G3, 2025, **15(11)**, jkaf202
https://doi.org/10.1093/g3journal/jkaf202
Advance Access Publication Date: 8 September 2025
Investigation

Coalescent theory of the ψ directionality index

Egor Lappo ,* Noah A. Rosenberg

Department of Biology, Stanford University, Stanford, CA 94305, United States

*Corresponding author: Department of Biology, Stanford University, Stanford, CA 94305, United States. Email: elappo@stanford.edu.

The ψ directionality index was introduced by Peter and Slatkin (*Evolution 67: 3274–3289, 2013*) to infer the direction of range expansions from single-nucleotide polymorphism variation. Computed from the joint site frequency spectrum for two populations, ψ uses shared genetic variants to measure the difference in the amount of genetic drift experienced by the populations, associating excess drift with greater distance from the origin of the range expansion. Although ψ has been successfully applied in natural populations, its statistical properties have not been well understood. In this article, we define Ψ as a random variable originating from a coalescent process in a two-population demography. For samples consisting of a pair of diploid genomes, one from each of two populations, we derive expressions for moments $\mathbb{E}[\Psi^k]$ for standard parameterizations of bottlenecks during a founder event. For the expectation $\mathbb{E}[\Psi]$, we identify parameter combinations that represent distinct demographic scenarios yet yield the same value of $\mathbb{E}[\Psi]$. We also show that the variance $\mathbb{V}[\Psi]$ increases with the time since the bottleneck and bottleneck severity, but does not depend on the size of the ancestral population; the ancestral population size affects ψ computed from many biallelic loci only through its contribution to the total number of loci available for the computation. Finally, we analyze the values of $\mathbb{E}[\Psi]$ computed from existing demographic models of *Drosophila melanogaster* and compare them with empirically computed ψ. Our work builds the foundation for theoretical treatments of the ψ index and can help in evaluating its behavior in empirical applications.

Keywords: bottleneck; coalescent theory; directionality index; founder effect; range expansion

Introduction

Inference of the demographic history of populations—including their population-size changes and relationships with other populations—is a major objective of statistical population genetics (e.g. Marchi et al. 2021). The combination of statistical methods based on the coalescent theory with extensive genetic data has enabled researchers to investigate diverse features of demographic histories (e.g. Pool et al. 2010).

One of the most fundamental ways in which genetic data can be summarized for statistical analysis is by the site frequency spectrum (SFS), which counts the numbers of sites—typically single-nucleotide polymorphisms (SNPs)—that are present in different multiplicities in a sample (e.g. Wakeley and Hey 1997; Achaz 2009). Comparisons of the empirical SFS in a population to predictions of a coalescent model can detect phenomena, such as bottlenecks, expansions, or selective sweeps (e.g. Ferretti et al. 2010; Ronen et al. 2013). The SFS has received extensive theoretical treatment under many demographic scenarios and has often been applied for inference in real populations (e.g. Nielsen et al. 2005; Thornton and Andolfatto 2006).

In data from *multiple* populations, a joint SFS can be defined that records SNP allele frequencies in each population (Gutenkunst et al. 2009). A joint SFS enables inference of processes, such as admixture, migration, and differences in selection between populations (Caicedo et al. 2007; Nielsen et al. 2009; Excoffier et al. 2013; Zhan et al. 2014; Arguello et al. 2019; Liu and Fu 2020). In the setting of population pairs, Peter and Slatkin (2013) proposed a statistic, the ψ directionality index, which is computed from the joint SFS for the two populations. This index was designed for characterizing the process of range expansion, in which a population sequentially settles locations increasingly distant from its origin (e.g. Ramachandran et al. 2005; Excoffier and Ray 2008; Excoffier et al. 2009).

In a range expansion, the leading edge of the expansion experiences stronger genetic drift relative to the point of origin (e.g. Hallatschek and Nelson 2008; Slatkin and Excoffier 2012; Peter and Slatkin 2015; Peischl and Excoffier 2016). In the genetic history of individuals at the edge of the expansion, the range expansion process can manifest as a sequence of population size bottlenecks, as increasingly distant geographic locations are settled (e.g. DeGiorgio et al. 2009; Deshpande et al. 2009; DeGiorgio et al. 2011).

For two populations that are part of the expansion, the ψ directionality index seeks to identify the direction of the expansion. The approach relies on the fact that if a given derived allele is shared between the two populations, then its frequency is expected to be higher in the derived population at the edge of the range expansion than in the source population (e.g. Edmonds et al. 2004; Klopfstein et al. 2006; Excoffier and Ray 2008; Schlichta et al. 2022). Alleles at low frequency in the source population are likely to be lost during the expansion and therefore would not be shared. The derived population has a smaller founding population size than the source population, so that alleles—if they are not entirely absent—tend to possess greater frequencies. The ψ index considers the population differences of allele

Received on 18 June 2025; accepted on 23 August 2025

frequencies specifically in the shared genetic variation between the two populations.

Among pairwise quantities that can be computed as summary statistics useful for interpreting population-genetic data (e.g. F_{ST}), the ψ index stands out as a signed quantity. For two populations A and B, the order of the populations matters, with $\psi(A, B) = -\psi(B, A)$. Therefore, whereas F_{ST} is often seen as a genetic measure of distance, ψ is akin to a vector directed from one population to another (see also Peter and Slatkin 2013, Figs. 5 and 7).

The ψ index was first defined by Peter and Slatkin (2013), who developed a method that integrates information about pairwise ψ with geographic distances between sampling locations to identify coordinates of the expansion origin. They then applied it to simulated scenarios including isolation-by-distance and range expansion on a grid of populations, as well as to complex configurations involving migration barriers.

Peter and Slatkin (2015) then studied theoretical properties of ψ in a discrete time-expansion model. The model consisted of a linearly arranged set of demes with equal population size, with a single deme d_0 settled initially and the rest of the demes empty. At an integer timepoint t, a new deme d_t is settled by individuals from the previous deme d_{t-1}. The quantity of interest was $\psi(d_0, d_t)$ at time t between the origin deme and the leading edge of the expansion. Peter and Slatkin (2015) showed that in the model, the expected value of ψ between the source and the leading edge of the expansion—which has experienced a sequence of founder events—depends on the relative founder sizes of settlement events (the fraction of individuals selected from deme d_{t-1} to settle d_t) and the number of founder events, equal to t in their scaling of time. Peter and Slatkin (2015) used ψ to identify the expansion origin for natural populations of *Arabidopsis thaliana*, which they presumed to have expanded spatially in a manner compatible with a linear arrangement of demes.

Several recent uses of ψ have since sought to examine scenarios where, instead of an expansion over a linear spatial dimension, an expansion involves pairwise computations for a small number of discrete demes, as few as two. For example, Zhan et al. (2014) examined the expansion of monarch butterflies from North America to South America, the Pacific, and Europe, computing ψ between a source population in North America and a destination population elsewhere. Puckett and Munshi-South (2019) examined the expansion of brown rats from Eastern Asia to the Middle East, the Middle East to Europe, and Europe to North America, computing ψ between pairs of populations in two different geographic regions. Ioannidis et al. (2021) similarly used pairwise values of ψ between pairs of human populations of different Pacific islands to understand sequences of events in the human settlement of the region.

In this article, building from the interest in using ψ for expansions involving small numbers of discrete populations rather than many demes along a spatial continuum, we examine the ψ statistic theoretically in the simplest discrete-deme structured population: a pair of populations. We define Ψ as a random variable arising from the coalescent process and derive expressions for moments of Ψ under the coalescent. We focus on the scenario in which a single diploid individual is sampled in each of a pair of populations. Next, we consider specific commonly used parameterizations of range expansions in the setting of population pairs, explicitly incorporating exponential growth, bottlenecks, and instantaneous bottlenecks (Fig. 1). We then explore theoretical predictions for the expectation $\mathbb{E}[\Psi]$ and variance $\mathbb{V}[\Psi]$, interpreting them in terms of the reliability of inferences and the identifiability of demographic scenarios. We use the central limit theorem to analyze the sample variance of the ψ index computed from many SNPs. Finally, we show how our results can be used in the evaluation of empirical inferences of demographic parameters for real populations.

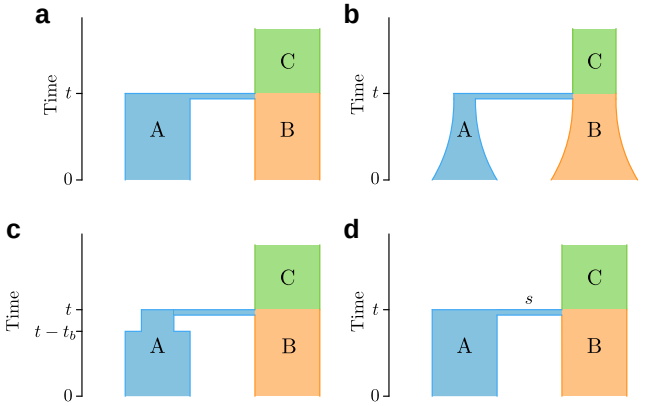

Fig. 1. Four demographic scenarios. a) Population split: an ancestral population C splits into two populations A and B at time t. b) Exponential growth: after the split, populations A and B grow at rates r_A and r_B, respectively. c) Bottleneck: after the split, population A goes through a bottleneck of population size N_b and duration t_b. d) Instantaneous bottleneck: after the split, population A goes through a burst of coalescences of strength s. Times t and $t - t_b$ are measured in generations back from the present.

