Available online at www.sciencedirect.com
ScienceDirect
Theoretical Population Biology 71 (2007) 502–523
Theoretical Population Biology
www.elsevier.com/locate/tpb

The probability distribution under a population divergence model of the number of genetic founding lineages of a population or species

Mattias Jakobsson*, Noah A. Rosenberg

Bioinformatics Program, Department of Human Genetics, University of Michigan, 2017 Palmer Commons, 100 Washtenaw Ave, Ann Arbor, MI 48109-2218, USA

Received 9 August 2006
Available online 14 January 2007

Abstract

The composition of genetic variation in a population or species is shaped by the number of events that led to the founding of the group. We consider a neutral coalescent model of two populations, where a derived population is founded as an offshoot of an ancestral population. For a given locus, using both recursive and nonrecursive approaches, we compute the probability distribution of the number of genetic founding lineages that have given rise to the derived population. This number of genetic founding lineages is defined as the number of ancestral individuals that contributed at the locus to the present-day derived population, and is formulated in terms of interspecific coalescence events. The effects of sample size and divergence time on the probability distribution of the number of founding lineages are studied in detail. For 99.99% of the loci in the derived population to each have one founding lineage, the two populations must be separated for $\geqslant 9.9N$ generations. However, only $\sim 0.87N$ generations must pass since divergence for 99.99% of the loci to have <6 founding lineages. Our results are useful as a prior expectation on the number of founding lineages in scenarios that involve the evolution of one population from the splitting of an ancestral group, such as in the colonization of islands, the formation of polyploid species, and the domestication of crops and livestock from wild ancestors.

Keywords: Coalescence; Founding events; Lineage sorting; Ancestral population; Speciation

1. Introduction

The composition of genetic variation in a population that recently diverged from an ancestral population is shaped by the nature of its founders. That a few founding individuals of a new offshoot population can strongly influence the properties of genetic variation of the population has been known for a long time (Wright, 1931; Nei et al., 1975; Hedrick, 2000). Founder effects are expected to cause an offshoot population to have fewer rare alleles in comparison with its ancestral population, as well as decreased genetic variation. After a founder event, the new population may diverge genetically. For example, colonization of the small Galapagos island Daphne Major in 1982 by five large ground finches caused this newly established population to diverge, both morphologically and genetically, from its ancestral population (Grant et al., 2001). Another example comes from the remote south Atlantic island of Tristan de Cunha, which was settled by humans in the early 1800s. Genealogical records show that 15 settler women contributed to a population of ~300 individuals at the time that the settlement was evacuated. However, only five founding mitochondrial DNA lineages remained in a sample of 161 individuals taken from the modern population (Soodyall et al., 1997).

Several studies have attempted to estimate the number of founding individuals in diverse evolutionary contexts. For example, investigations of the initial sizes of populations of domesticated crops and livestock have contributed to the understanding of human agricultural expansion (Luikart et al., 2001; Matsuoka et al., 2002; Harter et al., 2004; Wright et al., 2005). The sizes of founding populations have also been studied in humans themselves; for example, Hey (2005) estimated that the effective size of the founding population of the Americas may have been fewer than 80 individuals.

*Corresponding author. Fax: +1 734 615 6553.
E-mail address: mjakob@umich.edu (M. Jakobsson).

0040-5809/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2007.01.004

An alternative to studying the number of founding individuals of a population or species is to study the number of ancestral genetic lineages that have contributed to the group, since the actual number of founding individuals of a present-day population may be more difficult to infer. The number of founding lineages that *contributed* to the present-day population has a natural interpretation in the coalescent framework, and provides an informative quantity that can be studied in a reasonably straightforward manner. For example, the number of founding lineages has been studied in polyploid species, which are founded by discrete polyploidization events. For an (at least partly) outbreeding polyploid species, each polyploidization event can be expected to have contributed to a fraction of the genome of the contemporary polyploid species. In this context, the number of founding lineages that contributed to the species can be considered to be the number of distinct origins that have contributed to the present-day polyploid species. This fact has been used to infer from molecular data the number of origins in several polyploid species (e.g. Segraves et al., 1999; Ainouche et al., 2004; Evans et al., 2005; Jakobsson et al., 2006). Another example of founding lineages is the identification of five main mtDNA haplogroups—A, B, C, D and X—among Native Americans (Schurr et al., 1990; Brown et al., 1998; Smith et al., 1999), a result that has been used to examine alternative scenarios for the peopling of the Americas (Bonatto and Salzano, 1997; Crawford, 1998; Mulligan et al., 2004; Schurr and Sherry, 2004).

Consider two populations A and B that have been separated for some length of time. Population B is considered to be ancestral, and population A was founded as an offshoot from population B. For a particular genomic region, a lineage of population A may coalesce with other lineages from population A starting at the present and going backwards in time. When the separation time is passed, lineages of population A may also coalesce with lineages of population B. We are interested in the number of times that a lineage from population A coalesces with a lineage from population B—in other words, the number of founding lineages of population A that contributed to the present population A. This quantity is related to other aspects of interspecific coalescence that have previously been investigated (e.g. Takahata, 1989; Rosenberg, 2002); however, previous studies have not focused on the *number* of interspecific coalescences.

In this article, using a coalescent model of population divergence, we derive the probability distribution of the number of founding lineages of a population or species. We solve this problem by generalizing results in Rosenberg (2003) on the probability distribution of genealogical shape for two species. Because the exact expression becomes cumbersome to evaluate for large sample sizes, a recursion expression is also given to simplify the calculations. We then explore the effects of sample size, divergence time and differing population sizes on the number of founding lineages. We also consider implications of our results for samples of lineages at a single locus for the whole population, and implications for the number of founding lineages across the whole genome. The results are useful for describing under a simple null model the properties of the number of founding lineages that contributed to a population or a species.

