (i.e. within-locus recombination rate of 0.0), a mutation rate of $1e-8$, and a transition rate of 0.33. Odd-numbered blocks reflected the neutral genomic background, each contained $n_n = 3,500$ loci and an admixture pulse of $\alpha_n = 0.05$. Conversely, even-numbered blocks

reflected loci under selection. While all three selected blocks shared parameters in one simulation, we varied the number of loci n_s and migration rates α_s across different simulations.

We generated 10 replicates for each parameter combination and used a custom script to transform the generated output files into standard VCF files and concatenating the seven blocks corresponding to a single chromosome. We then applied a minimum allele frequency filter of $maf = 0.05$ with VCFtools (Danecek et al., 2011). These filtered VCFs served as input for estimating sliding window F_{st} for simulated data only consisting of two or three populations as well as for running `D-suite Dinvestigate` (Malinsky et al., 2021) with varying window sizes $s = (10, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500)$, a sliding locus of 1, and the true trio and corresponding outgroup for demographic scenarios with more than three populations. Concurrently, we executed `TreeSwirl` using the same filtered data and the expected tree topology.

We employed a receiver operating characteristic (ROC) curve analysis to assess the area under the curve (AUC), which summarizes the performance of the method in distinguishing introgression from the "neutral state" ix. For the ROC analysis, we used the estimated mean posteriors obtained from `TreeSwirl`, along with the computed values of F_{st}, Patterson's D, f_d, f_{dM}, and d_f for various window sizes. For each comparison, we then used the statistics and window-size that resulted in the best AUC.

2.6. Data Processing

2.6.1. Anopheles gambiae species complex

We downloaded the mosquito dataset from `https://datadryad.org/stash/dataset/doi:10.5061/dryad.f4114`. The VCF file contains data for chromosome 3La, encompassing eight populations and a total of 71 samples. When converting the data into allele counts, we excluded the *Anopheles gambiae* and *Anopheles coluzzii* populations and only kept sites at which all populations had data and with a minimum allele frequency (maf) of 0.05. The resulting dataset consisted of 295,017 SNPs across six populations with a total of 37 samples. The admixture graph was derived from Figure 1C in Fontaine et al. (2015).

2.7. Data Availability Statement

No new data were generated or analysed in support of this research.

3. Results

3.1. Comparison to related D- and F-statistic methods

We used `fastsimcoal2` to extensively generate coalescent simulations from five demographic histories of population splits and mixtures. The simulation parameters were chosen to be reasonable. We used an effective population size of $N_e = 10,000$, a sample size of $N = 100$ and a shared common ancestor for all populations dating back approximately 2000 generations (Fig 2, first column). Each simulated chromosome involved seven genomic blocks with variable lengths and migration rates. To evaluate our method, we applied `TreeSwirl` to the simulated data derived from the five models with distinct histories (Fig 2, second column), estimated F_{st} and computed summary statistics using `D-suite Dinvestigate` for all applicable simulation scenarios.

Compared to the best-performing summary statistic and window size, `TreeSwirl` demonstrated a higher power to estimate introgression across all simulations. Our method showcases higher sensitivity and specificity, allowing for the identification of a greater number of true introgressed loci while maintaining an exceptionally low false-positive rate (Fig 2, third and fourth column). The underperformance of related D- and F-statistic methods may be attributed to the effect of recombination, as our simulations assume no recombination. It has been reported that these methods are more accurate as recombination rate increases, which can be explained by the growth in the number of independent sites within an analyzed region. In the case of `TreeSwirl`, exploiting information from linked sites to detect introgression can substantially enhance power, particularly when linkage spans numerous loci.

`TreeSwirl` also exhibits consistency in identifying introgressed loci across all demographic models, even for models featuring two- and three-taxon topologies. This presents a significant advantage over f_4-stat methods, which are constrained to four-taxon configurations and defining an outgroup. Intriguingly, `TreeSwirl` encounters difficulties in accurately inferring mixture proportions for very short introgressed regions (approximately 100 loci) in graphs with two sister lineages (Fig 2, first row). This pronounced pattern is not observed when the length of the introgressed region increases, although the actual mixture proportions are incorrectly estimated in some instances compared to F_{st} results. This suggests that, in a two-taxon topology, our method may exhibit limitations in detecting regions under selection, particularly when they encompass a minimal number of loci.

While the power of inference for all methods is comparable in cases of strong introgression, `TreeSwirl` clearly demonstrates superior performance across simulations with low migration rates and short lengths, even for regions of approximately 100 loci where related D- and F-statistics yield the lowest AUC values. However, it is worth noting that, despite the ability of our method to detect weak signals of introgression, `TreeSwirl` also struggles to accurately infer mixture proportions when the introgression rate is very close to the attractor state (here 0.05). For instance, our method has the most false positives when the migration rate is between 0.1 and 0.15, regardless of the model and the length of the introgressed region, suggesting that there may be insufficient power to differentiate regions under selection. In such cases, it could be beneficial to increase the number of discretized migration rates when running `TreeSwirl` (by default 21 states). By doing so, our method may gain increased power to discern weak signals that are close to the attractor state.

3.2. Applications

3.2.1. Anopheles gambiae species complex

To showcase the performance of `TreeSwirl` with real data, we applied it to the *Anopheles gambiae* species complex. This complex represents a medically significant group of Afrotropical mosquito sibling species, as they serve as primary vectors of human malaria. The population genetic history of this Afrotropical complex was recently explored, revealing that traits enhancing vectorial capacity may be acquired through extensive introgression events (Fontaine et al., 2015). Among the most remarkable introgressed regions was a continuous segment aligned with the 3L arm chromosomal inversion. In this region, the original sequence found in ancestral populations of *An. quadriannulatus* has been completely supplanted by the corresponding sequence from *An. merus*.