Coalescent-based definition of the ψ index

ψ for a pair of genomes

The directionality index ψ is a two-population statistic computed from allele frequencies for a set of biallelic SNPs. For the rest of this article, we assume that the derived and ancestral alleles are known for each SNP, and we call the SNP *shared* between two populations if the derived allele is present in at least one copy in both populations and the SNP is polymorphic in the pooled pair of populations.

Suppose now that we know allele frequencies for a set of SNPs in two populations A and B. In its most general form, the value of the ψ index is then defined as

$$\psi(A, B) = \frac{1}{|S|}\sum_{j \in S}(f_{A,j} - f_{B,j}), \quad (1)$$

where S is the set of SNPs *shared* between the two populations, $f_{A,j}$ is the frequency of the derived allele of SNP j in population A, and $f_{B,j}$ is its frequency in population B (Peter and Slatkin 2015, Equation 1).

We proceed by focusing on the simplest case in which ψ can be meaningfully studied in two populations. In particular, if allele frequencies are computed using a single diploid individual sampled from each population, and if shared SNPs are identified based on this pair of individuals, then the expression for ψ reduces to

$$\psi(A, B) = \frac{n_{21} - n_{12}}{n_{11} + n_{12} + n_{21}}, \quad (2)$$

where n_{ij} is the number of SNPs that have i copies of the derived allele in the individual from population A and j copies in the individual from population B, and i and j can each equal 0, 1, or 2.

The random variable Ψ under the coalescent

We now analyze the ψ index as a random variable under the coalescent. We use the notation Ψ to distinguish the theoretical random variable for the directionality index from the empirical ψ computed from data.

We assume that all SNPs are unlinked, so that coalescent trees for different SNPs are independent. We also assume that SNPs obey the standard infinitely-many-sites mutation model (Durrett 2008, p. 29), such that each SNP results from a single mutation on a coalescent tree. Finally, we assume that we have specified a demographic history for populations A and B (Fig. 1), and that a single diploid individual is sampled from each population. Such a sample configuration—one diploid individual with sample size 2 alleles in each population—allows us to use the simplified Equation (2).

Conditional on the demography and a sample of size 2 from each of a pair of populations, the coalescent model defines a probability distribution over the genealogies of lineages from A and B. In this framework, we can determine the expectation $\mathbb{E}[\Psi](A, B)$ of the directionality index under the coalescent model for a single SNP shared between populations A and B. We use $\mathbb{E}[\Psi^k]$ with $k > 1$ to denote higher moments of the random variable Ψ under the coalescent.

To compute $\mathbb{E}[\Psi]$, we consider probabilities under the coalescent model of entries in the joint site frequency spectrum for populations A and B, conditional on the demography:

$$\mathcal{S}_{A,B} = \begin{pmatrix} 0 & s_{01} & s_{02} \\ s_{10} & s_{11} & s_{12} \\ s_{02} & s_{21} & 0 \end{pmatrix}, \tag{3}$$

where s_{ij} is the probability that a randomly sampled mutation—that is, the derived variant of a random SNP on the genealogy of four lineages—occurs in i copies in population A and in j copies in population B. For example, s_{12} is the probability that for a random SNP, the diploid sample from population A has one ancestral and one derived allele, and the sample from population B is homozygous with two copies of the derived allele. As a shorthand, we will say that such a SNP has "type 12," and we indicate other elements of the site frequency spectrum similarly.

Suppose now that we have sampled a single *shared* SNP. First, the probabilities of a SNP having a specific type are obtained from the site frequency spectrum $\mathcal{S}$ by conditioning on being shared,

$$\mathbb{P}[\text{type 21} \mid \text{shared SNP}] = \frac{s_{21}}{s_{11} + s_{12} + s_{21}}, \tag{4a}$$

$$\mathbb{P}[\text{type 11} \mid \text{shared SNP}] = \frac{s_{11}}{s_{11} + s_{12} + s_{21}}, \tag{4b}$$

$$\mathbb{P}[\text{type 12} \mid \text{shared SNP}] = \frac{s_{12}}{s_{11} + s_{12} + s_{21}}. \tag{4c}$$

Further, taking the total number of sampled SNPs to be 1 in Equation (2), we know that the value of Ψ is constant for all SNPs of the same type, with

$$\Psi \mid \text{shared SNP of type 21} = 1, \tag{5a}$$

$$\Psi \mid \text{shared SNP of type 11} = 0, \tag{5b}$$

$$\Psi \mid \text{shared SNP of type 12} = -1. \tag{5c}$$

Combining Equation (4) with Equation (5), we can write the definition of random variable Ψ for a single shared SNP:

$$\Psi(A, B) = \begin{cases} 1, & \text{probability } \dfrac{s_{21}}{s_{11} + s_{12} + s_{21}}, \\ 0, & \text{probability } \dfrac{s_{11}}{s_{11} + s_{12} + s_{21}}, \\ -1, & \text{probability } \dfrac{s_{12}}{s_{11} + s_{12} + s_{21}}. \end{cases} \tag{6}$$

The expectation of Ψ can then be straightforwardly computed as

$$\mathbb{E}[\Psi] = \frac{s_{21} - s_{12}}{s_{11} + s_{12} + s_{21}}. \tag{7}$$

The second moment of Ψ is

$$\mathbb{E}[\Psi^2] = \frac{s_{21} + s_{12}}{s_{11} + s_{12} + s_{21}}. \tag{8}$$

The higher moments of Ψ can be computed similarly, with

$$\mathbb{E}[\Psi] = \mathbb{E}[\Psi^3] = \mathbb{E}[\Psi^5] = \cdots, \tag{9}$$

$$\mathbb{E}[\Psi^2] = \mathbb{E}[\Psi^4] = \mathbb{E}[\Psi^6] = \cdots. \tag{10}$$

In the remainder of this article, we discuss only the expectation $\mathbb{E}[\Psi]$ and the variance $\mathbb{V}[\Psi] = \mathbb{E}[\Psi^2] - \mathbb{E}[\Psi]^2$, as the other moments can be obtained from these cases.

The only remaining quantities we need are the s_{ij}: the probabilities that a randomly chosen SNP has i copies of the derived allele in the sample of size 2 from population A and j copies in the sample of size 2 from population B, with $(i, j) = (1, 1)$, $(1, 2)$, or $(2, 1)$. In other words, under a coalescent-based demographic model with infinitely-many-sites mutation, we seek to compute, as a fraction of all SNPs, the number that occur on genealogical branches ancestral to i copies in population A and j copies in population B.

In a random genealogy, the expected total number of SNPs with type ij is $\Theta\mathbb{E}[L_{ij}]/2$, where L_{ij} is the total length of branches ancestral to i lineages from population A and j lineages from population B, and $\Theta/2$ is the Poisson mutation rate along a branch. $\mathbb{E}[L_{ij}]$ is computed by considering each topology separately:

$$\mathbb{E}[L_{ij}] = \sum_{\text{topology } \tau} p_\tau \mathbb{E}[L_{\tau,ij}], \tag{11}$$

where p_τ is the probability that topology τ occurs and $L_{\tau,ij}$ is the length of branches ancestral to i lineages from A and j lineages from B in genealogies with topology τ. The value of s_{ij} is then proportional to

$$s_{ij} \propto \sum_{\text{topology } \tau} p_\tau \Theta \mathbb{E}[L_{\tau,ij}]/2. \tag{12}$$

Because we regard lineages within populations as exchangeable—so that we do not distinguish between two lineages from the same population—six topologies must be considered in Equation (12) (Fig. 2). We denote the six topologies $\alpha, \beta, \gamma, \delta, \varepsilon, \zeta$. The topology probabilities p_τ and the expected branch lengths $\mathbb{E}[L_{\tau,ij}]$ can be computed for various demographic models, so that Equation (12) can be calculated and hence also Equations (7) and (8). In the next section, we compute these quantities for simple models representing a founder effect.

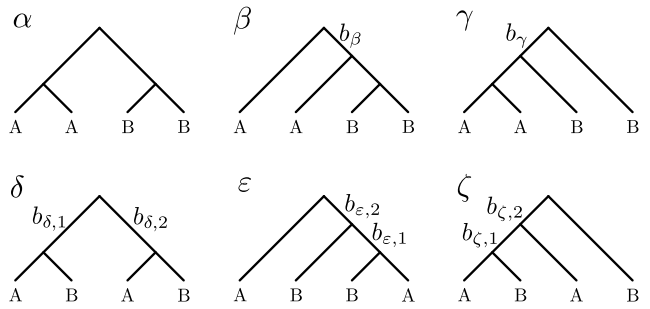

Fig. 2. The six tree topologies possible for samples of two lineages each from two populations A and B. Trees are labeled by Greek letters. Branches relevant to the calculation of $\mathbb{E}[\Psi]$—branches that are ancestral to lineages from both populations—are labeled by b_β, b_γ, etc.

Expectation and variance of Ψ for specific demographic models

The exact expressions for topology probabilities and branch lengths in Equation (12) depend on specific parameterizations of the demographic history. In this section, we derive expressions for s_{ij} and moments of Ψ in Equations (7) and (8) for the four demographies shown in Fig. 1.

