Demographic Inference Using Spectral Methods on SNP Data, with an Analysis of the Human Out-of-Africa Expansion

Sergio Lukić*,†,1 and Jody Hey*

*Department of Genetics, Rutgers University, Piscataway, New Jersey 08854, and †School of Natural Sciences, Institute for Advanced Study, Princeton, New Jersey 08540

ABSTRACT We present an implementation of a recently introduced method for estimating the allele-frequency spectrum under the diffusion approximation. For single-nucleotide polymorphism (SNP) frequency data from multiple populations, the method computes numerical solutions to the allele-frequency spectrum (AFS) under a complex model that includes population splitting events, migration, population expansion, and admixture. The solution to the diffusion partial differential equation (PDE) that mimics the evolutionary process is found by means of truncated polynomial expansions. In the absence of gene flow, our computation of frequency spectra yields exact results. The results are compared to those that use a finite-difference method and to forward diffusion simulations. In general, all the methods yield comparable results, although the polynomial-based approach is the most accurate in the weak-migration limit. Also, the economical use of memory attained by the polynomial expansions makes the study of models with four populations possible for the first time. The method was applied to a four-population model of the human expansion out of Africa and the peopling of the Americas, using the Environmental Genome Project (EGP) SNP database. Although our confidence intervals largely overlapped previous analyses of these data, some were significantly different. In particular, estimates of migration among African, European, and Asian populations were considerably lower than those in a previous study and the estimated time of migration out of Africa was earlier. The estimated time of founding of a human population outside of Africa was 52,000 years (95% confidence interval: 36,000–80,800 years).

THE study of demographic history from genetic data is important for understanding how populations have diverged and come to be in their present state. In the case of human populations, genetic studies of demographic history can be a great complement to archaeological studies of human prehistory. Having a model of demographic history is also important for facilitating the identification of genomic regions that have been evolving under selective pressures. The inference of signatures of natural selection from DNA sequence data sets requires accounting for different demographic forces that contribute to shaping such patterns (Risch and Merikangas 1996; Nielsen 2001; Goldstein and Chikhi 2002; Shriver *et al.* 2003; Laberge *et al.* 2005; Schaffner *et al.* 2005; Lao *et al.* 2006; Chen 2012).

The extraction of information from patterns of variation in genetic data requires mathematical models that can capture the diversity and richness of population histories, as well as efficient statistical tools for fitting models to data. The demands on models and on tools are especially great for questions about human populations, which have a complex history and for which large amounts of data can often be brought to bear. An important approach to modeling demographic history and comparing models to genetic data focuses on the allele-frequency spectrum (AFS). Given DNA sequence data from several individuals in K populations, the resulting single-nucleotide polymorphism (SNP) joint AFS is a K-dimensional matrix in which each cell specifies the number of derived alleles that were found at a particular set of K frequencies in the data. Inference of a demographic model of history for a data set consists of finding a model and a set of parameter values that correspond to an expected AFS that closely resembles the AFS that was observed in the data.

doi: 10.1534/genetics.112.141846
Manuscript received May 7, 2012; accepted for publication July 12, 2012
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.141846/-/DC1.

[1] Corresponding author: Institute for Advanced Study, Einstein Dr., Princeton, NJ 08540. E-mail: lukic@ias.edu

In this article we present a software implementation of a recently introduced spectral method (Lukic *et al.* 2011) for estimating the AFS under the diffusion approximation. The implementation allows for the study of a broad class of demographic models for multiple closely related populations. In this Introduction we review the history of methods for solving these types of problems in population genetics, as well as some of the limitations of the different approaches.

A brief history of the use of diffusion processes for approximating the AFS

The classical approach to computing the AFS was developed by Fisher (1930), Wright (1931), and Kimura (1964), who introduced forward diffusion processes and irreversible mutations in a single population to model the evolutionary process. The theory was extended by Kimura (1969) to study many nucleotide positions by introducing the infinite-sites mutation model. Coalescent models can also be used to develop exact solutions for the AFS under the infinite-sites assumption, and these are particularly amenable for some models with multiple populations (Wakeley and Hey 1997).

Until relatively recently, most applications of classical forward diffusion theory to demographic inference were limited to models that assumed some form of equilibrium (mutation/drift equilibrium, mutation/selection equilibrium, etc.) and could be solved exactly (Sawyer and Hartl 1992; Ewens 2004). However, the recent introduction of numerical methods to integrate arbitrary diffusion equations has allowed for the relaxing of the assumption of equilibrium so that general nonequilibrium scenarios can be studied. In Williamson *et al.* (2005) a finite-difference scheme was used to numerically solve the equations associated with a model that combined the effects of selection and population-size growth in one population. The finite-difference method is a classical technique for numerically solving partial differential equations in which the density of allele frequencies is approximated by a piecewise linear function, and the derivatives are approximated by means of finite subtractions. The piecewise linear approximation relies upon a grid defined on frequency space, and the accuracy of the method increases as the grid becomes finer. Later in Gutenkunst *et al.* (2009), these techniques were implemented in the program ∂a∂i and extended to consider models with two and three simultaneous populations that undergo random drift, migration events, arbitrary population size changes, admixture events, and directional selection (*e.g.*, Xing *et al.* 2010; Albert *et al.* 2011; Jensen and Bachtrog 2011).

Despite the superficial differences between these approaches to population genetics, it is known that both coalescent and forward diffusion processes are dual processes in an analytic sense (Griffiths and Spano 2010). The connection between these dual processes can be made explicit through the application of orthogonal polynomial theory to Wright–Fisher processes (see Griffiths and Spano 2010 for a recent review on the topic). Previous studies that applied orthogonal polynomial theory to the two-allele neutral Wright–Fisher process made it possible to solve the associated diffusion equations for one population (Myers *et al.* 2008). Also, a recently introduced spectral expansion makes it possible to exactly integrate the same diffusion equations with arbitrary diploid selection (Song and Steinrücken 2011). In this study, we employ series of orthogonal polynomials to solve multipopulation Wright–Fisher processes (Lukic *et al.* 2011), and we use these solutions to infer models of demographic history from genetic data. In general, these methods provide approximate solutions for finite numbers of polynomials and approach the exact solution in the limit as the number of polynomials goes to infinity. In the particular case of neutral models wherein each population size is fixed for the duration that a population persists, and there is no gene exchange between populations, our polynomial-based approach provides the exact solution of the AFS, which was first described using a coalescent approach by Wakeley and Hey (1997) (see supporting information, File S1, section 1).

The curse of dimensionality

Another important motivation for the use of orthogonal polynomials is to be able to tackle models in which the high dimensionality of the frequency spectra becomes an important limitation. As the number of variables used to approximate the density of population allele frequencies grows exponentially, the total number of simultaneous populations that one can study becomes limited. In particular, if we use a grid of G grid points per population to approximate the density of allele frequencies ϕ as a piecewise linear function on the grid, the number of variables used in the approximation is G^K, with K the total number of populations.

The use of truncated polynomial expansions to approximate the density of population allele frequencies has two main advantages with respect to the piecewise linear approximations on grids that are used in finite-difference methods:

First, the contributions to an allele-frequency spectrum built from C haploid genomes sampled per population can be located in the first $(C - 2)^K$ terms of lowest degree of a polynomial expansion of the density of population allele frequencies ϕ. More precisely, given an AFS $f_{i_1,i_2,\dots i_K}$ built from C sampled chromosomes in K populations, a particular demographic scenario θ, and a joint density of population frequencies $\phi(x|\theta)$ associated with θ, we know that the AFS and the joint density of frequencies are related (Sawyer and Hartl 1992) as

$$f_{i_1,\,i_2,\dots i_K}(\theta) = \frac{(C!)^K}{i_1!(C-i_1)!\cdots i_K!(C-i_K)!} \times \int_{[0,1]^K} \phi(x|\theta) x_1^{i_1}\,(1-x_1)^{C-i_1}\cdots x_K^{i_K}\,(1-x_K)^{C-i_K}\,dx_1\cdots dx_K. \quad (1)$$

If $\phi(x|\theta)$ is expanded in the basis of shifted Jacobi polynomials $T_n(x) = \sqrt{(n+2)(2n+3)/(n+1)}P_n^{(1,1)}(2x-1)$ (see *Appendix*),

$$\phi(x|\theta) = \sum_{n_1,\ldots,n_K=0}^{\infty} a_{n_1,\ldots,n_K}(\theta)T_{n_1}(x_1)\cdots T_{n_K}(x_K),$$

the following integrals,

$$\int_{[0,1]^K} T_{n_1}(x_1)\cdots T_{n_K}(x_K)x_1^{i_1}(1-x_1)^{C-i_1}\cdots x_K^{i_K}(1-x_K)^{C-i_K}dx_1\cdots dx_K = 0, \tag{2}$$

vanish for $n_1 > C-2$ or $n_2 > C-2 \ldots$ or $n_K > C-2$. Therefore, only the first $(C-2)^K$ terms of lowest degree in the polynomial expansion yield nonzero contributions to the AFS in Equation 1, for $0 < i_1 < C, \ldots, 0 < i_K < C$. This means that the information in $\phi(x, t)$ required to compute the AFS can be represented as a vector in the vector space spanned by $\{T_i(x)\}_{i=0}^{i=C-2}$, *i.e.*, the vector space of polynomials of degree bounded by $C-2$ (see *Appendix*).

Second, it is known that polynomial approximations of smooth functions exhibit exponential convergence (Hesthaven *et al.* 2007). In particular, the amount of computational resources needed to numerically solve the multipopulation Wright–Fisher equations depends on the number of variables needed to approximate $\phi(x)$ (*e.g.*, number of floating-point values). We denote this number as n_ϕ. Any other relevant quantity in the algorithm, such as the number of variables needed to approximate the diffusion operator or the number of operations needed to evaluate the AFS, will be a function of n_ϕ. Therefore, any efficient algorithm that solves the multipopulation Wright–Fisher diffusion process needs to be designed with the goal of minimizing n_ϕ given a fixed bound for the numerical error. If one uses a finite-difference algorithm, then $n_\phi = G^K$ with G the number of grid points per population. On the other hand, if one uses spectral methods, then $n_\phi = (\Lambda + 2)^K - 2$. Here, Λ is the number of polynomials per population and the term $\Lambda + 2$ comes from the fact that each boundary component, except the ancestral and derived vertices of the K-cube, contributes its own polynomial expansion (Lukic *et al.* 2011). In both algorithms the diffusion operator is approximated as a sparse matrix of size $n_\phi \times n_\phi$, and the number of operations needed to evaluate Equation 1 is the same function of n_ϕ, K, and C. As the polynomial approximations of smooth functions exhibit exponential convergence, we can use lower values of Λ to accurately approximate the solutions of the diffusion equations. This allows us to use a larger number of populations.

As a simple illustration of the convergence properties of polynomial expansions, we approximated the equilibrium density of allele frequencies in one population [$\phi(x) = dx/x$ with $1/2\,N \leq x \leq 1$] by means of polynomial expansions $\phi_\Lambda(x)$. Also, we considered piecewise linear approximations $\phi_G(x)$ on grids of size G. For $0.005 \leq x \leq 1$, the error function decayed exponentially as

$$\|\phi(x)-\phi_\Lambda(x)\|_{L^1} = \int_{1/2N}^{1}|\phi(x)-\phi_\Lambda(x)|dx \sim 72.16e^{-0.65\Lambda},$$

with Λ the number of polynomials in the polynomial approximation of the density of allele frequencies; see Figure 1. However, in the case of piecewise linear approximations, the error function decayed as a power law

$$\|\phi(x)-\phi_G(x)\|_{L^1} = \int_{1/2N}^{1}|\phi(x)-\phi_G(x)|dx \sim 101.85G^{-1.92},$$

with G the number of grid points; see Figure 1. The parameters of the exponential and power law functions were estimated by means of the least-squares method. Therefore, given any fixed truncation $n_\phi = \Lambda = G$, a polynomial expansion gives a much more accurate approximation of the true density $\phi(x)$.

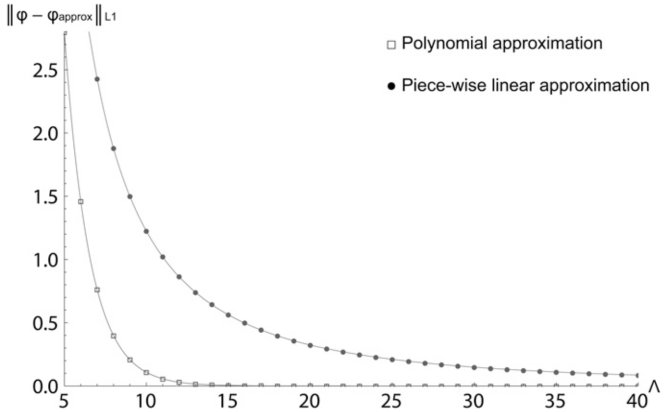

Figure 1 Decay of the error function between the equilibrium density of allele frequencies and its polynomial approximation and piecewise linear approximation. The horizontal axis denotes the number of polynomials for the lower curve and the number of grid points for the upper curve.

Outline of the article

In this Introduction we have reviewed some of the history of methods for modeling demographic history by means of approximations of the AFS. Additionally, we have described two numerical methods, the finite-difference method and the spectral method, that can be used to solve different diffusion equations that arise in the computation of AFS. The benefits of the latter approach include exact solutions for a large family of multipopulation models and rapidly converging approximations for a larger class of models. In the remainder of the article we describe the method in detail and assess how well it performs both based on simulated data sets and in comparison to a grid-based approach. Finally, we report the analyses of a four-population model of human history, using a SNP data set that has previously been analyzed under three-population models, using the grid-based approach.

Materials and Methods

Background

The observable object that we aim to reproduce theoretically is the allele-frequency spectrum. In this section, we review the definition of the AFS in a multipopulation context and how one can approximate it with numerical solutions of diffusion-based models that use truncated expansions by orthogonal polynomials (Lukic *et al.* 2011).

The joint AFS is defined as a K-dimensional matrix built from the allele counts observed in a sample of individuals from K different populations. Each value in the matrix is an expected number (in the case of an AFS calculated under a theoretical model) or an observed number (in the case of data) of diallelic polymorphisms that fall into a particular frequency class. We denote as $n_{i_1,i_2,\ldots,i_K}$ an entry of the observed joint AFS that specifies the number of SNPs in which their derived state occurs $i_1 \in [0, C_1]$ times in the first population, $i_2 \in [0, C_2]$ times in the second population, etc. Here, C_a is the total number of chromosomes sampled from the a^{th} population ($a = 1, \ldots, K$). For simplicity in the notation, we assume $C_a = C$ for all a throughout this article. Here, while $n_{i_1,i_2,\ldots,i_K}$ denotes an entry of the empirical AFS, we denote as $f_{i_1,i_2,\ldots,i_K}(\theta)$ the analogous entry of the theoretical AFS.

The AFS can be seen as an object derived from the distribution of population allele frequencies $\phi(x)$ on $[0, 1]^K$. In particular, if the derived allele frequencies of a SNP taken at random consist of a vector $\{x_a\}_{a=1}^{K}$, where x_a is the frequency of the SNP in population a, independently and identically distributed with respect to the distribution $\phi(x)$, the AFS consists of a finite sample of population alleles as defined in Equation 1. In our model-based approach $\phi(x)$ is interpreted as a present-time density that has been shaped by a historical Wright–Fisher process on a population tree specified by the parameters θ. We denote the resulting model-dependent joint density by $\phi(x|\theta)$. The parameters depend on the particular model and usually involve effective population sizes, migration rates, splitting times, admixture coefficients, population growth rates, etc. In the diffusion approximation to multipopulation Wright–Fisher processes exchanging migrants, the time evolution of $\phi(x, t)$ obeys a partial differential equation (PDE) of the type

$$\begin{aligned}\frac{\partial}{\partial t}\phi(x,t) = \sum_{a,b}\frac{1}{2}\frac{\partial^2}{\partial x_a \partial x_b}&\left(\delta^{ab}\,\frac{x_a(1-x_a)}{2N_{e,a}(t)}\phi(x,t)\right)\\ &-\frac{\partial}{\partial x_a}(m_{ab}(t)(x_b-x_a)\phi(x,t)) + \rho(x,t).\end{aligned} \qquad (3)$$

Here, $\{N_{e,a}(t)\}_{a=1}^{K}$ denotes the effective population sizes, $\{m_{ab}(t)\}_{a,b=1}^{K}$ denotes the fraction of chromosomes that population a receives from b, and the nonhomogeneous term $\rho(x, t)$ describes the total incoming/outgoing flow of SNPs per generation into the K-cube from different boundary components of the K-cube and from *de novo* mutations. These boundary conditions are treated in more detail later.

Approximate solutions by means of polynomial expansions: Our approach to approximate the solutions of Equation 3 assumes that the density solution ϕ can be expanded in a polynomial basis with time-dependent coefficients that can be determined numerically. The expansion consists of a contribution associated with the bulk of the K-cube and other different contributions associated with each boundary component. The expansion can be expressed compactly as

$$\phi(x,t) = \sum_{n_1=0}^{\Lambda+2}\cdots\sum_{n_K=0}^{\Lambda+2} a_{n_1\ldots n_K}(t)R_{n_1}(x_1)\cdots R_{n_K}(x_K), \qquad (4)$$

for a truncation parameter Λ. The set of functions $\{R_n(x)\}_{n=0}^{\Lambda+2}$ is defined as

$$R_n(x) = T_n(x) = \sqrt{\frac{(n+2)(2n+3)}{n+1}}P_n^{(1,1)}(2x-1), \quad n \leq \Lambda, \qquad (5)$$

$$R_n(x) = \delta(x), \quad n = \Lambda + 1, \qquad (6)$$

$$R_n(x) = \delta(1-x), \quad n = \Lambda + 2, \qquad (7)$$

where $P_n^{(1,1)}(z)$ are the classical Jacobi polynomials defined on the interval $-1 \leq z \leq 1$ with weight $w(z) = (1 - z)(1 + z)$, and $\delta(x)$ is the Dirac delta function (for more details on the basis of polynomials see the *Appendix*). When the migration coefficients m_{ab} in Equation 3 vanish, and $\Lambda = C - 2$ (with C the number of haploid genomes sampled in the definition of the AFS), the truncated expansion Equation 4 yields an exact formula for the AFS (see File S1, section 1 for a detailed derivation). However, in general the coefficients $a_{n_1\ldots n_K}(t)$ obey a numerically integrable ordinary differential equation, and the truncated polynomial expansion in Equation 4 gives rise to approximations of the AFS that have the potential to become more accurate as the truncation parameter Λ increases.

Some of the most important differences between scenarios with and without migration originate in the delicate dynamics of ϕ at the boundary of the K-cube. In the following paragraphs we review these boundary conditions. In the next subsection we examine how to deal with numerical artifacts (known as Gibbs phenomena) that appear when we consider approximations with finite Λ.

By the boundary dynamics of ϕ we mean the dynamics of those SNPs that have an allele fixed in some populations but that remain polymorphic in others and of those new SNPs that arise by the influx of mutations. This class of SNPs is described by those terms that are multiplied by Dirac deltas in Equation 4. For illustrative purposes we examine in detail the particular case of two populations. When $K = 2$, we can decompose Equation 4 as

$$\begin{aligned}\phi(x_1, x_2, t) &\\ = \phi^A(x_1, x_2, t) &+ \phi^B_{(x_2=0)}(x_1, t)\delta(x_2)\\ &+ \phi^B_{(x_2=1)}(x_1, t)\delta(1-x_2) + \phi^B_{(x_1=0)}(x_2, t)\delta(x_1)\\ &+ \phi^B_{(x_1=1)}(x_2, t)\delta(1-x_1)\\ &+ \phi^C_{(x_1=1, x_2=0)}(t)\delta(1-x_1)\delta(x_2)\\ &+ \phi^C_{(x_1=0, x_2=1)}(t)\delta(x_1)\delta(1-x_2).\end{aligned} \tag{8}$$

The terms that are multiplied by Dirac deltas represent the density of allele frequencies of those SNPs that are localized in the different boundary components. In particular, the A term is localized in the bulk of the square, the four B terms are localized in the edges of the square, and finally the two C terms are localized in the two vertices of the square that are not absorbing. The ancestral vertex ($x_1 = 0, x_2 = 0$) and the derived vertex ($x_1 = 1, x_2 = 1$) are absorbing and hence do not contribute SNPs to the density $\phi(x_1, x_2, t)$. Now, we can recover Equation 4 if we expand each term in Equation 8, using the basis of shifted Jacobi polynomials $T_n(x)$. In particular, we write the polynomial expansion of each term in Equation 8 as

$$\begin{aligned}\phi^A(x_1, x_2, t) &= \sum_{n,m=0}^{\infty} a^A_{nm}(t)T_n(x_1)T_m(x_2),\\ \phi^B_{(x_2=0)}(x_1, t) &= \sum_{n=0}^{\infty} a^B_{(x_2=0),n}(t)T_n(x_1),\\ \phi^B_{(x_2=1)}(x_1, t) &= \sum_{n=0}^{\infty} a^B_{(x_2=1),n}(t)T_n(x_1),\\ \phi^B_{(x_1=0)}(x_2, t) &= \sum_{m=0}^{\infty} a^B_{(x_1=0),m}(t)T_m(x_2),\\ \phi^B_{(x_1=1)}(x_2, t) &= \sum_{m=0}^{\infty} a^B_{(x_1=1),m}(t)T_m(x_2),\\ \phi^C_{(x_1=1, x_2=0)}(t) &= a^C_{(x_1=1, x_2=0)}(t),\\ \phi^C_{(x_1=0, x_2=1)}(t) &= a^C_{(x_1=0, x_2=1)}(t).\end{aligned} \tag{9}$$

The inflow/outflow of polymorphisms affects the dynamics of each term differently. For instance, in every generation, *de novo* mutations contribute a mass $2N_1u\delta(x_1 - 1/2N_1)$ to $\phi^B_{(x_2=0)}(x_1, t)$ and a mass $2N_2u\delta(x_2 - 1/2N_2)$ to $\phi^B_{(x_1=0)}(x_2, t)$, where u is the mutation rate. This means that a total of $2N_a u$ new SNPs at frequency $x_a = 1/2N_a$ appear each generation in population a due to *de novo* mutation events. On the other hand, random drift can fix some variants that were polymorphic in populations 1 and 2. This means that a SNP with allele frequencies distributed as $\phi^A(x_1, x_2, t)$ becomes a SNP with allele frequencies distributed as $\phi^B_{(x_2=0)}(x_1, t)$, $\phi^B_{(x_2=1)}(x_1, t)$, $\phi^B_{(x_1=0)}(x_2, t)$, or $\phi^B_{(x_1=1)}(x_2, t)$, depending on which frequency class becomes fixed ($x_2 = 0$, $x_2 = 1$, $x_1 = 0$, or $x_1 = 1$). Finally, variants segregating on any of the edges of the square can also become fixed because of random drift. Therefore, the density of SNPs at the edge ($x_1 = 1, x_2 = 0$), $\phi^C_{(x_1=1, x_2=0)}(t)$, receives SNPs that reach the fixed frequency $x_1 = 1$ from $\phi^B_{(x_2=0)}(x_1, t)$ and SNPs that reach $x_2 = 0$ from $\phi^B_{(x_1=1)}(x_2, t)$. Similarly, $\phi^C_{(x_1=0, x_2=1)}(t)$ receives SNPs that reach $x_1 = 0$ from $\phi^B_{(x_2=1)}(x_1, t)$ and SNPs that reach $x_2 = 1$ from $\phi^B_{(x_1=0)}(x_2, t)$. When the migration coefficients are zero, this dynamic can be integrated exactly (see File S1, section 1).