2. Theory

2.1. The basic model

Consider two populations (or species) A and B that have been separated for T coalescent units of time (Fig. 1). Population A is founded by population B in such a way that an ancestral lineage of an A lineage and a B lineage is defined to be a B lineage. In other words, when an A lineage and a B lineage coalesce, the single ancestral lineage is defined as a B lineage. The ancestral lineage of two A lineages is an A lineage and the ancestral lineage of two B lineages is a B lineage. The numbers of sampled A and B lineages are denoted r_A and r_B. The numbers of A and B lineages ancestral at time T (backwards in time) to the samples of r_A and r_B lineages are denoted q_A and q_B, respectively. We define the number of genetic founding lineages of a sample from population A as the number of coalescences that involve an A lineage and a B lineage. The random number of such interpopulation coalescence events is denoted K. The probability of k interpopulation coalescences given r_A and r_B is then $P(K=k|r_A,r_B,T)$, or for convenience, $P(k|r_A,r_B,T)$. The sizes of populations A and B (total number of haploid lineages) are denoted N_A and N_B. The numbers of lineages that are ancestral to the full populations of N_A and N_B lineages at time T are denoted s_A and s_B. The random number of founding lineages that contributed to the whole present-day A population (not just the sample of r_A lineages) is denoted L. The probability of l interpopulation coalescence events given s_A and s_B is then $P(l|s_A,s_B,T)$. In most practical scenarios, the investigator will be interested in the number of founding lineages of the whole population, but will only have access to samples from this population. We are therefore interested in both probability distributions $P(l|s_A,s_B,T)$ and $P(k|r_A,r_B,T)$. We henceforth use the term *founding lineages* when referring to the founding individuals that contributed to present-day lineages.

Time, measured in generations, t, is rescaled to coalescent time T by dividing by the haploid effective population size, so that $T=t/N_e$. We make use of the standard approximation to finite populations of coalescent results that are based on the assumption of $N_e\to\infty$ (Nordborg, 2001). In our model we assume N_A and N_B are finite with $N_B/N_A=\beta$. If we assume that the generation times in populations A and B are equal, then $T_A/T_B=N_B/N_A=\beta$. We begin by assuming that $\beta=1$. Cases with $\beta\neq1$ will be considered after we have developed the general theory.

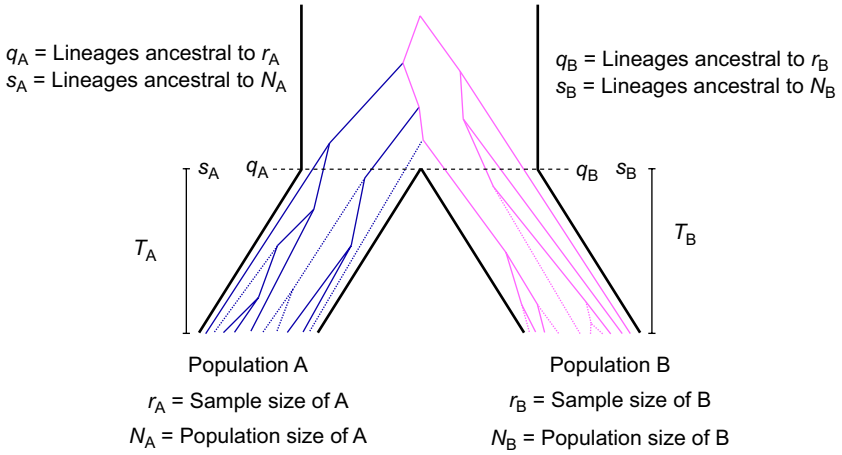

Fig. 1. A model of two populations (A and B) that diverged from an ancestral population. The (effective) population sizes are N_A and N_B for populations A and B. The sample sizes are denoted r_A and r_B. The numbers of lineages ancestral to the sample at divergence are denoted q_A and q_B, and the numbers of lineages ancestral to the whole population are denoted s_A and s_B. Time is measured in coalescent units for each population, T_A and T_B. Sampled lineages (and lineages ancestral to these) are shown as solid lines. In this example, $r_A = 6$, $r_B = 5$, $N_A = 10$, $N_B = 10$, $q_A = 3$, $q_B = 4$, $s_A = 4$ and $s_B = 4$.

We are here considering both recently separated species as well as recently diverged populations that do not exchange individuals after divergence. We henceforth make no distinction between these cases, and we use the term "population" throughout this article.

2.2. Labeled histories

Under the coalescent model all pairs of lineages have equal probabilities of coalescing, and the particular lineages that coalesce do so independently of when the coalescence event occurs. For n lineages, the coalescent process randomly joins pairs of lineages until only one lineage remains—the most recent common ancestor (MRCA). The number of ways to coalesce n lineages is equivalent to the number of labeled histories of n taxa, and can be obtained by considering the number of possible choices for each coalescence event. There are ${}_nC_2$ (n choose 2) possible pairs of lineages at the first coalescence event, ${}_{n-1}C_2$ pairs at the second coalescence, and so on. For n lineages, the total number H_n of coalescence sequences, or labeled histories, is (Edwards, 1970)

$$H_n = \binom{n}{2}\binom{n-1}{2}\cdots\binom{3}{2}\binom{2}{2} = \frac{n!(n-1)!}{2^{n-1}}. \tag{1}$$

The total number of "k-truncated labeled histories" (Rosenberg, 2006) for n lineages coalescing to k lineages ($k \leqslant n$) is

$$I_{n,k} = \binom{n}{2}\binom{n-1}{2}\cdots\binom{k+2}{2}\binom{k+1}{2} = \frac{n!(n-1)!}{2^{n-k}k!(k-1)!}. \tag{2}$$