Population split

We first consider a simple population split demography (Fig. 1a). A single ancestral population C of size N_C splits into two populations t generations ago. The two resulting populations A and B have sizes N_A and N_B individuals, respectively.

To compute $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$, we compute topology probabilities and relevant branch length expectations for each topology in Fig. 2. Computations with topology α are not needed because this topology cannot generate shared polymorphisms. For topology probabilities p_β and p_γ, we must further distinguish the tree topologies based on the population in which the "cherry coalescence" happens. We denote by $p_{\beta,\mathrm{B}}$ the probability of the topology β in which the coalescence (B,B) happens in population B; if the coalescence (B,B) happens in the ancestral population C, then we denote the probability by $p_{\beta,\mathrm{C}}$, with similar notation for topology γ. These additional topology labels are depicted in Fig. 3.

First, we compute $p_{\beta,\mathrm{B}}$. The two lineages from population B must coalesce in population B; the probability of this event is $1-e^{-t/(2N_B)}$. The two lineages from population A must not coalesce until they enter population C; the probability of that event is $e^{-t/(2N_A)}$. As a result, two lineages from A and one lineage from B enter population C. In population C, lineages from A and B must coalesce first; the probability of this event is $\frac{2}{3}$; in the remaining $\frac{1}{3}$ of cases, the two A lineages coalesce first. As a result, we obtain $p_{\beta,\mathrm{B}}=\frac{2}{3}e^{-t/(2N_A)}[1-e^{-t/(2N_B)}]$. This derivation modifies the two-population calculation of Tajima (1983) by allowing for different population sizes for populations A and B rather than assuming their exchangeability. Using the same logic for other tree topologies, we obtain the following probabilities:

$$p_{\beta,\mathrm{B}}=\frac{2}{3}e^{-t/(2N_A)}\left[1-e^{-t/(2N_B)}\right], \tag{13a}$$

$$p_{\gamma,\mathrm{A}}=\frac{2}{3}\left[1-e^{-t/(2N_A)}\right]e^{-t/(2N_B)}, \tag{13b}$$

$$p_{\beta,\mathrm{C}}=p_{\gamma,\mathrm{C}}=\frac{1}{9}e^{-t/(2N_A)}e^{-t/(2N_B)}, \tag{13c}$$

$$p_\delta=p_\varepsilon=p_\zeta=\frac{2}{9}e^{-t/(2N_A)}e^{-t/(2N_B)}. \tag{13d}$$

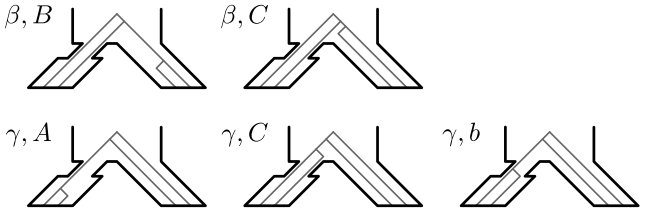

Fig. 3. Distinguishing locations of the cherry coalescence for topologies β and γ. The ancestral population is C and the descendant populations are A (left) and B (right).

For β and γ, summing the probabilities of the two cases, we get

$$p_\beta=\frac{2}{3}e^{-t/(2N_A)}-\frac{5}{9}e^{-t/(2N_A)}e^{-t/(2N_B)}, \tag{14a}$$

$$p_\gamma=\frac{2}{3}e^{-t/(2N_B)}-\frac{5}{9}e^{-t/(2N_A)}e^{-t/(2N_B)}. \tag{14b}$$

We now compute expected lengths of relevant branches for specific sample configurations. All branches below the gene tree root that are shared by at least one pair of lineages from different populations are labeled in Fig. 2. For example, branch b_β is ancestral to two lineages from population B and one lineage from population A for topology β; similarly, branch $b_{\zeta,1}$ is ancestral to one lineage from population A and one lineage from population B if the genealogy has topology ζ.

With Equation (12), we obtain equations for entries of the expected joint site frequency spectrum $\mathcal{S}_{\mathrm{A,B}}$:

$$s_{12}\propto\frac{\Theta}{2}\left(p_\beta\mathbb{E}[b_\beta]+p_\varepsilon\mathbb{E}[b_{\varepsilon,2}]\right), \tag{15a}$$

$$s_{21}\propto\frac{\Theta}{2}\left(p_\gamma\mathbb{E}[b_\gamma]+p_\zeta\mathbb{E}[b_{\zeta,2}]\right), \tag{15b}$$

$$s_{11}\propto\frac{\Theta}{2}\left[p_\delta\left(\mathbb{E}[b_{\delta,1}]+\mathbb{E}[b_{\delta,2}]\right)+p_\varepsilon\mathbb{E}[b_{\varepsilon,1}]+p_\zeta\mathbb{E}[b_{\zeta,1}]\right]. \tag{15c}$$

In the final expressions for moments of Ψ (Equations (7) and (8)), the mutation rate cancels because all branches that can generate shared sites can only appear in population C, so that $\Theta=4N_C\mu$ in all parts of Equation (15).

We are now left with calculating the expected branch lengths. Because polymorphisms shared between populations A and B can only result from mutations in the ancestral population C, our branch length calculations need only consider coalescent theory in a single population of size N_C individuals. In particular, expectations of branch lengths b_β and b_γ are equal to the expectation of the time $\mathbb{E}[T_2]$ to coalescence of two lineages in the diploid population of size N_C, so that

$$\mathbb{E}[b_\beta]=\mathbb{E}[b_\gamma]=2N_C. \tag{16}$$

Similar logic applies to trees ε and ζ, with expectations of $b_{\varepsilon,1}$ and $b_{\zeta,1}$ equaling the expectation of the time T_3 to the first coalescence with three lineages,

$$\mathbb{E}[b_{\varepsilon,1}]=\mathbb{E}[b_{\zeta,1}]=\frac{2N_C}{3}. \tag{17}$$

The lengths of $b_{\varepsilon,2}$ and $b_{\zeta,2}$ are again proportional to $\mathbb{E}[T_2]$ as in Equation (16),

$$\mathbb{E}[b_{\varepsilon,2}]=\mathbb{E}[b_{\zeta,2}]=2N_C. \tag{18}$$

The expected length of branches $b_{\delta,1}$ and $b_{\delta,2}$ together is equal to $2\mathbb{E}[T_2] + \mathbb{E}[T_3]$,

$$\mathbb{E}[b_{\delta,1}] + \mathbb{E}[b_{\delta,2}] = 4N_C + \frac{2N_C}{3}. \tag{19}$$

Finally, we can substitute expressions for the branch lengths (Equations (16)–(19)) and the topology probabilities (Equation (13)) into Equation (15) to obtain expressions for SFS entries s_{ij}. We then plug these quantities into Equations (7) and (8) to obtain

$$\mathbb{E}[\Psi] = \frac{e^{t/(2N_A)} - e^{t/(2N_B)}}{e^{t/(2N_A)} + e^{t/(2N_B)}}, \tag{20a}$$

$$\mathbb{E}[\Psi^2] = 1 - \frac{1}{e^{t/(2N_A)} + e^{t/(2N_B)}}. \tag{20b}$$

The variance of Ψ is then

$$\mathbb{V}[\Psi] = \frac{4e^{t/(2N_A)}e^{t/(2N_B)} - e^{t/(2N_A)} - e^{t/(2N_B)}}{\left[e^{t/(2N_A)} + e^{t/(2N_B)}\right]^2}. \tag{20c}$$

Examining the expressions in Equations (20a) and (20c), we see that both the mean of Ψ and the variance of Ψ do not depend on the size N_C of the ancestral population. We also observe that if the population sizes are equal, $N_A = N_B$, then $\mathbb{E}[\Psi] = 0$. Moreover, examination of Equation (20) can directly bound the expectation and variance. As $\lim_{N_B \to 0} \mathbb{E}[\Psi] = -1$ and $\lim_{N_A \to 0} \mathbb{E}[\Psi] = 1$, we have $-1 < \mathbb{E}[\Psi] < 1$; for variance, we have $0 < \mathbb{V}[\Psi] < 1$ because, on one hand, $\lim_{t \to \infty} \mathbb{V}[\Psi] = 0$, and on the other hand,

$$\begin{aligned}\mathbb{V}[\Psi] &= 1 - \frac{1}{e^{t/(2N_A)} + e^{t/(2N_B)}} - \left(\mathbb{E}[\Psi]\right)^2 \\ &\leq 1 - \frac{1}{e^{t/(2N_A)} + e^{t/(2N_B)}} < 1\end{aligned}$$

for $N_A, N_B > 0$ and $t \geq 0$.

Informally, we can write Equation (20a) as

$$\begin{aligned}\mathbb{E}[\Psi] &= \tanh\left(\frac{t}{4N_A} - \frac{t}{4N_B}\right) \\ &= \tanh(\text{“drift in A”} - \text{“drift in B”}).\end{aligned} \tag{21}$$

In this simple model, the “amount of drift” is that of a neutral population of size N_A (or N_B) evolving for t generations. However, by treating N_A and N_B as effective population sizes, a variety of demographic scenarios that include population growth or bottlenecks can be considered. In subsequent subsections, we explicitly parameterize models with population size changes and present modified versions of Equation (20).

Exponential growth

We next consider populations A and B evolving under the classic exponential growth model. A and B begin exponential growth immediately after splitting from the ancestral population C, as shown in Fig. 1b.