We, thus, used the admixture graph from Figure 1C in Fontaine et al. (2015) to infer the mixture proportions from *An. merus* into *An. quadriannulatus*, particularly concentrating on the 3L arm. As depicted in Figure 3, our analysis uncovered multiple candidate regions for strong introgression within the 3La inversion, along with a limited number of outliers outside this region, even when using a highly conservative false discovery rate (FDR) of 0.0001. Our findings not only support the robust introgression signal on the 3L arm chromosomal inversion, as previously reported in (Fontaine et al., 2015; Pfeifer and Kapan, 2019), but also provide a more fine-grained resolution, as smaller genome regions experiencing introgression are detected. Hence, this may contribute to elucidate signals of adaptive introgression, such as insecticide resistance and an increased ability to transmit malaria within human populations.

3.3. Runtime considerations

The computational performance of `TreeSwirl` is influenced by multiple factors, such as the number of discrete states J, the number of matrices Σ, and the total number of sites and admixture events. Computation times scales linearly with the number of loci, making it less practical for whole-genome applications in a single run. However, the computations can be efficiently distributed across multiple computer nodes by dividing the genome into independent segments, such as individual chromosomes or chromosome arms. This approach is valid because linkage does not persist across chromosome boundaries and is typically weak across the centromere. Moreover, it should be noted that the computation time grows exponentially with an increasing number of migration edges i and states J.

4. Discussion

One approach to infer historical relationships among populations is to model allele frequency changes along a phylogenetic tree as a Gaussian process (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1981). This rather old concept was recently revived by extending the model to a graph with migration edges and by providing a user-friendly tool to infer parameters under such a graph (Pickrell and Pritchard, 2012). However, this model assumes migration rates to be constant along the genome, an assumption that may not hold in the face of selection or strong genetic drift. Indeed, theory predicts variation in the rate of effective gene flow along the genome (Harrison, 1993), in which local barriers to gene flow are anticipated to emerge from the random accumulation of Dobzhansky-Muller incompatibilities, both under models of secondary contact after isolation (Barton and Gale, 1993) as well as under models of continuous gene flow during speciation (Wu, 2001). In the case of gene flow between highly divergent gene pools, selection is likely to act as the primary driving force for variation in effective gene flow along the genome, with rates of introgression being particularly low in genomic regions involved in adaptation, so called islands of speciation, but potentially much higher in regions free from the selection pressure (Dasmahapatra et al., 2012).

In light of these considerations, we here present `TreeSwirl`, an extension of the model described in Pickrell and Pritchard (2012) that allows for mixture proportions to vary along the genome in an auto-correlated way that reflects the effect of linkage. We evaluated the performance of our model to identify such variation in comparison to existing methods related to D- and F-statistics, such as F_{st}, Patterson's D (Patterson et al., 2012), f_d (Martin et al.,

2015), f_{dM} (Malinsky et al., 2015), and d_f (Pfeifer and Kapan, 2019), which have been frequently applied to identify signatures of introgression using arbitrary genomic window sizes. As we show using extensive simulations, our method had superior accuracy and sensitivity in detecting retrogressed loci under a wide range of demographic histories characterized by single admixture pulses.

The approach presented here also addresses numerous constraints inherent to the use of related D- and F-statistics. First, these summary statistics are limited to bi-bifurcating four-population topologies. In cases involving graphs of five or more populations, the simplest option is to subsample a section of the graph in the appropriate configuration, as done in `Dsuite` (Malinsky et al., 2021) and replicated in our simulation tests involving six-population topologies (and five?). In cases involving two- or three-population topologies, one would need to resort to F_{ST}-based metrics. In contrast, the method presented here is not constraint by topology, working well with any number of populations and also under topologies that include polytomies.

Second, our HMM-based approach to model linkage eliminates the need to specify window sizes. Instead, the parameters governing auto-correlation are directly inferred from the data along with introgression rates. In our simulations, the choice of window sizes, as well as the choice of the specific statistics to use, had a big impact on power. To ensure a fair comparison between methods, we thus tested all available summary statistics for a wide range of window sizes and only report the results of the combination of summary statistics and window size that was optimal for each individual case. In applications to real data, however, such explorations are not possible, likely leading to an even larger difference in power between `TreeSwirl` and these summary statistics.

Third, and although not explored here, `TreeSwirl` supports graphs with multiple migration edges for which introgression rates are learned simultaneously. However, it is important to note that the performance of `TreeSwirl` likely dependent on the quality of the tree topology used as input and may not perform well if the tree topology is poorly resolved or incorrect.

We also reexamined datasets from mosquito populations, which hold significant economic and ecological importance and have been reported to experience introgression. Our analysis indeed identified multiple introgressed loci within these populations, consistent with previous findings, which further validates our model. The broader implications of introgression in species evolution, however, remain a subject of debate and are not yet thoroughly documented, primarily due to the challenges associated with accurately inferring introgressed loci. In fact, the potential for adaptive introgression to serve as a source of adaptation in response to ongoing global changes has often been underestimated (Suarez-Gonzalez et al., 2018). With our tool, we anticipate facilitating a deeper understanding of complex genetic histories within populations and shedding light on the processes that have shaped the genetic diversity patterns observed today.

Data and code availability

The authors affirm that all data required to validate the conclusions of this article are either included within the article itself or accessible through the indicated repositories. The source code for `TreeSwirl` can be found in the following Git repository: `https://bitbucket.org/wegmannlab/treeswirl2/`, which also contains a user manual. Addi-

tional scripts utilized for simulations are available upon request. This study did not generate any new data.

Acknowledgments

This work was supported by Swiss National Science Foundation grants 31003A_173062 and 310030_200420 to DW.

Author contributions

DW conceived the idea; DW, CL and CSRB developed the model; CSRB implemented the method in collaboration with MC and MG; CSRB conducted all simulations and data analyses; CSRB and DW led the writing of the manuscript. All authors contributed critically to the draft and gave final approval for publication.

Declaration of interests

The authors declare no competing interests

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J.W., Baird, S.J., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C.A., Buggs, R., Butlin, R.K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S.H., Hermansen, J.S., Hewitt, G., Hudson, A.G., Jiggins, C., Jones, J., Keller, B., Marczewski, T., Mallet, J., Martinez-Rodriguez, P., Möst, M., Mullen, S., Nichols, R., Nolte, A.W., Parisod, C., Pfennig, K., Rice, A.M., Ritchie, M.G., Seifert, B., Smadja, C.M., Stelkens, R., Szymura, J.M., Väinölä, R., Wolf, J.B., Zinner, D., 2013. Hybridization and speciation. Journal of Evolutionary Biology 26, 229–246. doi:10.1111/J.1420-9101.2012.02599.X.