Note that $I_{n,k}H_k = H_n$. More generally, if $a_1 \geqslant a_2 \geqslant \cdots \geqslant a_n$, then

$$I_{a_1,a_2}I_{a_2,a_3}\cdots I_{a_n,1} = I_{a_1,1} = H_{a_1}. \tag{3}$$

We also need two identities, which follow directly from Eqs. (1) and (2):

$$nI_{n,k} = \frac{2}{n+1}I_{n+1,k}, \tag{4}$$

$$I_{n,k}H_{k-1} = \frac{2}{k(k-1)}H_n. \tag{5}$$

Using Eq. (4) we have ($n_i > 1$)

$$I_{a,n_1}\left(\prod_{i=1}^{k-1} I_{n_i-1,n_{i+1}}\right)H_{n_k-1} = \frac{2^k H_a}{\prod_{i=1}^{k} n_i(n_i-1)}. \tag{6}$$

Consider two sequences of coalescence events, or nodes of the coalescent tree. Suppose the first sequence has s_1 nodes and the second sequence of coalescences has s_2 nodes (Fig. 2). There are then ${}_{s_1+s_2}C_{s_1}$ ways to order these s_1 and s_2 nodes, keeping intact the order of events within each of the two sequences. The number of ways that s_1 and s_2 nodes can be ordered is denoted

$$W_2(s_1,s_2) = \binom{s_1+s_2}{s_1}. \tag{7}$$

In our model, populations A and B have been separated for some time T, and they behave according to the standard coalescent model from the present back to T. The probability that n lineages have j ancestors T units of coalescent time in the past is given by (Tavaré, 1984, Eq. (6.1))

$$g_{n,j}(T) = \sum_{k=j}^{n} e^{-k(k-1)T/2}\frac{(2k-1)(-1)^{k-j}j_{(k-1)}n_{[k]}}{j!(k-j)!n_{(k)}}, \tag{8}$$

where $a_{(k)} = a(a+1)\cdots(a+k-1)$ and $a_{[k]} = a(a-1)\cdots(a-k+1)$ for $k \geqslant 1$, with $a_{(0)} = a_{[0]} = 1$. Except when

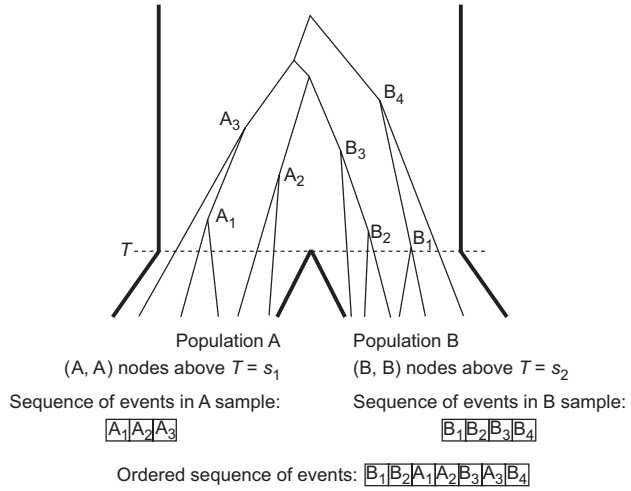

Fig. 2. Above time T and more recently than the first coalescence event of an A lineage and a B lineage, there is a sequence of s_1 coalescence events of A lineages, and a sequence of s_2 coalescence events of B lineages. For these two sequences of events, there are ${}_{s_1+s_2}C_{s_1}$ possible ways to order them. In this example, $s_1 = 3$ and $s_2 = 4$. There are then ${}_{3+4}C_3 = 35$ ways to order the s_1 and s_2 nodes.

$1 \leqslant j \leqslant n$, $g_{n,j}(T) = 0$. If $n \to \infty$, we have (Tavaré, 1984, Eq. (6.3)):

$$g_{\infty,j}(T) = \sum_{k=j}^{\infty} e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)}}{j!(k-j)!}. \tag{9}$$

For $T \geqslant 0.1$, $g_{n,j}(T)$ becomes negligible for large n and j. In our numerical computations, similarly to Rosenberg (2003), we assume $g_{n,j}(T) = 0$ if $j \geqslant 50$, $n \geqslant 90$ and $T \geqslant 0.1$.

2.3. The probability of one founding lineage given a sample of one A lineage

First, we consider the history of the populations that is above the divergence time T (see Fig. 1), that is, the case of $T = 0$ ($r_A = q_A$ and $r_B = q_B$). We will then add the population history below time T to describe the full model in Fig. 1. To compute the probability distribution of the number of founding lineages conditional on q_A and q_B, we employ a general approach of counting labeled histories that satisfy a required condition, and dividing by the total number of labeled histories for $q_A + q_B$ lineages, $H_{q_A+q_B}$. This approach can be used because each of the labeled histories is equally likely to occur.

To illustrate the approach, we begin with the simplest case, the probability of one founding lineage given a sample of one A lineage, $P(K = 1|q_A = 1, q_B)$. We start by counting the number of labeled histories that lead to one founding lineage (given $q_A = 1$). This case is trivial because if only one A lineage is sampled then there must be exactly one (A,B) event. This can be verified easily using the strategy that we will use to derive the main results of this article. Let m_1 be the number of ancestral B lineages at the time of the (first) coalescence between an A lineage and a B lineage. Let E_{m_1} be the event that there were exactly m_1 ($m_1 \geqslant 1$) ancestral B lineages when the first interpopulation coalescence occurs. If $K = 1$ then there are I_{q_B,m_1} ways for q_B B lineages to coalesce to m_1 lineages before the (A,B) event. The single A lineage may then coalesce with any of the m_1 remaining B lineages. There are H_{m_1} ways to coalesce the m_1 remaining B lineages. Thus,

$$P(E_{m_1}) = \frac{I_{q_B,m_1} m_1 H_{m_1}}{H_{q_A+q_B}} = \frac{H_{q_B} m_1}{H_{q_A+q_B}}. \tag{10}$$