Let population A have size $N_{A,0}$ at the present time, such that its population size over time is

$$N_{A,\tau} = N_{A,0}e^{-r_A\tau}, \tag{22}$$

where τ is time, measured in generations from the present into the past, and r_A is the growth rate. Equation (22) is defined such that if $r_A > 0$, then population A is increasing in size forward in time. If population A has size $N_{A,t}$ immediately after the split, then the growth rate can be computed from Equation (22) as

$$r_A = -\frac{1}{t}\ln\left(\frac{N_{A,t}}{N_{A,0}}\right). \tag{23}$$

Slatkin and Hudson (1991) showed that for a pair of lineages, the coalescent in a growing population of size $N_{A,\tau}$ is equivalent to the coalescent in the constant population of size $N_{A,0}$, with time re-scaled by

$$T = \frac{e^{r_A\tau} - 1}{r_A}. \tag{24}$$

Hence, the probability that two lineages coalesce in the first t generations in the population of size $N_{A,\tau}$ is

$$\mathbb{P}[T_A \leq t] = 1 - \exp\left(-\frac{e^{r_A t} - 1}{2N_{A,0}r_A}\right). \tag{25}$$

A corresponding equation holds for population B.

We can repeat the calculations of tree topology probabilities in Equations (13) and (14) by replacing the constant-size coalescence probability $1 - \exp[-t/(2N)]$ by the quantity in Equation (25). As a result, we obtain the following expressions for expectation and variance of Ψ under the exponential growth model:

$$\mathbb{E}[\Psi] = \tanh\left(\frac{t_A}{4N_{A,0}} - \frac{t_B}{4N_{B,0}}\right), \tag{26a}$$

$$\mathbb{E}[\Psi^2] = 1 - \frac{1}{e^{t_A/(2N_{A,0})} + e^{t_B/(2N_{B,0})}}, \tag{26b}$$

$$\mathbb{V}[\Psi] = \frac{4e^{t_A/(2N_{A,0})}e^{t_B/(2N_{B,0})} - e^{t_A/(2N_{A,0})} - e^{t_B/(2N_{B,0})}}{\left[e^{t_A/(2N_{A,0})} + e^{t_B/(2N_{B,0})}\right]^2}, \tag{26c}$$

where we have introduced a shorthand notation

$$t_A = \frac{e^{r_A t} - 1}{r_A}, \tag{27a}$$

$$t_B = \frac{e^{r_B t} - 1}{r_B}. \tag{27b}$$

If only one population is subject to exponential growth, then expressions for $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$ can be found by taking the limits in Equation (26) as the growth rate approaches zero. For example, if $r_B = 0$, then we have

$$\mathbb{E}[\Psi] = \tanh\left(\frac{t_A}{4N_{A,0}} - \frac{t}{4N_B}\right), \tag{28a}$$

$$\mathbb{V}[\Psi] = \frac{4e^{t_A/(2N_{A,0})}e^{t/(2N_B)} - e^{t_A/(2N_{A,0})} - e^{t/(2N_B)}}{\left[e^{t_A/(2N_{A,0})} + e^{t/(2N_B)}\right]^2}, \tag{28b}$$

where again t_A is defined by Equation (27a).

Bottleneck

For our third model, we assume that immediately after the split, population A goes through a bottleneck of length t_b with constant population size N_b, as shown in Fig. 1c. This type of model has

been used in studies of human expansion from Africa (DeGiorgio et al. 2009, 2011).

The calculations of $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$ are similar to those for the population split demography, except that the topology probabilities differ: we consider special cases for topologies β and γ (Fig. 2). For β, we distinguish between coalescent genealogies in which the node (B,B) is located in population B ($p_{\beta,\mathrm{B}}$) and in population C ($p_{\beta,\mathrm{C}}$). For γ, we distinguish between genealogies in which the node (A,A) occurs in population A after the bottleneck ($p_{\gamma,\mathrm{A}}$), during the bottleneck ($p_{\gamma,\mathrm{bot}}$), and in population C ($p_{\gamma,\mathrm{C}}$). The probabilities are:

$$p_{\beta,\mathrm{B}} = \frac{2}{3} e^{-(t-t_b)/(2N_A)} e^{-t_b/(2N_b)} \left[1 - e^{-t/(2N_B)}\right], \tag{29a}$$

$$p_{\beta,\mathrm{C}} = p_{\gamma,\mathrm{C}} = \frac{1}{9} e^{-(t-t_b)/(2N_A)} e^{-t_b/(2N_b)} e^{-t/(2N_B)}, \tag{29b}$$

$$p_{\gamma,\mathrm{A}} = \frac{2}{3}\left[1 - e^{-(t-t_b)/(2N_A)}\right] e^{-t/(2N_B)}, \tag{29c}$$

$$p_{\gamma,\mathrm{bot}} = \frac{2}{3} e^{-(t-t_b)/(2N_A)} \left[1 - e^{-t_b/(2N_b)}\right] e^{-t/(2N_B)}, \tag{29d}$$

$$p_{\delta} = p_{\varepsilon} = p_{\zeta} = \frac{2}{9} e^{-(t-t_b)/(2N_A)} e^{-t_b/(2N_b)} e^{-t/(2N_B)}. \tag{29e}$$

The expressions for the moments of Ψ are

$$\mathbb{E}[\Psi] = \tanh\left(\frac{t - t_b}{4N_A} + \frac{t_b}{4N_b} - \frac{t}{4N_B}\right), \tag{30a}$$

$$\mathbb{E}\left[\Psi^2\right] = 1 - \frac{1}{e^{(t-t_b)/(2N_A)} + e^{t_b/(2N_b)} e^{t/(2N_B)}}, \tag{30b}$$

$$\mathbb{V}[\Psi] = \frac{e^{\frac{t_b}{2N_A}}\left[4e^{\frac{t}{2N_A}} e^{\frac{t}{2N_B}} e^{\frac{t_b}{2N_b}} - e^{\frac{t}{2N_B}} e^{\frac{t_b}{2N_b}} - e^{\frac{t}{2N_B}} e^{\frac{t_b}{2N_A}}\right]}{\left[e^{\frac{t}{2N_B}} e^{\frac{t_b}{2N_A}} + e^{\frac{t}{2N_A}} e^{\frac{t_b}{2N_b}}\right]^2}. \tag{30c}$$

If $N_b = N_A$, then Equation (30) reduces to Equation (20). If $t_b = 0$, then expressions in Equation (30) match Equation (20) irrespective of the value of N_b.

Founder effect

The final model that we consider is a model that has been proposed for simplifying the modeling of founder effects. Instead of a prolonged bottleneck, we introduce an *instantaneous bottleneck* into population A (Fig. 1d). An instantaneous bottleneck is defined as a burst of coalescences; mathematically, two lineages going through an instantaneous bottleneck of strength s behave as if going through s (imaginary) generations of drift in the population of final size N_A. Instantaneous bottlenecks are typically used in situations where the bottleneck is short enough such that the possibility of mutations happening *during* the bottleneck can be disregarded (Galtier et al. 2000; Bunnefeld et al. 2015). In practice, this scenario could correspond to a low number of lineages from population C settling the whole population A that exists after the split.

Similarly to the bottleneck demography scenario, we adjust the tree topology probabilities to reflect the demography in Fig. 1d:

$$p_{\beta,\mathrm{B}} = \frac{2}{3} e^{-(t+s)/(2N_A)}\left[1 - e^{-t/(2N_B)}\right], \tag{31a}$$

$$p_{\beta,\mathrm{C}} = p_{\gamma,\mathrm{C}} = \frac{1}{9} e^{-(t+s)/(2N_A)} e^{-t/(2N_B)}, \tag{31b}$$

$$p_{\gamma,\mathrm{A}} = \frac{2}{3}\left[1 - e^{-t/(2N_A)}\right] e^{-t/(2N_B)}, \tag{31c}$$

$$p_{\gamma,\mathrm{bot}} = \frac{2}{3} e^{-t/(2N_A)}\left[1 - e^{-s/(2N_A)}\right] e^{-t/(2N_B)}, \tag{31d}$$

$$p_{\delta} = p_{\varepsilon} = p_{\zeta} = \frac{2}{9} e^{-(t+s)/(2N_A)} e^{-t/(2N_B)}. \tag{31e}$$

The expressions for the moments of Ψ in this case are:

$$\mathbb{E}[\Psi] = \tanh\left(\frac{t+s}{4N_A} - \frac{t}{4N_B}\right), \tag{32a}$$

$$\mathbb{E}\left[\Psi^2\right] = 1 - \frac{1}{e^{(t+s)/(2N_A)} + e^{t/(2N_B)}}, \tag{32b}$$

$$\mathbb{V}[\Psi] = \frac{\left[4e^{(t+s)/(2N_A)} e^{t/(2N_B)} - e^{(t+s)/(2N_A)} - e^{t/(2N_B)}\right]}{\left[e^{(t+s)/(2N_A)} + e^{t/(2N_B)}\right]^2}. \tag{32c}$$

These expressions reflect the fact that the "strength" of the bottleneck depends on its duration t_b and population size N_b only through the ratio $t_b/(2N_b)$, as captured by the parameter s. If $s = 0$, then Equation (32) reduces to Equation (20).