Anderson, E., 1949. Introgressive hybridization. J. Wiley, New York. URL: https://www.biodiversitylibrary.org/bibliography/4553, doi:10.5962/bhl.title.4553.

Anderson, E., Hubricht, L., 1938. Hybridization in Tradescantia. III. The Evidence for Introgressive Hybridization. American Journal of Botany 25, 396. doi:10.2307/2436413.

Barton, N., Gale, K., 1993. Genetic analysis of hybrid zones, in: Harrison, R.G. (Ed.), Hybrid Zones and the Evolutionary Process. Oxford University Press, pp. 13–45.

Barton, N.H., 2001. The role of hybridization in evolution. Molecular Ecology 10, 551–568. doi:10.1046/J.1365-294X.2001.01216.X.

Barton, N.H., Hewitt, G.M., 1985. Analysis of Hybrid Zones. Annual Review of Ecology, Evolution, and Systematics , 113–148doi:10.1146/ANNUREV.ES.16.110185.000553.

Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. The Annals of Mathematical Statistics 41, 164–171. doi:10.1214/AOMS/1177697196.

Beech, M., Shepherd, E., 2001. Archaeobotanical evidence for early date consumption on Dalma Island, United Arab Emirates. Antiquity 75, 83–89. URL: https://www.cambridge.org/core/journals/antiquity/article/archaeobotanical-evidence-for-early-date-consumption-on-dalma-island-united-arab-emir FA80E36C36D1BE2D10AAD0286F1BFAE5, doi:10.1017/S0003598X00052765.

Browning, S.R., Browning, B.L., Zhou, Y., Tucci, S., Akey, J.M., 2018. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell 173, 53–61.e9. doi:10.1016/J.CELL.2018.02.031.

Burgarella, C., Barnaud, A., Kane, N.A., Jankowski, F., Scarcelli, N., Billot, C., Vigouroux, Y., Berthouly-Salazar, C., 2019. Adaptive introgression: An untapped evolutionary mechanism for crop adaptation. Frontiers in Plant Science 10, 4. doi:10.3389/FPLS.2019.00004/BIBTEX.

Cavalli-Sforza, L.L., Edwards, A.W., 1967. Phylogenetic analysis. Models and estimation procedures. American Journal of Human Genetics 19, 233. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1706274/, doi:10.2307/2406616.

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., McVean, G., Durbin, R., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. URL: https://academic.oup.com/bioinformatics/article/27/15/2156/402296, doi:10.1093/BIOINFORMATICS/BTR330.

Dasmahapatra, K.K., Walters, J.R., Briscoe, A.D., Davey, J.W., Whibley, A., Nadeau, N.J., Zimin, A.V., Salazar, C., Ferguson, L.C., Martin, S.H., Lewis, J.J., Adler, S., Ahn, S.J., Baker, D.A., Baxter, S.W., Chamberlain, N.L., Ritika, C., Counterman, B.A., Dalmay, T., Gilbert, L.E., Gordon, K., Heckel, D.G., Hines, H.M., Hoff, K.J., Holland, P.W., Jacquin-Joly, E., Jiggins, F.M., Jones, R.T., Kapan, D.D., Kersey, P., Lamas, G., Lawson, D., Mapleson, D., Maroja, L.S., Martin, A., Moxon, S., Palmer, W.J., Papa, R., Papanicolaou, A., ick Pauchet, Y., Ray, D.A., Rosser, N., Salzberg, S.L., Supple, M.A., Surridge, A., Tenger-Trolander, A., Vogel, H., Wilkinson, P.A., Wilson, D., Yorke, J.A., Yuan, F., Balmuth, A.L., Eland, C., Gharbi, K., Thomson, M., Gibbs, R.A., Han, Y., Jayaseelan, J.C., Kovar, C., Mathew, T., Muzny, D.M., Ongeri, F., Pu, L.L., Qu, J., Thornton, R.L., Worley, K.C., Wu, Y.Q., Linares, M., Blaxter, M.L., Ffrench-Constant, R.H., Joron, M., Kronforst, M.R., Mullen, S.P., Reed, R.D., Scherer, S.E., Richards, S., Mallet, J., Mc Millan, W.O., Jiggins, C.D., 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487, 94–98. URL: https://pubmed.ncbi.nlm.nih.gov/22722851/, doi:10.1038/NATURE11041.

Durand, E.Y., Patterson, N., Reich, D., Slatkin, M., 2011. Testing for ancient admixture between closely related populations. Molecular Biology and Evolution 28, 2239–2252. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/21325092/?tool=EBIhttps://europepmc.org/article/PMC/PMC3144383, doi:10.1093/MOLBEV/MSR048.

Durvasula, A., Sankararaman, S., 2019. A statistical model for reference-free inference of archaic local ancestry. PLOS Genetics 15, e1008175. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008175, doi:10.1371/JOURNAL.PGEN.1008175.

Durvasula, A., Sankararaman, S., 2020. Recovering signals of ghost archaic introgression in African populations. Science Advances 6. URL: https://www.science.org/doi/10.1126/sciadv.aax5097, doi:10.1126/SCIADV.AAX5097/SUPPL_FILE/AAX5097_SM.PDF.

Eaton, D.A., Ree, R.H., 2013. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). Systematic Biology 62, 689–706. URL: https://academic.oup.com/sysbio/article/62/5/689/1684460, doi:10.1093/SYSBIO/SYT032.

Ellstrand, N.C., 2014. Is gene flow the most important evolutionary force in plants? American Journal of Botany 101, 737–753. URL: https://onlinelibrary.wiley.com/doi/full/10.3732/ajb.1400024https://onlinelibrary.wiley.com/doi/abs/10.3732/ajb.1400024https://bsapubs.onlinelibrary.wiley.com/doi/10.3732/ajb.1400024, doi:10.3732/AJB.1400024.