The boundary dynamics become more complicated when the migration rates are nonzero. This is due to the fact that when a SNP reaches fixation in one population and remains polymorphic in the other, it can become polymorphic again in the first population because of potential migration events. This contrasts with the zero-migration scenario, in which the number of populations where a SNP is polymorphic decreases as a function of time. The contributions due to migration events in the different components of Equation 8 follow a complicated formula that was previously analyzed in the literature. We refer to Lukic *et al.* (2011) for more detailed information on these terms and how to integrate the full dynamics using a numerical method such as a Runge–Kutta method. Here, for illustrative purposes, we write only the contribution to ϕ^A due to migration events,

$$\begin{aligned}&\frac{\Delta_{\mathrm{m}}\phi^A(x_1, x_2, t)}{\Delta t}\\ &= \phi^B_{(x_2=0)}(x_1, t)\delta(x_2 - m_{21}x_1)\\ &\quad + \phi^B_{(x_2=1)}(x_1, t)\delta(1 - m_{21}x_1 - x_2) + \phi^B_{(x_1=0)}(x_2, t)\delta(x_1 - m_{12}x_2)\\ &\quad + \phi^B_{(x_1=1)}(x_2, t)\delta(1 - m_{12}x_2 - x_1),\end{aligned}$$

where $\Delta_{\mathrm{m}}\phi/\Delta t$ denotes the change in the density of SNPs per unit of time due to migration events.

Gibbs phenomena: In general, it is not possible to work with infinite sums such as the ones introduced in Equation 9. This is why proper truncated expansions such as Equation 4 are used instead. Although truncated polynomial expansions can sometimes yield exact results, in general scenarios with migration we are faced with the task of approximating the influx of mutations $2Nu\delta(x - 1/2N)$, the contributions due to migration events [*e.g.*, $\phi^B_{(x_2=0)}(x_1, t)\delta(x_2 - m_{21}x_1)$], and the time evolution of the density ϕ by means of truncated polynomial expansions. Technically, using a truncated polynomial expansion to approximate a generalized function such as a Dirac delta is a far from perfect approximation. In particular, the approximations tend to exhibit oscillatory behaviors and a slow convergence rate. The circle of phenomena associated with imperfect polynomial approximations of nonsmooth functions is known as Gibbs phenomena.

A particular way to deal with this limitation consists of using smooth exponential functions with proper normalization constants, instead of plain Dirac deltas located near the boundary. For instance, the influx of mutations can be approximated by the term $c_k \exp(-k(\Lambda)x)$ (see section 3 in File S1 for a derivation of c_k), and terms due to migration events such as $\phi^B_{(x_2=0)}(x_1, t)\delta(x_2 - m_{21}x_1)$ can be approximated by

$\phi^{B}_{(x_2=0)}(x_1,t)c_k\exp(-k(\Lambda)(x_2-m_{21}x_1))$. Here, $k(\Lambda)$ is a function of the truncation parameter, and it is defined as the largest real number k that satisfies a bound for the truncation error between $\exp(-k(\Lambda)x)$ and its truncated polynomial approximation.

This treatment of the boundary conditions is superficially different from the conditions used by Gutenkunst *et al.* (2009). In the case of one population, one can prove that our solution and that of Gutenkunst *et al.* (2009) converge to the same exact solution (see section 3 in File S1 for a mathematical proof of this statement). The case of two or more populations with migration is significantly more difficult and we could not demonstrate that both treatments of the boundary conditions yield the same exact solution in the limits of infinite Λ and an infinitely fine grid. Our choice of exponential functions to approximate the Dirac deltas associated with migration events at the boundary is inspired by the approximation of $\delta(x-1/2N)$ in the one-population case. Although we demonstrate that this approximation converges to the exact solution in the one-population case, we do not have an equivalent demonstration for the case of the migration events at boundary. Therefore, our use of exponential functions in the case of migration events is justified by a heuristic argument and is not based on a mathematical proof that the associated approximations converge to the exact solution (for evidence using simulated data that both approaches converge to the same solution see *Comparison of different diffusion theory-based approaches* in *Results*).

Maximum-likelihood inference: We used a maximum-likelihood approach to infer the model parameters and a nonparametric bootstrap resampling approach to estimate confidence intervals and confidence regions. In particular, given the theoretical AFS $f_{i_1,i_2,\ldots,i_K}(\theta)$ as defined in Equation 1, we used the likelihood function $L(\theta|x)$ of a random Poisson field as defined in Sawyer and Hartl (1992),

$$L(\theta|x)=\prod_{i_1,i_2,\ldots,i_K}\frac{\exp\Big(-f_{i_1,i_2,\ldots,i_K}(\theta)\Big)f_{i_1,i_2,\ldots,i_K}(\theta)^{n_{i_1,i_2,\ldots,i_K}}}{n_{i_1,i_2,\ldots,i_K}!},\tag{10}$$

with $n_{i_1,i_2,\ldots,i_K}$ the number of counts observed in the specified bin of the empirical AFS. We also used the associated log-likelihood function

$$\ell(\theta)=\sum_{i_1,i_2,\ldots,i_K}n_{i_1,i_2,\ldots,i_K}\log\Big(f_{i_1,i_2,\ldots,i_K}(\theta)\Big)-f_{i_1,i_2,\ldots,i_K}(\theta)-\log n_{i_1,\ldots,i_K}!.\tag{11}$$

Simulations

To compare the different approaches to solving the diffusion equations, we used simulated data. To simulate the AFS we used an algorithm that mimics the forward diffusion process on population trees. Although there are very efficient coalescent-based tools to simulate data, we preferred to use the forward simulation approach because it exactly models the solution to the forward diffusion Equation 3 and because it can be adapted to incorporate the effects of natural selection more easily than coalescent-based models. As a quality check, we compared our Monte Carlo simulations with coalescent-based simulations [we used the program ms (Hudson 2002)]. As expected, both approaches produced nearly identical frequency spectra.

The basic approach is a standard one (Glasserman 2003) and consists of three steps:

1. Specify the sample sizes in the AFS and a demographic model. The model includes a population tree topology and relevant demographic parameters such as effective sizes, migration rates, and splitting times.
2. Generate Π stochastic sampling paths of SNP frequencies. These will yield population allele frequencies associated with Π independent SNPs.
3. Map each population allele frequency to the AFS matrix by using binomial sampling formulas for the target sample sizes and sum over all the Π contributions. The variance of each entry in the simulated AFS is inversely proportional to Π, as is standard in Monte Carlo simulation.

The simulation of a single sampling path begins by generating a random initial condition (*i.e.*, initial time of the sampling path and initial frequency), followed by sampling a sequence of allele frequencies using the Euler approximation

$$\begin{aligned}X_a(t+\Delta t)&=X_a(t)+\sum_b m_{ab}\,(X_b(t)-X_a(t))\Delta t\\&\quad+\sqrt{\frac{X_a(t)(1-X_a(t))}{2N_a}}\epsilon_a\sqrt{\Delta t}.\end{aligned}\tag{12}$$

In this approximation we replace Δt with a short time interval and we randomly sample K values $\{\epsilon_a\}_{a=1}^{a=K}$ from the standard normal distribution $N(0,\,1)$ in each time step (Matsumoto and Nishimura 1998).

The space of initial conditions depends on the particular model, and it consists of the following:

The frequencies in the ancestral population at drift–mutation equilibrium at time zero: These frequencies are distributed as $4N_A\mu/x$.

De novo mutations on the different branches of the population tree: In this case, the frequencies are fixed ($x_i=1/2N_i$ with i the population where the mutation event arises, and $x_{j\neq 1}=0$ for the rest) and the time coordinate is a random variable uniformly distributed on the time interval.

As an example, let us consider a two-population model in more detail (see Figure 2). An ancestral population with size N_A increases its size to N_B at time $t=0$; the population size remains constant and equal to N_B up to time $t=t_1$; at time

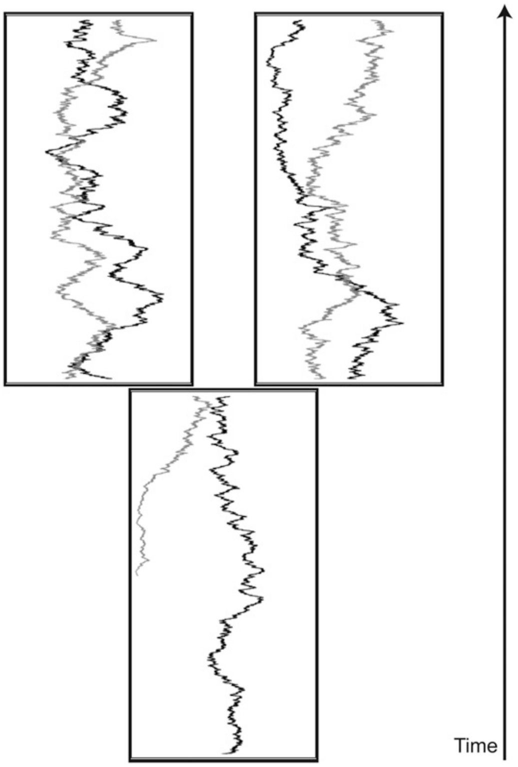

Figure 2 Simulation of two Brownian paths on a two-population tree. The plot illustrates different types of initial conditions for the paths. The initial condition can be either a random allele frequency in the ancestral population at mutation–drift equilibrium (solid lines) or a *de novo* mutation that arises in one of the populations after the ancestral population leaves the state of equilibrium (shaded lines).

$t = t_1$ the population splits into two populations with sizes N_1 and N_2; the present time is reached at time $t = t_2$. Time is measured in units of generations. In this model one can define the space of initial conditions as the union of sets

$$\begin{aligned}&\left\{(2N_A)^{-1} \le x < 1,\ t = 0\right\} \cup \left\{x = (2N_B)^{-1},\ 0 < t \le t_1\right\}\\ &\cup \left\{x_1 = (2N_1)^{-1},\ x_2 = 0,\ t_1 < t \le t_2\right\}\\ &\cup \left\{x_1 = 0,\ x_2 = (2N_2)^{-1},\ t_1 < t \le t_2\right\}.\end{aligned} \tag{13}$$

The probability density on this space of initial conditions is

$$\left[\frac{4N_A\mu dx}{x},\ 2N_B\mu dt, 2N_1\mu dt, 2N_2\mu dt\right].$$

Here, x is the random variable in the first set, and t is the random variable in the remaining three sets. The total number of initial states is

$$Z = 4N_A\mu\ \log(2N_A) + 2N_B\mu t_1 + 2N_1\mu(t_2 - t_1) + 2N_2\mu(t_2 - t_1).$$

Therefore, a random initial state is a random variable on the space of initial conditions associated with the probability density function

$$\left[\frac{4N_A\mu dx}{Zx},\ \frac{2N_B\mu dt}{Z},\ \frac{2N_1\mu dt}{Z},\ \frac{2N_2\mu dt}{Z}\right].$$

Finally, the simulated AFS n_{ij} associated with a sample of C chromosomes per population is

$$n_{ij} = \frac{Z}{\Pi}\frac{(C!)^2}{i!(C-i)!j!(C-j)!}\sum_{p=1}^{\Pi} x_{p,1}^{i}\left(1-x_{p,1}\right)^{C-i} x_{p,2}^{j}\left(1-x_{p,2}\right)^{C-j}, \tag{14}$$

with $0 \le i \le C$, $0 \le j \le C$, $0 < i + j < 2C$, and $\left\{(x_{1,p}, x_{2,p})\right\}_{p=1}^{\Pi}$ the simulated population allele frequencies.

Demographic scenarios: We simulated data under seven different demographic histories for two and three populations to compare different approaches to forward diffusions. We generated 50,000 population allele frequencies using Monte Carlo simulations in each of the seven demographic scenarios. We sampled 20 chromosomes in the scenarios that involved three populations and 50 chromosomes in the scenarios that involved two populations. For each demographic scenario, we computed the AFS with our polynomial-based approach (MultiPop) and with the finite-difference method (∂a∂i) and compared each AFS with the AFS computed with Monte Carlo simulations. The different models and parameters used in the simulations are described below.

Two populations: In all the two-population scenarios the ancestral population size N_A was 4000. The time from the population splitting event up to the present was $T = 150$ generations. The population sizes after the splitting event were $N_1 = 4000$ and $N_2 = 1815.1$. Finally, the migration matrices were

$$m = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

for the $2Nm = 0$ scenario (model 1),

$$m = \begin{pmatrix} 0 & 0.0000375 \\ 0.00011 & 0 \end{pmatrix}$$

for the $2Nm < 0.5$ scenario (model 2),

$$m = \begin{pmatrix} 0 & 0.0001 \\ 0.00022 & 0 \end{pmatrix}$$

for the $2Nm \sim 1$ scenario (model 3), and

$$m = \begin{pmatrix} 0 & 0.00014 \\ 0.0003 & 0 \end{pmatrix}$$

for the $2Nm > 1$ scenario (model 4).

Three populations: In all the three-population scenarios the ancestral population size N_A was 5000. The time from the first population splitting event up to the second population splitting event was $T = 220$ generations. The time from the second population splitting event up to the present was $T =$

130 generations. The population sizes after the first splitting event were $N_1 = 1815.1$ and $N_2 = 815.1$. The second population splitting event occurred in population 2. The population sizes after the second splitting event were $N_1 = 1815.1$, $N_2 = 315.1$, and $N_3 = 815.1$. Finally, the migration matrices were

$$m^a = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$m^b = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

for the $2Nm = 0$ scenario (model 5),

$$m^a = \begin{pmatrix} 0 & 0.000005 \\ 0.0005 & 0 \end{pmatrix}$$

and

$$m^b = \begin{pmatrix} 0 & 0.0002 & 0.00003 \\ 0.0000005 & 0 & 0.00003 \\ 0.0003 & 0.0003 & 0 \end{pmatrix}$$

for the $2Nm \sim 1$ scenario (model 6), and

$$m^a = \begin{pmatrix} 0 & 0.000005 \\ 0.0012 & 0 \end{pmatrix}$$

and

$$m^b = \begin{pmatrix} 0 & 0.0006 & 0.00003 \\ 0.0000005 & 0 & 0.00003 \\ 0.0003 & 0.0003 & 0 \end{pmatrix}$$

for the $2Nm > 1$ scenario (model 7).

Data

We used the Environmental Genome Project (EGP) SNP database (Environmental Genome Project 2010) to determine the observed joint AFS in four human populations. The sampled populations consist of 12 individuals of West African ancestry (YRI), 22 individuals of northern European ancestry (CEU), 24 individuals of East Asian ancestry (CHB), and 22 individuals of Mexican ancestry (MEX). These data are the result of direct Sanger resequencing (with a low error rate) of environmental response genes and have been the subject of several previous studies (Akey *et al.* 2004; Williamson *et al.* 2005; Gutenkunst *et al.* 2009). We used this data set to compare our method with that of Gutenkunst *et al.* (2009). The number of environmentally responsive genes sequenced as part of the EGP has been steadily increasing since the project started in 2001 (Environmental Genome Project 2010), and the EGP database is now larger than when Gutenkunst *et al.* (2009) performed their study. However, the number of individuals in the data set has not changed. As we were not able to reconstruct the original set of SNPs used in Gutenkunst *et al.* (2009), we instead used all the available loci. The difference between the data sets turned out to be small; while Gutenkunst *et al.* (2009) used 27,824 SNPs of noncoding DNA, we used 28,875 SNPs.

We considered all 28,875 diallelic SNPs located in 4.07 Mb of noncoding DNA resequenced from 197 different autosomal genic regions. Each SNP was ordered into an ancestral and a derived state, using the pairwise alignment from chimpanzee to human available in the UCSC panTro2 draft of the chimpanzee genome (Chimpanzee-Sequencing-Consortium 2005). We computed the context-dependent probability of misidentification of the ancestral state and introduced the associated corrections in the allele-frequency spectrum (see Hernandez *et al.* 2007 and File S1, section 2). As in Gutenkunst *et al.* (2009), we assumed a divergence time of 6 million years between human and chimpanzee and a generation time of 25 years. We estimated a mutation rate of $\mu = 2.35 \times 10^{-8}$ per site per generation from sequence divergence present in the data. This is identical to the estimate by Gutenkunst *et al.* (2009) and comparable to other estimates (*e.g.*, Nachman and Crowell 2000).

As the number of chromosomes in the data varies depending on the particular SNP, we projected the AFS down to a fixed number of chromosomes. If the number of chromosomes sampled in each population is the same and equal to C, the total number of bins in the associated AFS will be $(C + 1)^4 - 2$. Also, as the theoretical value of each bin $f_{i_1,i_2,i_3,i_4}(\theta)$ in the AFS is computed as the four-dimensional integral defined by Equation 1, the computational time needed to evaluate a single bin of the theoretical AFS is larger than in the cases of two or three populations. To reduce the computational burden of evaluating Equation 1 many times, we used an adaptive allele-frequency spectrum in which we decomposed the AFS into bins of different sizes, depending on the fraction of SNPs that occupy a particular region of the frequency space. We fixed the bin sizes by sampling 19, 9, or 4 chromosomes per population. We chose 19 as the maximum number of chromosomes because 19 + 1 can be divided twice by two, which allows us to easily build an adaptive histogram using three different bin sizes. To this end we first projected the empirical AFS down to 4 chromosomes per population. In our adaptive construction the coarsest AFS possible has $(4 + 1)^4$ bins, which we further refine by considering bins of size 1/9 and 1/19 within each bin of size 1/4. We computed the fraction of SNPs present in each bin of size 1/4, and if this fraction exceeded a certain bound b, we further refined the adaptive AFS to consider bins of size 1/9. Recursively, we isolated those bins of size 1/9 that contained a fraction of SNPs $>b$ and further refined them into smaller bins of size 1/19. Using all the SNP data and a bound parameter of $b = 5 \times 10^{-4}$, the total number of bins of the adaptive AFS becomes 5078 (452 bins of size 1/4, 2644 bins of size 1/9, and 1982 bins of size 1/19). This is a significant reduction in the total number of bins for a four-population

AFS relative to a full representation for 19 chromosomes, which has $20^4 - 2 = 159{,}998$ bins.

Implementation

In our software implementation we use two different bases of the vector space spanned by polynomials on $0 < x < 1$ with degree $\leq \Lambda$. We use the shifted Gegenbauer polynomials $T_n(x) = \sqrt{(n+2)(2n+3)/(n+1)}P_n^{(1,1)}(2x-1)$ and the shifted Chebyshev polynomials $C_n(x) = ((1/\sqrt{\pi} - \sqrt{2/\pi})\delta_{0,n} + \sqrt{2/\pi}) \times \cos(n\ \arccos(2x-1)$ (see the *Appendix* for more details). We inject mutations at the boundary, using the term $c_k \exp(-kx)$. Similarly, the associated drift–mutation equilibrium density that we use as the initial condition in our demographic models is (see section 3 of File S1 for more details)

$$\begin{aligned}&\phi_{\mathrm{eq}}(x)\\&= 4N_A\mu \frac{1-x+((1+k(1-x))\exp(-k)-\exp(-kx))/(1-\exp(-k)-k\ \exp(-k))}{x(1-x)}.\end{aligned}$$

We evaluated the integrals that appear in Equation 1, such as $I(y) = \int_0^y T_n(x)x^i(1-x)^j dx$, by means of the Runge–Kutta four method. In particular, $I(y)$ obeys the ordinary differential equation (ODE)

$$\frac{dI}{dy} = T_n(y)y^i(1-y)^j,$$

and $I(1)$ is obtained as the integral of the ODE between 0 and 1.

Population splitting events were modeled by assuming that the distributions of allele frequencies in the two daughter populations are identical; *i.e.*, $\phi(x, x_{K+1}) = \delta(x_i - x_{K+1})\phi(x)$, with i and $K+1$ the populations that arise after the divergence of i. The Dirac delta was approximated by a Gaussian function peaked at the diagonal $x_i = x_{K+1}$ with a user-defined standard deviation that we call a thickening parameter in the software implementation. Such a smooth Gaussian approximation allows us to use a truncated polynomial expansion to accurately approximate the density after the splitting event. The larger the truncation parameter used, the smaller the standard deviation that can be used and the closer the approximation will resemble $\phi(x, x_{K+1}) = \delta(x_i - x_{K+1})\phi(x)$.

The computer implementation was written in the C++ language, and the source code is freely available in Google Code (http://www.code.google.com/p/multipop/). We compared the results found using the MultiPop program to those found using a different class of numerical techniques that estimate the time evolution of ϕ, using grid approximations and a finite-difference method to integrate the PDE. The latter method was implemented in the computer program ∂a∂i (Gutenkunst *et al.* 2009).

Nonlinear optimization: When inferring the demographic parameters of a model given the simulated data or the human data set, we have to maximize the likelihood-function Equation 11. Maximizing Equation 11 on a high-dimensional model parameter space is a challenging nonlinear optimization problem. To this end, we used three classical algorithms in nonlinear optimization: simulated annealing, the downhill simplex method, and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method. When inferring the parameters from the simulated data, we used only local optimization techniques (downhill simplex). We used the true value as the initial point in the local optimization algorithm. We checked that a local maximum had been reached by numerically evaluating the Hessian of the minus log-likelihood at the critical point and confirmed that its eigenvalues were positive. This technique allowed us to determine the local maximum of the likelihood surface closest to the true value.

In the case of the human data, we performed an initial exploration of the parameter space by means of simulated annealing to find the global maximum for a fixed empirical AFS with all the SNPs. Subsequent maxima of the likelihood function associated with bootstrapped samples were computed by means of local optimization algorithms (*e.g.*, downhill simplex or quasi-Newton). When using a local optimization algorithm, we used as initial seed the global maximum initially determined by means of simulated annealing.

Statistical inference and bootstrap strategy

The theoretical AFS can be considered as both the density of the allele frequency for a single diallelic locus and the expected distribution for a very large set of independent diallelic loci. Therefore, the calculation of the likelihood using Equation 10 is accurate only if each SNP has been sampled independently from other SNPs. This is indeed the approach that we follow with the simulated data, where each SNP is independent from the others.

One approach to dealing with nonindependent (*i.e.*, linked) SNPs is to use Equation 10 with all linked SNPs and treat the result as a composite likelihood. In the case of the human data set, the EGP database consists of just 170 independent loci with 80% of the SNPs occurring in 30% of the loci. As the SNPs are tightly linked within each locus, unusual histories of a DNA segment with many tightly linked SNPs can overrepresent the inferred demographic history when the number of independent loci is small. In other words, although composite-likelihood estimators are known to be consistent (Wiuf 2006), they can be biased when the sample size is small.

Because the human EGP data set consists of a relatively small number of SNPs, many of which lie close to one another, we have tried to minimize the effects of linkage in our bootstrap approach to estimate parameters and confidence intervals. For each sample, 170 SNPs were selected at random, conditioned on their being separated from each other by at least 200 kb. Multiple sets of 170 SNPs were sampled with replacement from the data set, which yielded different maximum-likelihood points $\{\hat{\theta}_i^*\}_{i=1}^B$ in the parameter space. Here, B is the number of bootstrap samples. Ninety-five percent confidence intervals and regions were calculated using the percentile approximation (DiCiccio and Efron 1996). We used the marginal distribution of maximum-likelihood values for each parameter to estimate the confidence intervals by considering the 2.5th percentile and the 97.5th percentile.