Illustrations of $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$

To illustrate our theoretical expressions, we plot $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$ for a range of parameter values. For these plots, we use the instantaneous bottleneck formulation, as it has only five parameters t, s, N_A, N_B, and N_C instead of six, as in the exponential growth and bottleneck scenarios.

Figure 4 shows $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$ for varying t and s and fixed population sizes $N_A = 400$, $N_B = 600$, and $N_C = 1{,}000$. The behavior of $\mathbb{E}[\Psi]$ confirms the intuition in Equation (21): as the bottleneck strength s increases, population A accumulates larger amounts of drift due to increased probability of coalescence in the bottleneck (Equation (32a)), leading to higher positive values of $\mathbb{E}[\Psi]$. Similarly, increasing t leads to more drift in population A because $N_A < N_B$, and A accumulates drift at a higher rate than B. Note that if we were to set $N_A = N_B$ in Equation (32a), then the value of $\mathbb{E}[\Psi]$ would not depend on the time t since the bottleneck.

The variance $\mathbb{V}[\Psi]$ also increases with both the time t since the bottleneck and the bottleneck strength s, with stronger dependence on t compared to $\mathbb{E}[\Psi]$. As s or t increases, the probability of observing a SNP of type 11, s_{11}, decreases, as a type 11 SNP requires all four lineages from A and B to persist into population C without coalescing, whereas type 12 and type 21 SNPs can be produced with only three lineages persisting into population C. SNPs of type 11 can appear *only* in genealogies δ, ε, and ζ where all coalescences happen in ancestral population C, and the probability of observing these genealogies is small for large s and t. The second moment $\mathbb{E}\left[\Psi^2\right] = (s_{21} + s_{12})/(s_{12} + s_{21} + s_{11})$ then grows large, increasing the variance.

We can use our theoretical results to analyze identifiability of demographic scenarios with the ψ index. In analyses of genetic

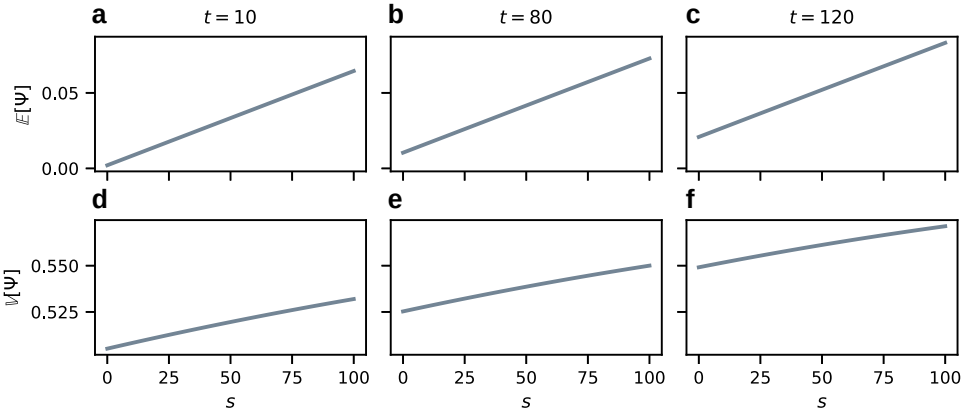

Fig. 4. Values of $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$ under the founder effect demography of Fig. 1d. The population sizes are constant, with $N_A = 400$, $N_B = 600$, and $N_C = 1{,}000$. The values are computed using Equation (32a) for $\mathbb{E}[\Psi]$ and Equation (32c) for $\mathbb{V}[\Psi]$. a) $\mathbb{E}[\Psi]$ for varying s, $t = 10$. b) $\mathbb{E}[\Psi]$ for varying s, $t = 80$. c) $\mathbb{E}[\Psi]$ for varying s, $t = 120$. d) $\mathbb{V}[\Psi]$ for varying s, $t = 10$. e) $\mathbb{V}[\Psi]$ for varying s, $t = 80$. f) $\mathbb{V}[\Psi]$ for varying s, $t = 120$.

data, a positive ψ is used to claim that the population A is located further from the source of the range expansion (Peter and Slatkin 2013), with the population with more drift experiencing bottlenecks during founder events. However, this logic does not account for the possibility that other demographic scenarios could generate an identical value of ψ. Figure 5 provides an example of this phenomenon by showing that if the population size N_B is sufficiently small relative to N_A, then $\mathbb{E}[\Psi]$ could be negative even in the presence of a bottleneck in population A. Figure 6 shows the dependence of $\mathbb{E}[\Psi]$ on N_B and t. If N_B is small enough in relation to N_A, then the value of $\mathbb{E}[\Psi]$ can decrease with increasing t and can even reverse its sign. The negative value of the directionality index then obscures the ancient bottleneck in A.

Sampling theory of Ψ

We have demonstrated that the expectation $\mathbb{E}[\Psi]$ and variance $\mathbb{V}[\Psi]$ do not depend on the ancestral population size N_C. In this section, we show that our confidence in the value of ψ computed from SNP data does depend on N_C through sample variance.

The random variable Ψ and its associated quantities $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$ refer to the directionality index for a single SNP under the coalescent. In a data analysis, ψ is computed using many shared SNPs across the genome, say n. Denote by $\overline{\Psi}$ the (random) many-SNP ψ index, signifying that this quantity can be seen as the mean of many single-SNP observations of ψ. For n sampled independent shared SNPs, the central limit theorem states that the resulting $\overline{\Psi}$ approaches a Gaussian distribution with variance proportional to $\frac{1}{n}$,

$$\overline{\Psi} \overset{d}{\to} \mathcal{N}\left(\mathbb{E}[\Psi], \frac{\mathbb{V}[\Psi]}{n}\right). \tag{33}$$

As the number of sampled shared SNPs increases, the probability that $\overline{\Psi}$ is close to its mathematical expectation $\mathbb{E}[\Psi]$ increases.

Under the infinitely-many-sites model, the number of shared SNPs n is itself a random variable that depends on the mutation rate μ and the ancestral population size N_C, as shared SNPs reflect mutations in the ancestral population. More precisely, $\mathbb{E}[n] = \Theta\mathbb{E}[L]/2$, where $\Theta = 4N_C\mu$ is the scaled mutation rate in a diploid population of N_C individuals, and $\mathbb{E}[L]$ is the expected length of branches that can yield shared SNPs, in units of $2N_C$ generations. All branches that can generate shared SNPs were identified in Fig. 2, so we can use Equations (16)–(19) to write an equation for $\mathbb{E}[L]$:

$$\begin{aligned} 2N_C\mathbb{E}[L] &= p_\beta\mathbb{E}[b_\beta] + p_\gamma\mathbb{E}[b_\gamma] + p_\delta\left(\mathbb{E}[b_{\delta,1}] + \mathbb{E}[b_{\delta,2}]\right) \\ &\quad + p_\varepsilon\left(\mathbb{E}[b_{\varepsilon,1}] + \mathbb{E}[b_{\varepsilon,2}]\right) + p_\zeta\left(\mathbb{E}[b_{\zeta,1}] + \mathbb{E}[b_{\zeta,2}]\right), \\ \mathbb{E}[L] &= p_\beta + p_\gamma + \frac{7}{3}p_\delta + \frac{4}{3}p_\varepsilon + \frac{4}{3}p_\zeta. \end{aligned} \tag{34}$$

For example, with the founder effect model, we get

$$\mathbb{E}[L] = \frac{2}{3}\left[e^{-(t+s)/(2N_A)} + e^{-t/(2N_B)}\right], \tag{35}$$

$$\mathbb{E}[n] = \frac{4N_C\mu}{3}\left[e^{-(t+s)/(2N_A)} + e^{-t/(2N_B)}\right]. \tag{36}$$

The expected number of shared SNPs depends linearly on the ancestral population size, and N_C then affects the number of shared SNPs available for empirical analyses using the directionality index.

Equation (36) specifies the dependence of the random variable n on the demographic parameters. For example, we can see that stronger bottlenecks—higher values of s—lead to an increase not just in the variance $\mathbb{V}[\Psi]$ (Fig. 4), but also in the sample variance by decreasing the number of available shared SNPs, $\mathbb{E}[n]$.

Application to Out-of-Africa expansion of *Drosophila melanogaster*

To test our coalescent-based predictions for ψ, we compute ψ for a specific demographic event in two ways. First, we use Equation (1) to compute ψ directly from genotypes in natural populations. Second, we make use of existing estimates of demographic parameters to evaluate our equations for $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$.

The demographic scenario we consider here is the Out-of-Africa expansion of *D. melanogaster*. It is generally agreed that the modern European populations trace to a small founding population as the species range expanded from Africa (e.g. Stephan and Li 2007 ; Arguello et al. 2019). As the founder event was directed from Africa to Europe, we expect to see ψ(Europe, Africa) > 0.