If we sum over all possible values of m_1, then

$$\begin{aligned} P(K = 1|q_A = 1, q_B) &= \sum_{m_1=1}^{q_B} P(E_{m_1}) = \frac{H_{q_B}}{H_{q_A+q_B}} \sum_{m_1=1}^{q_B} m_1 \\ &= \frac{2}{(q_B+1)q_B} \frac{(q_B+1)q_B}{2} = 1. \end{aligned} \tag{11}$$

2.4. The probability of one founding lineage

A more general case, the probability of one founding lineage, $P(K = 1|q_A, q_B)$, was studied by Rosenberg (2003) in the context of the probability of monophyly of the lineages of a species. As we will see, this previously studied result is a special case of the general calculation described in this paper. We first note that to have $K = 1$, all q_A A lineages must coalesce to one lineage before coalescing with any B lineage. In other words the q_A A lineages must be monophyletic. There are H_{q_A} ways to coalesce q_A lineages to one lineage. There are I_{q_B,m_1} ways to coalesce q_B lineages to m_1 lineages. There are $W_2(q_A - 1, q_B - m_1)$ ways to order the $q_A - 1$ and $q_B - m_1$ coalescence events. The single remaining A lineage may now coalesce with any of the m_1 remaining B lineages. For the m_1 remaining B lineages there are H_{m_1} ways to coalesce. This leads to (this corrects an error on the first line of Eq. (10) in Rosenberg, 2003)

$$\begin{aligned} P(E_{m_1}) &= \frac{H_{q_A} I_{q_B,m_1} W_2(q_A - 1, q_B - m_1) m_1 H_{m_1}}{H_{q_A+q_B}} \\ &= \frac{2m_1 \binom{q_B}{m_1}}{\binom{q_A+q_B}{q_A} \binom{q_A+q_B-1}{m_1} q_B}. \end{aligned} \tag{12}$$

We retrieve the result of Rosenberg (2003, Eq. (11)) for the probability of one founding lineage:

$$P(K = 1|q_A, q_B) = \sum_{m_1=1}^{q_B} P(E_{m_1}) = \frac{2(q_A + q_B)}{\binom{q_A+q_B}{q_A} q_A(q_A+1)}. \tag{13}$$

2.5. The probability of two founding lineages

For the probability of two founding lineages, $P(K=2|q_A,q_B)$, m_1 is again defined as the number of B lineages at the time that the first (A,B) event occurs. We define m_2 as the number of B lineages at the second (A,B) event (going backwards in time). By definition, $m_2 \leqslant m_1$. We also define the number of A lineages at the first (A,B) event as n_1. Note that if $q_A<2$, then $P(K=2|q_A,q_B)=0$. In order that at least one A lineage remains available to participate in the second (A,B) event, n_1 must be at least 2. Thus, we assume here that $q_A \geqslant 2$. There are I_{q_B,m_1} ways to coalesce q_B B lineages to m_1 lineages. There are I_{q_A,n_1} ways to coalesce q_A A lineages to n_1 lineages. There are $W_2(q_A-n_1,q_B-m_1)$ ways to order the nodes up until the first (A,B) event, and n_1m_1 ways to choose this event from n_1 A lineages and m_1 B lineages. After the first (A,B) event, there are H_{n_1-1} ways to coalesce the remaining n_1-1 A lineages and I_{m_1,m_2} ways to coalesce m_1 B lineages to m_2 lineages. There are then n_1-2 nodes for n_1-1 A lineages coalescing to one lineage, and m_1-m_2 nodes for m_1 B lineages coalescing down to m_2 lineages. These nodes can be ordered in $W_2(n_1-2,m_1-m_2)$ ways. The single remaining A lineage has m_2 choices of B lineages with which to coalesce. Finally, there are H_{m_2} ways for the remaining m_2 B lineages to coalesce. Let E_{n_1,m_1,m_2} be the event that there were exactly n_1 ancestral A lineages and m_1 ancestral B lineages at the first (A,B) event, and exactly m_2 ancestral B lineages at the second (A,B) event. Using (3) and (4) to simplify, we have

$$\begin{aligned}P(E_{m_1,m_2;n_1}) &= [I_{q_A,n_1}I_{q_B,m_1}W_2(q_A-n_1,q_B-m_1)m_1n_1\\ &\quad\times H_{n_1-1}I_{m_1,m_2}W_2(n_1-2,m_1-m_2)m_2H_{m_2}]/H_{q_A+q_B}\\ &= \frac{2H_{q_A}H_{q_B}}{H_{q_A+q_B}(n_1-1)}m_1W_2(q_A-n_1,q_B-m_1)\\ &\quad\times m_2W_2(n_1-2,m_1-m_2).\end{aligned}$$

By noting that

$$\begin{aligned}W_2(n_1-2,m_1-m_2) &= \binom{n_1-2+m_1-m_2}{n_1-2}\\ &= \frac{\binom{m_1}{m_2}\binom{n_1-2+m_1}{m_1}}{\binom{n_1-2+m_1}{m_2}},\end{aligned} \tag{14}$$

we obtain

$$\begin{aligned}P(E_{m_1,m_2;n_1}) &= \frac{2H_{q_A}H_{q_B}}{H_{q_A+q_B}(n_1-1)}\\ &\quad\times m_1W_2(q_A-n_1,q_B-m_1)m_2\\ &\quad\times\frac{\binom{m_1}{m_2}\binom{n_1-2+m_1}{m_1}}{\binom{n_1-2+m_1}{m_2}}.\end{aligned} \tag{15}$$