Results

Comparison of different diffusion theory-based approaches

To study the biases introduced by the different numerical approximations, we computed the AFS in seven different demographic scenarios with varying numbers of populations and intensities of gene flow (models 1–7). We used spectral methods, finite-difference methods, and Monte Carlo simulations of the forward diffusion process. The different frequency spectra were compared by means of the chi-square statistic (Table 1 and Figures 3 and 4). We found that using 35 polynomials per population gives rise to very good approximations of the true AFS in the seven scenarios. In the limit of very small gene flow, the chi-square statistic associated with the spectral method approach converged much faster to zero as the truncation parameter Λ was increased. This behavior, which was observed in two- and three-population models, is the expected one as the polynomial-based approach yields exact solutions of the AFS in the absence of migration. As the intensity of migration increases, the rate of convergence to the exact AFS in the spectral method approach worsens. This behavior is also the expected one (Lukic *et al.* 2011), as the polynomial expansion does not yield exact solutions of the diffusion PDEs with nonzero gene flow. Also, it becomes more difficult to implement exactly the boundary conditions because of the emission of polymorphisms to other populations. However, as the truncation parameter Λ increases, the quality of the approximations of the AFS increases in all scenarios.

Similarly, the finite-difference approach gave rise to approximations of the AFS of a comparable quality to those of the approach that uses polynomial series expansions. In particular, it produced better approximations when the intensity of gene flow was strong. However, the rates of convergence in the two-population models were very different from the ones in the three-population models. In the limit of zero gene flow the rate of convergence in the two-population model was significantly faster than that in the three-population model. There is not a simple way to explain these results, because among other things we do not know how the numerical discretization used in the finite-difference scheme relates to the exact AFS in any subset of the parameter space for a finite grid size.

Although using the chi-square statistics (*e.g.*, Figures 3 and 4) is helpful for quantifying the numerical error in the AFS, this measure of error is not informative enough to estimate the optimal numerical error given a certain level of statistical uncertainty when inferring demographic parameters. This is because the "numerical error" (here, error measures the deviation of the approximated AFS from the exact AFS) can be very different from the "propagated numerical error" (here, error measures the deviation of the numerically approximated likelihood peak from the exact likelihood peak), even if both decay as the truncation parameter increases. For instance, the appearance of nearly flat directions of the likelihood function on the parameter space might amplify the numerical errors that arise in the numerical solution of the PDE, giving rise to large propagated numerical errors. To estimate correctly the optimal error given a certain level of statistical uncertainty, one should compare the location of the maximum-likelihood peaks in the numerical approximation with the true location of the peaks. We perform this analysis in the following subsection.

The inverse problem: maximum-likelihood estimates: To study these propagated numerical errors we inferred the maxima of the likelihood functions. In this case, the effective population sizes, splitting times, and migration rates were free parameters, and the observed frequency spectra were constructed using Monte Carlo simulations. We computed the maximum-likelihood peaks associated with the polynomial-based and grid-based approximations of the maximum-likelihood function in each demographic model (see Table 2). In our polynomial approximation we used 40 polynomials per population in the two-population scenarios and 35 polynomials per population in the three-population scenarios. For the finite-difference method we used 100 grid points per population in the two-population and three-population scenarios.

We found that in models with few parameters both approaches yield maximum-likelihood peaks that are close to the true values (see Table 2). As a general trend, the finite-difference method tends to overestimate the amount of gene flow while the spectral method tends to overestimate the largest effective population size. We also observe that as the number of model parameters increases, many inferred migration rates deviate significantly from the true values (for instance, see models 6 and 7 in Table 2). As we simulated large sets of SNP allele frequencies for each scenario, the statistical noise was very small and the main source of bias that we observe can be attributed to numerical errors.

These biases are caused by two main factors: the propagation of numerical errors in the evaluation of the likelihood function and the geometry of the likelihood function associated with a particular model. In particular, models in which a set of parameters yields flat directions around the likelihood peak will be particularly prone to propagate small numerical errors toward large errors in the inferred parameters. One can interpret the biases obtained in the migration rates of models 6 and 7 in Table 2 along these lines. Here, the large number of migration parameters introduced in the models gave rise to many flat directions that amplified numerical errors associated with the evaluation of the likelihood function.

Table 1 Comparison of numerical approximations of the AFS and the simulated AFS

Model no.	Intensity of migration (2*Nm*)	MPop *vs.* Monte Carlo (chi-square statistic)	∂a∂i *vs.* Monte Carlo (chi-square statistic)
1	0	0.001265	0.002479
2	<0.5	0.00894	0.005955
3	1	0.01076	0.006457
4	>1	0.01140	0.006286
5	< 0.5	0.007558	0.04684
6	1	0.01511	0.02882
7	>1	0.02979	0.01791

Chi-square statistics associated with seven demographic scenarios are shown. The AFS computed by MultiPop (MPop) corresponds to the Λ = 35 AFS, and the AFS computed by ∂a∂i corresponds to a grid size of 40 grid points per population. The frequency spectra were normalized in all the cases such that the total number of SNPs was 1.

Computational performance: Our current implementation of the spectral method to study demography with diffusion approximations is optimized to use little memory to tackle more than three populations. This economical use of memory is attained by increasing the number of operations in the algorithm and hence reducing its speed. ∂a∂i is optimized to work with two and three populations and is significantly faster than our current implementation (see Table 3).

One way to increase the speed of the method is to reconsider the memory model used for cases with two and three populations. In particular, the diffusion operator is a matrix of size $[(\Lambda + 2)^K - 2]^2$, whose storage requires a very large amount of memory. Our present implementation needs only four matrices of size Λ^2 for any K, which are used to recover the full diffusion operator at running time by exploiting its tensorial structure. This implementation is very economical from the point of view of memory use. However, it makes the algorithm significantly slower. An implementation that uses sparse matrices to approximate the full diffusion operator will significantly reduce the number of operations and increase the speed of the algorithm when $K < 4$.

Worldwide human expansion out of Africa and peopling of the Americas

We considered a four-population model with 18 free parameters to model the human expansion out of Africa and peopling of the Americas (see Figure 5). The model is inspired by several studies reported in the literature (*e.g.*, see Gutenkunst *et al.* 2009; Gravel *et al.* 2011). The root of the population tree consists of an ancestral human population in Africa at mutation–drift equilibrium. Such a population experiences a sudden increase of its effective population size at some time before the out-of-Africa event. The divergence of non-African populations after the out-of-Africa event is further modeled by population splits with gene flow. These population splits describe the European–Asian split and the bottleneck associated with the peopling of the Americas. An exponential growth model is then used to describe the population growths of Europeans, Asians, and Native Americans after they become independent populations, and recent population admixture is introduced to model high European gene flow into the ancestral Amerindian population associated with the Mexican population. To reduce the number of parameters, we considered symmetric migration rates except during the first stage of the out-of-Africa event ($m_{\mathrm{AF}\rightarrow B} \neq m_{B\rightarrow \mathrm{AF}}$). We did not assume that this migration rate was symmetrical because $m_{\mathrm{AF}\rightarrow B}$ might be significantly larger than $m_{B\rightarrow \mathrm{AF}}$ as one indeed infers from the data (see Table 4). We used a basis of polynomials up to degree 20 (Λ = 20) to approximate the density of population frequencies. Taking into account boundary contributions, the dimension of the space of densities was $22^4 - 2 = 234{,}254$.

The inferred parameters and confidence intervals are shown in Table 4. Our estimate of the time at which the ancestral Amerindian population split from the ancestral East Asian population is earlier than previous estimates for the time of settling of the Americas. This is compatible with the fact that the ancestral population of the people of the Americas should have shared a common ancestor with East Asians some time before the Americas were peopled. The other inferred parameters are consistent with those of many previous studies. For instance, we infer that the human dispersal out-of-Africa event took place ~52,000 years ago (95% confidence interval: 36–81 KYA) followed by a high migration rate. This agrees with previous studies that infer a separation of Africans and non-Africans ~60,000 years ago followed by significant genetic exchange up until 20,000−40,000 years ago (see Reich 2001; Keinan *et al.* 2007; Gravel *et al.* 2011; Li and Durbin 2011). Our estimates are also in broad agreement with previously reported values using the diffusion approximation in demographic inference. For instance, Gravel *et al.* (2011) find a time of split between Africans and Eurasians of 51,000 years ago (95% confidence interval: 45–69 KYA) by applying a similar demographic model to the 1000 genomes project data set. In the case of Gutenkunst *et al.* (2009), the inferred parameters are broadly consistent with our estimates, although our resulting confidence intervals are substantially narrower than the intervals determined by the authors using conventional bootstrap. In particular, Gutenkunst *et al.* (2009) use a combination of two three-population models very similar to our four-population model and apply it to the EGP data set. In this case the time of the out-of-Africa event was inferred to be 140,000 years ago (95% confidence interval: 40–270 KYA).

The differences in the width of the confidence intervals are due to a combination of different factors. The most important factor is that we use a four-population model instead of two three-population models, and this choice limits the number of polynomials that we can use to solve the diffusion PDEs associated with the model. In particular, we used only 20 polynomials per population. As we discussed in the previous subsection using simulated data (see

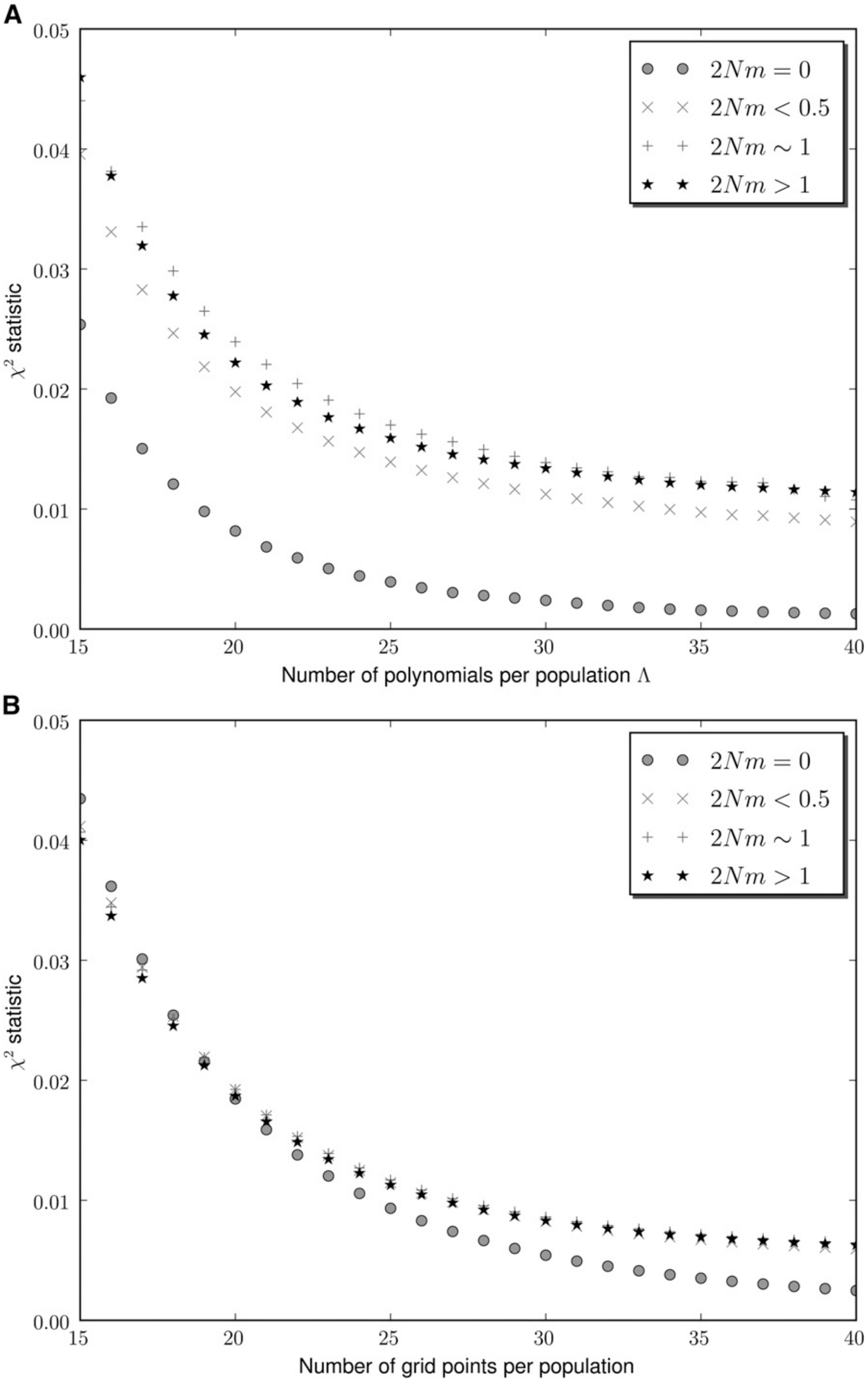

Figure 3 Decay of the chi-square statistic in MultiPop (top) and ∂a∂i (bottom). Four different demographic scenarios with two simultaneous populations and 50 chromosomes sampled per population are considered. For simplicity, the average scaled migration rate is used to label each scenario. The observed AFS were constructed using Π = 50,000 independent loci produced with Monte Carlo simulations.

Figures 3 and 4), for any fixed value of Λ the quality of the approximations of the frequency spectra worsens as the intensity of gene flow increases. Similarly, for any fixed values of the migration rates the quality of the approximations worsens as the number of polynomials used decreases. Therefore, choosing $\Lambda = 20$ has given rise to poor approximations of the AFS in regions of the parameter space that involve high intensities of gene flow. The associated distortions of the AFS have yielded artificially low likelihoods in those regions of the parameter space. Hence, this numerical artifact is largely responsible for producing narrower confidence intervals in our study than those in the study by Gutenkunst *et al.* (2009). For instance, for the time out of Africa we inferred smaller confidence intervals (36–81 KYA) than those in Gutenkunst *et al.* (2009) (40–270 KYA). While Gutenkunst *et al.* (2009) infer very large values for the gene flow between populations after the out-of-Africa event, our confidence intervals for the migration rates are significantly smaller and closer to the zero-migration limit. Other differences between this study and the one by Gutenkunst *et al.* (2009) are in the bootstrap strategy and the particular data set. In our bootstrap strategy we sample one SNP per locus

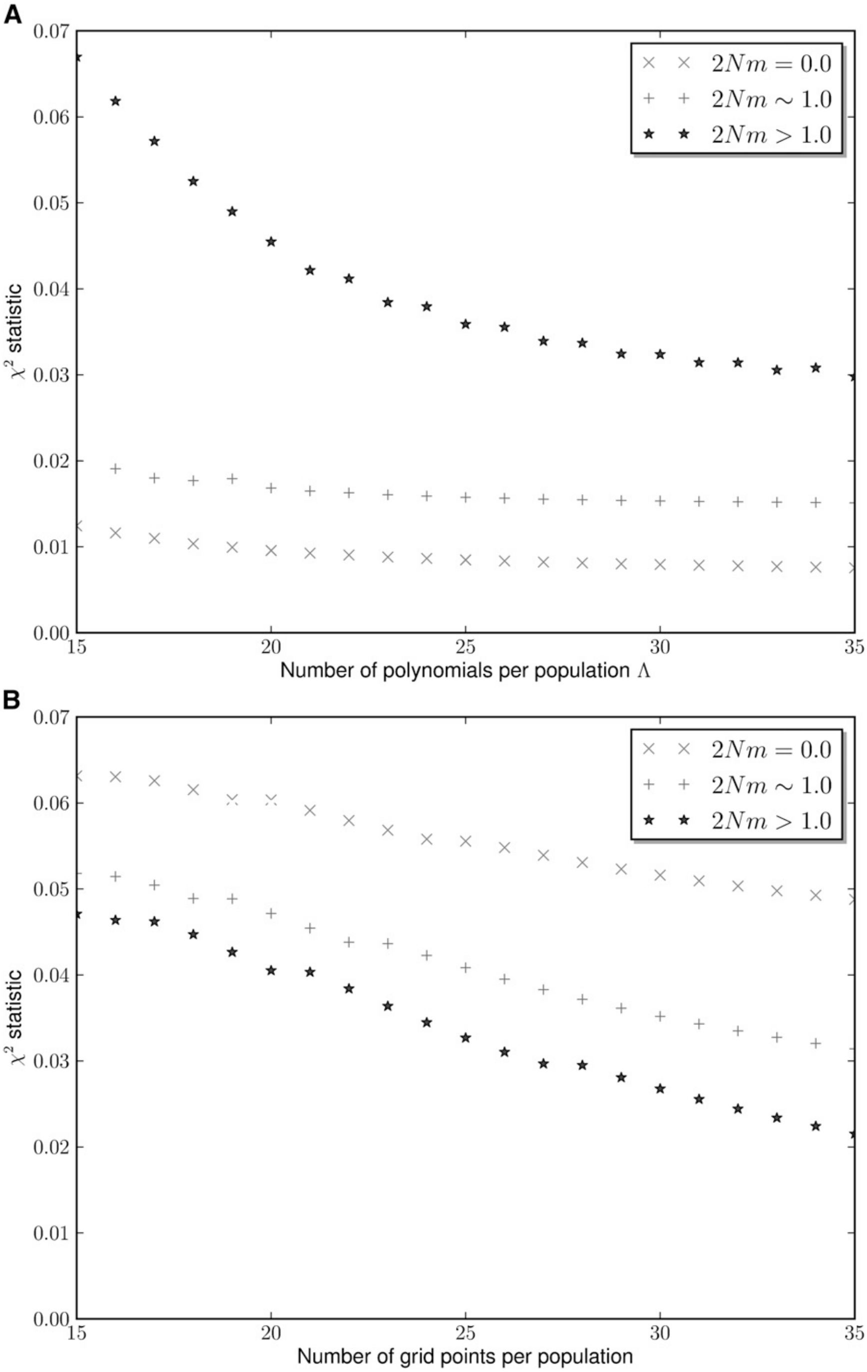

Figure 4 Decay of the chi-square statistic in MultiPop (top) and ∂a∂i (bottom). Three different demographic scenarios with three simultaneous populations and 20 chromosomes sampled per population are considered. For simplicity, the average scaled migration rate is used to label each scenario. The observed AFS were constructed using Π = 50,000 independent loci produced with Monte Carlo simulations.

each time, and each pair of SNPs is set to be separated by at least 200 kb. In the analysis by Gutenkunst *et al.* (2009) the statistical artifacts due to linked SNPs were corrected by considering simulations with linked loci in a parametric bootstrap approach to compute confidence intervals. However, no constraint for the distance between pairs of SNPs was imposed in their nonparametric bootstrap approach. If no constraints are imposed on the minimal distance between any two SNPs in each bootstrap, SNPs from loci with high SNP density will be sampled more often. Therefore, the demographic history of a locus with a high density of SNPs will be overrepresented in the overall inference. Although both bootstrap strategies should converge to the same result in the limit of a large number of loci, in the case of a data set with a small number of loci, such as the EGP data set, the differences might be more important. Indeed, the study by Gravel *et al.* (2011) inferred significantly narrower confidence intervals by applying ∂a∂i to genome-wide sequence data. Finally, although the EGP data set contains a few more sequenced loci since the study by Gutenkunst *et al.* (2009) and this fact affects the width of the confidence intervals, this contribution should be small.

Table 2 Comparison of maximum-likelihood estimates using different numerical approximations

Model	Model parameter	True value	θ MPop	θ ∂a∂i
1	$4N_Au$	1	1	1
1	N_1/N_A	1	1.015	1.042
1	N_2/N_A	0.4538	0.4385	0.4421
1	$T/2N_A$	0.01875	0.01788	0.01840
2	$4N_Au$	1	1	1
2	N_1/N_A	1	1.08	1.042
2	N_2/N_A	0.4538	0.6467	0.6379
2	$T/2N_A$	0.01875	0.02304	0.02720
2	$2N_Am_{1\rightarrow2}$	0.3	0.5706	3.529
2	$2N_Am_{2\rightarrow1}$	0.88	0.8735	3.228
3	$4N_Au$	1	1	1
3	N_1/N_A	1	1.099	1.044
3	N_2/N_A	0.4538	0.6648	0.6497
3	$T/2N_A$	0.01875	0.02287	0.02783
3	$2N_Am_{1\rightarrow2}$	0.8	0.5873	4.022
3	$2N_Am_{2\rightarrow1}$	1.76	0.9469	3.965
4	$4N_Au$	1	1	1
4	N_1/N_A	1	1.086	1.068
4	N_2/N_A	0.4538	0.6712	0.6493
4	$T/2N_A$	0.01875	0.02258	0.02592
4	$2N_Am_{1\rightarrow2}$	1.12	0.6190	2.574
4	$2N_Am_{2\rightarrow1}$	2.64	0.9910	3.798
5	$4N_Au$	1	1	1
5	N_1/N_A	0.3630	0.3713	0.3522
5	N_2/N_A	0.1630	0.1540	0.1440
5	N_3/N_A	0.06302	0.05673	0.05569
5	$T^a/2N_A$	0.022	0.02158	0.02041
5	$T^b/2N_A$	0.013	0.01225	0.01152
6	$4N_Au$	1	1	1
6	N_1/N_A	0.3630	0.3713	0.3344
6	N_2/N_A	0.1630	0.17855	0.1527
6	N_3/N_A	0.06302	0.05950	0.06112
6	$T^a/2N_A$	0.022	0.02996	0.01749
6	$T^b/2N_A$	0.013	0.01211	0.01249
6	$2N_Am^a_{1\rightarrow2}$	0.05	0.04562	0.0001828
6	$2N_Am^a_{2\rightarrow1}$	5	0.02345	0.0009895
6	$2N_Am^b_{1\rightarrow2}$	2	0.002115	0.003311
6	$2N_Am^b_{1\rightarrow3}$	0.3	0.3897	1.505
6	$2N_Am^b_{2\rightarrow1}$	0.005	0.007982	0.09358
6	$2N_Am^b_{2\rightarrow3}$	0.3	0.0007637	4.4e-09
6	$2N_Am^b_{3\rightarrow1}$	3	0.6831	3.398
6	$2N_Am^b_{3\rightarrow2}$	3	1.788	2.640
7	$4N_Au$	1	1	1
7	N_1/N_A	0.3630	0.3874	0.3138
7	N_2/N_A	0.1630	0.1724	0.1445
7	N_3/N_A	0.06302	0.05843	0.05589
7	$T^a/2N_A$	0.022	0.02415	0.01296
7	$T^b/2N_A$	0.013	0.01155	0.01115
7	$2N_Am^a_{1\rightarrow2}$	0.05	0.02982	0.7007
7	$2N_Am^a_{2\rightarrow1}$	12	0.007439	0.004225
7	$2N_Am^b_{1\rightarrow2}$	6	0.0007166	0.0003699
7	$2N_Am^b_{1\rightarrow3}$	0.3	0.00090192	3.1354
7	$2N_Am^b_{2\rightarrow1}$	0.005	0.0003413	0.0020024
7	$2N_Am^b_{2\rightarrow3}$	0.3	0.0007713	0.0001095
7	$2N_Am^b_{3\rightarrow1}$	3	0.007362	3.232
7	$2N_Am^b_{3\rightarrow2}$	3	0.4090	0.01191

Maximum-likelihood estimates of seven different demographic scenarios are shown. Many results are biased due to numerical errors in the calculation of the frequency spectra (see subsection *Numerical errors and bias-corrected confidence intervals* for a detailed discussion on how numerical errors affect the location of the maximum-likelihood peak). In the case of MultiPop, 40 polynomials were used in the two-population models and 35 polynomials in the three-population models. We used 100 grid points in all models approximated with finite-difference schemes. The observed AFS were constructed using Π = 50,000 independent loci produced with Monte Carlo simulations. The maximum-likelihood peak was found by means of the BFGS method, using the coordinates of the true peak as the initial point in this local optimization algorithm.