ψ computed from sequence data

For our empirical computations, we evaluated ψ from the sequences of intronic and intergenic X-chromosomal loci used for

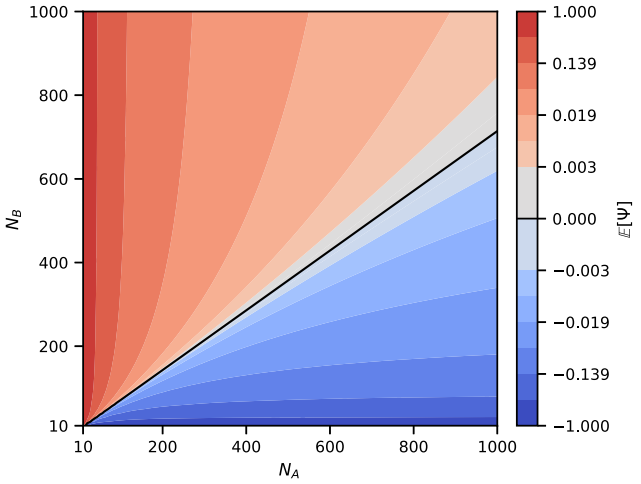

Fig. 5. Theoretical values of $\mathbb{E}[\Psi]$ (Equation (32a)) for varying values of N_A and N_B in the founder effect demography of Fig. 1d. Parameters s and t are fixed, with $s = 20$ and $t = 50$. The black line shows parameter sets (N_A, N_B) for which $\mathbb{E}[\Psi] = 0$.

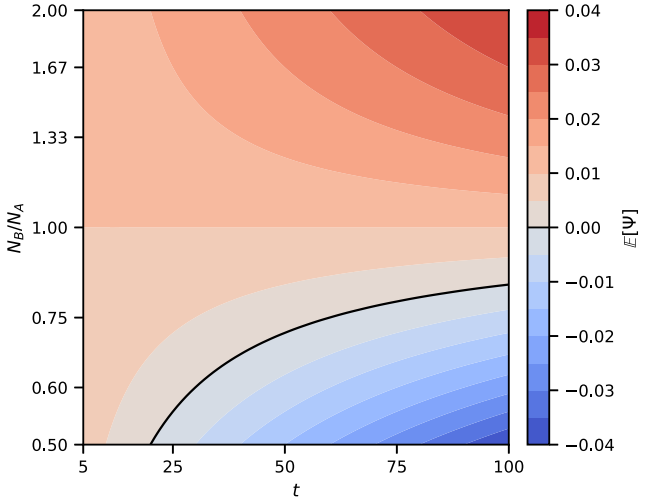

Fig. 6. Theoretical values of $\mathbb{E}[\Psi]$ (Equation (32a)) for varying values of t and N_B in the founder effect demography of Fig. 1d. Parameters s and N_A are fixed, with $s = 20$ and $N_A = 500$. The black line shows parameter sets (t, N_B) for which $\mathbb{E}[\Psi] = 0$.

demographic inference by Li and Stephan (2006), Laurent et al. (2011), and Duchen et al. (2013), originally obtained by Glinka et al. (2003) and Ometto et al. (2005). We downloaded sequences of X-chromosomal loci from the European Nucleotide Archive (`ebi.ac.uk/ena`), sequence IDs AJ568984 to AJ571588 and AM000058 to AM003900 (originally deposited by Glinka et al. 2003; Ometto et al. 2005).

Each locus had nucleotide data in the form of a single (haploid) genotype for a set of inbred lines from European (the Netherlands, NTH) and African (Zimbabwe, ZW) populations of *D. melanogaster*, as well as for a single line from a North American population of *Drosophila simulans*. The genetic sequence for each line was haploid due to the sequencing being performed with homozygous inbred lines. The total number of X-chromosomal loci was 229, with locus sequence lengths ranging from 210 to 784 nucleotides (median 563).

The *D. simulans* sequence was used in place of the ancestral genotype in the analysis. Across the 229 loci, the maximum number of lines sequenced for the Netherlands population was 12 and the minimum number of lines sequenced was 10. For the Zimbabwe population, the maximum was 12 and the minimum was 9.

Separately for each locus, we used *MUSCLE v5.1* (Edgar 2004) with default settings (`-perturb 0 -perm none -consiters 2 -refineiters 100`) to perform a joint multiple sequence alignment for the lines from the NTH and ZW populations of *D. melanogaster* as well as the *D. simulans* line.

To compute ψ for a set of loci, we generated a sample of 1,000 sets of four lines, two from the NTH population and two from the ZW population. The sets of four were sampled with replacement, but each set had two distinct NTH lines and two distinct ZW lines ("distinct" here refers to distinct sample labels, not to distinctness of the genotypes).

For each set of four lines together with the *D. simulans* line, we discarded sites that had insertions or deletions in the alignment of five sequences. We next discarded invariable sites as well as sites with three or more distinct alleles. Next, we discarded sites that failed to meet a sharing criterion. In particular, we kept only those shared (biallelic) sites in the sense of the definition in Equation (1), requiring the derived allele to be present in at least one copy in both NTH and ZW populations and to be polymorphic in the pooled pair of populations. We then computed ψ using Equation (2) by sampling a single site from the final set of shared sites.

To understand the uncertainty in the ψ computation that arises from differences in evolutionary history across loci, we analyzed 1,000 bootstrap replicate datasets, where each bootstrap replicate involves a resample of 229 loci. In particular, in each bootstrap replicate, we first sampled 229 loci with replacement. Next, we generated 1,000 sets of four lines and computed ψ, as described in the previous two paragraphs.

For each bootstrap replicate, we averaged ψ over the 1,000 values, each obtained from a random set of four lineages, obtaining a mean ψ for that replicate. We also obtained a variance across the 1,000 values.

Considering the 1,000 bootstrap replicates, the median of the mean ψ values was

$$\psi(\text{NTH, ZW}) = 0.5280, \tag{37}$$

with 95% of mean ψ values lying in the interval (0.4860, 0.5710).

The median across 1,000 bootstrap replicates of the variance of ψ was equal to

$$\text{Var}\big[\psi(\text{NTH, ZW})\big] = 0.5244, \tag{38}$$

and the interval containing the variance values from 95% of the replicates was (0.4811, 0.5666).

$\mathbb{E}[\Psi]$ computed from demographic estimates

We now compare empirical ψ values with the ψ values predicted by demographic models; we use demographic models that have been inferred in studies of the European founder event in *D. melanogaster*.

Multiple studies have estimated population sizes and divergence times for *D. melanogaster*. In particular, Li and Stephan (2006) used a maximum likelihood method based on the joint SFS, and Laurent et al. (2011) and Duchen et al. (2013) used approximate Bayesian computation. All three studies used the same set of X-chromosomal sequences from Glinka et al. (2003),

Table 1. Inferred demographic parameters reported for European and African *D. melanogaster* populations in the previous studies that used X-chromosomal loci, and corresponding model-predicted values of $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$.

Article	t	N_{AFR}	European bottleneck	$\mathbb{E}[\Psi]$(EUR, AFR)	$\mathbb{V}[\Psi]$(EUR, AFR)
Li and Stephan (2006)	158,000	8,603,000	$N_{EUR,b} = 2{,}200$ for $t_b = 3{,}400$ generations, then $N_{EUR} = 1{,}075{,}000$	0.3950	0.5372
Laurent et al. (2011)	168,490	3,589,770	Exponential growth from $N_{EUR,t} = 16{,}550$ to $N_{EUR,0} = 1{,}224{,}378$	0.5165	0.4970
Duchen et al. (2013)	194,984	4,975,360	Exponential growth from $N_{EUR,t} = 16{,}982$ to $N_{EUR,0} = 3{,}122{,}470$	0.4912	0.5092

In all studies, ten generations per year are assumed; we report times in generations.

with the Netherlands representing Europe and Zimbabwe representing Africa.

The study of Laurent et al. (2011) incorporates Asian samples, and the study of Duchen et al. (2013) adds North American samples, but here we focus on subsets of the inferred demographic parameters in these studies, specifically on the divergence of African and European populations shared by all three studies.

The modes of parameter estimates from the three articles are summarized in Table 1. The model of Li and Stephan (2006) assumes a prolonged bottleneck of constant size in the European population. Laurent et al. (2011) and Duchen et al. (2013) instead assumed exponential growth in Europe. All three models assume constant population size in the African population after the split with the European population.

For Li and Stephan (2006), the values for the model in their Fig. 1a appear on p. 1,582–1,583; African population size is labeled $\hat{N}_{A0}$ by Li and Stephan (2006, p. 1,582), current European population size is $\widehat{N}_{E0}$ (p. 1,583), bottleneck European population size is $\hat{N}_{E1}$ (p. 1,583), and time variables are $\hat{t}_{E0}$ for bottleneck length and $\hat{t}_{E1}$ for post-bottleneck interval length (p. 1,583). In their inference algorithm, although the data are from the X chromosome, the authors use a coalescent process with N diploid individuals and a parameterization $\Theta = 4N\mu$. In our calculations in the section on "Expectation and variance of Ψ for specific demographic models," we have also assumed that the coalescent process has N diploid individuals and $2N$ lineages. Because the parameterization of Li and Stephan (2006) matches our parameterization in Equation (30), we use the values from Li and Stephan (2006) directly.

For Duchen et al. (2013), we extracted parameter values for their model C (defined in their Table S2) for the African population from their Table 4 (N_{Ac}) and for the European population from their Table 5 (N_{Ea} and N_{Ec} immediately after the split and at the present time, respectively, and time T_{AE} since the split), exponentiating values reported logarithmically. The tables of Duchen et al. (2013) report numbers of diploid individuals in a population; hence, we use their population size values directly to calculate $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$.