Then by summing over all possible values of n_1, m_1 and m_2, $P(K=2|q_A,q_B)$ becomes

$$\begin{aligned}P(K=2|q_A,q_B) &= \sum_{n_1=2}^{q_A}\sum_{m_1=1}^{q_B}\sum_{m_2=1}^{m_1}P(E_{m_1,m_2,n_1})\\ &= \frac{2H_{q_A}H_{q_B}}{H_{q_A+q_B}}\sum_{n_1=2}^{q_A}\sum_{m_1=1}^{q_B}\frac{m_1}{n_1-1}\\ &\quad\times W_2(q_A-n_1,q_B-m_1)\\ &\quad\times\binom{n_1-2+m_1}{m_1}\sum_{m_2=1}^{m_1}m_2\frac{\binom{m_1}{m_2}}{\binom{n_1-2+m_1}{m_2}}.\end{aligned}$$

Using combinatorial Identity 1 (Appendix A) on the sum over m_2,

$$\begin{aligned}P(K=2|q_A,q_B) &= \frac{2H_{q_A}H_{q_B}}{H_{q_A+q_B}}\sum_{n_1=2}^{q_A}\sum_{m_1=1}^{q_B}\frac{m_1}{n_1-1}\\ &\quad\times W_2(q_A-n_1,q_B-m_1)\binom{n_1-2+m_1}{m_1}\\ &\quad\times\frac{m_1(n_1-1+m_1)}{n_1(n_1-1)}\\ &= \frac{2H_{q_A}H_{q_B}}{H_{q_A+q_B}}\sum_{n_1=2}^{q_A}\frac{1}{(n_1-1)^2n_1}\\ &\quad\times\left[\sum_{m_1=1}^{q_B}m_1^2(n_1-1+m_1)\binom{n_1-2+m_1}{m_1}\right.\\ &\quad\left.\times\binom{q_A-n_1+q_B-m_1}{q_B-m_1}\right].\end{aligned}$$

We can then use Identity 2 (Appendix A) on the part within the brackets, which yields

$$\begin{aligned}&P(K=2|q_A,q_B)\\ &= \frac{4(q_A-1)!(q_B-1)!}{(q_A+q_B-1)!\binom{q_A+q_B}{q_A}}\sum_{n_1=2}^{q_A}\frac{1}{(n_1-1)^2n_1}\\ &\quad\times\left[\frac{n_1(n_1-1)(n_1q_B-n_1+q_B+q_A+1)q_B}{(q_A+2)(q_A+1)}\binom{q_A+q_B}{q_A}\right]\\ &= \frac{4(q_A+q_B)}{(q_A+2)(q_A+1)q_A\binom{q_A+q_B}{q_A}}\\ &\quad\times\left[\sum_{n_1=2}^{q_A}\frac{n_1q_B-n_1-q_B+1}{n_1-1}+(2q_B+q_A)\sum_{n_1=2}^{q_A}\frac{1}{n_1-1}\right]\\ &= \frac{4(q_A+q_B)}{(q_A+2)(q_A+1)q_A\binom{q_A+q_B}{q_A}}\\ &\quad\times\left[(q_A-1)(q_B-1)+(q_A+2q_B)\sum_{n_1=2}^{q_A}\frac{1}{n_1-1}\right].\end{aligned} \tag{16}$$

The last sum is approximated by $\gamma + \ln(q_A - 1)$ for large q_A, where $\gamma \approx 0.577216$ is Euler's constant. Thus, for large q_A

$$P(K=2|q_A,q_B) \approx \frac{4(q_A+q_B)}{(q_A+2)(q_A+1)q_A\binom{q_A+q_B}{q_A}} \times[(q_A-1)(q_B-1)+(q_A+2q_B) \times[\gamma+\ln(q_A-1)]]. \quad (17)$$

2.6. The probability of three founding lineages

We now derive an expression for the probability of three founding lineages, $P(K=3|q_A,q_B)$, in order to find a pattern that can be used to set up the general case of the probability of k founding lineages. For $K=3$, we retain the definitions of m_1, m_2 and n_1 from above. We define n_2 as the number of A lineages at the second (A,B) event and m_3 as the number of B lineages at the third (A,B) event. It must be true that $m_3 \leqslant m_2 \leqslant m_1 \leqslant q_B$, $q_A \geqslant n_1 \geqslant 3$ and $n_1 - 1 \geqslant n_2 \geqslant 2$. This case reduces to an expression with two sums (over n_1 and n_2). To simplify the expression we follow the strategy used for solving $P(K=2|q_A,q_B)$, and we use Identity 1 followed by Identities 2 and 3 (Appendix A). Simplifying using (3)–(5) we obtain

$$\begin{aligned}
&P(E_{m_1,m_2,m_3;n_1,n_2}) \\
&\quad = I_{q_A,n_1} I_{n_1-1,n_2} H_{n_2-1} I_{q_B,m_1} I_{m_1,m_2} I_{m_2,m_3} H_{m_3} \\
&\qquad \times W_2(q_A-n_1, q_B-m_1) \\
&\qquad \times W_2(n_1-1-n_2, m_1-m_2) \\
&\qquad \times W_2(n_2-2, m_2-m_3) m_1 m_2 m_3 n_1 n_2 / H_{q_A+q_B} \\
&\quad = \frac{4H_{q_A}H_{q_B}m_1m_2m_3}{H_{q_A+q_B}(n_1-1)(n_2-1)} \\
&\qquad \times W_2(q_A-n_1, q_B-m_1) W_2(n_1-1-n_2, m_1-m_2) \\
&\qquad \times W_2(n_2-2, m_2-m_3) \\
&\quad = \frac{4H_{q_A}H_{q_B}m_1m_2m_3}{H_{q_A+q_B}(n_1-1)(n_2-1)} \binom{q_A-n_1+q_B-m_1}{q_B-m_1} \\
&\qquad \times \binom{n_1-1-n_2+m_1-m_2}{m_1-m_2} \frac{\binom{m_2}{m_3}\binom{n_2-2+m_2}{m_2}}{\binom{n_2-2+m_2}{m_3}}, \quad (18)
\end{aligned}$$