One of the inferred parameters that is most different between our study and that of Gutenkunst *et al.* (2009) is the proportion of European ancestry in Mexicans. We infer an admixture proportion of 20.4% (95% C.I.: 3.2–41%), while the three-population model used by ∂a∂i inferred 48% (95% C.I.: 42–60%) (Gutenkunst *et al.* 2009). In this case the confidence intervals do not overlap. Again, explanations of this discrepancy range from the presence of more sequence data in the EGP data set since the study by Gutenkunst *et al.* (2009) was done to small differences in the statistical analyses made. Other studies have pointed out the difficulty of estimating admixture proportions when data of one of the pre-admixture ancestral populations are missing (Alexander *et al.* 2009). Therefore, it is not surprising that slight differences in the data set and in the statistical analyses can yield significant differences in the inferred parameters. In particular, Alexander *et al.* (2009) found that the European admixture proportion in the individuals of Mexican ancestry genotyped by HapMap III was ~20% if inferred by the algorithm ADMIXTURE, while it was inferred to be ~50% by the algorithm STRUCTURE using the same data set.

Discussion

Forward diffusion equations played an important role during the development of classical population genetics, as they were originally introduced by R. Fisher and S. Wright to model the evolutionary process (Kimura 1964). With the arrival of modern DNA sequencing technologies, forward diffusion processes have been applied to the inference of demographic parameters and the effects of natural selection (Williamson *et al.* 2005; Boyko *et al.* 2008; Gutenkunst *et al.* 2009). These studies have been limited to scenarios with one, two, and three simultaneous populations. In this article, we have introduced a different approach to solving the forward diffusion equations by means of truncated polynomial expansions. These methods yield exact solutions of the AFS in the absence of migration; they can be used equally to study demographic models in one, two, and three populations with gene flow, and furthermore they can be applied to study models with four simultaneous populations.

We have applied our method to the study of the human expansion out of Africa and peopling of the Americas by means of a model with four simultaneous populations. Our four-population model can be seen as a combination of two three-population models that were studied before by means of diffusion-theory–based techniques (Gutenkunst *et al.* 2009). Similarly, we used the Environmental Genome Project SNP database. The demographic parameters that we have inferred in this model are similar to many recent

Table 3 Comparison of computing time

Model	No. polynomials (MPop)	Computing time of MPop (sec)	Grid size (∂a∂i)	Computing time of ∂a∂i (sec)
2	15	0.27	15	0.07
2	20	0.61	50	0.10
2	25	1.10	100	0.13
2	30	1.86	500	1.43
2	35	2.92	1000	5.68
6	15	3.1	15	0.16
6	20	7.95	50	0.86
6	25	17.31	100	6.33
6	30	33.3	200	68.41
6	35	57.41	300	253.85

Computing times required to evaluate an allele-frequency spectrum using MultiPop and ∂a∂i are shown. The demographic scenarios involved two populations and 50 chromosomes sampled per population or three populations and 20 chromosomes sampled per population. The CPU used to measure the computing times was an Intel Core(TM)2 Duo P8600 with speed 2.40 GHz.

results (see Reich 2001; Keinan *et al.* 2007; Gravel *et al.* 2011; Li and Durbin 2011). However, some of our inferred parameters are substantially different from the parameters inferred in another study that used numerical solutions to forward diffusion equations (Gutenkunst *et al.* 2009).

In this article we have also studied the behavior of different numerical solutions of the diffusion PDEs that approximate the AFS under a specified demographic model. In particular, we have compared the polynomial-based approach introduced in Lukic *et al.* (2011) with the finite-difference approach implemented in Gutenkunst *et al.* (2009). Although the methods exhibit comparable behaviors, we found that the polynomial-based approach obtains better results in the regime where it yields exact solutions of the AFS, *i.e.*, in the zero-migration limit (see Table 1 and Figures 3 and 4). Also, we confirmed that the polynomial-based approach allows us to broadly predict the magnitude of the numerical error as a function of the model parameters for a given truncation parameter Λ. The finite-difference approach exhibited a better behavior in the models with strong intensity of migration that we have considered in this work. However, in the case of the method introduced in Gutenkunst *et al.* (2009) there is not a general theory that allows us to predict how the numerical error behaves as a function of the parameter space.

Numerical errors and bias-corrected confidence intervals

A common limitation that both approaches sometimes exhibit is that small numerical errors in the computation of the AFS can propagate to large biases in the parameter space when we search for the maxima of the numerical approximation of the likelihood function. Hence, even if biases due to numerical errors can be minimized, in some cases small but significant

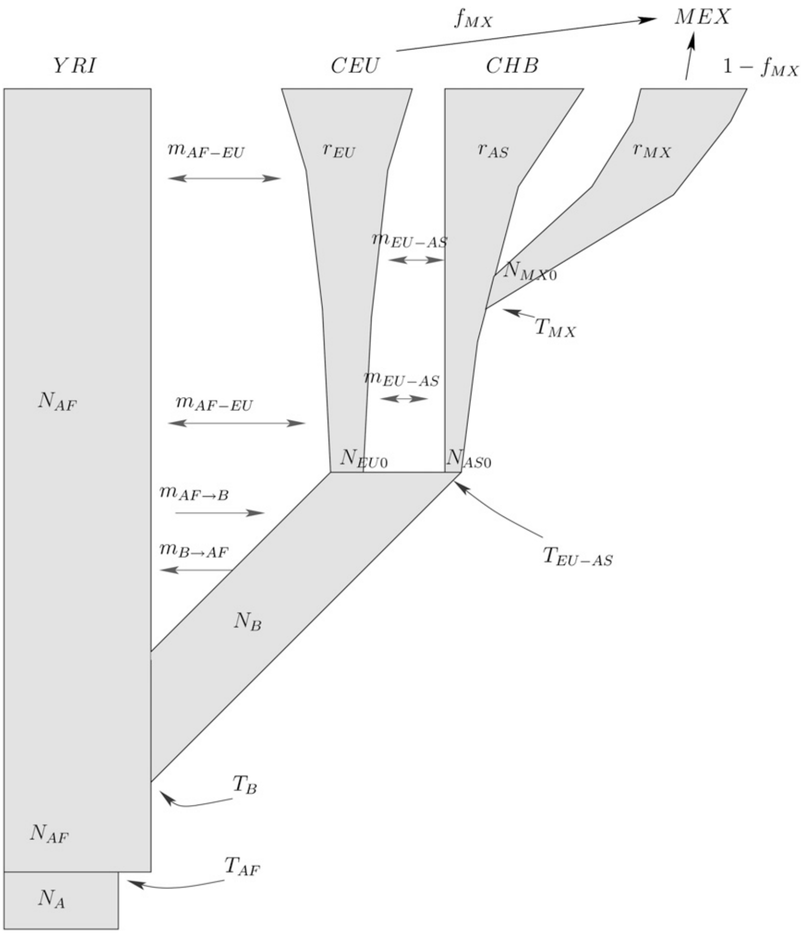

Figure 5 A graphical representation of a four-population model for the human expansion out of Africa and peopling of the Americas. The nonconstancy of the population sizes of CEU, CHB, and MEX is modeled by means of an exponential growth model with growth rates r_{EU}, r_{AS}, and r_{MX}.

Table 4 Inference of a four-population model for the human expansion out of Africa and peopling of the Americas

Model parameters	θ MPop	95% C.I.	θ ∂a∂i, 1	95% C.I.	θ ∂a∂i, 2	95% C.I.
N_A	10,400	8,670–12,200	7,300	4,400–10,100		
N_{AF}	17,300	10,900–29,300	12,300	11,500–13,900		
N_B	2,060	346–5,070	2,100	1,400–2,900		
N_{EU0}	1,710	1,030–3,020	1,000	500–1,900	1,500	700–2,100
r_{EU} (generation^{-1})	0.0055	0.00351–0.0103	0.004	0.0015–0.0066	0.0023	0.0008–0.0045
N_{AS0}	453	210–800	510	310–910	590	320–800
r_{AS} (generation^{-1})	0.016	0.0102–0.0301	0.0055	0.0023–0.0088	0.0037	0.0016–0.006
$m_{\mathrm{AF}\rightarrow B}$ (×10^{-5})	6.06	0.0–13.6	25	15–34		
$m_{\mathrm{AF}\rightarrow B}$ (×10^{-5})	1.63	0.0–3.47	3.0	2.0–6.0		
$m_{\mathrm{AF}\rightarrow B}$ (×10^{-5})	0.487	0.0–1.07	25	15–34		
$m_{\mathrm{AF}\rightarrow B}$ (×10^{-5})	1.54	0.0–2.96	9.6	2.3–17.4		
T_{AF} (yr)	125,400	54,300–250,000	220,000	100,000–510,000		
T_B (yr)	52,400	36,000–80,800	140,000	40,000–270,000		
$T_{\mathrm{EU}\rightarrow\mathrm{AS}}$ (yr)	29,500	23,500–38,000	21,200	17,200–26,500	26,400	18,100–43,100
N_{MX0}	3,200	1,100–6,100			800	160–1,800
r_{MX} (generation^{-1})	0.0071	0.0043–0.011			0.005	0.0014–0.0117
T_{MX} (yr)	29,300	23,000–37,500			21,600	16,300–26,900
f_{MX} (%)	20.4	3.2–41			48	42–60

Inference of parameters by means of maximum likelihood is shown. Confidence intervals were computed by means of nonparametric bootstrap. The estimated parameters θ with MultiPop correspond to the mean of the bootstrap distribution. The estimates by MultiPop are compared with the estimates by ∂a∂i with two three-population models. ∂a∂i 1 denotes the three-population model for the out-of-Africa event described in Gutenkunst *et al.* (2009). ∂a∂i 2 denotes the three-population model for the peopling of the Americas studied in Gutenkunst *et al.* (2009).

numerical sources of error that affect the statistical accuracy of the inferred demographic parameters will remain. For example, Table 2 exhibits several cases where the bias is very large. These biases are not expected to diminish as the sample size grows, or as the number of SNPs increases, because they are due to numerical artifacts. One could minimize these biases by choosing a larger truncation Λ. However, numerical floating-point errors in the evaluation of polynomials become important sources of numerical error when the degree of the polynomial is large enough. In our implementation, we found that for values of $\Lambda > 40$ these sources of numerical error became larger than the truncation error due to a finite choice of Λ.

To minimize the impact of such biases one can either consider models with fewer parameters or introduce statistical corrections to the propagated numerical errors. Here, we apply standard bootstrap methods to correct for bias in the estimators and confidence intervals (see Efron and Tibshirani 1994; DiCiccio and Efron 1996) of some of the models studied above. If $\hat{\theta}^*_\Lambda$ is the maximum-likelihood estimator of a demographic model, where Λ denotes the truncation parameter of the numerical approximation, the bias is defined as

$$\beta\left(\hat{\theta}^*_\Lambda\right) = \mathbb{E}\left(\hat{\theta}^*_\Lambda\right) - \mathbb{E}\left(\hat{\theta}^*\right).$$

Here, $\mathbb{E}(\hat{\theta}^*_\Lambda)$ is the expected biased estimator and $\mathbb{E}(\hat{\theta}^*)$ is the expected unbiased estimator. Although we do not know the unbiased estimator, we can estimate the bias by means of the parametric bootstrap. In particular, we simulate a large number of SNP allele frequencies under the estimated parameters $\hat{\theta}^*_\Lambda$ and consider the associated AFS for each set of simulated SNPs. We denote by $\hat{\theta}^*(b)_\Lambda$ the maximum-likelihood estimate associated with the bth simulated AFS (where $1 \leq b \leq B$ and B is the number of bootstraps). Therefore, the bootstrap estimate of bias is

$$\hat{\beta}\left(\hat{\theta}^*_\Lambda\right) \approx \sum_{b=1}^{B} \frac{\hat{\theta}^*(b)_\Lambda}{B} \; - \; \hat{\theta}^*_\Lambda,$$

where $\hat{\theta}^*_\Lambda$ is the original maximum-likelihood estimate. This approximation will be valid as long as the bias is small and the number of bootstraps B is large enough. Given $\hat{\beta}$, we can now compute the bias-corrected 95% confidence intervals using nonparametric bootstraps (see DiCiccio and Efron 1996) as $(\hat{\theta}^*_\Lambda - \hat{\beta}(\hat{\theta}^*_\Lambda)) \pm \Delta\hat{\theta}^*_\Lambda(\alpha)$, where $\Delta\hat{\theta}^*_\Lambda(\alpha)$ is the $100 \cdot \alpha$th percentile of the nonparametric bootstrap distribution.

As an example, in model 3 we used the parameters inferred by maximum likelihood that are shown in Table 2 to simulate 10,000 SNPs per bootstrap by means of Monte Carlo methods. After estimating the bias with the parametric bootstrap, the 95% bias-corrected confidence intervals for the parameters of model 3 are $0.988-1.04$ for $4N_Au$, 0.722–1.26 for N_1/N_A, 0.255–0.488 for N_2/N_A, 0.0136–0.0197 for $T/2N_A$, 0.479–0.83 for $2N_Am_{1\rightarrow2}$, and 1.69–2.24 for $2N_Am_{2\rightarrow1}$.

Overcoming current limitations

Several limitations exist in the use of joint allele-frequency spectra with many populations. The most important one is that given the joint density of population frequencies $\phi(x|\theta)$, the time needed to compute an AFS grows exponentially with the number of populations. Therefore, the only way to extract demographic information from such a high number of populations requires reducing the number of cells in the AFS to be computed. Because of this, future applications of our method for $K > 3$ will require the use of either adaptive frequency spectra or projections of high-dimensional AFS into triplets or couplets of populations. For instance, by integrating out populations we can compute the joint density of frequencies associated with every triplet of populations from the higher joint density as

$$\tilde{\phi}(x_1, x_2, x_3|\theta) = \int_0^1 \cdots \int_0^1 \phi(x_1, \ldots, x_K|\theta)dx_4 \cdots x_K.$$

The associated three-population AFS is derived from Equation 1 and the density $\tilde{\phi}(x_1, x_2, x_3|\theta)$.

A second question that remains to be explored concerns the bias of the estimator. For a given observed AFS, our method yields a sequence of maximum-likelihood peaks $\{\hat{\theta}^*_\Lambda\}$ labeled by the number of polynomials Λ used. The convergence of the numerical approximation to the exact AFS in the limit of an infinite number of polynomials guarantees that $\hat{\theta}^*_{\Lambda=\infty}$ is an unbiased estimator. It is important to understand the asymptotic behavior of the sequence of peaks $\{\hat{\theta}^*_\Lambda\}$ to estimate $\hat{\theta}^*_{\Lambda=\infty}$ given a few finite values of the sequence $[\{\hat{\theta}^*_\Lambda\}, \{\hat{\theta}^*_{\Lambda+1}\}, \ldots]$. In this study we computed different estimators for several values of Λ, to confirm that our estimators were converging toward the unbiased true estimator. In practice, this is an elaborate approach that requires running nonlinear optimization algorithms for different values of the parameters used in the approximation. A better understanding of this asymptotic behavior will allow the simplification of the analysis of propagated numerical errors associated with finite values of Λ. Similarly, it is important to identify the maximum sample size (number of SNPs) used in each bootstrap that allows us to estimate accurate confidence intervals for a given Λ and a demographic model. The importance of choosing the right sample size for a given Λ lies in that statistical error decreases as sample size increases and propagated numerical error decreases as Λ increases. Therefore, for any given Λ there exists a large enough sample size that, if used to estimate confidence intervals, will yield significantly biased intervals that are difficult to correct via standard statistical methods. This is due to the fact that the numerical error produced by a given Λ will be significantly larger than the statistical error produced by a large enough sample size. In our four-population model we used only 170 SNPs per bootstrap, which yields a conservative estimate of the confidence intervals as we confirmed in simulations. However, in the general case when more data are available and it is not clear what sample size to use in the bootstrap for a given Λ, one can combine parametric and nonparametric bootstrap techniques to estimate the sample size and the bias. In the parametric bootstrap, one generates simulated data with specified parameters that one uses to estimate the bias due to numerical errors. Also, one can determine which sample sizes are small enough to produce larger statistical errors than numerical errors. Then one can use this knowledge to estimate accurate confidence intervals by applying corrections for bias in the nonparametric bootstrap. This approach, although elaborate, will help to estimate accurate confidence intervals for any value of Λ in a wide variety of models.

Acknowledgments

We thank K. Chen, A. Grelaud, R. Gutenkunst, A. Naqvi, A. Sengupta, V. Sousa, and S. Sunyaev for critical comments and the Kavli Institute of Theoretical Physics for supporting the visit of one author (S.L.), during which time part of this project was completed. This work was partially supported by the National Institutes of Health (grant R01GM078204 to J.H.) and the Rita Allen Foundation (Institute for Advanced Study fellowship to S.L.).

Literature Cited

Akey, J., M. Eberle, M. Rieder, C. Carlson, and M. Shriver, 2004 Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. 2: e286.

Albert, F., E. Hodges, J. Jensen, F. Besnier, Z. Xuan *et al.*, 2011 Targeted resequencing of a genomic region influencing tameness and aggression reveals multiple signals of positive selection. Heredity 107: 205–214.

Alexander, D., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19: 1655–1664.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4: e1000083.

Chen, H., 2012 The joint allele frequency spectrum of multiple populations: a coalescent theory approach. Theor. Popul. Biol. 81: 179–195.

Chimpanzee-Sequencing-Consortium, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69–87.

DiCiccio, T., and B. Efron, 1996 Bootstrap confidence intervals. Stat. Sci. 11: 189228.

Efron, B., and R. Tibshirani, 1994 *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York.

Environmental Genome Project, 2010 *NIEHS Environmental Genome Project*. National Institute of Environmental Health Sciences, Research Triangle Park, NC.

Ewens, W., 2004 *Mathematical Population Genetics. I. Theoretical Introduction*, Interdisciplinary Applied Mathematics, Vol. 27. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

Fisher, R., 1930 The distribution of gene ratios for rare mutations. Proc. R. Soc. Edinb. 50: 205–220.

Glasserman, P., 2003 *Monte Carlo Methods in Financial Engineering*, Stochastic Modelling and Applied Probability Edition. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

Goldstein, D. B., and L. Chikhi, 2002 Human migrations and population structure: what we know and why it matters. Annu. Rev. Genomics Hum. Genet. 3: 129–152.

Gravel, S., B. Henn, and R. Gutenkunst, 2011 Demographic history and rare alleles sharing among human populations. Proc. Natl. Acad. Sci. USA 108: 11983–11988.

Griffiths, R., and D. Spano, 2010 *Diffusion Processes and Coalescent Trees*, London Mathematical Society Lecture Notes Series Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman. Cambridge University Press, Cambridge, UK.

Gutenkunst, R., R. Hernandez, S. Williamson, and C. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5: e100695.

Hernandez, R. D., S. H. Williamson, and C. D. Bustamante, 2007 Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol. Biol. Evol. 24: 1792–1800.

Hesthaven, J. S., S. Gottlieb, and D. Gottlieb, 2007 *Spectral Methods for Time-Dependent Problems* (Cambridge Monographs on Applied and Computational Mathematics No. 21). Cambridge University Press, Cambridge/London/New York.

Hudson, R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Jensen, J., and D. Bachtrog, 2011 Characterizing the influence of effective population size on the rate of adaptation: Gillespie's Darwin domain. Genome Biol. Evol. 3: 687–701.

Keinan, A., J. Mullikin, J. Patterson, and D. Reich, 2007 Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat. Genet. 39: 1251–1255.

Kimura, M., 1964 Diffusion models in population genetics. J. Appl. Probab. 1: 177–232.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893–903.

Laberge, A.-M., J. Michaud, A. Richter, E. Lemyre, M. Lambert *et al.*, 2005 Population history and its impact on medical genetics in Quebec. Clin. Genet. 68: 287–301.

Lao, O., K. van Duijn, P. Kersbergen, P. de Knijff, and M. Kayser, 2006 Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. Am. J. Hum. Genet. 78: 680–690.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

Lukic, S., J. Hey, and K. Chen, 2011 Non-equilibrium allele frequency spectra via spectral methods. Theor. Popul. Biol. 79: 203–219.

Matsumoto, M., and T. Nishimura, 1998 Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. Simul. 8: 3–30.

Myers, S., C. Fefferman, and N. Patterson, 2008 Can one learn history from the allelic spectrum? Theor. Popul. Biol. 73: 342–348.

Nachman, M., and S. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297–304.

Nielsen, R., 2001 Statistical tests of selective neutrality in the age of genomics. Heredity 86: 641–647.

Reich, D., 2001 Linkage disequilibrium in the human genome. Nature 411: 199–204.

Risch, N., and K. Merikangas, 1996 The future of genetic studies of complex human diseases. Science 273: 1516–1517.

Sawyer, S., and D. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15: 1576–1583.

Shriver, M. D., E. J. Parra, S. Dios, C. Bonilla, H. Norton *et al.*, 2003 Skin pigmentation, biogeographical ancestry and admixture mapping. Hum. Genet. 112: 387–399.

Song, Y. S., and M. Steinrücken, 2011 A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. Genetics 190: 1117–1129.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics 145: 847–855.

Williamson, S., R. Hernandez, A. Fledel-Alon, L. Zhu, and R. Nielsen, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA 102: 7882–7887.

Wiuf, C., 2006 Consistency of estimators of population scaled parameters using composite likelihood. J. Math. Biol. 53: 821–841.

Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.

Xing, J., W. Watkins, Y. Hu, C. Huff, A. Sabo *et al.*, 2010 Genetic diversity in India and the inference of Eurasian population expansion. Genome Biol. 11: R113.