Ellstrand, N.C., Barrett, S.C., Linington, S., Stephenson, A.G., Comai, L., 2003. Current knowledge of gene flow in plants: implications for transgene flow. Philosophical

Transactions of the Royal Society of London. Series B: Biological Sciences 358, 1163–1170. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2003.1299, doi:10.1098/RSTB.2003.1299.

Excoffier, L., Marchi, N., Marques, D.A., Matthey-Doret, R., Gouy, A., Sousa, V.C., 2021. fastsimcoal2: demographic inference under complex evolutionary scenarios. Bioinformatics 37, 4882–4885. URL: https://academic.oup.com/bioinformatics/article/37/24/4882/6308558, doi:10.1093/BIOINFORMATICS/BTAB468.

Felsenstein, J., 1981. Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. Evolution; international journal of organic evolution 35, 1229–1242. URL: https://pubmed.ncbi.nlm.nih.gov/28563384/, doi:10.1111/J.1558-5646.1981.TB04991.X.

Flowers, J.M., Hazzouri, K.M., Gros-Balthazard, M., Mo, Z., Koutroumpa, K., Perrakis, A., Ferrand, S., Khierallah, H.S., Fuller, D.Q., Aberlenc, F., Fournaraki, C., Purugganan, M.D., 2019. Cross-species hybridization and the origin of North African date palms. Proceedings of the National Academy of Sciences of the United States of America 116, 1651–1658. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1817453116, doi:10.1073/PNAS.1817453116/SUPPL_FILE/PNAS.1817453116.SD01.PDF.

Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I.V., Jiang, X., Hall, A.B., Catteruccia, F., Kakani, E., Mitchell, S.N., Wu, Y.C., Smith, H.A., Rebecca Love, R., Lawniczak, M.K., Slotman, M.A., Emrich, S.J., Hahn, M.W., Besansky, N.J., 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347, 1258524. URL: https://www.science.org/doi/10.1126/science.1258524, doi:10.1126/SCIENCE.1258524/SUPPL_FILE/1258524_FONTAINE_SM_REVISION1.PDF.

Galimberti, M., Leuenberger, C., Wolf, B., Szilágyi, S.M., Foll, M., Wegmann, D., 2020. Detecting Selection from Linked Sites Using an F-Model. Genetics 216, 1205–1215. URL: https://pubmed.ncbi.nlm.nih.gov/33067324/, doi:10.1534/GENETICS.120.303780.

Grant, P.R., Grant, B.R., 1992. Hybridization of Bird Species. Science 256, 193–197. URL: https://www.science.org/doi/10.1126/science.256.5054.193, doi:10.1126/SCIENCE.256.5054.193.

Grant, P.R., Grant, B.R., 1994. Phenotypic and genetic effects of hybridization in Darwin's finches. Evolution 48, 297–316. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1558-5646.1994.tb01313.x, doi:10.1111/J.1558-5646.1994.TB01313.X.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., De La Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R.,

Kelso, J., Lachmann, M., Reich, D., Pääbo, S., 2010. A draft sequence of the neandertal genome. Science 328, 710–722. URL: https://www.science.org/doi/10.1126/science.1188021, doi:10.1126/SCIENCE.1188021/SUPPL_FILE/GREEN_SOM.PDF.

Gros-Balthazard, M., Galimberti, M., Kousathanas, A., Newton, C., Ivorra, S., Paradis, L., Vigouroux, Y., Carter, R., Tengberg, M., Battesti, V., Santoni, S., Falquet, L., Pintaud, J.C., Terral, J.F., Wegmann, D., 2017. The Discovery of Wild Date Palms in Oman Reveals a Complex Domestication History Involving Centers in the Middle East and Africa. Current Biology 27, 2211–2218.e8. doi:10.1016/J.CUB.2017.06.045.

Harrison, R.G., 1993. Hybrid Zones and the Evolutionary Process. Oxford University Press.

Kozak, K.M., Joron, M., McMillan, W.O., Jiggins, C.D., 2021. Rampant Genome-Wide Admixture across the Heliconius Radiation. Genome Biology and Evolution 13. URL: https://academic.oup.com/gbe/article/13/7/evab099/6263859, doi:10.1093/GBE/EVAB099.

Kronforst, M.R., Hansen, M.E., Crawford, N.G., Gallant, J.R., Zhang, W., Kulathinal, R.J., Kapan, D.D., Mullen, S.P., 2013. Hybridization Reveals the Evolving Genomic Architecture of Speciation. Cell reports 5, 666. URL: /pmc/articles/PMC4388300//pmc/articles/PMC4388300/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4388300/, doi:10.1016/J.CELREP.2013.09.042.

Kulathinal, R.J., Stevison, L.S., Noor, M.A., 2009. The Genomics of Speciation in Drosophila: Diversity, Divergence, and Introgression Estimated Using Low-Coverage Genome Sequencing. PLOS Genetics 5, e1000550. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000550, doi:10.1371/JOURNAL.PGEN.1000550.

Lange, K., 2010. Numerical Analysis for Statisticians. Statistics and Computing, Springer New York, New York, NY. URL: https://link.springer.com/10.1007/978-1-4419-5945-4, doi:10.1007/978-1-4419-5945-4.

Lawson, D.J., Hellenthal, G., Myers, S., Falush, D., 2012. Inference of Population Structure using Dense Haplotype Data. PLOS Genetics 8, e1002453. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002453, doi:10.1371/JOURNAL.PGEN.1002453.

Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A., Wegmann, D., 2017. ATLAS: Analysis Tools for Low-depth and Ancient Samples. bioRxiv URL: https://www.biorxiv.org/content/early/2017/03/24/105346, doi:10.1101/105346.

Lipson, M., Loh, P.R., Levin, A., Reich, D., Patterson, N., Berger, B., 2013. Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. Molecular Biology and Evolution 30, 1788–1802. URL: https://academic.oup.com/mbe/article/30/8/1788/1017431, doi:10.1093/MOLBEV/MST099, arXiv:1212.2555.

