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 $-1 \le z \le 1$ 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:

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

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

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

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.

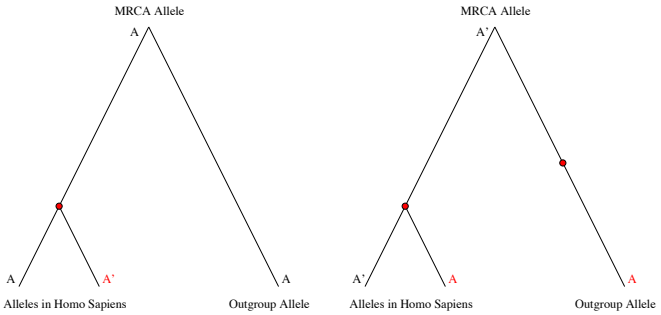

FIGURE 1. Most probable histories of a diallelic locus (with alleles A and A'). In red we denote the derived allele that arises as a mutation since the split with the most recent common ancestor. Here we assume that one of the alleles is identical to the orthologous base in an outgroup species that shares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

write the probability of mutation as

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

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

$p(xAy \rightarrow xA'y|\tau)$ 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 :

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

Here, $p(xwy \rightarrow xwy|\tau)$ equals $1 - \sum_{z \in S} p(xwy \rightarrow xzy|\tau)$, 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 $p_M^0(w|w,z) = 1$, one can solve Eq. (26) and recompute $p_M^1(w|w,z)$ using Eq. (30). This yields an iterative mechanism that produces a quickly convergent sequence of probabilities $p_M^n(w|w,z)$ 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* $\theta = 4N_0u$ *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,*

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

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,*

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

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

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

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

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

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