Communicating editor: Y. S. Song

Appendix

Here, we describe some basic relationships between the orthogonal polynomials that we use in our implementation of the spectral method. In our numerical solutions of the diffusion equations, we make great use of the vector space of polynomials on the interval [0, 1] with degree bounded by the truncation parameter Λ. In this *Appendix*, we denote such a vector space as V_Λ. We use two different orthonormal bases on V_Λ: the basis of Gegenbauer polynomials and the basis of Chebyshev polynomials. In particular, the vector space spanned by the Gegenbauer polynomials of degree $\leq\Lambda$ and the vector space spanned by the Chebyshev polynomials of degree $\leq\Lambda$ is the same vector space V_Λ. They are different orthonormal bases with respect to different inner products. This implies that any truncated expansion in terms of Gegenbauer polynomials can be exactly written as a truncated expansion of Chebyshev polynomials bounded by the same degree. More precisely, if $T_n(x)=\sqrt{(n+2)(2n+3)/(n+1)}P_n^{(1,1)}(2x-1)$ denotes the normalized Gegenbauer polynomials and $C_n(x)=((1/\sqrt{\pi}-\sqrt{2/\pi})\delta_{0,n}+\sqrt{2/\pi})\times\cos(n\ \arccos(2x-1))$ denotes the normalized Chebyshev polynomials on the interval [0, 1], the corresponding orthonormality relationships are

$$\langle T_n,T_m\rangle_{L^2(T)}=\int_0^1 T_n(x)T_m(x)x(1-x)\,dx=\delta_{nm},$$

$$\langle C_n,C_m\rangle_{L^2(C)}=\int_0^1 C_n(x)C_m(x)\,\frac{dx}{\sqrt{x(1-x)}}=\delta_{nm}.$$

One can write any Gegenbauer polynomial $T_n(x)$ in the basis of Chebyshev polynomials as the linear combination

$$T_n(x)=\sum_{i=0}^{n}C_i(x)\int_0^1 T_n(y)C_i(y)\,\frac{dy}{\sqrt{y(1-y)}},$$

with the coefficients $\int_0^1 T_n(y)C_i(y)(dy/\sqrt{y(1-y)})=0$ for all $i>n$. Analogously, one can write any Chebyshev polynomial $C_n(x)$ in the basis of Gegenbauer polynomials as the linear combination

$$C_n(x)=\sum_{i=0}^{n}T_i(x)\int_0^1 C_n(y)T_i(y)y(1-y)dy,$$

with the coefficients $\int_0^1 C_n(y)T_i(y)y(1-y)\,dy=0$ for all $i>n$. We can summarize the changes of basis by introducing the square matrices M and L, defined as

$$M_{mn}=\int_0^1 T_m(y)C_n(y)\frac{dy}{\sqrt{y(1-y)}},$$

and

$$L_{mn}=\int_0^1 C_m(y)T_n(y)y(1-y)\,dy,$$

being both related as $M^T=L^{-1}$ (*i.e.*, the transpose matrix of M equals the inverse matrix of L), as is expected. As an example, the first 6×6 coefficients of M and L are

$$M=\begin{pmatrix}
\sqrt{15\pi} & 0 & 0 & 0 & 0 & 0\\
0 & \frac{5}{4}\sqrt{\frac{21\pi}{2}} & 0 & 0 & 0 & 0\\
\frac{27}{8}\sqrt{\frac{5\pi}{2}} & 0 & \frac{21}{8}\sqrt{\frac{5\pi}{2}} & 0 & 0 & 0\\
0 & \frac{7\sqrt{165\pi}}{16} & 0 & \frac{21\sqrt{165\pi}}{64} & 0 & 0\\
\frac{25\sqrt{273\pi}}{64} & 0 & \frac{45\sqrt{273\pi}}{128} & 0 & \frac{33\sqrt{273\pi}}{128} & 0\\
0 & \frac{675\sqrt{105\pi}}{1024} & 0 & \frac{297\sqrt{105\pi}}{512} & 0 & \frac{429\sqrt{105\pi}}{1024}
\end{pmatrix},$$

and

$$L=\begin{pmatrix}
\frac{1}{\sqrt{15\pi}} & 0 & 0 & 0 & 0 & 0\\
0 & \frac{4}{5}\sqrt{\frac{2}{21\pi}} & 0 & 0 & 0 & 0\\
-\frac{3}{7}\sqrt{\frac{3}{5\pi}} & 0 & \frac{8}{21}\sqrt{\frac{2}{5\pi}} & 0 & 0 & 0\\
0 & -\frac{16}{15}\sqrt{\frac{2}{21\pi}} & 0 & \frac{64}{21\sqrt{165\pi}} & 0 & 0\\
\frac{1}{21}\sqrt{\frac{5}{3\pi}} & 0 & -\frac{8\sqrt{\frac{10}{\pi}}}{77} & 0 & \frac{128}{33\sqrt{273\pi}} & 0\\
0 & \frac{4}{55}\sqrt{\frac{6}{7\pi}} & 0 & -\frac{128}{91}\sqrt{\frac{3}{55\pi}} & 0 & \frac{1024}{429\sqrt{105\pi}}
\end{pmatrix}.$$

The reason we use two different bases of the same vector space is for the sake of convenience. The basis of Gegenbauer polynomials is convenient because the diffusion operator $\Delta\phi=(d^2/dx^2)[x(1-x)\phi]$ is diagonal in such a basis and the integrals that define the AFS become zero when the degree of the polynomial is larger than the number of chromosomes that define the AFS, as defined in Equation 2. The basis of Chebyshev polynomials is convenient to project a given density such as the mutation density $\mu(x)=\exp(-kx)$ or the associated equilibrium density described in Lukic *et al.* (2011), onto the vector space V_Λ. The advantage of using Chebyshev polynomials to project the mutation density rather than the Gegenbauer polynomials is that the weight associated with the Chebyshev polynomials $w=1/\sqrt{x(1-x)}$ is very high near the boundaries $x=0,\ 1$, where most of the information contained in the mutation density is located. On the other hand, the weight associated with the Gegenbauer polynomials $w=x(1-x)$ vanishes on the boundaries. There exist several bases of orthogonal polynomials associated with weights that are high at the boundaries $x=0,\ 1$; however, we chose the basis of Chebyshev polynomials because of the simplicity associated with their implementation.

Being more precise, the two different L^2 products that we use allow us to project the mutation density $\mu(x)$ onto V_Λ in two different ways:

$$\mathcal{P}_{\Lambda,\mathcal{C}}\mu(x)=\sum_{n=0}^{\Lambda}C_n(x)\int_0^1\mu(y)C_n(y)\frac{dy}{\sqrt{y(1-y)}},$$

and

$$\mathcal{P}_{\Lambda,\mathcal{T}}\mu(x)=\sum_{n=0}^{\Lambda}T_n(x)\int_0^1\mu(y)T_n(y)y(1-y)dy.$$

Generically, $\mathcal{P}_{\Lambda,C}\mu(x)$ and $\mathcal{P}_{\Lambda,T}\mu(x)$ are two different polynomials in V_Λ. Even if $\mu(x)$ is a smooth function, and $|\mathcal{P}_{\Lambda,C}\mu(x) - \mathcal{P}_{\Lambda,T}\mu(x)| \to 0$ in the limit of large Λ, both finite polynomials will be different. In our work we use the projection $\mathcal{P}_{\Lambda,\mathrm{C}}\mu(x)$, which is implemented by means of the Gauss–Chebyshev quadrature. The implementation of the Gauss quadrature also illustrates the benefits of using the Chebyshev projection. In particular, Gauss quadratures are evaluated by means of the roots $\{x_i\}$ of an orthogonal polynomial of high degree and a set of weights $\{w_i\}$:

$$\int_0^1 f(x)w(x)dx \simeq \sum_{i=1}^{d} w_i f(x_i).$$

Given a degree d, the roots of the Chebyshev polynomial are closer to the boundaries $x = 0, 1$ than the roots of the Gegenbauer polynomial. For instance, given $d = 10$, the roots of $C_{10}(x)$ are 0.994, 0.946, 0.854, 0.727, 0.578, 0.422, 0.273, 0.146, 0.054, and 0.006, while the roots of $T_{10}(x)$ are 0.972, 0.910, 0.816, 0.699, 0.568, 0.432, 0.3, 0.184, 0.090, and 0.028.

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.141846/-/DC1

Demographic Inference Using Spectral Methods on SNP Data, with an Analysis of the Human Out-of-Africa Expansion

Sergio Lukić and Jody Hey

DOI: 10.1534/genetics.112.141846

File S1
Supporting Material

1. Multi-population Wright-Fisher processes with no migration

In this section we compute the solution to the diffusion equations that describe the time evolution of the density of population allele frequencies under random drift, mutational influx and no migration between populations. First, we review the solution given by Kimura in [1] when the number of populations is $K = 1$. Second, we consider $K = 2$ populations. To this end we use the boundary conditions introduced in [2], solve the associated equations and finally, we show how this solution can be extended to an arbitrary number of populations K.

1.1. **One population.** When the number of populations is one, the density of population allele frequencies $\phi(x,t)$ satisfies the diffusion equation:

$$\frac{\partial \phi(x,t)}{\partial t} = \frac{1}{4N}\frac{\partial^2}{\partial x^2}\left[x(1-x)\phi(x,t)\right] + 2Nu\delta(x-1/2N), \tag{1}$$

where N is the effective population size of a diploid panmictic population, $\delta(x-1/2N)$ is the Dirac delta peaked at $x = 1/2N$, and $\phi(x,t)$ satisfies absorbing boundary conditions at $x = 0$ and $x = 1$. In more general scenarios we can use an effective mutation density $\mu(x)$ instead of the Dirac delta term, [2].

Kimura showed in [1] how Eq. (1) can be solved explicitly by expressing $\phi(x,t)$ as a polynomial expansion. In particular, he used the basis of Gegenbauer polynomials in which the diffusion operator can be expressed as an infinite diagonal matrix. The shifted Gegenbauer polynomials are a class of classical polynomials on the interval $[0,1]$ defined as

$$T_n(x) = \sqrt{\frac{(n+2)(2n+3)}{n+1}}P_n^{(1,1)}(2x-1), \qquad \int_0^1 T_n(x)T_m(x)x(1-x)dx = \delta_{nm} \tag{2}$$

where $P_n^{(1,1)}(z)$ are the classical Jacobi polynomials defined on the interval $-1 \leq z \leq 1$ with weight $w(z) = (1-z)(1+z)$. These polynomials satisfy the associated Jacobi equation:

$$\frac{\partial^2}{\partial x^2}\left[x(1-x)T_n(x)\right] = -(n+1)(n+2)T_n(x). \tag{3}$$

Date: July 27, 2012.

Thus, if we expand the density of population frequencies in this polynomial basis

$$\phi(x,t) = \sum_{n=0}^{\infty} a_n(t) T_n(x),$$

the diffusion equation in Eq. (1) can be written as

(4)

$$\sum_{n=0}^{\infty} \frac{da_n(t)}{dt} T_n(x) = -\sum_{n=0}^{\infty} \frac{(n+1)(n+2)}{4N} a_n(t) T_n(x) + 2Nu \sum_{n=0}^{\infty} T_n(1/2N) \frac{1-1/2N}{2N} T_n(x).$$

For simplicity and to shorten the notation, we denote as μ_n the contribution due to mutational influx $\mu_n = 2NuT_n(1/2N)\frac{1-1/2N}{2N}$. Using this notation, the Ordinary Differential Equation that obeys the coefficients $a_n(t)$ can be written as:

$$\frac{da_n(t)}{dt} = -\frac{(n+1)(n+2)}{4N} a_n(t) + \mu_n. \tag{5}$$

Eq. (5) is a linear differential equation of first order with an inhomogeneous term; this class of equations have a known simple solution which can be written as

$$a_n(t) = \left[a_n(0) - \frac{4N\mu_n}{(n+1)(n+2)} \right] \exp\left(-\frac{(n+1)(n+2)}{4N} t \right) + \frac{4N\mu_n}{(n+1)(n+2)}. \tag{6}$$

Here, $a_n(0)$ are the coefficients associated with the polynomial expansion of the initial density of population frequencies, which can be computed as

$$a_n(0) = \int_0^1 \phi(x,0) T_n(x) x(1-x) dx.$$

Therefore, given *any* density of population frequencies $\phi(x,0)$ at time $t = 0$, we can compute the resulting density $\phi(x,t)$ after t generations evolving under random drift and mutational influx by means of the Gegenbauer expansion $\phi(x,t) = \sum_{n=0}^{\infty} a_n(t) T_n(x)$. The time-dependent coefficients $a_n(t)$ determined in Eq. (6), are a function of the coefficients at initial time and other population genetic parameters such as population size, mutation rate and time. Given the solution $\phi(x,t)$, the Allele Frequency Spectrum associated with a sample of C chromosomes is easily computed by introducing the binomial distribution with parameters C and x as:

$$f_i(t) = \frac{C!}{(C-i)!i!} \sum_{n=0}^{\infty} a_n(t) \int_0^1 x^i (1-x)^{C-i} T_n(x) dx, \quad 0 < i < C,$$

where f_i is the expected number of SNPs that have the derived state in exactly i chromosomes (out of a sample of C chromosomes). Properties of the Jacobi polynomials show that all terms of this sum vanish for $n > C - 2$, thus the AFS can be computed exactly as the finite sum

$$f_i(t) = \frac{C!}{(C-i)!i!} \sum_{n=0}^{C-2} a_n(t) \int_0^1 x^i (1-x)^{C-i} T_n(x) dx. \tag{7}$$

This exact solution can be generalized to an arbitrary number of populations. In the next subsection we show how to compute the solution to the time-evolution of the density of allele frequencies when the number of populations is two.

1.2. **Two populations.** The diffusion equation that describes the dynamics of the density of allele frequencies in two isolated populations is a natural generalization of the one-population case studied above. In particular, if x_1 and x_2 are the derived allele frequencies in population 1 and 2, N_1 and N_2 are the effective population sizes of both populations and $\phi(x_1, x_2, t)$ is the joint density of population frequencies, $\phi(x_1, x_2, t)$ satisfies the following forward diffusion equation

$$\begin{aligned}\frac{\partial \phi}{\partial t} &= \frac{1}{4N_1}\frac{\partial^2}{\partial x_1^2}\left[x_1(1-x_1)\phi\right] + 2N_1 u\delta(x_1 - 1/2N_1)\delta(x_2)\\ &+\frac{1}{4N_2}\frac{\partial^2}{\partial x_2^2}\left[x_2(1-x_2)\phi\right] + 2N_2 u\delta(x_1)\delta(x_2 - 1/2N_2).\end{aligned} \tag{8}$$

As was shown in [2], the solution to Eq. (8) can be expressed as a generalized density with contributions from the different boundary components of the square $[0,1]\times[0,1]$:

$$\begin{aligned}\phi(x_1,x_2,t) = \phi^A(x_1,x_2,t) + \phi^B_{(x_2=0)}(x_1,t)\delta(x_2) +\\ \phi^B_{(x_2=1)}(x_1,t)\delta(1-x_2) + \phi^B_{(x_1=0)}(x_2,t)\delta(x_1) + \phi^B_{(x_1=1)}(x_2,t)\delta(1-x_1) +\\ \phi^C_{(x_1=1,x_2=0)}(t)\delta(1-x_1)\delta(x_2) + \phi^C_{(x_1=0,x_2=1)}(t)\delta(x_1)\delta(1-x_2).\end{aligned} \tag{9}$$

The terms that are multiplied by Dirac deltas represent the contributions to the density that are localized in the different boundary components. In particular, the A-term is localized in the bulk of the square, the four B-terms are localized in the edges of the square and finally, the two C-terms are localized in the two vertices of the square that are not absorbing. The Ancestral vertex $(x_1 = 0, x_2 = 0)$ and the Derived vertex $(x_1 = 1, x_2 = 1)$ are absorbing and hence do not contribute SNPs to the density $\phi(x_1, x_2, t)$.

As Eq. (8) is the natural extension of the one-population process and the one-population diffusion equation can be solved by means of polynomials expansions, we expand each term in Eq. (9) using the same basis of Jacobi polynomials $T_n(x)$ defined in Eq. (2). As we will see at the end of this section, such a polynomial expansion will allow us to find the exact solution of the two-population process. In particular, we write the polynomial expansion of each term in Eq. (9) as:

$$\begin{aligned}\phi^A(x_1,x_2,t) &= \sum_{n,m=0}^{\infty} a^A_{nm}(t)T_n(x_1)T_m(x_2),\\ \phi^B_{(x_2=0)}(x_1,t) &= \sum_{n=0}^{\infty} a^B_{(x_2=0),n}(t)T_n(x_1),\\ \phi^B_{(x_2=1)}(x_1,t) &= \sum_{n=0}^{\infty} a^B_{(x_2=1),n}(t)T_n(x_1),\\ \phi^B_{(x_1=0)}(x_2,t) &= \sum_{m=0}^{\infty} a^B_{(x_1=0),m}(t)T_m(x_2),\\ \phi^B_{(x_1=1)}(x_2,t) &= \sum_{m=0}^{\infty} a^B_{(x_1=1),m}(t)T_m(x_2),\\ \phi^C_{(x_1=1,x_2=0)}(t) &= a^C_{(x_1=1,x_2=0)}(t),\\ \phi^C_{(x_1=0,x_2=1)}(t) &= a^C_{(x_1=0,x_2=1)}(t).\end{aligned} \tag{10}$$

In this polynomial basis, Eq. (8) requires that the a-variables satisfy a set of Ordinary Differential Equations (ODE) that can be integrated exactly. The associated ODEs can be determined by taking into account the different contributions to the dynamics of the a-variables (random drift, influx of polymorphisms in the boundary components due to fixation events, and influx of polymorphisms due to mutations). Following [2] we know that the dynamics of the $a^A_{nm}(t)$-terms is just governed by random drift (there is no influx of polymorphisms). On the other hand, the dynamics of the terms $a^B_{(x_1=1),m}(t)$ and $a^B_{(x_2=1),n}(t)$ depend on both random drift and the influx of polymorphisms that reach fixation at either $x_1 = 1$ or $x_2 = 1$. The terms $a^B_{(x_2=0),n}(t)$ and $a^B_{(x_1=0),m}(t)$ furthermore receive the constant influx of polymorphisms due to de novo mutations at the population level. Finally, the time evolution of the terms $a^C_{(x_1=1,x_2=0)}(t)$ and $a^C_{(x_1=0,x_2=1)}(t)$ is described by the influx of polymorphisms that reach fixation from $\phi^B_{(x_2=0)}(x_1, t)$ and $\phi^B_{(x_1=1)}(x_2, t)$, in the case of $a^C_{(x_1=1,x_2=0)}(t)$, or from $\phi^B_{(x_1=0)}(x_2, t)$ and $\phi^B_{(x_2=1)}(x_1, t)$ in the case of $\phi^C_{(x_1=0,x_2=1)}(t)$.

The dynamics of the a-coefficients can be made quantitatively explicit in the following system of linear differential equations:

$$\frac{da^A_{nm}}{dt} = -\left(\frac{(n+1)(n+2)}{4N_1} + \frac{(m+1)(m+2)}{4N_2}\right) a^A_{nm}, \tag{11}$$

$$\frac{da^B_{(x_2=0),n}}{dt} = -\frac{(n+1)(n+2)}{4N_1} a^B_{(x_2=0),n} + \mu^1_n + \sum_{m=0}^{\infty} \frac{a^A_{nm} T_m(0)}{4N_2}, \tag{12}$$

here, $\mu^1_n = 2N_1 u \times T_n(1/2N_1)\frac{1-1/2N_1}{2N_1}$ is the contribution due to mutational influx in population 1,

$$\frac{da^B_{(x_1=0),m}}{dt} = -\frac{(m+1)(m+2)}{4N_2} a^B_{(x_1=0),m} + \mu^2_m + \sum_{n=0}^{\infty} \frac{a^A_{nm} T_n(0)}{4N_1}, \tag{13}$$

here, $\mu^2_m = 2N_2 u \times T_m(1/2N_2)\frac{1-1/2N_2}{2N_2}$ is the contribution due to mutational influx in population 2,

$$\frac{da^B_{(x_2=1),n}}{dt} = -\frac{(n+1)(n+2)}{4N_1} a^B_{(x_2=1),n} + \sum_{m=0}^{\infty} \frac{a^A_{nm} T_m(1)}{4N_2}, \tag{14}$$

$$\frac{da^B_{(x_1=1),m}}{dt} = -\frac{(m+1)(m+2)}{4N_2} a^B_{(x_1=1),m} + \sum_{n=0}^{\infty} \frac{a^A_{nm} T_n(1)}{4N_1}, \tag{15}$$

$$\frac{da^C_{(x_1=1,x_2=0)}}{dt} = \sum_{n=0}^{\infty} \frac{a^B_{(x_2=0),n} T_n(1)}{4N_1} + \sum_{m=0}^{\infty} \frac{a^B_{(x_1=1),m} T_m(0)}{4N_2}, \tag{16}$$

and

$$\frac{da^C_{(x_1=0,x_2=1)}}{dt} = \sum_{n=0}^{\infty} \frac{a^B_{(x_2=1),n} T_n(0)}{4N_1} + \sum_{m=0}^{\infty} \frac{a^B_{(x_1=0),m} T_m(1)}{4N_2}. \tag{17}$$

This system of coupled linear differential equations can be solved by integrating first the uncoupled equation Eq. (11), using the corresponding solution to solve Eqs. (12), (13), (14), and (15), and finally using those solutions to solve Eq. (16) and Eq. (17). At each step, one has to integrate a set of linear ODEs of first order whose solutions are known.

The solution of Eq. (11) is:

$$a^A_{nm}(t) = a^A_{nm}(0)\exp\left[-\left(\frac{(n+1)(n+2)}{4N_1} + \frac{(m+1)(m+2)}{4N_2}\right)t\right], \tag{18}$$

with $a^A_{nm}(0)$ the coefficients associated with $\phi^A(x_1, x_2, 0)$ at initial time:

$$a^A_{nm}(0) = \int_0^1\int_0^1 \phi^A(x_1, x_2, 0)T_n(x_1)T_m(x_2)x_1(1-x_1)x_2(1-x_2)dx_1dx_2.$$

Now, we can use the solution Eq. (18) to integrate Eqs. (12), (13), (14), and (15). Hence, we can write the solution of Eq. (12) as

(19)

$$a^B_{(x_2=0),n}(t) = b^B_{(x_2=0),n}\exp\left(-\frac{(n+1)(n+2)}{4N_1}t\right) + \frac{4N_1\mu^1_n}{(n+1)(n+2)} - \sum_{m=0}^{\infty}\frac{a^A_{nm}(t)T_m(0)}{(m+1)(m+2)},$$

with $b^B_{(x_2=0),n}$ a time-independent function defined as

$$b^B_{(x_2=0),n} = a^B_{(x_2=0),n}(0) - \frac{4N_1\mu^1_n}{(n+1)(n+2)} + \sum_{m=0}^{\infty}\frac{a^A_{nm}(0)T_m(0)}{(m+1)(m+2)}.$$

The coefficients $a^B_{(x_2=0),n}(0)$ are associated with the initial-time density as

$$a^B_{(x_2=0),n}(0) = \int_0^1 \phi^B_{(x_2=0)}(x_1, 0)T_n(x_1)x_1(1-x_1)dx_1.$$

Similarly, the solution of (13) is

(20)

$$a^B_{(x_1=0),m}(t) = b^B_{(x_1=0),m}\exp\left(-\frac{(m+1)(m+2)}{4N_2}t\right) + \frac{4N_2\mu^2_m}{(m+1)(m+2)} - \sum_{n=0}^{\infty}\frac{a^A_{nm}(t)T_n(0)}{(n+1)(n+2)},$$

with $b^B_{(x_1=0),m}$ defined as

$$b^B_{(x_1=0),m} = a^B_{(x_1=0),m}(0) - \frac{4N_2\mu^2_m}{(m+1)(m+2)} + \sum_{n=0}^{\infty}\frac{a^A_{nm}(0)T_n(0)}{(n+1)(n+2)}.$$

The solution of (14) is

$$a^B_{(x_2=1),n}(t) = b^B_{(x_2=1),n}\exp\left(-\frac{(n+1)(n+2)}{4N_1}t\right) - \sum_{m=0}^{\infty}\frac{a^A_{nm}(t)T_m(1)}{(m+1)(m+2)}, \tag{21}$$

with $b^B_{(x_2=1),n}$ defined as

$$b^B_{(x_2=1),n} = a^B_{(x_2=1),n}(0) + \sum_{m=0}^{\infty}\frac{a^A_{nm}(0)T_m(1)}{(m+1)(m+2)}.$$

And finally, for this class of solutions, the solution of (15) is

$$a^B_{(x_1=1),m}(t) = b^B_{(x_1=1),m}\exp\left(-\frac{(m+1)(m+2)}{4N_2}t\right) - \sum_{n=0}^{\infty}\frac{a^A_{nm}(t)T_n(1)}{(n+1)(n+2)}, \tag{22}$$

with $b^B_{(x_1=1),m}$ defined as

$$b^B_{(x_1=1),m} = a^B_{(x_1=1),m}(0) + \sum_{n=0}^{\infty}\frac{a^A_{nm}(0)T_n(1)}{(n+1)(n+2)}.$$

The solutions to Eqs. (16) and (17) are frequency-independent functions of time which can be obtained by integrating Eqs. (19), (20), (21), and (22):
(23)

$$\Delta a^C_{(x_1=1,x_2=0)}(t) = \sum_{n=0}^{\infty} \frac{T_n(1)}{4N_1} \int_0^t a^B_{(x_2=0),n}(u)du + \sum_{m=0}^{\infty} \frac{T_m(0)}{4N_2} \int_0^t a^B_{(x_1=1),m}(u)du,$$

and
(24)

$$\Delta a^C_{(x_1=0,x_2=1)}(t) = \sum_{n=0}^{\infty} \frac{T_n(0)}{4N_1} \int_0^t a^B_{(x_2=1),n}(u)du + \sum_{m=0}^{\infty} \frac{T_m(1)}{4N_2} \int_0^t a^B_{(x_1=0),m}(u)du,$$

where the Δa terms are defined as:

$$\Delta a^C_{(x_1=1,x_2=0)}(t) := a^C_{(x_1=1,x_2=0)}(t) - a^C_{(x_1=1,x_2=0)}(0),$$

and

$$\Delta a^C_{(x_1=0,x_2=1)}(t) := a^C_{(x_1=0,x_2=1)}(t) - a^C_{(x_1=0,x_2=1)}(0).$$

In summary, the solution of Eq. (8) can be written as a generalized density with seven components (as in Eq. (9)). Each of these seven boundary-specific densities can be expanded by means of a polynomial expansion (as in Eq. (10)). The time-dependent coefficients associated with these expansions were obtained in Eqs. (18)-(24).