where the last step follows from (14) with n_2, m_2, and m_3 in place of n_1, m_1, and m_2. Then the desired probability is

$$\begin{aligned}
P(K=3|q_A,q_B) &= \sum_{n_1=3}^{q_A}\sum_{n_2=2}^{n_1-1}\sum_{m_1=1}^{q_B}\sum_{m_2=1}^{m_1}\sum_{m_3=1}^{m_2} P(E_{m_1,m_2,m_3;n_1,n_2}) \\
&= \frac{4H_{q_A}H_{q_B}}{H_{q_A+q_B}} \sum_{n_1=3}^{q_A}\frac{1}{n_1-1}\sum_{n_2=2}^{n_1-1}\frac{1}{n_2-1}\sum_{m_1=1}^{q_B} m_1 \\
&\quad \times \binom{q_A-n_1+q_B-m_1}{q_B-m_1} \\
&\quad \times \sum_{m_2=1}^{m_1} m_2 \binom{n_1-1-n_2+m_1-m_2}{m_1-m_2} \\
&\quad \times \binom{n_2-2+m_2}{m_2} \sum_{m_3=1}^{m_2} \frac{m_3\binom{m_2}{m_3}}{\binom{n_2-2+m_2}{m_3}}. \quad (19)
\end{aligned}$$

We now use Identity 1 on the last sum of (19) to obtain

$$\begin{aligned}
P(K=3|q_A,q_B) &= \frac{4H_{q_A}H_{q_B}}{H_{q_A+q_B}} \sum_{n_1=3}^{q_A}\frac{1}{n_1-1}\sum_{n_2=2}^{n_1-1}\frac{1}{n_2(n_2-1)^2} \\
&\quad \times \sum_{m_1=1}^{q_B} m_1 \binom{q_A-n_1+q_B-m_1}{q_B-m_1} \\
&\quad \times \left[\sum_{m_2=1}^{m_1} m_2^2(n_2-1+m_2)\binom{n_2-2+m_2}{m_2} \right. \\
&\quad \left. \times \binom{n_1-1-n_2+m_1-m_2}{m_1-m_2}\right]. \quad (20)
\end{aligned}$$

Applying Identity 2 on the part within the brackets in (20) yields

$$\begin{aligned}
&P(K=3|q_A,q_B) \\
&\quad = \frac{4H_{q_A}H_{q_B}}{H_{q_A+q_B}} \sum_{n_1=3}^{q_A}\frac{1}{n_1-1}\sum_{n_2=2}^{n_1-1}\frac{1}{n_2(n_2-1)^2} \\
&\qquad \times \sum_{m_1=1}^{q_B} m_1 \binom{q_A-n_1+q_B-m_1}{q_B-m_1} \\
&\qquad \times \left[\frac{n_2(n_2-1)(n_2m_1-n_2+m_1+n_1-1+1)m_1}{(n_1+1)n_1} \right. \\
&\qquad \left. \times \binom{n_1-1+m_1}{m_1}\right] \\
&\quad = \frac{4H_{q_A}H_{q_B}}{H_{q_A+q_B}} \sum_{n_1=3}^{q_A}\frac{1}{(n_1+1)n_1(n_1-1)}\sum_{n_2=2}^{n_1-1}\frac{1}{n_2-1} \\
&\qquad \times \left[\sum_{m_1=1}^{q_B} m_1^2(n_2m_1-n_2+m_1+n_1) \right. \\
&\qquad \left. \times \binom{n_1-1+m_1}{m_1}\binom{q_A-n_1+q_B-m_1}{q_B-m_1}\right]. \quad (21)
\end{aligned}$$

Using Identity 3 on the part within the brackets in expression (21)

$$\begin{aligned}
P(K=3|q_A,q_B) &= \frac{4H_{q_A}H_{q_B}}{H_{q_A+q_B}} \sum_{n_1=3}^{q_A}\frac{1}{(n_1+1)n_1(n_1-1)} \\
&\quad \times \sum_{n_2=2}^{n_1-1}\frac{1}{n_2-1}\left[\frac{(n_1+1)(q_A+q_B)!}{(q_A+3)!(q_B-1)!} \right. \\
&\quad \times [n_1q_Aq_B - n_1q_A + q_B^2n_1 - n_1 + 1 \\
&\quad + 2q_A + 3q_B + q_A^2 + 2q_B^2 + 3q_Aq_B \\
&\quad + n_2(2n_1 - 3n_1q_B + n_1q_B^2 + 2q_Aq_B \\
&\quad \left. - 2 - 2q_A + 2q_B^2)]\right]
\end{aligned}$$

$$= \frac{8(q_A+q_B)}{\binom{q_A+q_B}{q_A}(q_A+3)(q_A+2)(q_A+1)q_A} \times \sum_{n_1=3}^{q_A} \frac{1}{n_1(n_1-1)}\left[D + E\sum_{n_2=1}^{n_1-2}\frac{1}{n_2}\right], \quad (22)$$

where

$$D = (q_B-1)(n_1-2)(n_1q_B - 2n_1 + 2q_B + 2q_A + 2)$$

and

$$E = n_1q_Aq_B - n_1q_A + 2n_1q_B^2 + n_1 + 3q_B + q_A^2 + 4q_B^2 + 5q_Aq_B - 3n_1q_B - 1.$$