Lipson, M., Loh, P.R., Patterson, N., Moorjani, P., Ko, Y.C., Stoneking, M., Berger, B., Reich, D., 2014. Reconstructing Austronesian population history in Island Southeast Asia. Nature Communications 2014 5:1 5, 1–7. URL: https://www.nature.com/articles/ncomms5689, doi:10.1038/ncomms5689.

Malinsky, M., Challis, R.J., Tyers, A.M., Schiffels, S., Terai, Y., Ngatunga, B.P., Miska, E.A., Durbin, R., Genner, M.J., Turner, G.F., 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. Science 350, 1493–1498. URL: https://www.science.org/doi/10.1126/science.aac9927, doi:10.1126/SCIENCE.AAC9927/SUPPL_FILE/AAC9927S2.MOV.

Malinsky, M., Matschiner, M., Svardal, H., 2021. Dsuite - Fast D-statistics and related admixture evidence from VCF files. Molecular Ecology Resources 21, 584–595. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13265, doi:10.1111/1755-0998.13265.

Mallet, J., 2005. Hybridization as an invasion of the genome. Trends in Ecology & Evolution 20, 229–237. doi:10.1016/J.TREE.2005.02.010.

Maples, B.K., Gravel, S., Kenny, E.E., Bustamante, C.D., 2013. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. The American Journal of Human Genetics 93, 278–288. doi:10.1016/J.AJHG.2013.06.020.

Martin, S.H., Dasmahapatra, K.K., Nadeau, N.J., Salazar, C., Walters, J.R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., Jiggins, C.D., 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome research 23, 1817–1828. URL: https://pubmed.ncbi.nlm.nih.gov/24045163/, doi:10.1101/GR.159426.113.

Martin, S.H., Davey, J.W., Jiggins, C.D., 2015. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Molecular Biology and Evolution 32, 244–257. URL: https://academic.oup.com/mbe/article/32/1/244/2925550, doi:10.1093/MOLBEV/MSU269.

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. URL: https://research.google/pubs/pub38136/.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. The Computer Journal 7, 308–313. URL: https://typeset.io/papers/a-simplex-method-for-function-minimization-2jcvs0pxrq, doi:10.1093/COMJNL/7.4.308.

Nocedal, J., Wright, S., 2006. Numerical Optimization. Springer Science & Business Media. URL: https://books.google.com/books/about/Numerical{_}Optimization.html?hl=de{&}id=VbHYoSyelFcC.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., Reich, D., 2012. Ancient admixture in human history. Genetics 192, 1065–1093. URL: https://academic.oup.com/genetics/article/192/3/1065/5935193, doi:10.1534/GENETICS.112.145037/-/DC1.

Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S., Reich, D., 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 2006 441:7097 441, 1103–1108. URL: https://www.nature.com/articles/nature04789, doi:10.1038/nature04789.

Pease, J.B., Hahn, M.W., 2015. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. Systematic Biology 64, 651–662. URL: https://academic.oup.com/sysbio/article/64/4/651/1650669, doi:10.1093/SYSBIO/SYV023.

Pfeifer, B., Kapan, D.D., 2019. Estimates of introgression as a function of pairwise distances. BMC Bioinformatics 20, 1–11. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2747-z, doi:10.1186/S12859-019-2747-Z.

Pickrell, J.K., Pritchard, J.K., 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLOS Genetics 8, e1002967. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002967, doi:10.1371/JOURNAL.PGEN.1002967, arXiv:1206.2332.

Plagnol, V., Wall, J.D., 2006. Possible ancestral structure in human populations. Plos Genetics 2, e105–e105. URL: https://europepmc.org/articles/PMC1523253https://europepmc.org/article/PMC/1523253, doi:10.1371/JOURNAL.PGEN.0020105.

Price, A.L., Helgason, A., Palsson, S., Stefansson, H., St. Clair, D., Andreassen, O.A., Reich, D., Kong, A., Stefansson, K., 2009. The Impact of Divergence Time on the Nature of Population Structure: An Example from Iceland. PLOS Genetics 5, e1000505. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000505, doi:10.1371/JOURNAL.PGEN.1000505.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., De Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J.C., Vohr, S.H., Green, R.E., Hellmann, I., Johnson, P.L., Blanche, H., Cann, H., Kitzman, J.O., Shendure, J., Eichler, E.E., Lein, E.S., Bakken, T.E., Golovanova, L.V., Doronichev, V.B., Shunkov, M.V., Derevianko, A.P., Viola, B., Slatkin, M., Reich, D., Kelso, J., Pääbo, S., 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 2013 505:7481 505, 43–49. URL: https://www.nature.com/articles/nature12886, doi:10.1038/nature12886.

Rabiner, L.R., Juang, B.H., 1986. An Introduction to Hidden Markov Models. IEEE ASSP Magazine 3, 4–16. doi:10.1109/MASSP.1986.1165342.

Racimo, F., Marnetto, D., Huerta-Sánchez, E., 2017. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. Molecular Biology and Evolution 34, 296–317. URL: https://academic.oup.com/mbe/article/34/2/296/2633371, doi:10.1093/MOLBEV/MSW216.

Racimo, F., Sankararaman, S., Nielsen, R., Huerta-Sánchez, E., 2015. Evidence for archaic adaptive introgression in humans. Nature Reviews Genetics 2015 16:6 16, 359–371. URL: https://www.nature.com/articles/nrg3936https://www.nature.com/articles/nrg3936/, doi:10.1038/nrg3936.

Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N., García, L.F., Triana, O., Blair, S., Maestre, A., Dib, J.C., Bravi, C.M., Bailliet, G., Corach, D., Hünemeier, T., Bortolini, M.C., Salzano, F.M., Petzl-Erler, M.L., Acuña-Alonzo, V., Aguilar-Salinas, C., Canizales-Quinteros, S., Tusié-Luna, T., Riba, L., Rodríguez-Cruz, M., Lopez-Alarcón, M., Coral-Vazquez, R., Canto-Cetina, T., Silva-Zolezzi, I., Fernandez-Lopez, J.C., Contreras, A.V., Jimenez-Sanchez, G., Gómez-Vázquez, M.J., Molina, J., Carracedo, Á., Salas, A., Gallo, C., Poletti, G., Witonsky, D.B., Alkorta-Aranburu, G., Sukernik, R.I., Osipova, L., Fedorova, S.A., Vasquez, R., Villena, M., Moreau, C., Barrantes, R., Pauls, D., Excoffier, L., Bedoya, G., Rothhammer, F., Dugoujon, J.M., Larrouy, G., Klitz, W., Labuda, D., Kidd, J., Kidd, K., Di Rienzo, A., Freimer, N.B., Price, A.L., Ruiz-Linares, A., 2012. Reconstructing Native American population history. Nature 488, 370–374. URL: https://www.nature.com/articles/nature11258, doi:10.1038/nature11258.

Reich, D., Thangaraj, K., Patterson, N., Price, A.L., Singh, L., 2009. Reconstructing Indian population history. Nature 461, 489–494. URL: https://www.nature.com/articles/nature08365, doi:10.1038/nature08365.

Rheindt, F.E., Fujita, M.K., Wilton, P.R., Edwards, S.V., 2014. Introgression and phenotypic assimilation in Zimmerius flycatchers (Tyrannidae): population genetic and phylogenetic inferences from genome-wide SNPs. Systematic biology 63, 134–152. URL: https://pubmed.ncbi.nlm.nih.gov/24304652/, doi:10.1093/SYSBIO/SYT070.

Rieseberg, L.H., Wendel, J.F., 1993. lntrogression and Its Consequences in Plants, in: Harrison, R.G. (Ed.), Hybrid zones and the evolutionary process. Oxford University Press. chapter 4, pp. 70–109.

Sankararaman, S., 2020. Methods for detecting introgressed archaic sequences. Current Opinion in Genetics & Development 62, 85–90. doi:10.1016/J.GDE.2020.05.026.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., Reich, D., 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 2014 507:7492 507, 354–357. URL: https://www.nature.com/articles/nature12961, doi:10.1038/nature12961.

Seguin-Orlando, A., Korneliussen, T.S., Sikora, M., Malaspinas, A.S., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., Goebel, T., Westaway, M., Lambert, D., Khartanovich, V., Wall, J.D., Nigst, P.R., Foley, R.A., Lahr, M.M., Nielsen, R., Orlando, L., Willerslev, E., 2014. Genomic structure in Europeans dating back at least 36,200 years. Science 346, 1113–1118. URL: https://www.science.org/doi/10.1126/science.aaa0114, doi:10.1126/SCIENCE.AAA0114/SUPPL_FILE/SEGUIN-ORLANDO.SM.PDF.

Skov, L., Hui, R., Shchur, V., Hobolth, A., Scally, A., Schierup, M.H., Durbin, R., 2018. Detecting archaic introgression using an unadmixed outgroup. PLOS Genetics 14, e1007641. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007641, doi:10.1371/JOURNAL.PGEN.1007641.

Slatkin, M., 1985a. Gene Flow in Natural Populations. Source: Annual Review of Ecology and Systematics 16, 393–430. URL: https://www.jstor.org/stable/2097054.

Slatkin, M., 1985b. Rare alleles indicators of gene flow. Evolution; international journal of organic evolution 39, 53–65. URL: https://pubmed.ncbi.nlm.nih.gov/28563643/, doi:10.1111/J.1558-5646.1985.TB04079.X.

Slatkin, M., 1987. Gene Flow and the Geographic Structure of Natural Populations. Science 236, 787–792. URL: https://www.science.org/doi/10.1126/science.3576198, doi:10.1126/SCIENCE.3576198.

Smith, J., Kronforst, M.R., 2013. Do Heliconius butterfly species exchange mimicry alleles? Biology letters 9. URL: https://pubmed.ncbi.nlm.nih.gov/23864282/, doi:10.1098/RSBL.2013.0503.

Sousa, V.C., Fritz, M., Beaumont, M.A., Chikhi, L., 2009. Approximate Bayesian Computation Without Summary Statistics: The Case of Admixture. Genetics 181, 1507–1519. URL: https://academic.oup.com/genetics/article/181/4/1507/6081151, doi:10.1534/GENETICS.108.098129.

Steinrücken, M., Spence, J.P., Kamm, J.A., Wieczorek, E., Song, Y.S., 2018. Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. Molecular Ecology 27, 3873–3888. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/mec.14565https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14565https://onlinelibrary.wiley.com/doi/10.1111/mec.14565, doi:10.1111/MEC.14565.

Suarez-Gonzalez, A., Lexer, C., Cronk, Q.C., 2018. Adaptive introgression: a plant perspective. Biology Letters 14. URL: https://royalsocietypublishing.org/doi/10.1098/rsbl.2017.0688, doi:10.1098/RSBL.2017.0688.

Tang, H., Coram, M., Wang, P., Zhu, X., Risch, N., 2006. Reconstructing Genetic Ancestry Blocks in Admixed Individuals. The American Journal of Human Genetics 79, 1–12. doi:10.1086/504302.

Tung, J., Barreiro, L.B., 2017. The contribution of admixture to primate evolution. Current Opinion in Genetics & Development 47, 61–68. doi:10.1016/J.GDE.2017.08.010.

Vernot, B., Akey, J.M., 2014. Resurrecting surviving Neandertal lineages from modern human genomes. Science 343, 1017–1021. URL: https://www.science.org/doi/10.1126/science.1245938, doi:10.1126/SCIENCE.1245938/SUPPL_FILE/VERNOT.SM.PDF.

Vernot, B., Tucci, S., Kelso, J., Schraiber, J.G., Wolf, A.B., Gittelman, R.M., Dannemann, M., Grote, S., McCoy, R.C., Norton, H., Scheinfeldt, L.B., Merriwether, D.A., Koki, G., Friedlaender, J.S., Wakefield, J., Pääbo, S., Akey, J.M., 2016. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science 352, 235–239. URL: https://www.science.org/doi/10.1126/science.aad9416, doi:10.1126/SCIENCE.AAD9416/SUPPL_FILE/VERNOT-SM.PDF.