Given an explicit solution $\phi(x_1, x_2, t)$, one can make connections with measurable quantities by computing the theoretical prediction of some of them. For instance, one can compute the Allele Frequency Spectrum associated with a sample of C chromosomes by introducing the binomial distribution as:
(25)

$$f_{ij}(t) = \frac{C!}{(C-i)!i!} \frac{C!}{(C-j)!j!} \int_0^1 \int_0^1 x_1^i (1-x_1)^{C-i} x_2^j (1-x_2)^{C-j} \phi(x_1, x_2, t) dx_1 dx_2,$$

for $0 \leq i \leq C, 0 \leq j \leq C$ and $0 < i+j < 2C$. Here, f_{ij} is the expected number of SNPs in which the derived state is found in i chromosomes in population one and j chromosomes in population two. In general, evaluating Eq. (25) requires integrating $\phi(x_1, x_2, t)$, which involves computing several infinite sums. However, this formula becomes particularly simple when $0 < i < C$ and $0 < j < C$:

$$f_{ij}(t) = \frac{C!}{(C-i)!i!} \frac{C!}{(C-j)!j!} \sum_{n,m=0}^{\infty} a^A_{nm}(t) \times$$

$$\int_0^1 \int_0^1 x_1^i (1-x_1)^{C-i} x_2^j (1-x_2)^{C-j} T_n(x_1) T_m(x_2) dx_1 dx_2,$$

and because of properties of the Jacobi polynomials this simplifies to the finite sum

$$f_{ij}(t) = \frac{C!}{(C-i)!i!} \frac{C!}{(C-j)!j!} \sum_{n,m=0}^{C-2} a^A_{nm}(t) \times$$

$$\int_0^1 \int_0^1 x_1^i (1-x_1)^{C-i} x_2^j (1-x_2)^{C-j} T_n(x_1) T_m(x_2) dx_1 dx_2.$$

This resembles the simple formula Eq. (7) derived in the one-population case. Hence, after including the contributions from every boundary component, the solution of the two-population diffusion equation describing the time evolution of the density of allele frequencies is a natural extension of the one-population solution. One can also generalize the two-population case studied here, to a scenario with an arbitrary number of populations.

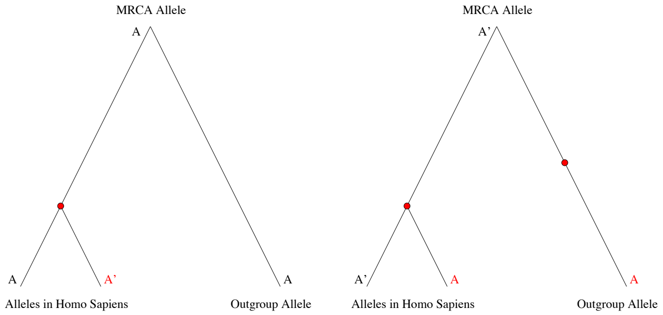

FIGURE 1. Most probable histories of a diallelic locus (with alleles A and A'). In red we denote the derived allele that arises as a mutation since the split with the most recent common ancestor. Here we assume that one of the alleles is identical to the orthologous base in an outgroup species that shares a recent common ancestor, such as Pan troglodytes or Rhesus macaque in the case of homo sapiens.

write the probability of mutation as

$$\begin{aligned} p(xAy \to xA'y|\tau) &= p(xAy \text{ has diverged since MRCA}|\tau) \\ &\times \frac{p(xAy \to xA'y|xAy \text{ has diverged since MRCA};\ \tau)}{p(xAy \text{ has diverged since MRCA}|xAy \to xA'y;\ \tau)} \end{aligned} \tag{26}$$

Here, x and y are the flanking nucleotides that define the context, and τ is the time of divergence between the species under consideration. Eq. (26) allows to estimate the mutation rates using genome wide data on the divergence between species. More explicitly, each term in (26) can be computed as:

$$p(xAy \text{ has diverged since MRCA}|\tau) = 64 \times r_{div} \times \pi_{xAy} \tag{27}$$

$$\begin{aligned} p(xAy \to xA'y|xAy &\text{ has diverged since MRCA};\ \tau) = \\ (\pi_{A;A',A}p_M(A|A,A') &+ \pi_{A';A',A}(1-p_M(A'|A',A))) \times \\ (\pi_{A;A',A}p_M(A|A,A') &+ \pi_{A';A',A}(1-p_M(A'|A',A)) + \\ \pi_{A;B,A}p_M(A|A,B) &+ \pi_{B;B,A}(1-p_M(B|B,A)) + \\ \pi_{A;B',A}p_M(A|A,B') &+ \pi_{B';B',A}(1-p_M(B'|B',A)))^{-1} \end{aligned} \tag{28}$$

$$p(xAy \text{ has diverged since MRCA}|xAy \to xA'y;\ \tau) = 1.0 \tag{29}$$

In Eq. (27), r_{div} is the probability that two random homologous nucleotides are different, which is estimated to be $1.57/100$ between human and chimp. π_{xAy} is the genome-wide average frequency of trinucleotides xAy, and $64 = 4^3$ is a normalization constant. In Eq. (28), $\pi_{w;z,w}$ is the genome-wide frequency of trinucleotides xwy in the outgroup species whose orthologous has polymorphisms xwy and xzy in the species under consideration. The probability $p_M(w|w,z)$ is a shorthand for

$$p(xwy \text{ is MRCA}|\text{Outgroup} = xwy, \text{Alleles} = xzy, xwy).$$

And finally, B and B' are the two nucleotides in g,t,a,c, which are not A nor A'; i.e. B and B' span the complementary set to A and A' in $\{g,t,a,c\}$. Therefore, all parameters that appear in Eq. (26) can be estimated using genomic and polymorphic data, except

$p(xAy \to xA'y|\tau)$ and $p_M(w|w,z)$. The probability functions $1 - p_M(w|w,z)$ are exactly the quantities that define the probability of ancestral allele misidentification using the outgroup base. Such probabilities also satisfy :

$$\begin{aligned} p_M(w|w,z) = {} & p(xwy \to xzy|\tau)p(xwy \to xwy|\tau) \times \\ & (p(xwy \to xzy|\tau)p(xwy \to xwy|\tau) + \\ & p(xzy \to xwy|\tau)p(xzy \to xwy|\tau))^{-1}. \end{aligned} \tag{30}$$

Here, $p(xwy \to xwy|\tau)$ equals $1 - \sum_{z\in S} p(xwy \to xzy|\tau)$, with S the set $\{g,t,a,c\}\backslash w$. In other words, $p_M(w|w,z)$ is approximately equal to the probability that the history represented in the left tree of Fig. 1 actually happened, given that the left and right trees represent the most probable events.

Thus, by substituting Eq. (26) into Eq. (30), one gets a system of equations in the unknown variables $p_M(w|w,z)$, which can be solved easily.

We estimated the probabilities of ancestral allele misidentification in humans, using the chimp as the outgroup species. Using the human and chimp genomes, plus the EGP SNP data, we estimated all the parameters in Eqs. (27), (28) and (29). By starting with initial values $p^0_M(w|w,z) = 1$, one can solve Eq. (26) and recompute $p^1_M(w|w,z)$ using Eq. (30). This yields an iterative mechanism that produces a quickly convergent sequence of probabilities $p^n_M(w|w,z)$ towards a unique fixed point, solution of the system of equations. We found that the resulting probabilities $1 - p_M(w|w,z)$ can be broken down into CpG and non-CpG contexts. In the non-CpG context, i.e. mutations which are not of the type CG to TG nor CT to CA, all the probabilities $1 - p_M(w|w,z)$ are smaller than 0.006. However, for mutations of the type CG to TG or CT to CA, the probabilities $1 - p_M(w|w,z)$ range between a maximum of 0.16 and a minimum of 0.06. This result is very similar to the one given in [4].

3. Comparison of the different boundary conditions used in this study

The one-population two-allele Wright-Fisher diffusion with influx of mutations can be defined by means of the PDE

$$\frac{\partial \phi}{\partial t} = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}\left[x(1-x)\phi(x,t)\right] + 2Nu\delta(x - 1/2N). \tag{31}$$

Here, N_e denotes the effective population size, N is the census population size and u the mutation rate. The boundary conditions at $x = 0$ and $x = 1$ are absorbing, and the term $2Nu\delta(x - 1/2N)$ denotes the source of new mutations that arise at frequency $x = 1/2N$ for large N. It is very important to understand how to regularize $\delta(x - 1/2N)$ in any finite approximation that one applies to numerically solve Eq. (31). In particular, experience with different numerical solutions of Eq. (31) suggests that small changes in the finite regularization of the Dirac delta might have large effects on the numerical solution of Eq. (31).

In this section, we study the convergence properties of the finite-difference method used in [6] and the spectral method used in this paper for the particular case of Eq. (31). Although several sources of numerical error exist (e.g. either the truncated spectral expansion or the finite-difference approximation of $\phi(x,t)$), here we only consider the contribution to error due to the finite regularization of $\delta(x - 1/2N)$.

In particular, the finite regularizations of $\delta(x - 1/2N)$ that we consider here can be described using the diffusion equation

(32) $$\frac{\partial \phi}{\partial t} = \frac{1}{4N_e} \frac{\partial^2}{\partial x^2} \left[x(1-x)\phi(x,t) \right] + u\mu(x),$$

with $\mu(x)$ a function of the frequency that depends on the particular choice of numerical method. The standard diffusion in Eq. (31) is recovered when $\mu(x) = 2N\delta(x - 1/2N)$. In general, we denote this function as

(33) $$\mu_N(x) = 2N\delta(x - 1/2N).$$

In the case of the spectral method (see [2]) we instead use the function

(34) $$\mu_k(x) = c_k \exp(-kx),$$

with

$$c_k = \frac{k^2}{1 - \exp(-k) - k\exp(-k)}.$$

Here, k is a positive real number that depends monotonically on the truncation parameter Λ. In particular, k is chosen such that the truncated polynomial approximation of Eq. (34) is accurate enough. Thus, the limit of large Λ corresponds with the limit of large k.

In the case of the finite-difference method, one approximates $\phi(x,t)$ as a piece-wise linear function. More precisely, if $\{x_j\}_{j=0}^{G}$ are the grid points on $[0,\,1]$ that we use in the finite-difference scheme, we introduce a basis of functions $\{f_j(x)\}_{j=0}^{G}$ with

$$f_j(x) = \theta(x - x_{j-1})\theta(x_{j+1} - x) \left(\frac{x - x_{j-1}}{x_j - x_{j-1}} \theta(x_j - x) + \frac{x_{j+1} - x}{x_{j+1} - x_j} \theta(x - x_j) \right),$$

for $0 < j < G$,

$$f_0(x) = \theta(x_1 - x) \left(\frac{x_1 - x}{x_1 - x_0} \theta(x - x_0) \right),$$

and

$$f_G(x) = \theta(x - x_{G-1}) \left(\frac{x - x_{G-1}}{x_G - x_{G-1}} \theta(x_G - x) \right),$$

such that the finite-difference approximation of $\phi(x,t)$ can be written as

$$\phi(x,t) \simeq \sum_{j=0}^{j=G} \phi_j^t f_j(x).$$

Here, $x_0 = 0$, $x_G = 1$ and $\theta(x)$ is the Heaviside step function (defined as $\theta(x) = 0$ for $x < 0$ and $\theta(x) = 1$ for $x > 0$). In Gutenkunst et al. [6] the authors use an adaptive grid on $[0,\,1]$ that is uniform near $x = 0$. Therefore, for $f_0(x)$ and $f_1(x)$ we assume that $x_1 - x_0 = x_2 - x_1 = \Delta$, $x_0 = 0$, and the corresponding basis functions are

$$f_0(x) = \theta(\Delta - x) \left(\frac{\Delta - x}{\Delta} \theta(x) \right),$$

and

$$f_1(x) = \theta(2\Delta - x) \left(\frac{x}{\Delta} \theta(\Delta - x) + \frac{2\Delta - x}{\Delta} \theta(x - \Delta) \right).$$

Gutenkunst et al. [6] inject new mutations at each time-step by updating the value of ϕ_1^t as (see Eq. (S9) in [6])

(35) $$\frac{\phi_1^{t+dt} - \phi_1^t}{dt} = \frac{u}{\Delta^2}.$$

Remark 1. *Note that Gutenkunst et al.* [6] *write Eq.* (32) *using different units. In particular, they introduce a reference population size N_0 with $\theta = 4N_0 u$ and write Eq.* (32) *as*

$$\frac{\partial \phi}{\partial \tau} = \frac{1}{2\nu} \frac{\partial^2}{\partial x^2} \left[x(1-x)\phi(x,\tau) \right] + \frac{\theta}{2}\mu(x), \tag{36}$$

with $\tau = t/2N_0$ and $\nu = N_e/N_0$. In their notation the value of ϕ_1^τ is updated as

$$\frac{\phi_1^{\tau+d\tau} - \phi_1^\tau}{d\tau} = \frac{\theta}{2\Delta^2}.$$

Updating the value of ϕ_1^t, as in Eq. (35), when solving the diffusion equations is equivalent to using Eq. (32) and the function

$$\mu_\Delta(x) = c_\Delta \Delta f_1(x) = c_\Delta [x\theta(\Delta - x) + (2\Delta - x)\theta(x - \Delta)\theta(2\Delta - x)], \tag{37}$$

with $c_\Delta = \Delta^{-3}$. Observe that $\theta(\Delta - x)\theta(2\Delta - x) = \theta(\Delta - x)$ and that $\theta(x)$ denotes here the Heaviside step function.

It is not obvious that $\mu_\Delta(x)$ in Eq. (37) or $\mu_k(x)$ in Eq. (34) converge to $\mu_N(x) = 2N\delta(x - 1/2N)$ in the limits $N \to \infty$, $k \to \infty$ and $\Delta \to 0$. Hence, it is not obvious that the solutions associated with each finite regularization converge to the exact solution of Eq. (31). However, in the remainder of this section we demonstrate how both approximate solutions actually converge to the exact solution.

Proposition 1. *Let $\phi_N(x,t)$, $\phi_k(x,t)$, and $\phi_\Delta(x,t)$ be the solutions of Eq.* (32) *corresponding to the functions $\mu(x)$ defined in Eq.* (33) *for $\phi_N(x,t)$, Eq.* (34) *for $\phi_k(x,t)$ and Eq.* (37) *for $\phi_\Delta(x,t)$. Additionally, let the initial condition be the same arbitrary density $\varphi(x)$ in all of the three cases:*

$$\phi_N(x, t=0) = \phi_k(x, t=0) = \phi_\Delta(x, t=0) = \varphi(x).$$

Then, iff c_k in Eq. (34) *is defined as*

$$c_k = \frac{k^2}{1 - \exp(-k) - k\exp(-k)},$$

$\phi_k(x,t)$ converges to the exact solution $\phi_N(x,t)$ in the limits $k \to \infty$, $N \to \infty$ and finite N_e. In particular,

$$\parallel \phi_{N\to\infty}(x,t) - \phi_k(x,t) \parallel_{L^1} \leq \frac{4N_e u}{k}(1 + \exp(-t/2N_e)), \quad t \geq 0.$$

Similarly, iff c_Δ in Eq. (37) *is defined as $c_\Delta = \Delta^{-3}$, $\phi_\Delta(x,t)$ converges to the exact solution $\phi_N(x,t)$ in the limits $\Delta \to 0$, $N \to \infty$ and finite N_e. In particular,*

$$\parallel \phi_{N\to\infty}(x,t) - \phi_\Delta(x,t) \parallel_{L^1} \leq \frac{7}{3} N_e u \Delta (1 + \exp(-t/2N_e)), \quad t \geq 0.$$

Proof. The proof consists of three parts. First, we describe the solution of Eq. (32) for an arbitrary choice of $\mu(x)$; second, we derive a general bound for the L^1-norm of the difference of two solutions associated with different choices of $\mu(x)$ (see Eq. (42)); and third, we apply this general argument to the particular cases of $\phi_N(x,t)$, $\phi_k(x,t)$, and $\phi_\Delta(x,t)$ and the L^1-norms

$$\parallel \phi_{N\to\infty}(x,t) - \phi_k(x,t) \parallel_{L^1} = \int_0^1 |\phi_{N\to\infty}(x,t) - \phi_k(x,t)| x(1-x) dx,$$

and

$$\| \phi_{N\to\infty}(x,t) - \phi_\Delta(x,t) \|_{L^1} = \int_0^1 |\phi_{N\to\infty}(x,t) - \phi_\Delta(x,t)| x(1-x)dx.$$

Any solution of Eq. (32) can be described as the sum of a homogeneous solution and an inhomogeneous solution. In particular, if $\phi_{e,\mu}(x)$ is the steady state solution that satisfies

$$0 = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}\left[x(1-x)\phi_{e,\mu}(x)\right] + u\mu(x), \tag{38}$$

$\phi_\mu(x, t = 0) = \varphi(x)$ is the initial condition, and $\gamma(x,t) = \exp(tL_{FP})\gamma(x,0)$ is the solution to the homogenous ($\mu(x) = 0$) problem

$$\frac{\partial \gamma(x,t)}{\partial t} = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}\left[x(1-x)\gamma(x,t)\right],$$

then one can write the solution of Eq. (32) as

$$\phi_\mu(x,t) = \exp(tL_{FP})(\varphi(x) - \phi_{e,\mu}(x)) + \phi_{e,\mu}(x).$$

Here, $\exp(tL_{FP})$ denotes the time evolution operator, and L_{FP} denotes the Fokker-Planck diffusion operator. Therefore, if $\phi_{\mu_1}(x,t)$ and $\phi_{\mu_2}(x,t)$ are solutions of Eq. (32) associated with the functions $\mu_1(x)$ and $\mu_2(x)$, the difference $\phi_{\mu_1} - \phi_{\mu_2}$ satisfies

$$\phi_{\mu_1}(x,t) - \phi_{\mu_2}(x,t) = \exp(tL_{FP})(\phi_{e,\mu_2}(x) - \phi_{e,\mu_1}(x)) + \phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x).$$

In order to bound $\| \phi_{\mu_1} - \phi_{\mu_2} \|_{L^1}$ we apply the Minkowski inequality as follows:

$$\| \phi_{\mu_1}(x,t) - \phi_{\mu_2}(x,t) \|_{L^1} = \| \exp(tL_{FP})(\phi_{e,\mu_2}(x) - \phi_{e,\mu_1}(x)) + \phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x) \|_{L^1} \leq$$

$$\| \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x)) \|_{L^1} + \| \phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x) \|_{L^1}. \tag{39}$$

In our particular case (in which $\mu_1(x) = \mu_{N\to\infty}(x)$ and $\mu_2(x) = \mu_k(x)$ or $\mu_2(x) = \mu_\Delta(x)$), $\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x)$ is non-negative for all $x \in (0, 1)$. As the time-evolution operator $\exp(tL_{FP})$ preserves the non-negativity of the density, we can write Eq. (39) as

$$\| \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x)) \|_{L^1} + \| \phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x) \|_{L^1} =$$

$$\int_0^1 \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))x(1-x)dx + \int_0^1 (\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))x(1-x)dx.$$

The operator $\exp(tL_{FP})$ is diagonal in the basis spanned by the Gegenbauer polynomials (see Eq. (2)). In particular, we can write $\exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))$ as

$$\exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x)) = \sum_{n=0}^{\infty} a_n \exp(-t(n+1)(n+2)/4N_e)T_n(x),$$

with

$$a_n = \int_0^1 T_n(x)(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))x(1-x)dx.$$

Now, using this expansion and the fact that $T_0(x) = \sqrt{6}$, we can write

$$\int_0^1 \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))x(1-x)dx =$$

$$\frac{1}{\sqrt{6}}\int_0^1 \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))T_0(x)x(1-x)dx =$$

$$\frac{1}{\sqrt{6}}\sum_{n=0}^{\infty} a_n \exp(-t(n+1)(n+2)/4N_e)\int_0^1 T_n(x)T_0(x)x(1-x)dx = \frac{a_0}{\sqrt{6}}\exp(-t/2N_e).$$

Therefore, if we define I_{μ_1,μ_2} as

$$I_{\mu_1,\mu_2} = \int_0^1 (\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))x(1-x)dx, \tag{40}$$

then $a_0 = \sqrt{6}I_{\mu_1,\mu_2}$ and the sum of L^1-norms is

$$\| \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x)) \|_{L^1} + \| \phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x) \|_{L^1} =$$

$$I_{\mu_1,\mu_2}(1 + \exp(-t/2N_e)). \tag{41}$$

Now, from Eq. (39) it follows that

$$\| \phi_{\mu_1}(x,t) - \phi_{\mu_2}(x,t) \|_{L^1} \leq I_{\mu_1,\mu_2}(1 + \exp(-t/2N_e)). \tag{42}$$

In order to determine the bound in Eq. (42) one needs only to evaluate the integral in Eq. (40). This requires solving Eq. (38) to obtain a closed-form expression for $\phi_{e,\mu_1}(x)$ and $\phi_{e,\mu_2}(x)$. One can solve Eq. (38) simply by integrating the equation twice

$$\int_0^x \int_0^y \frac{d^2\psi(z)}{dz^2} dzdy = -4N_e u \int_0^x \int_0^y \mu(z)dzdy,$$

$$\psi(x) = \psi(0) + \psi'(0)x - 4N_e u \int_0^x \int_0^y \mu(z)dzdy,$$

with $\psi(x) = x(1-x)\phi_{e,\mu}(x)$ and $\psi'(x) = d\psi/dx$. We require $\phi_{e,\mu}(x)$ to be finite at the boundaries $x = 0$ and $x = 1$, i.e. $\psi(0) = \psi(1) = 0$. Therefore, for the particular functions $\mu(x)$ that we consider here (Eq. (33), Eq. (34) and Eq. (37)) we find the following solutions of Eq. (38):

$$\phi_{e,N}(x) = \frac{4N_e u}{x(1-x)} \left[(2N-1)x - 2N(x - 1/2N)\theta(x - 1/2N)\right], \tag{43}$$

$$\phi_{e,k}(x) = \frac{4N_e u}{x(1-x)} \frac{c_k}{k^2} \left[x(\exp(-k) - 1) - \exp(-kx) + 1\right]. \tag{44}$$

and

$$\begin{aligned}\phi_{e,\Delta}(x) \quad = \quad & \frac{4N_e u}{x(1-x)} c_\Delta \Delta^3 \Bigg[(\Delta^{-1} - 1)x - \frac{x^3}{6\Delta^3}\theta(\Delta - x) + \\ & \left(\frac{x^3}{6\Delta^3} - \frac{x^2}{\Delta^2} + \frac{x}{\Delta} - \frac{1}{3}\right)\theta(x - \Delta)\theta(2\Delta - x) + (1 - \Delta^{-1}x)\theta(x - 2\Delta)\Bigg].\end{aligned} \tag{45}$$

Note that Eq. (43) yields $\phi_{e,N}(x) = 4N_e u/x$ for $x > 1/2N$. Thus, the limit $N \to \infty$ of Eq. (43) corresponds with $\phi_{e,N}(x) = 4N_e u/x$ for $0 < x \leq 1$. Note also that only if $c_k = k^2/(1 - \exp(-k) - k\exp(-k))$ then $\phi_{e,k}(x)$ converges to $4N_e u/x$ near $x = 1$. Similarly, only if $c_\Delta = \Delta^{-3}$ then $\phi_{e,\Delta}(x)$ converges to $4N_e u/x$ near $x = 1$.