2.7. The probability of k founding lineages

For the general case, the probability of k founding lineages, $P(K=k|q_A,q_B)$, we need some additional definitions. Let n_i be the number of A lineages at the ith (A,B) event. For $i=2,\ldots,k-1$ it is true that $1+k-i \leqslant n_i < n_{i-1}$. Also $k \leqslant n_1 \leqslant q_A$ and $n_k = 1$. Let m_i be the number of B lineages at the ith (A,B) event. For each i, $m_i \leqslant m_{i-1}$. Then

$$\begin{aligned}
&P(E_{m_1,\ldots,m_i,\ldots,m_k;n_1,\ldots,n_i,\ldots,n_{k-1}}) \\
&= \left[I_{q_A,n_1}\left(\prod_{i=1}^{k-2} I_{n_{i-1},n_{i+1}}\right) H_{n_{k-1}-1} I_{q_B,m_1} \right. \\
&\quad \times \left(\prod_{i=1}^{k-1} I_{m_i,m_{i+1}}\right) H_{m_k} W_2(q_A-n_1, q_B-m_1) \\
&\quad \times W_2(n_1-1-n_2, m_1-m_2) \\
&\quad \times \cdots \times W_2(n_i-1-n_{i+1}, m_i-m_{i+1}) \\
&\quad \times \cdots \times W_2(n_{k-2}-1-n_{k-1}, m_{k-2}-m_{k-1}) \\
&\quad \left. \times W_2(n_{k-1}-2, m_{k-1}-m_k) m_k \prod_{i=1}^{k-1} n_i m_i \right] \Big/ H_{q_A+q_B}.
\end{aligned} \quad (23)$$

From Eqs. (6) and (3),

$$I_{q_A,n_1}\left(\prod_{i=1}^{k-2} I_{n_{i-1},n_{i+1}}\right) H_{n_{k-1}-1} = \frac{2^{k-1} H_{q_A}}{\prod_{i=1}^{k-1} n_i(n_i-1)},$$

$$I_{q_B,m_1}\left(\prod_{i=1}^{k-1} I_{m_i,m_{i+1}}\right) H_{m_k} = H_{q_B}.$$

We can therefore rewrite (23) so that

$$\begin{aligned}
&P(E_{m_1,\ldots,m_i,\ldots,m_k;n_1,\ldots,n_i,\ldots,n_{k-1}}) \\
&= \frac{2^{k-1} H_{q_A} H_{q_B}}{H_{q_A+q_B}} \frac{m_k \prod_{i=1}^{k-1} n_i m_i}{\prod_{i=1}^{k-1} n_i(n_i-1)} W_2(q_A-n_1, q_B-m_1) \\
&\quad \times W_2(n_1-1-n_2, m_1-m_2) \\
&\quad \times \cdots \times W_2(n_i-1-n_{i+1}, m_i-m_{i+1}) \\
&\quad \times \cdots \times W_2(n_{k-2}-1-n_{k-1}, m_{k-2}-m_{k-1}) \\
&\quad \times W_2(n_{k-1}-2, m_{k-1}-m_k)
\end{aligned}$$

$$\begin{aligned}
&= \frac{2^k}{q_B\binom{q_A+q_B}{q_A}\binom{q_A+q_B-1}{q_B}} m_k \left(\prod_{i=1}^{k-1}\frac{m_i}{n_i-1}\right) \\
&\quad \times W_2(q_A-n_1, q_B-m_1) \\
&\quad \times \left[\prod_{i=1}^{k-2} W_2(n_i-1-n_{i+1}, m_i-m_{i+1})\right] \\
&\quad \times W_2(n_{k-1}-2, m_{k-1}-m_k).
\end{aligned} \quad (24)$$

By summing over the indices n_1 to n_{k-1} and m_1 to m_k we obtain

$$\begin{aligned}
P(K=k|q_A,q_B) &= \sum_{n_1=1+k-1}^{q_A} \cdots \sum_{n_i=1+k-i}^{n_{i-1}-1} \cdots \sum_{n_{k-1}=2}^{n_{k-2}-1} \\
&\quad \sum_{m_1=1}^{q_B} \cdots \sum_{m_i=1}^{m_{i-1}} \cdots \sum_{m_k=1}^{m_{k-1}} \\
&\quad P(E_{m_1,\ldots,m_i,\ldots,m_k;n_1,\ldots,n_i,\ldots,n_{k-1}}) \\
&= \frac{2^k}{q_B\binom{q_A+q_B}{q_A}\binom{q_A+q_B-1}{q_B}} \sum_{n_1=1+k-1}^{q_A} \\
&\quad \cdots \sum_{n_i=1+k-i}^{n_{i-1}-1} \cdots \sum_{n_{k-1}=2}^{n_{k-2}-1} \sum_{m_1=1}^{q_B} \cdots \sum_{m_i=1}^{m_{i-1}} \cdots \sum_{m_k=1}^{m_{k-1}} \\
&\quad \left[m_k\left(\prod_{i=1}^{k-1}\frac{m_i}{n_i-1}\right) W_2(q_A-n_1, q_B-m_1) \right. \\
&\quad \times \left[\prod_{i=1}^{k-2} W_2(n_i-1-n_{i+1}, m_i-m_{i+1})\right] \\
&\quad \left. \times W_2(n_{k-1}-2, m_{k-1}-m_k)\right].
\end{aligned} \quad (25)$$

If $k=2$ or 3, (25) reduces to (16) or (22), respectively. If $k=1$, (25) reduces to (13) by noting that $n_k = n_1 = 1$ and letting empty sums and products and W_2 of negative integers equal 1.