Wall, J.D., Lohmueller, K.E., Plagnol, V., 2009. Detecting Ancient Admixture and Estimating Demographic Parameters in Multiple Human Populations. Molecular Biology and Evolution 26, 1823–1827. URL: https://academic.oup.com/mbe/article/26/8/1823/980556, doi:10.1093/MOLBEV/MSP096.

Wegmann, D., Kessner, D.E., Veeramah, K.R., Mathias, R.A., Nicolae, D.L., Yanek, L.R., Sun, Y.V., Torgerson, D.G., Rafaels, N., Mosley, T., Becker, L.C., Ruczinski, I., Beaty, T.H., Kardia, S.L., Meyers, D.A., Barnes, K.C., Becker, D.M., Freimer, N.B., Novembre, J., 2011. Recombination rates in admixed individuals identified by ancestry-based inference. Nature Genetics 2011 43:9 43, 847–853. URL: https://www.nature.com/articles/ng.894, doi:10.1038/ng.894.

Wu, C.I., 2001. The genic view of the process of speciation. Journal of Evolutionary Biology 14, 851–865. URL: https://onlinelibrary.wiley.com/doi/full/10.1046/j.1420-9101.2001.00335.xhttps://onlinelibrary.wiley.com/doi/abs/10.1046/j.1420-9101.2001.00335.xhttps://onlinelibrary.wiley.com/doi/10.1046/j.1420-9101.2001.00335.x, doi:10.1046/J.1420-9101.2001.00335.X.

Yamamichi, M., Innan, H., 2012. Estimating the migration rate from genetic variation data. Heredity 108, 362–363. doi:10.1038/HDY.2011.83.

Yang, M.A., Malaspinas, A.S., Durand, E.Y., Slatkin, M., 2012. Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. Molecular biology and evolution 29, 2987–2995. URL: https://pubmed.ncbi.nlm.nih.gov/22513287/, doi:10.1093/MOLBEV/MSS117.

Zhang, X., Witt, K.E., Bañuelos, M.M., Ko, A., Yuan, K., Xu, S., Nielsen, R., Huerta-Sanchez, E., 2021. The history and evolution of the Denisovan-EPAS1 haplotype in Tibetans. Proceedings of the National Academy of Sciences of the United States of America 118, e2020803118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2020803118, doi:10.1073/PNAS.2020803118/SUPPL_FILE/PNAS.2020803118.SAPP.PDF.

Figures

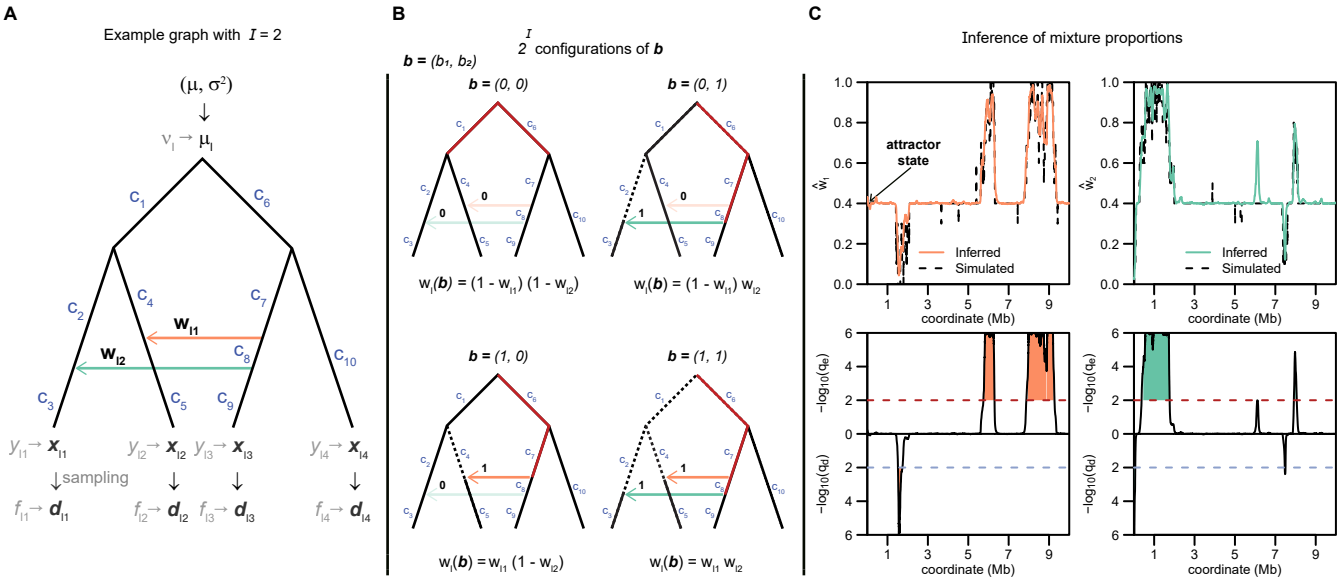

Figure 1: Inference example. A: admixture graph with two migration edges marked in different colors. Parameters of interest are shown on the graph (root priot and branch lengths) as well as the untransformed and transformed ancient, sampling and population allele frequency variables. B: All possible configurations of **b** for two migration events when they are open or closed. C: Example of inference under our `TreeSwirl` model for each migration event. The top panel shows the posterior mean mixture proportions $\hat{w}_l$ compared to simulated estimates and the bottom panel shows the identified candidate regions under possible selection, where the false discovery rates (FDR) for excess (q_e) and dearth (q_d) introgression was determined for each locus as explained in the "Inference" section.