Now we can evaluate the integral in Eq. (40) for $\mu_1(x) = \mu_N(x)$, $\mu_2(x) = \mu_k(x)$ and $c_k = k^2/(1 - \exp(-k) - k\exp(-k))$, as

$$\int_0^1 \left(\frac{4N_e u}{x} - \phi_{e,k}(x)\right) x(1-x)dx = 4N_e u \frac{1 + k/2 + (1 - \exp(k))/k}{1 + k - \exp(k)}, \tag{46}$$

which in the limit of large k converges to

$$I_{\mu_{N\to\infty},\mu_k} = \frac{4N_e u}{k}. \tag{47}$$

Similarly, for $\mu_1(x) = \mu_N(x)$, $\mu_2(x) = \mu_\Delta(x)$ and $c_\Delta = \Delta^{-3}$, we find

$$I_{\mu_{N\to\infty},\mu_\Delta} = \int_0^1 \left(\frac{4N_e u}{x} - \phi_{e,\Delta}(x)\right) x(1-x)dx = \frac{7}{3}N_e u\Delta. \tag{48}$$

By using Eq. (47) and Eq. (48) in Eq. (41) we obtain the bounds that are stated in Proposition 1. □

References

[1] M Kimura, *Solution of a process of random genetic drift with a continuous model*. PNAS, 41 (1955), 144150.

[2] S Lukic, J Hey and K Chen, *Non-equilibrium allele frequency spectra via spectral methods*. Theoretical Population Biology, **79**, 203-219 (2011).

[3] S Myers, C Fefferman and N Patterson, *Can one learn history from the allelic spectrum?*. Theoretical Population Biology **73**, 342-348 (2008).

[4] The Chimpanzee Sequencing and Analysis Consortium, *Initial sequence of the chimpanzee genome and comparison with the human genome*. Nature **437**, 69-87 (2005).

[5] R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, *Context dependence, ancestral misidentification, and spurious signatures of natural selection*. Mol. Biol. Evol. **24**, 1792-1800 (2007).

[6] R.N. Gutenkunst, R.D. Hernandez, S.H. Williamson and C.D. Bustamante, *Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data*, PLoS Genetics 5, (2009).

[1] School of Natural Sciences, Institute for Advanced Study, Princeton NJ 08540, USA.

[2] Department of Genetics, Rutgers University, Piscataway NJ 08854, USA

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.141846/-/DC1

Demographic Inference Using Spectral Methods on SNP Data, with an Analysis of the Human Out-of-Africa Expansion

Sergio Lukić and Jody Hey

DOI: 10.1534/genetics.112.141846

File S1
Supporting Material

1. Multi-population Wright-Fisher processes with no migration

In this section we compute the solution to the diffusion equations that describe the time evolution of the density of population allele frequencies under random drift, mutational influx and no migration between populations. First, we review the solution given by Kimura in [1] when the number of populations is $K = 1$. Second, we consider $K = 2$ populations. To this end we use the boundary conditions introduced in [2], solve the associated equations and finally, we show how this solution can be extended to an arbitrary number of populations K.

1.1. **One population.** When the number of populations is one, the density of population allele frequencies $\phi(x,t)$ satisfies the diffusion equation:

$$\frac{\partial \phi(x,t)}{\partial t} = \frac{1}{4N}\frac{\partial^2}{\partial x^2}\left[x(1-x)\phi(x,t)\right] + 2Nu\delta(x-1/2N), \tag{1}$$

where N is the effective population size of a diploid panmictic population, $\delta(x-1/2N)$ is the Dirac delta peaked at $x = 1/2N$, and $\phi(x,t)$ satisfies absorbing boundary conditions at $x = 0$ and $x = 1$. In more general scenarios we can use an effective mutation density $\mu(x)$ instead of the Dirac delta term, [2].

Kimura showed in [1] how Eq. (1) can be solved explicitly by expressing $\phi(x,t)$ as a polynomial expansion. In particular, he used the basis of Gegenbauer polynomials in which the diffusion operator can be expressed as an infinite diagonal matrix. The shifted Gegenbauer polynomials are a class of classical polynomials on the interval $[0,1]$ defined as

$$T_n(x) = \sqrt{\frac{(n+2)(2n+3)}{n+1}}P_n^{(1,1)}(2x-1), \qquad \int_0^1 T_n(x)T_m(x)x(1-x)dx = \delta_{nm} \tag{2}$$

where $P_n^{(1,1)}(z)$ are the classical Jacobi polynomials defined on the interval $-1 \leq z \leq 1$ with weight $w(z) = (1-z)(1+z)$. These polynomials satisfy the associated Jacobi equation:

$$\frac{\partial^2}{\partial x^2}\left[x(1-x)T_n(x)\right] = -(n+1)(n+2)T_n(x). \tag{3}$$

Date: July 27, 2012.

Thus, if we expand the density of population frequencies in this polynomial basis

$$\phi(x,t) = \sum_{n=0}^{\infty} a_n(t) T_n(x),$$

the diffusion equation in Eq. (1) can be written as

(4)
$$\sum_{n=0}^{\infty} \frac{da_n(t)}{dt} T_n(x) = -\sum_{n=0}^{\infty} \frac{(n+1)(n+2)}{4N} a_n(t) T_n(x) + 2Nu \sum_{n=0}^{\infty} T_n(1/2N) \frac{1-1/2N}{2N} T_n(x).$$

For simplicity and to shorten the notation, we denote as μ_n the contribution due to mutational influx $\mu_n = 2NuT_n(1/2N)\frac{1-1/2N}{2N}$. Using this notation, the Ordinary Differential Equation that obeys the coefficients $a_n(t)$ can be written as:

(5)
$$\frac{da_n(t)}{dt} = -\frac{(n+1)(n+2)}{4N} a_n(t) + \mu_n.$$

Eq. (5) is a linear differential equation of first order with an inhomogeneous term; this class of equations have a known simple solution which can be written as

(6)
$$a_n(t) = \left[a_n(0) - \frac{4N\mu_n}{(n+1)(n+2)} \right] \exp\left(-\frac{(n+1)(n+2)}{4N} t \right) + \frac{4N\mu_n}{(n+1)(n+2)}.$$

Here, $a_n(0)$ are the coefficients associated with the polynomial expansion of the initial density of population frequencies, which can be computed as

$$a_n(0) = \int_0^1 \phi(x,0) T_n(x) x(1-x) dx.$$

Therefore, given *any* density of population frequencies $\phi(x,0)$ at time $t=0$, we can compute the resulting density $\phi(x,t)$ after t generations evolving under random drift and mutational influx by means of the Gegenbauer expansion $\phi(x,t) = \sum_{n=0}^{\infty} a_n(t) T_n(x)$. The time-dependent coefficients $a_n(t)$ determined in Eq. (6), are a function of the coefficients at initial time and other population genetic parameters such as population size, mutation rate and time. Given the solution $\phi(x,t)$, the Allele Frequency Spectrum associated with a sample of C chromosomes is easily computed by introducing the binomial distribution with parameters C and x as:

$$f_i(t) = \frac{C!}{(C-i)!i!} \sum_{n=0}^{\infty} a_n(t) \int_0^1 x^i (1-x)^{C-i} T_n(x) dx, \quad 0 < i < C,$$

where f_i is the expected number of SNPs that have the derived state in exactly i chromosomes (out of a sample of C chromosomes). Properties of the Jacobi polynomials show that all terms of this sum vanish for $n > C-2$, thus the AFS can be computed exactly as the finite sum

(7)
$$f_i(t) = \frac{C!}{(C-i)!i!} \sum_{n=0}^{C-2} a_n(t) \int_0^1 x^i (1-x)^{C-i} T_n(x) dx.$$

This exact solution can be generalized to an arbitrary number of populations. In the next subsection we show how to compute the solution to the time-evolution of the density of allele frequencies when the number of populations is two.

1.2. **Two populations.** The diffusion equation that describes the dynamics of the density of allele frequencies in two isolated populations is a natural generalization of the one-population case studied above. In particular, if x_1 and x_2 are the derived allele frequencies in population 1 and 2, N_1 and N_2 are the effective population sizes of both populations and $\phi(x_1, x_2, t)$ is the joint density of population frequencies, $\phi(x_1, x_2, t)$ satisfies the following forward diffusion equation

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= \frac{1}{4N_1} \frac{\partial^2}{\partial x_1^2} \left[x_1(1-x_1)\phi \right] + 2N_1 u \delta(x_1 - 1/2N_1)\delta(x_2) \\ &+ \frac{1}{4N_2} \frac{\partial^2}{\partial x_2^2} \left[x_2(1-x_2)\phi \right] + 2N_2 u \delta(x_1)\delta(x_2 - 1/2N_2). \end{aligned} \tag{8}$$

As was shown in [2], the solution to Eq. (8) can be expressed as a generalized density with contributions from the different boundary components of the square $[0, 1] \times [0, 1]$:

$$\begin{aligned} \phi(x_1, x_2, t) = \phi^A(x_1, x_2, t) + \phi^B_{(x_2=0)}(x_1, t)\delta(x_2) + \\ \phi^B_{(x_2=1)}(x_1, t)\delta(1-x_2) + \phi^B_{(x_1=0)}(x_2, t)\delta(x_1) + \phi^B_{(x_1=1)}(x_2, t)\delta(1-x_1) + \\ \phi^C_{(x_1=1,x_2=0)}(t)\delta(1-x_1)\delta(x_2) + \phi^C_{(x_1=0,x_2=1)}(t)\delta(x_1)\delta(1-x_2). \end{aligned} \tag{9}$$

The terms that are multiplied by Dirac deltas represent the contributions to the density that are localized in the different boundary components. In particular, the A-term is localized in the bulk of the square, the four B-terms are localized in the edges of the square and finally, the two C-terms are localized in the two vertices of the square that are not absorbing. The Ancestral vertex $(x_1 = 0, x_2 = 0)$ and the Derived vertex $(x_1 = 1, x_2 = 1)$ are absorbing and hence do not contribute SNPs to the density $\phi(x_1, x_2, t)$.

As Eq. (8) is the natural extension of the one-population process and the one-population diffusion equation can be solved by means of polynomials expansions, we expand each term in Eq. (9) using the same basis of Jacobi polynomials $T_n(x)$ defined in Eq. (2). As we will see at the end of this section, such a polynomial expansion will allow us to find the exact solution of the two-population process. In particular, we write the polynomial expansion of each term in Eq. (9) as:

$$\begin{aligned} \phi^A(x_1, x_2, t) &= \sum_{n,m=0}^{\infty} a^A_{nm}(t) T_n(x_1) T_m(x_2), \\ \phi^B_{(x_2=0)}(x_1, t) &= \sum_{n=0}^{\infty} a^B_{(x_2=0),n}(t) T_n(x_1), \\ \phi^B_{(x_2=1)}(x_1, t) &= \sum_{n=0}^{\infty} a^B_{(x_2=1),n}(t) T_n(x_1), \\ \phi^B_{(x_1=0)}(x_2, t) &= \sum_{m=0}^{\infty} a^B_{(x_1=0),m}(t) T_m(x_2), \\ \phi^B_{(x_1=1)}(x_2, t) &= \sum_{m=0}^{\infty} a^B_{(x_1=1),m}(t) T_m(x_2), \\ \phi^C_{(x_1=1,x_2=0)}(t) &= a^C_{(x_1=1,x_2=0)}(t), \\ \phi^C_{(x_1=0,x_2=1)}(t) &= a^C_{(x_1=0,x_2=1)}(t). \end{aligned} \tag{10}$$

In this polynomial basis, Eq. (8) requires that the a-variables satisfy a set of Ordinary Differential Equations (ODE) that can be integrated exactly. The associated ODEs can be determined by taking into account the different contributions to the dynamics of the a-variables (random drift, influx of polymorphisms in the boundary components due to fixation events, and influx of polymorphisms due to mutations). Following [2] we know that the dynamics of the $a^A_{nm}(t)$-terms is just governed by random drift (there is no influx of polymorphisms). On the other hand, the dynamics of the terms $a^B_{(x_1=1),m}(t)$ and $a^B_{(x_2=1),n}(t)$ depend on both random drift and the influx of polymorphisms that reach fixation at either $x_1 = 1$ or $x_2 = 1$. The terms $a^B_{(x_2=0),n}(t)$ and $a^B_{(x_1=0),m}(t)$ furthermore receive the constant influx of polymorphisms due to de novo mutations at the population level. Finally, the time evolution of the terms $a^C_{(x_1=1,x_2=0)}(t)$ and $a^C_{(x_1=0,x_2=1)}(t)$ is described by the influx of polymorphisms that reach fixation from $\phi^B_{(x_2=0)}(x_1,t)$ and $\phi^B_{(x_1=1)}(x_2,t)$, in the case of $a^C_{(x_1=1,x_2=0)}(t)$, or from $\phi^B_{(x_1=0)}(x_2,t)$ and $\phi^B_{(x_2=1)}(x_1,t)$ in the case of $\phi^C_{(x_1=0,x_2=1)}(t)$.

The dynamics of the a-coefficients can be made quantitatively explicit in the following system of linear differential equations:

$$\frac{da^A_{nm}}{dt} = -\left(\frac{(n+1)(n+2)}{4N_1} + \frac{(m+1)(m+2)}{4N_2}\right) a^A_{nm}, \tag{11}$$

$$\frac{da^B_{(x_2=0),n}}{dt} = -\frac{(n+1)(n+2)}{4N_1} a^B_{(x_2=0),n} + \mu^1_n + \sum_{m=0}^{\infty} \frac{a^A_{nm} T_m(0)}{4N_2}, \tag{12}$$

here, $\mu^1_n = 2N_1 u \times T_n(1/2N_1)\frac{1-1/2N_1}{2N_1}$ is the contribution due to mutational influx in population 1,

$$\frac{da^B_{(x_1=0),m}}{dt} = -\frac{(m+1)(m+2)}{4N_2} a^B_{(x_1=0),m} + \mu^2_m + \sum_{n=0}^{\infty} \frac{a^A_{nm} T_n(0)}{4N_1}, \tag{13}$$

here, $\mu^2_m = 2N_2 u \times T_m(1/2N_2)\frac{1-1/2N_2}{2N_2}$ is the contribution due to mutational influx in population 2,

$$\frac{da^B_{(x_2=1),n}}{dt} = -\frac{(n+1)(n+2)}{4N_1} a^B_{(x_2=1),n} + \sum_{m=0}^{\infty} \frac{a^A_{nm} T_m(1)}{4N_2}, \tag{14}$$

$$\frac{da^B_{(x_1=1),m}}{dt} = -\frac{(m+1)(m+2)}{4N_2} a^B_{(x_1=1),m} + \sum_{n=0}^{\infty} \frac{a^A_{nm} T_n(1)}{4N_1}, \tag{15}$$

$$\frac{da^C_{(x_1=1,x_2=0)}}{dt} = \sum_{n=0}^{\infty} \frac{a^B_{(x_2=0),n} T_n(1)}{4N_1} + \sum_{m=0}^{\infty} \frac{a^B_{(x_1=1),m} T_m(0)}{4N_2}, \tag{16}$$

and

$$\frac{da^C_{(x_1=0,x_2=1)}}{dt} = \sum_{n=0}^{\infty} \frac{a^B_{(x_2=1),n} T_n(0)}{4N_1} + \sum_{m=0}^{\infty} \frac{a^B_{(x_1=0),m} T_m(1)}{4N_2}. \tag{17}$$

This system of coupled linear differential equations can be solved by integrating first the uncoupled equation Eq. (11), using the corresponding solution to solve Eqs. (12), (13), (14), and (15), and finally using those solutions to solve Eq. (16) and Eq. (17). At each step, one has to integrate a set of linear ODEs of first order whose solutions are known.

The solution of Eq. (11) is:

$$a^A_{nm}(t) = a^A_{nm}(0)\exp\left[-\left(\frac{(n+1)(n+2)}{4N_1} + \frac{(m+1)(m+2)}{4N_2}\right)t\right], \tag{18}$$

with $a^A_{nm}(0)$ the coefficients associated with $\phi^A(x_1, x_2, 0)$ at initial time:

$$a^A_{nm}(0) = \int_0^1 \int_0^1 \phi^A(x_1, x_2, 0) T_n(x_1) T_m(x_2) x_1(1-x_1) x_2(1-x_2) dx_1 dx_2.$$

Now, we can use the solution Eq. (18) to integrate Eqs. (12), (13), (14), and (15). Hence, we can write the solution of Eq. (12) as

$$a^B_{(x_2=0),n}(t) = b^B_{(x_2=0),n}\exp\left(-\frac{(n+1)(n+2)}{4N_1}t\right) + \frac{4N_1\mu^1_n}{(n+1)(n+2)} - \sum_{m=0}^{\infty}\frac{a^A_{nm}(t)T_m(0)}{(m+1)(m+2)}, \tag{19}$$

with $b^B_{(x_2=0),n}$ a time-independent function defined as

$$b^B_{(x_2=0),n} = a^B_{(x_2=0),n}(0) - \frac{4N_1\mu^1_n}{(n+1)(n+2)} + \sum_{m=0}^{\infty}\frac{a^A_{nm}(0)T_m(0)}{(m+1)(m+2)}.$$

The coefficients $a^B_{(x_2=0),n}(0)$ are associated with the initial-time density as

$$a^B_{(x_2=0),n}(0) = \int_0^1 \phi^B_{(x_2=0)}(x_1, 0) T_n(x_1) x_1(1-x_1) dx_1.$$

Similarly, the solution of (13) is

$$a^B_{(x_1=0),m}(t) = b^B_{(x_1=0),m}\exp\left(-\frac{(m+1)(m+2)}{4N_2}t\right) + \frac{4N_2\mu^2_m}{(m+1)(m+2)} - \sum_{n=0}^{\infty}\frac{a^A_{nm}(t)T_n(0)}{(n+1)(n+2)}, \tag{20}$$

with $b^B_{(x_1=0),m}$ defined as

$$b^B_{(x_1=0),m} = a^B_{(x_1=0),m}(0) - \frac{4N_2\mu^2_m}{(m+1)(m+2)} + \sum_{n=0}^{\infty}\frac{a^A_{nm}(0)T_n(0)}{(n+1)(n+2)}.$$

The solution of (14) is

$$a^B_{(x_2=1),n}(t) = b^B_{(x_2=1),n}\exp\left(-\frac{(n+1)(n+2)}{4N_1}t\right) - \sum_{m=0}^{\infty}\frac{a^A_{nm}(t)T_m(1)}{(m+1)(m+2)}, \tag{21}$$

with $b^B_{(x_2=1),n}$ defined as

$$b^B_{(x_2=1),n} = a^B_{(x_2=1),n}(0) + \sum_{m=0}^{\infty}\frac{a^A_{nm}(0)T_m(1)}{(m+1)(m+2)}.$$

And finally, for this class of solutions, the solution of (15) is

$$a^B_{(x_1=1),m}(t) = b^B_{(x_1=1),m}\exp\left(-\frac{(m+1)(m+2)}{4N_2}t\right) - \sum_{n=0}^{\infty}\frac{a^A_{nm}(t)T_n(1)}{(n+1)(n+2)}, \tag{22}$$

with $b^B_{(x_1=1),m}$ defined as

$$b^B_{(x_1=1),m} = a^B_{(x_1=1),m}(0) + \sum_{n=0}^{\infty}\frac{a^A_{nm}(0)T_n(1)}{(n+1)(n+2)}.$$

The solutions to Eqs. (16) and (17) are frequency-independent functions of time which can be obtained by integrating Eqs. (19), (20), (21), and (22):

(23)

$$\Delta a^{C}_{(x_1=1,x_2=0)}(t) = \sum_{n=0}^{\infty} \frac{T_n(1)}{4N_1} \int_0^t a^{B}_{(x_2=0),n}(u)du + \sum_{m=0}^{\infty} \frac{T_m(0)}{4N_2} \int_0^t a^{B}_{(x_1=1),m}(u)du,$$

and

(24)

$$\Delta a^{C}_{(x_1=0,x_2=1)}(t) = \sum_{n=0}^{\infty} \frac{T_n(0)}{4N_1} \int_0^t a^{B}_{(x_2=1),n}(u)du + \sum_{m=0}^{\infty} \frac{T_m(1)}{4N_2} \int_0^t a^{B}_{(x_1=0),m}(u)du,$$

where the Δa terms are defined as:

$$\Delta a^{C}_{(x_1=1,x_2=0)}(t) := a^{C}_{(x_1=1,x_2=0)}(t) - a^{C}_{(x_1=1,x_2=0)}(0),$$

and

$$\Delta a^{C}_{(x_1=0,x_2=1)}(t) := a^{C}_{(x_1=0,x_2=1)}(t) - a^{C}_{(x_1=0,x_2=1)}(0).$$

In summary, the solution of Eq. (8) can be written as a generalized density with seven components (as in Eq. (9)). Each of these seven boundary-specific densities can be expanded by means of a polynomial expansion (as in Eq. (10)). The time-dependent coefficients associated with these expansions were obtained in Eqs. (18)-(24).

Given an explicit solution $\phi(x_1, x_2, t)$, one can make connections with measurable quantities by computing the theoretical prediction of some of them. For instance, one can compute the Allele Frequency Spectrum associated with a sample of C chromosomes by introducing the binomial distribution as:

(25)

$$f_{ij}(t) = \frac{C!}{(C-i)!i!}\frac{C!}{(C-j)!j!} \int_0^1 \int_0^1 x_1^i (1-x_1)^{C-i} x_2^j (1-x_2)^{C-j} \phi(x_1, x_2, t) dx_1 dx_2,$$

for $0 \leq i \leq C, 0 \leq j \leq C$ and $0 < i+j < 2C$. Here, f_{ij} is the expected number of SNPs in which the derived state is found in i chromosomes in population one and j chromosomes in population two. In general, evaluating Eq. (25) requires integrating $\phi(x_1, x_2, t)$, which involves computing several infinite sums. However, this formula becomes particularly simple when $0 < i < C$ and $0 < j < C$:

$$f_{ij}(t) = \frac{C!}{(C-i)!i!}\frac{C!}{(C-j)!j!} \sum_{n,m=0}^{\infty} a^{A}_{nm}(t) \times$$

$$\int_0^1 \int_0^1 x_1^i (1-x_1)^{C-i} x_2^j (1-x_2)^{C-j} T_n(x_1) T_m(x_2) dx_1 dx_2,$$

and because of properties of the Jacobi polynomials this simplifies to the finite sum

$$f_{ij}(t) = \frac{C!}{(C-i)!i!}\frac{C!}{(C-j)!j!} \sum_{n,m=0}^{C-2} a^{A}_{nm}(t) \times$$

$$\int_0^1 \int_0^1 x_1^i (1-x_1)^{C-i} x_2^j (1-x_2)^{C-j} T_n(x_1) T_m(x_2) dx_1 dx_2.$$

This resembles the simple formula Eq. (7) derived in the one-population case. Hence, after including the contributions from every boundary component, the solution of the two-population diffusion equation describing the time evolution of the density of allele frequencies is a natural extension of the one-population solution. One can also generalize the two-population case studied here, to a scenario with an arbitrary number of populations.