For Laurent et al. (2011), we extracted estimates of parameter values from the X-chromosome column in their Table 3 for the model in their Fig. 1. Laurent et al. (2011) assumed equal proportions of males and females in the population, explicitly considering the X chromosome, so that the total number of lineages in a population of size N is $\frac{3}{2}$N, and $\Theta = 3N\mu$. To match the diploid autosomal parameterizations under which we derived our theoretical expressions in the section "Expectation and variance of Ψ for specific demographic models," we re-scaled population sizes reported in Table 3 of Laurent et al. (2011) by multiplying them by $\frac{3}{4}$, and we then used them to calculate $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$; this procedure is equivalent to rederiving Equation (28) with $3N$ in place of $4N$ and then inserting the N values from Laurent et al. (2011) directly.

For the demographic parameters of Li and Stephan (2006) that used a simple bottleneck model, we used Equation (30) with $N_A = N_{EUR}$, $N_b = N_{EUR,b}$, $N_B = N_{AFR}$, and $t_b = t_b$ from Table 1. For the demographic parameters of Laurent et al. (2011) and Duchen et al. (2013), which used an exponential growth model, we used Equation (28) with $N_{A,0} = N_{EUR,0}$, $N_{A,t} = N_{EUR,t}$, $N_B = N_{AFR}$, and the exponential growth rate r_A computed from European population sizes using Equation (23).

The values of $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$ computed for each set of parameters are shown in Table 1. For the Laurent et al. (2011) and Duchen et al. (2013) demographies, the value we expect from the coalescent theory—$\mathbb{E}[\Psi]$ in Table 1—lies inside the 95% bootstrap interval for the value obtained directly from data in Equation (37).

The variance in our empirical calculation (Equation (38)) closely matches the values of $\mathbb{V}[\Psi]$ implied in Table 1 by the demographic models, with all three values lying in the 95% bootstrap interval.

Discussion

Summary

We have examined the directionality index Ψ as a random variable under coalescent models of two populations with a shared demographic history. Using this formulation, we have derived exact values for the expectation $\mathbb{E}[\Psi]$ and variance $\mathbb{V}[\Psi]$ of the directionality index for four parameterizations of a population split demography. We have explored the behavior of the expectation and variance, showing the dependence of Ψ on demographic parameters and identifying parameter regions for which a positive value of ψ does not necessarily mean that the "A" population is more distant from the source of a range expansion. Our expression for $\mathbb{V}[\Psi]$ also allowed us to connect the sample variance of ψ across many sites to the size of the ancestral population. Finally, we showed how our theoretical results can be used to compare the predictions of demographic models with empirical observations.

Our explorations of the theoretical behavior of ψ show that in a sample of size 4 lineages, $\mathbb{E}[\Psi]$ tends to be more sensitive to changes in the bottleneck strength s and derived population sizes N_A and N_B than to the time t since the population split (Fig. 4a–c). The variance $\mathbb{V}[\Psi]$, however, increases quickly with increasing t (Fig. 4d–f). These results are informative for considering the effects of bottlenecks on empirical values of ψ. For example, in a model in which a bottleneck is ancient, the variance of Ψ would be larger compared to a model with a recent bottleneck.

Expressions for $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$ (Equations (20), (26), (30), and (32)) do not depend on N_C, the ancestral population size. This insight suggests that predictions about range expansions under the model are largely unaffected by events in the shared history of the two populations. However, N_C does affect the number of SNPs available for empirical calculation. When data from many sites are used to compute ψ, the expected number of shared SNPs in the calculation is proportional to $\Theta_C = 4N_C\mu$; for small N_C, we might not observe enough shared SNPs for the empirical computation of ψ to accurately reflect a model-based prediction.

Comparison of theoretical and empirical ψ for *D. melanogaster*

In an application to data from *D. melanogaster*, empirical evaluation of ψ revealed a positive value in a scenario with a European population in the role of the "A" population and an African population in the role of the "B" population. This observation is consistent with the higher level of drift in European populations of *D. melanogaster* than in African populations: the analysis accords with the general understanding of *D. melanogaster* demographic history. Further, for two of three *D. melanogaster* modeling studies, the empirical value of ψ matched the predictions for $\mathbb{E}[\Psi]$ from the demographic models (Table 1).

Connections

Our results recapitulate some of the insights of the branching process analysis of the discrete-time expansion model of Peter and Slatkin (2015). In that model, the expectation of Ψ was found to be $\mathbb{E}[\Psi] = \frac{1}{2}(\frac{N}{k} - 1)t$ where N is the population size of each deme, k is the founder population size during settlement of a new deme, and t is the settlement time of the tth deme, the integer time variable that counts sequential founder events from the origin to the most recently settled deme. This expression shows that $\mathbb{E}[\Psi]$ increases with smaller values for sizes of the founder populations (k) and with the number of founder events (t).

In our formulations, the founder population size corresponds to the initial population size after the split $N_{A,t}$ in the exponential growth model, the bottleneck population size N_b in the bottleneck model, and the reciprocal of the bottleneck strength s in the instantaneous bottleneck model. In these cases, we have observed a similar pattern in the magnitude of $\mathbb{E}[\Psi]$, which increases with smaller $N_{A,t}$ (Equation (26a)), smaller N_b (Equation (30a)), or larger s (Equation (32a), Fig. 4).

The role of the time variable t, however, differs between the model of Peter and Slatkin (2015) and our analysis. In particular, the linear chain of many populations by Peter and Slatkin (2015) gives rise to a linear dependence of $\mathbb{E}[\Psi]$ on t, whereas our two-population model produces a nonlinear dependence of $\mathbb{E}[\Psi]$ on t due to interactions among various demographic parameters (Fig. 6).

Our study follows a similar spirit to the work of DeGiorgio et al. (2011), who derived expressions for the distribution of pairwise coalescence times in a serial founder model with a sequence of multiple bottlenecks. Many population-genetic statistics are functions of expected pairwise coalescence times, among them F_{ST} (Slatkin 1991) and f_4 (Peter 2016). Because our study uses ratios of certain expected branch lengths rather than the pairwise coalescence times themselves, our expressions for $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$ are perhaps more closely connected to analyses that focus on internal and external branch length computations and other coalescent-based branch length ratios (Fu and Li 1993; Ferretti et al. 2017; Alimpiev and Rosenberg 2022).

An additional connection to other coalescent studies (Tajima 1983; Takahata and Nei 1985; Takahata and Slatkin 1990; Szpiech and Rosenberg 2011; Rosenberg 2013; Guerra and Nielsen 2022; Peter 2022) is that we have focused on the case of 4 sampled lineages. In many problems, the four-lineage analysis is the simplest non-trivial case, it can be studied analytically, and it provides insights useful for larger samples.

Further work

Our models are focused on pairs of populations, bottlenecks, and infinitely-many-sites mutation. Extended models could potentially consider additional phenomena; for example, recurrent and reverse mutation, the influence of natural selection on coalescence times for some sites, linkage among sites, and demographies that allow migration after population divergence. In the case of migration, a more recent study of than those that underlie Table 1 suggests a high rate of back-migration of *D. melanogaster* from Europe to Africa (Arguello et al. 2019), so that predictions for ψ in models that include migration would be meaningful. One approach to considering migration with ψ is an extension of the discrete-time expansion model of Peter and Slatkin (2015). Including migration between demes after founding events and exploring its impact on ψ is possible in a simulation-based extension of their model (Kemppainen et al. 2024).

In the framework of our theoretical analysis with four lineages and two populations, migration would allow for private mutations that appeared in population A to be introduced into population B, decreasing the value of ψ. Topologies that are more likely to generate shared mutations (such as δ, ε, and ζ in Fig. 2) would be observed more often, due to lineages being transported between populations in the time since the split between populations A and B, altering the values of $\mathbb{E}[\Psi]$ and $\mathbb{V}[\Psi]$. The theory could potentially be pursued by adding ψ to coalescent models that allow for post-divergence migration (Wakeley 1996 ; Rosenberg and Feldman 2002; Teshima and Tajima 2002; Wilkinson-Herbots 2008; Hobolth et al. 2011; Wilkinson-Herbots 2012).

Data availability

The genetic sequences for *D. melanogaster* X-chromosomal loci were obtained from the European Nucleotide Archive (ebi.ac.uk/ena), sequence IDs AJ568984 to AJ571588, and AM000058 to AM003900. Code to perform the computation of ψ from data and generate figures and tables in this article is available at github.com/EgorLappo/coalescent-psi.

Acknowledgments

We thank Ben Peter for help with a key aspect of the calculation.

Funding

We acknowledge National Institutes of Health grant R01 HG005855.

Conflicts of interest

The authors declare no conflict of interest.

Literature cited

Achaz G. 2009. Frequency spectrum neutrality tests: one for all and all for one. Genetics. 183:249–258. https://doi.org/10.1534/genetics.109.104042.

Alimpiev E, Rosenberg NA. 2022. A compendium of covariances and correlation coefficients of coalescent tree properties. Theor Popul Biol. 143:1–13. https://doi.org/10.1016/j.tpb.2021.09.008.

Arguello JR, Laurent S, Clark AG. 2019. Demographic history of the human commensal *Drosophila melanogaster*. Genome Biol Evol. 11:844–854. https://doi.org/10.1093/gbe/evz022.

Bunnefeld L, Frantz LAF, Lohse K. 2015. Inferring bottlenecks from genome-wide samples of short sequence blocks. Genetics. 201:1157–1169. https://doi.org/10.1534/genetics.115.179861.

Caicedo AL *et al.* 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. 3:e163. https://doi.org/10.1371/journal.pgen.0030163.