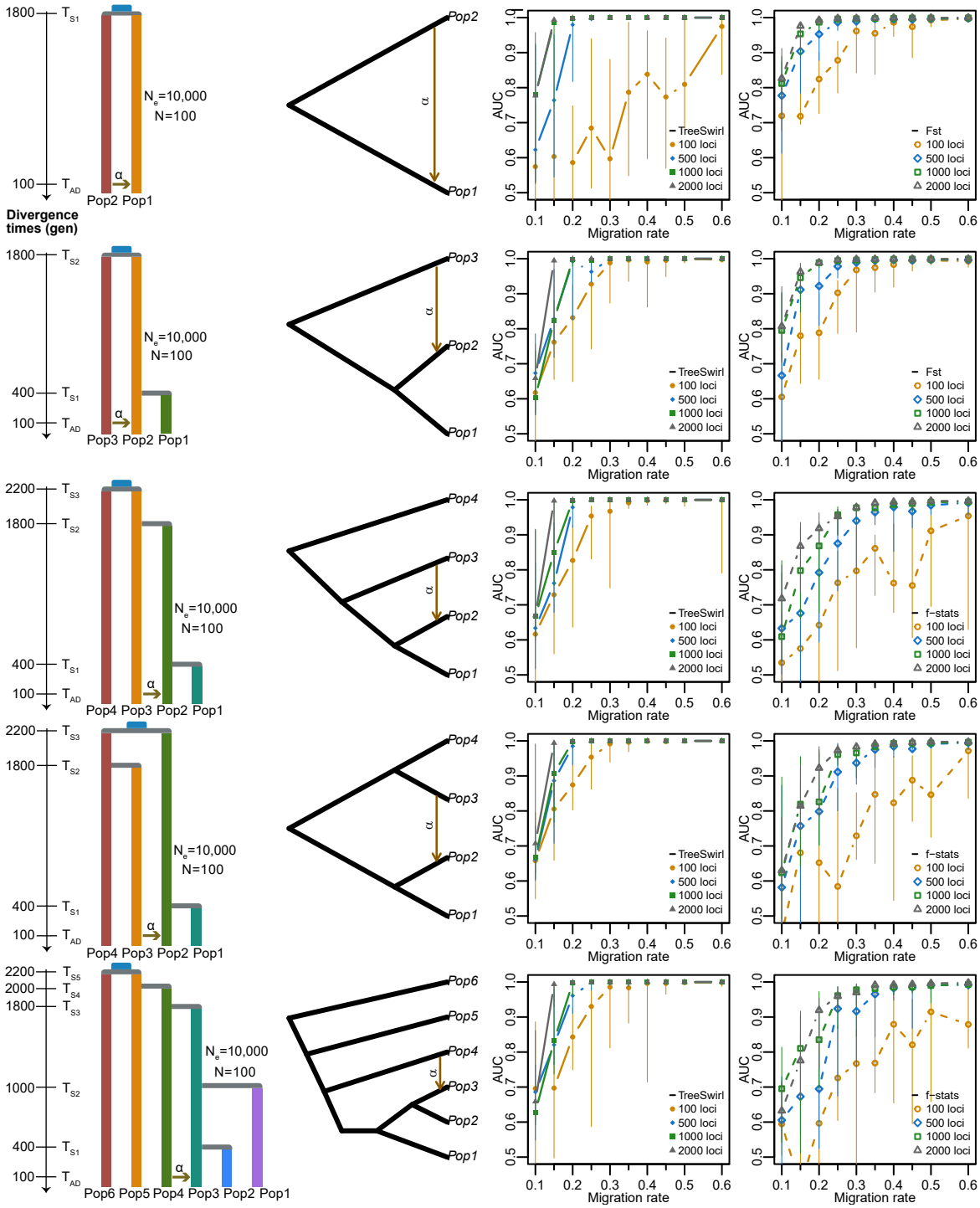

Figure 2: Performance of `TreeSwirl` and f_4-stats methods to measure the amount of introgression under different demographic histories with an background migration rate $\alpha = 0.05$. Firs column: simulated demographic histories. Second column: schematic of the topology models. Third and fourth column: AUC measures for `TreeSwirl` and F- and D-related stats (best summary statistic and window size were chosen). Different symbols are used for simulated blocks with lengths 100 to 2000 loci. Because of the minimum allele frequency filter, the sizes are relative.

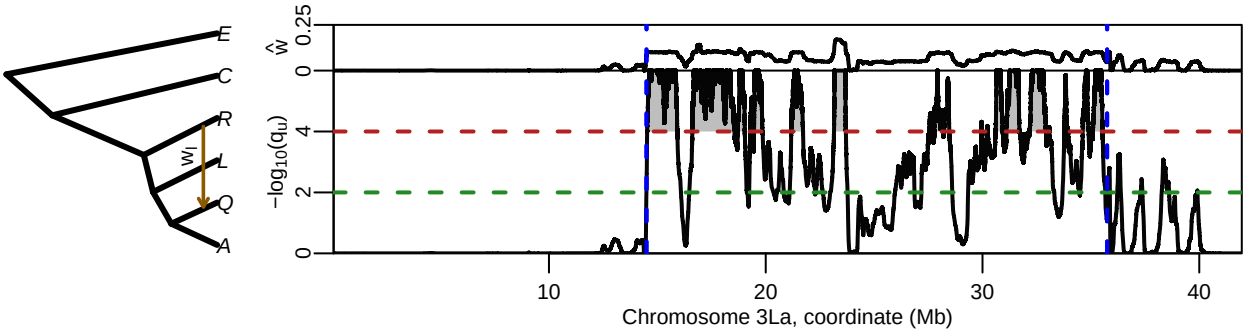

Figure 3: Inference of introgressed loci on the 3La inversion of *Anopheles gambiae*. First column: Topology of *Anopheles gambiae* rooted by *An. epiroticus* (E) and *An. christyi* (C), depicting one introgression event (orange arrow). The graph was taken from Figure 1C, Fontaine et al. (2015). Second column: confirmed signal of introgression on the 3L arm from *An. merus* (R) to *An. quadriannulatus* (Q). `TreeSwirl` was run with the depicted topology using 21 states and 10 Sigmas (Σ, sample size variance matrix). Estimated mean posteriors ($\hat{w}$) are shown on top. Candidate regions of introgression are shaded in gray at a false discovery rate (FDR) of 0.0001. The introgressed chromosomal inversion is delineated between the vertical dashed blue lines. Horizontal dashed lines indicate the 0.01 and 0.0001 FDR threshold.