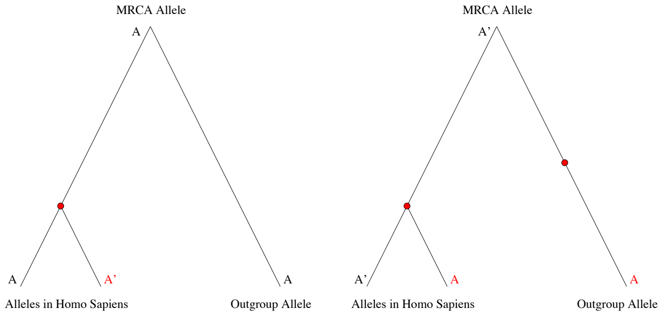

FIGURE 1. Most probable histories of a diallelic locus (with alleles A and A'). In red we denote the derived allele that arises as a mutation since the split with the most recent common ancestor. Here we assume that one of the alleles is identical to the orthologous base in an outgroup species that shares a recent common ancestor, such as Pan troglodytes or Rhesus macaque in the case of homo sapiens.

write the probability of mutation as

$$\begin{aligned} p(xAy \to xA'y|\tau) &= p(xAy \text{ has diverged since MRCA}|\tau) \\ &\times \frac{p(xAy \to xA'y|xAy \text{ has diverged since MRCA};\, \tau)}{p(xAy \text{ has diverged since MRCA}|xAy \to xA'y;\, \tau)} \end{aligned} \tag{26}$$

Here, x and y are the flanking nucleotides that define the context, and τ is the time of divergence between the species under consideration. Eq. (26) allows to estimate the mutation rates using genome wide data on the divergence between species. More explicitly, each term in (26) can be computed as:

$$p(xAy \text{ has diverged since MRCA}|\tau) = 64 \times r_{div} \times \pi_{xAy} \tag{27}$$

$$\begin{aligned} p(xAy \to xA'y|xAy \text{ has diverged since MRCA};\, \tau) &= \\ (\pi_{A;A',A} p_M(A|A,A') + \pi_{A';A',A}(1 - p_M(A'|A',A))) &\times \\ (\pi_{A;A',A} p_M(A|A,A') + \pi_{A';A',A}(1 - p_M(A'|A',A)) &+ \\ \pi_{A;B,A} p_M(A|A,B) + \pi_{B;B,A}(1 - p_M(B|B,A)) &+ \\ \pi_{A;B',A} p_M(A|A,B') + \pi_{B';B',A}(1 - p_M(B'|B',A)))^{-1} \end{aligned} \tag{28}$$

$$p(xAy \text{ has diverged since MRCA}|xAy \to xA'y;\, \tau) = 1.0 \tag{29}$$

In Eq. (27), r_{div} is the probability that two random homologous nucleotides are different, which is estimated to be $1.57/100$ between human and chimp. π_{xAy} is the genome-wide average frequency of trinucleotides xAy, and $64 = 4^3$ is a normalization constant. In Eq. (28), $\pi_{w;z,w}$ is the genome-wide frequency of trinucleotides xwy in the outgroup species whose orthologous has polymorphisms xwy and xzy in the species under consideration. The probability $p_M(w|w,z)$ is a shorthand for

$$p(xwy \,\text{is MRCA}|\text{Outgroup} = xwy, \text{Alleles} = xzy, xwy).$$

And finally, B and B' are the two nucleotides in g, t, a, c, which are not A nor A'; i.e. B and B' span the complementary set to A and A' in $\{g, t, a, c\}$. Therefore, all parameters that appear in Eq. (26) can be estimated using genomic and polymorphic data, except

$p(xAy \to xA'y|\tau)$ and $p_M(w|w,z)$. The probability functions $1 - p_M(w|w,z)$ are exactly the quantities that define the probability of ancestral allele misidentification using the outgroup base. Such probabilities also satisfy :

$$\begin{aligned} p_M(w|w,z) = p(xwy \to xzy|\tau)p(xwy \to xwy|\tau) \times \\ (p(xwy \to xzy|\tau)p(xwy \to xwy|\tau) + \\ p(xzy \to xwy|\tau)p(xzy \to xwy|\tau))^{-1}. \end{aligned} \tag{30}$$

Here, $p(xwy \to xwy|\tau)$ equals $1 - \sum_{z\in S} p(xwy \to xzy|\tau)$, with S the set $\{g,t,a,c\}\backslash w$. In other words, $p_M(w|w,z)$ is approximately equal to the probability that the history represented in the left tree of Fig. 1 actually happened, given that the left and right trees represent the most probable events.

Thus, by substituting Eq. (26) into Eq. (30), one gets a system of equations in the unknown variables $p_M(w|w,z)$, which can be solved easily.

We estimated the probabilities of ancestral allele misidentification in humans, using the chimp as the outgroup species. Using the human and chimp genomes, plus the EGP SNP data, we estimated all the parameters in Eqs. (27), (28) and (29). By starting with initial values $p^0_M(w|w,z) = 1$, one can solve Eq. (26) and recompute $p^1_M(w|w,z)$ using Eq. (30). This yields an iterative mechanism that produces a quickly convergent sequence of probabilities $p^n_M(w|w,z)$ towards a unique fixed point, solution of the system of equations. We found that the resulting probabilities $1 - p_M(w|w,z)$ can be broken down into CpG and non-CpG contexts. In the non-CpG context, i.e. mutations which are not of the type CG to TG nor CT to CA, all the probabilities $1 - p_M(w|w,z)$ are smaller than 0.006. However, for mutations of the type CG to TG or CT to CA, the probabilities $1 - p_M(w|w,z)$ range between a maximum of 0.16 and a minimum of 0.06. This result is very similar to the one given in [4].

3. Comparison of the different boundary conditions used in this study

The one-population two-allele Wright-Fisher diffusion with influx of mutations can be defined by means of the PDE

$$\frac{\partial \phi}{\partial t} = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}\left[x(1-x)\phi(x,t)\right] + 2Nu\delta(x - 1/2N). \tag{31}$$

Here, N_e denotes the effective population size, N is the census population size and u the mutation rate. The boundary conditions at $x = 0$ and $x = 1$ are absorbing, and the term $2Nu\delta(x - 1/2N)$ denotes the source of new mutations that arise at frequency $x = 1/2N$ for large N. It is very important to understand how to regularize $\delta(x - 1/2N)$ in any finite approximation that one applies to numerically solve Eq. (31). In particular, experience with different numerical solutions of Eq. (31) suggests that small changes in the finite regularization of the Dirac delta might have large effects on the numerical solution of Eq. (31).

In this section, we study the convergence properties of the finite-difference method used in [6] and the spectral method used in this paper for the particular case of Eq. (31). Although several sources of numerical error exist (e.g. either the truncated spectral expansion or the finite-difference approximation of $\phi(x,t)$), here we only consider the contribution to error due to the finite regularization of $\delta(x - 1/2N)$.

In particular, the finite regularizations of $\delta(x - 1/2N)$ that we consider here can be described using the diffusion equation

$$\frac{\partial \phi}{\partial t} = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}\left[x(1-x)\phi(x,t)\right] + u\mu(x), \tag{32}$$

with $\mu(x)$ a function of the frequency that depends on the particular choice of numerical method. The standard diffusion in Eq. (31) is recovered when $\mu(x) = 2N\delta(x - 1/2N)$. In general, we denote this function as

$$\mu_N(x) = 2N\delta(x - 1/2N). \tag{33}$$

In the case of the spectral method (see [2]) we instead use the function

$$\mu_k(x) = c_k \exp(-kx), \tag{34}$$

with

$$c_k = \frac{k^2}{1 - \exp(-k) - k\exp(-k)}.$$

Here, k is a positive real number that depends monotonically on the truncation parameter Λ. In particular, k is chosen such that the truncated polynomial approximation of Eq. (34) is accurate enough. Thus, the limit of large Λ corresponds with the limit of large k.

In the case of the finite-difference method, one approximates $\phi(x,t)$ as a piece-wise linear function. More precisely, if $\{x_j\}_{j=0}^{G}$ are the grid points on $[0, 1]$ that we use in the finite-difference scheme, we introduce a basis of functions $\{f_j(x)\}_{j=0}^{G}$ with

$$f_j(x) = \theta(x - x_{j-1})\theta(x_{j+1} - x)\left(\frac{x - x_{j-1}}{x_j - x_{j-1}}\theta(x_j - x) + \frac{x_{j+1} - x}{x_{j+1} - x_j}\theta(x - x_j)\right),$$

for $0 < j < G$,

$$f_0(x) = \theta(x_1 - x)\left(\frac{x_1 - x}{x_1 - x_0}\theta(x - x_0)\right),$$

and

$$f_G(x) = \theta(x - x_{G-1})\left(\frac{x - x_{G-1}}{x_G - x_{G-1}}\theta(x_G - x)\right),$$

such that the finite-difference approximation of $\phi(x,t)$ can be written as

$$\phi(x,t) \simeq \sum_{j=0}^{j=G} \phi_j^t f_j(x).$$

Here, $x_0 = 0$, $x_G = 1$ and $\theta(x)$ is the Heaviside step function (defined as $\theta(x) = 0$ for $x < 0$ and $\theta(x) = 1$ for $x > 0$). In Gutenkunst et al. [6] the authors use an adaptive grid on $[0, 1]$ that is uniform near $x = 0$. Therefore, for $f_0(x)$ and $f_1(x)$ we assume that $x_1 - x_0 = x_2 - x_1 = \Delta$, $x_0 = 0$, and the corresponding basis functions are

$$f_0(x) = \theta(\Delta - x)\left(\frac{\Delta - x}{\Delta}\theta(x)\right),$$

and

$$f_1(x) = \theta(2\Delta - x)\left(\frac{x}{\Delta}\theta(\Delta - x) + \frac{2\Delta - x}{\Delta}\theta(x - \Delta)\right).$$

Gutenkunst et al. [6] inject new mutations at each time-step by updating the value of ϕ_1^t as (see Eq. (S9) in [6])

$$\frac{\phi_1^{t+dt} - \phi_1^t}{dt} = \frac{u}{\Delta^2}. \tag{35}$$

Remark 1. *Note that Gutenkunst et al.* [6] *write Eq.* (32) *using different units. In particular, they introduce a reference population size N_0 with $\theta = 4N_0 u$ and write Eq.* (32) *as*

$$\frac{\partial \phi}{\partial \tau} = \frac{1}{2\nu} \frac{\partial^2}{\partial x^2} \left[x(1-x)\phi(x,\tau) \right] + \frac{\theta}{2}\mu(x), \tag{36}$$

with $\tau = t/2N_0$ and $\nu = N_e/N_0$. In their notation the value of ϕ_1^τ is updated as

$$\frac{\phi_1^{\tau+d\tau} - \phi_1^\tau}{d\tau} = \frac{\theta}{2\Delta^2}.$$

Updating the value of ϕ_1^t, as in Eq. (35), when solving the diffusion equations is equivalent to using Eq. (32) and the function

$$\mu_\Delta(x) = c_\Delta \Delta f_1(x) = c_\Delta [x\theta(\Delta - x) + (2\Delta - x)\theta(x - \Delta)\theta(2\Delta - x)], \tag{37}$$

with $c_\Delta = \Delta^{-3}$. Observe that $\theta(\Delta - x)\theta(2\Delta - x) = \theta(\Delta - x)$ and that $\theta(x)$ denotes here the Heaviside step function.

It is not obvious that $\mu_\Delta(x)$ in Eq. (37) or $\mu_k(x)$ in Eq. (34) converge to $\mu_N(x) = 2N\delta(x - 1/2N)$ in the limits $N \to \infty$, $k \to \infty$ and $\Delta \to 0$. Hence, it is not obvious that the solutions associated with each finite regularization converge to the exact solution of Eq. (31). However, in the remainder of this section we demonstrate how both approximate solutions actually converge to the exact solution.

Proposition 1. *Let $\phi_N(x,t)$, $\phi_k(x,t)$, and $\phi_\Delta(x,t)$ be the solutions of Eq.* (32) *corresponding to the functions $\mu(x)$ defined in Eq.* (33) *for $\phi_N(x,t)$, Eq.* (34) *for $\phi_k(x,t)$ and Eq.* (37) *for $\phi_\Delta(x,t)$. Additionally, let the initial condition be the same arbitrary density $\varphi(x)$ in all of the three cases:*

$$\phi_N(x,t=0) = \phi_k(x,t=0) = \phi_\Delta(x,t=0) = \varphi(x).$$

Then, iff c_k in Eq. (34) *is defined as*

$$c_k = \frac{k^2}{1 - \exp(-k) - k\exp(-k)},$$

$\phi_k(x,t)$ converges to the exact solution $\phi_N(x,t)$ in the limits $k \to \infty$, $N \to \infty$ and finite N_e. In particular,

$$\parallel \phi_{N\to\infty}(x,t) - \phi_k(x,t) \parallel_{L^1} \leq \frac{4N_e u}{k}(1 + \exp(-t/2N_e)), \quad t \geq 0.$$

Similarly, iff c_Δ in Eq. (37) *is defined as $c_\Delta = \Delta^{-3}$, $\phi_\Delta(x,t)$ converges to the exact solution $\phi_N(x,t)$ in the limits $\Delta \to 0$, $N \to \infty$ and finite N_e. In particular,*

$$\parallel \phi_{N\to\infty}(x,t) - \phi_\Delta(x,t) \parallel_{L^1} \leq \frac{7}{3} N_e u \Delta (1 + \exp(-t/2N_e)), \quad t \geq 0.$$

Proof. The proof consists of three parts. First, we describe the solution of Eq. (32) for an arbitrary choice of $\mu(x)$; second, we derive a general bound for the L^1-norm of the difference of two solutions associated with different choices of $\mu(x)$ (see Eq. (42)); and third, we apply this general argument to the particular cases of $\phi_N(x,t)$, $\phi_k(x,t)$, and $\phi_\Delta(x,t)$ and the L^1-norms

$$\parallel \phi_{N\to\infty}(x,t) - \phi_k(x,t) \parallel_{L^1} = \int_0^1 |\phi_{N\to\infty}(x,t) - \phi_k(x,t)| x(1-x) dx,$$

and

$$\parallel \phi_{N\to\infty}(x,t) - \phi_{\Delta}(x,t) \parallel_{L^1} = \int_0^1 |\phi_{N\to\infty}(x,t) - \phi_{\Delta}(x,t)| x(1-x)dx.$$

Any solution of Eq. (32) can be described as the sum of a homogeneous solution and an inhomogeneous solution. In particular, if $\phi_{e,\mu}(x)$ is the steady state solution that satisfies

$$0 = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}\left[x(1-x)\phi_{e,\mu}(x)\right] + u\mu(x), \tag{38}$$

$\phi_{\mu}(x,t=0) = \varphi(x)$ is the initial condition, and $\gamma(x,t) = \exp(tL_{FP})\gamma(x,0)$ is the solution to the homogenous ($\mu(x) = 0$) problem

$$\frac{\partial \gamma(x,t)}{\partial t} = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}\left[x(1-x)\gamma(x,t)\right],$$

then one can write the solution of Eq. (32) as

$$\phi_{\mu}(x,t) = \exp(tL_{FP})(\varphi(x) - \phi_{e,\mu}(x)) + \phi_{e,\mu}(x).$$

Here, $\exp(tL_{FP})$ denotes the time evolution operator, and L_{FP} denotes the Fokker-Planck diffusion operator. Therefore, if $\phi_{\mu_1}(x,t)$ and $\phi_{\mu_2}(x,t)$ are solutions of Eq. (32) associated with the functions $\mu_1(x)$ and $\mu_2(x)$, the difference $\phi_{\mu_1} - \phi_{\mu_2}$ satisfies

$$\phi_{\mu_1}(x,t) - \phi_{\mu_2}(x,t) = \exp(tL_{FP})(\phi_{e,\mu_2}(x) - \phi_{e,\mu_1}(x)) + \phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x).$$

In order to bound $\parallel \phi_{\mu_1} - \phi_{\mu_2} \parallel_{L^1}$ we apply the Minkowski inequality as follows:

$$\parallel \phi_{\mu_1}(x,t) - \phi_{\mu_2}(x,t) \parallel_{L^1} = \parallel \exp(tL_{FP})(\phi_{e,\mu_2}(x) - \phi_{e,\mu_1}(x)) + \phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x) \parallel_{L^1} \leq$$

$$\parallel \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x)) \parallel_{L^1} + \parallel \phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x) \parallel_{L^1}. \tag{39}$$

In our particular case (in which $\mu_1(x) = \mu_{N\to\infty}(x)$ and $\mu_2(x) = \mu_k(x)$ or $\mu_2(x) = \mu_{\Delta}(x)$), $\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x)$ is non-negative for all $x \in (0,\ 1)$. As the time-evolution operator $\exp(tL_{FP})$ preserves the non-negativity of the density, we can write Eq. (39) as

$$\parallel \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x)) \parallel_{L^1} + \parallel \phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x) \parallel_{L^1} =$$

$$\int_0^1 \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))x(1-x)dx + \int_0^1 (\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))x(1-x)dx.$$

The operator $\exp(tL_{FP})$ is diagonal in the basis spanned by the Gegenbauer polynomials (see Eq. (2)). In particular, we can write $\exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))$ as

$$\exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x)) = \sum_{n=0}^{\infty} a_n \exp(-t(n+1)(n+2)/4N_e)T_n(x),$$

with

$$a_n = \int_0^1 T_n(x)(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))x(1-x)dx.$$

Now, using this expansion and the fact that $T_0(x) = \sqrt{6}$, we can write

$$\int_0^1 \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))x(1-x)dx =$$

$$\frac{1}{\sqrt{6}}\int_0^1 \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))T_0(x)x(1-x)dx =$$

$$\frac{1}{\sqrt{6}}\sum_{n=0}^{\infty} a_n \exp(-t(n+1)(n+2)/4N_e)\int_0^1 T_n(x)T_0(x)x(1-x)dx = \frac{a_0}{\sqrt{6}}\exp(-t/2N_e).$$

Therefore, if we define I_{μ_1,μ_2} as

$$I_{\mu_1,\mu_2} = \int_0^1 (\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x))x(1-x)dx, \tag{40}$$

then $a_0 = \sqrt{6} I_{\mu_1,\mu_2}$ and the sum of L^1-norms is

$$\begin{gathered} \parallel \exp(tL_{FP})(\phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x)) \parallel_{L^1} + \parallel \phi_{e,\mu_1}(x) - \phi_{e,\mu_2}(x) \parallel_{L^1} = \\ I_{\mu_1,\mu_2}(1 + \exp(-t/2N_e)). \end{gathered} \tag{41}$$

Now, from Eq. (39) it follows that

$$\parallel \phi_{\mu_1}(x,t) - \phi_{\mu_2}(x,t) \parallel_{L^1} \leq I_{\mu_1,\mu_2}(1 + \exp(-t/2N_e)). \tag{42}$$

In order to determine the bound in Eq. (42) one needs only to evaluate the integral in Eq. (40). This requires solving Eq. (38) to obtain a closed-form expression for $\phi_{e,\mu_1}(x)$ and $\phi_{e,\mu_2}(x)$. One can solve Eq. (38) simply by integrating the equation twice

$$\int_0^x \int_0^y \frac{d^2\psi(z)}{dz^2} dz dy = -4N_e u \int_0^x \int_0^y \mu(z) dz dy,$$

$$\psi(x) = \psi(0) + \psi'(0)x - 4N_e u \int_0^x \int_0^y \mu(z) dz dy,$$

with $\psi(x) = x(1-x)\phi_{e,\mu}(x)$ and $\psi'(x) = d\psi/dx$. We require $\phi_{e,\mu}(x)$ to be finite at the boundaries $x = 0$ and $x = 1$, i.e. $\psi(0) = \psi(1) = 0$. Therefore, for the particular functions $\mu(x)$ that we consider here (Eq. (33), Eq. (34) and Eq. (37)) we find the following solutions of Eq. (38):

$$\phi_{e,N}(x) = \frac{4N_e u}{x(1-x)} \left[(2N-1)x - 2N(x - 1/2N)\theta(x - 1/2N)\right], \tag{43}$$

$$\phi_{e,k}(x) = \frac{4N_e u}{x(1-x)} \frac{c_k}{k^2} \left[x(\exp(-k) - 1) - \exp(-kx) + 1\right]. \tag{44}$$

and

$$\begin{aligned} \phi_{e,\Delta}(x) \quad = \quad & \frac{4N_e u}{x(1-x)} c_\Delta \, \Delta^3 \Bigg[(\Delta^{-1} - 1)x - \frac{x^3}{6\Delta^3}\theta(\Delta - x) + \\ & \left(\frac{x^3}{6\Delta^3} - \frac{x^2}{\Delta^2} + \frac{x}{\Delta} - \frac{1}{3}\right)\theta(x - \Delta)\theta(2\Delta - x) + (1 - \Delta^{-1}x)\theta(x - 2\Delta) \Bigg]. \end{aligned} \tag{45}$$

Note that Eq. (43) yields $\phi_{e,N}(x) = 4N_e u/x$ for $x > 1/2N$. Thus, the limit $N \to \infty$ of Eq. (43) corresponds with $\phi_{e,N}(x) = 4N_e u/x$ for $0 < x \leq 1$. Note also that only if $c_k = k^2/(1 - \exp(-k) - k\exp(-k))$ then $\phi_{e,k}(x)$ converges to $4N_e u/x$ near $x = 1$. Similarly, only if $c_\Delta = \Delta^{-3}$ then $\phi_{e,\Delta}(x)$ converges to $4N_e u/x$ near $x = 1$.

Now we can evaluate the integral in Eq. (40) for $\mu_1(x) = \mu_N(x)$, $\mu_2(x) = \mu_k(x)$ and $c_k = k^2/(1 - \exp(-k) - k\exp(-k))$, as

$$\int_0^1 \left(\frac{4N_e u}{x} - \phi_{e,k}(x)\right) x(1-x)dx = 4N_e u \frac{1 + k/2 + (1 - \exp(k))/k}{1 + k - \exp(k)}, \tag{46}$$

which in the limit of large k converges to

$$I_{\mu_{N\to\infty},\mu_k} = \frac{4N_e u}{k}. \tag{47}$$

Similarly, for $\mu_1(x) = \mu_N(x)$, $\mu_2(x) = \mu_\Delta(x)$ and $c_\Delta = \Delta^{-3}$, we find

$$I_{\mu_{N\to\infty},\mu_\Delta} = \int_0^1 \left(\frac{4N_e u}{x} - \phi_{e,\Delta}(x)\right) x(1-x)dx = \frac{7}{3}N_e u\Delta. \tag{48}$$

By using Eq. (47) and Eq. (48) in Eq. (41) we obtain the bounds that are stated in Proposition 1. □

References

[1] M Kimura, *Solution of a process of random genetic drift with a continuous model*. PNAS, 41 (1955), 144150.

[2] S Lukic, J Hey and K Chen, *Non-equilibrium allele frequency spectra via spectral methods*. Theoretical Population Biology, **79**, 203-219 (2011).

[3] S Myers, C Fefferman and N Patterson, *Can one learn history from the allelic spectrum?*. Theoretical Population Biology **73**, 342-348 (2008).

[4] The Chimpanzee Sequencing and Analysis Consortium, *Initial sequence of the chimpanzee genome and comparison with the human genome*. Nature **437**, 69-87 (2005).

[5] R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, *Context dependence, ancestral misidentification, and spurious signatures of natural selection*. Mol. Biol. Evol. **24**, 1792-1800 (2007).

[6] R.N. Gutenkunst, R.D. Hernandez, S.H. Williamson and C.D. Bustamante, *Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data*, PLoS Genetics 5, (2009).

[1] School of Natural Sciences, Institute for Advanced Study, Princeton NJ 08540, USA.

[2] Department of Genetics, Rutgers University, Piscataway NJ 08854, USA