DeGiorgio M, Degnan JH, Rosenberg NA. 2011. Coalescence-time distributions in a serial founder model of human evolutionary history. Genetics. 189:579–593. https://doi.org/10.1534/genetics.111.129296.

DeGiorgio M, Jakobsson M, Rosenberg NA. 2009. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. Proc Natl Acad Sci U S A. 106:16057–16062. https://doi.org/10.1073/pnas.0903341106.

Deshpande O, Batzoglou S, Feldman MW, Cavalli-Sforza LL. 2009. A serial founder effect model for human settlement out of Africa. Proc R Soc Lond B Biol Sci. 276:291–300. https://doi.org/10.1098/rspb.2008.0750.

Duchen P, Živković D, Hutter S, Stephan W, Laurent S. 2013. Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. Genetics. 193:291–301. https://doi.org/10.1534/genetics.112.145912.

Durrett R. 2008. Probability models for DNA sequence evolution. 2nd ed. New York: Springer.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797. https://doi.org/10.1093/nar/gkh340.

Edmonds CA, Lillie AS, Cavalli-Sforza LL. 2004. Mutations arising in the wave front of an expanding population. Proc Natl Acad Sci U S A. 101:975–979. https://doi.org/10.1073/pnas.0308064100.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet. 9:e1003905. https://doi.org/10.1371/journal.pgen.1003905.

Excoffier L, Foll M, Petit RJ. 2009. Genetic consequences of range expansions. Annu Rev Ecol Evol Syst. 40:481–501. https://doi.org/10.1146/ecolsys.2009.40.issue-1.

Excoffier L, Ray N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. Trend Ecol Evol. 23:347–351. https://doi.org/10.1016/j.tree.2008.04.004.

Ferretti L, Ledda A, Wiehe T, Achaz G, Ramos-Onsins SE. 2017. Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. Genetics. 207:229–240. https://doi.org/10.1534/genetics.116.188763.

Ferretti L, Perez-Enciso M, Ramos-Onsins S. 2010. Optimal neutrality tests based on the frequency spectrum. Genetics. 186:353–365. https://doi.org/10.1534/genetics.110.118570.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics. 133:693–709. https://doi.org/10.1093/genetics/133.3.693.

Galtier N, Depaulis F, Barton NH. 2000. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics. 155:981–987. https://doi.org/10.1093/genetics/155.2.981.

Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics. 165:1269–1278. https://doi.org/10.1093/genetics/165.3.1269.

Guerra G, Nielsen R. 2022. Covariance of pairwise differences on a multi-species coalescent tree and implications for F_{st}. Philos Trans R Soc B Biol Sci. 377:20200415. https://doi.org/10.1098/rstb.2020.0415.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5:e1000695. https://doi.org/10.1371/journal.pgen.1000695.

Hallatschek O, Nelson DR. 2008. Gene surfing in expanding populations. Theor Popul Biol. 73:158–170. https://doi.org/10.1016/j.tpb.2007.08.008.

Hobolth A, Andersen LN, Mailund T. 2011. On computing the coalescence time density in an isolation-with-migration model with few samples. Genetics. 187:1241–1243. https://doi.org/10.1534/genetics.110.124164.

Ioannidis AG *et al.* 2021. Paths and timings of the peopling of Polynesia inferred from genomic networks. Nature. 597:522–526. https://doi.org/10.1038/s41586-021-03902-8.

Kemppainen P, Schembri R, Momigliano P. 2024. Boundary effects cause false signals of range expansions in population genomic data. Mol Biol Evol. 41:msae091. https://doi.org/10.1093/molbev/msae091.

Klopfstein S, Currat M, Excoffier L. 2006. The fate of mutations surfing on the wave of a range expansion. Mol Biol Evol. 23:482–490. https://doi.org/10.1093/molbev/msj057.

Laurent SJ, Werzner A, Excoffier L, Stephan W. 2011. Approximate Bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. Mol Biol Evol. 28:2041–2051. https://doi.org/10.1093/molbev/msr031.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2:e166. https://doi.org/10.1371/journal.pgen.0020166.

Liu X, Fu YX. 2020. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. Genome Biol. 21:280. https://doi.org/10.1186/s13059-020-02196-9.

Marchi N, Schlichta F, Excoffier L. 2021. Demographic inference. Curr Biol. 31:R276–R279. https://doi.org/10.1016/j.cub.2021.01.053.

Nielsen R *et al.* 2005. Genomic scans for selective sweeps using SNP data. Genome Res. 15:1566–1575. https://doi.org/10.1101/gr.4252305.

Nielsen R *et al.* 2009. Darwinian and demographic forces affecting human protein coding genes. Genome Res. 19:838–849. https://doi.org/10.1101/gr.088336.108.

Ometto L, Glinka S, De Lorenzo D, Stephan W. 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol. 22:2119–2130. https://doi.org/10.1093/molbev/msi207.

Peischl S, Excoffier L. 2016. Expansion load: recessive mutations and the role of standing genetic variation. In: Barrett SCH, Colautti RI, Dlugosch KM, Rieseberg LH, editors. Invasion genetics. 1st ed. Chichester, UK: Wiley. p. 218–231.

Peter BM. 2016. Admixture, population structure, and *F*-statistics. Genetics. 202:1485–1501. https://doi.org/10.1534/genetics.115.183913.

Peter BM. 2022. A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis. Philos Trans R Soc B Biol Sci. 377:20200413. https://doi.org/10.1098/rstb.2020.0413.

Peter BM, Slatkin M. 2013. Detecting range expansions from genetic data. Evolution. 67:3274–3289. https://doi.org/10.1111/evo.12202.

Peter BM, Slatkin M. 2015. The effective founder effect in a spatially expanding population. Evolution. 69:721–734. https://doi.org/10.1111/evo.2015.69.issue-3.

Pool JE, Hellmann I, Jensen JD, Nielsen R. 2010. Population genetic inference from genomic sequence variation. Genome Res. 20:291–300. https://doi.org/10.1101/gr.079509.108.

Puckett EE, Munshi-South J. 2019. Brown rat demography reveals pre-commensal structure in eastern Asia before expansion into Southeast Asia. Genome Res. 29:762–770. https://doi.org/10.1101/gr.235754.118.

Ramachandran S *et al.* 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A. 102:15942–15947. https://doi.org/10.1073/pnas.0507611102.

Ronen R, Udpa N, Halperin E, Bafna V. 2013. Learning natural selection from the site frequency spectrum. Genetics. 195:181–193. https://doi.org/10.1534/genetics.113.152587.

Rosenberg NA. 2013. Discordance of species trees with their most likely gene trees: a unifying principle. Mol Biol Evol. 30: 2709–2713. https://doi.org/10.1093/molbev/mst160.

Rosenberg NA, Feldman MW. 2002. The relationship between coalescence times and population divergence times. In: Slatkin M, Veuille M, editors. Modern developments in theoretical population genetics. Oxford, UK: Oxford University Press. p. 130–164.

Schlichta F, Moinet A, Peischl S, Excoffier L. 2022. The impact of genetic surfing on neutral genomic diversity. Mol Biol Evol. 39: msac249. https://doi.org/10.1093/molbev/msac249.

Slatkin M. 1991. Inbreeding coefficients and coalescence times. Genet Res. 58:167–175. https://doi.org/10.1017/S0016672300029827.

Slatkin M, Excoffier L. 2012. Serial founder effects during range expansion: a spatial analog of genetic drift. Genetics. 191:171–181. https://doi.org/10.1534/genetics.112.139022.

Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics. 129:555–562. https://doi.org/10.1093/genetics/129.2.555.

Stephan W, Li H. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity. 98:65–68. https://doi.org/10.1038/sj.hdy.6800901.

Szpiech ZA, Rosenberg NA. 2011. On the size distribution of private microsatellite alleles. Theor Popul Biol. 80:100–113. https://doi.org/10.1016/j.tpb.2011.03.006.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105:437–460. https://doi.org/10.1093/genetics/105.2.437.

Takahata N, Nei M. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics. 110:325–344. https://doi.org/10.1093/genetics/110.2.325.

Takahata N, Slatkin M. 1990. Genealogy of neutral genes in two partially isolated populations. Theor Popul Biol. 38:331–350. https://doi.org/10.1016/0040-5809(90)90018-Q.

Teshima KM, Tajima F. 2002. The effect of migration during the divergence. Theor Popul Biol. 62:81–95. https://doi.org/10.1006/tpbi.2002.1580.

Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics. 172:1607–1619. https://doi.org/10.1534/genetics.105.048223.

Wakeley J. 1996. Pairwise differences under a general model of population subdivision. J Genet. 75:81–89. https://doi.org/10.1007/BF02931753.

Wakeley J, Hey J. 1997. Estimating ancestral population parameters. Genetics. 145:847–855. https://doi.org/10.1093/genetics/145.3.847.

Wilkinson-Herbots HM. 2008. The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. Theor Popul Biol. 73:277–288. https://doi.org/10.1016/j.tpb.2007.11.001.

Wilkinson-Herbots HM. 2012. The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. Theor Popul Biol. 82:92–108. https://doi.org/10.1016/j.tpb.2012.05.003.

Zhan S *et al.* 2014. The genetics of monarch butterfly migration and warning colouration. Nature. 514:317–321. https://doi.org/10.1038/nature13812.

Editor: A. Kern