2.8. The probability of q_A founding lineages

A special case is the probability of exactly q_A founding lineages, $P(K=q_A|q_A,q_B)$. To have exactly q_A founding lineages, the q_A A lineages cannot coalesce with each other: they must coalesce exclusively with B lineages. The number of A lineages will therefore decrease by one at each (A,B) event. Then $n_1 = q_A$, $n_2 = q_A - 1, \ldots, n_{k-1} = 2$, $n_k = 1$, and $I_{q_A,n_1} = 1$ and $I_{n_{i-1},n_{i+1}} = 1$. Note that all possible coalescence events [(A,B) and (B,B)] occur in one sequence and we do not have to worry about the ordering of nodes. We have

$$P(E_{m_1,\ldots,m_i,\ldots,m_k;n_1,\ldots,n_i,\ldots,n_{k-1}}) = \frac{H_{q_B}}{H_{q_A+q_B}}\left(\prod_{i=1}^{k} m_i\right)\left(\prod_{i=1}^{k-1} n_i\right).$$

Because in this case $\prod_{i=1}^{k-1} n_i = q_A!$,

$$P(E_{m_1,\ldots,m_i,\ldots,m_k;n_1,\ldots,n_i,\ldots,n_{k-1}}) = \frac{2^{q_A}(q_B-1)!\prod_{i=1}^{k} m_i}{\binom{q_A+q_B}{q_B}(q_A+q_B-1)!}. \quad (26)$$

By summing over the indices m_1 to m_k, we obtain

$$P(K=q_A|q_A,q_B)=\frac{2^{q_A}(q_B-1)!}{\binom{q_A+q_B}{q_B}(q_A+q_B-1)!}\times\sum_{m_1=1}^{q_B}m_1\cdots\sum_{m_i=1}^{m_{j-1}}m_i\cdots\sum_{m_k=1}^{m_{k-1}}m_k. \quad (27)$$

Note that when $q_B=1$, Eq. (27) reduces to

$$P(K=q_A|q_A,q_B=1)=\frac{q_A!}{H_{q_A+1}}=\frac{2^{q_A}q_A!}{(q_A+1)!q_A!}=\frac{2^{q_A}}{(q_A+1)!}.$$

In the limit as $q_B\to\infty$ (see Appendix B) we have

$$\lim_{q_B\to\infty}P(K=q_A|q_A,q_B)=1. \quad (28)$$

From this result it follows that when q_A is held constant, K converges in probability to q_A as $q_B\to\infty$, since for all $\varepsilon>0$

$$\lim_{q_B\to\infty}P(|K-q_A|<\varepsilon|q_A,q_B)\to 1. \quad (29)$$

Eqs. (28) and (29) provide information about the waiting times to coalescence for (A,A), (A,B), and (B,B) events. At a given time, the next coalescence event to take place is chosen as either an (A,A) event, an (A,B) event, or a (B,B) event. As q_B increases, the number of possible (B,B) lineage pairs increases, as does the number of (A,B) pairs, while the number of (A,A) pairs remains constant. Thus, the probability increases that (B,B) and (A,B) events will occur, rather than (A,A) events. The q_A A lineages will tend to be absorbed by (A,B) events as the waiting times between (A,B) events decrease in comparison with the times between (A,A) events.

2.9. Recursion

By inspecting $P(K=2|q_A,q_B)$ and $P(K=3|q_A,q_B)$ (Eqs. (16) and (22)), we see that

$$P(K=2|q_A,q_B)=\frac{2^2}{C_{q_A,q_B}}\sum_{n_1=2}^{q_A}\frac{1}{n_1-1}\sum_{m_1=1}^{q_B}m_1\times W_2(q_A-n_1,q_B-m_1)\times\sum_{m_2=1}^{m_1}m_2W_2(n_1-2,m_1-m_2),$$

where

$$C_{q_A,q_B}=q_B\binom{q_A+q_B}{q_A}\binom{q_A+q_B-1}{q_B},$$

and

$$P(K=3|q_A,q_B)=\frac{2^3}{C_{q_A,q_B}}\sum_{n_1=3}^{q_A}\frac{1}{n_1-1}\sum_{m_1=1}^{q_B}m_1\times W_2(q_A-n_1,q_B-m_1)\times\left[\sum_{n_2=2}^{n_1-1}\frac{1}{n_2-1}\sum_{m_2=1}^{m_1}m_2\times W_2(n_1-1-n_2,m_1-m_2)\times\sum_{m_3=1}^{m_2}m_3W_2(n_2-2,m_2-m_3)\right]. \quad (30)$$

By replacing m_2,m_3,n_2,m_1 and n_1-1 with m_1,m_2,n_1,q_B and q_A inside the brackets in Eq. (30), it can be observed that the part within the brackets has the same form as

$$P(K=2|n_1-1,m_1)\frac{m_1\binom{n_1-1+m_1}{n_1-1}\binom{n_1-2+m_1}{m_1}}{2^2}=P(K=2|n_1-1,m_1)\frac{C_{n_1-1,m_1}}{2^2}.$$

This suggests the following recursion for $k>2$:

$$P(K=k|q_A,q_B)=\frac{2}{C_{q_A,q_B}}\sum_{n_1=k}^{q_A}\frac{1}{n_1-1}\sum_{m_1=1}^{q_B}m_1\times W_2(q_A-n_1,q_B-m_1)C_{n_1-1,m_1}\times P(K=k-1|n_1-1,m_1). \quad (31)$$

To verify this recursion, we can obtain the probability of k founding lineages recursively from the probability of $k-1$ founding lineages as follows. At the time of the first (A,B) event, the number of A lineages, n_1, must be at least k, because after the first (A,B) event, at least $k-1$ lineages from population A must be available to participate in the remaining $k-1$ (A,B) events. The number of B lineages at the time of the first (A,B) event, m_1, can be any value in $[1,q_B]$.

The number of ways that q_A lineages can coalesce to n_1 lineages is I_{q_A,n_1}, and the number of ways that q_B lineages can coalesce to m_1 lineages is I_{q_B,m_1}. The number of ways of ordering the q_A-n_1 and q_B-m_1 events that occur more recently than the first (A,B) event is $W_2(q_A-n_1,q_B-m_1)$. The number of ways of choosing an A lineage and a B lineage to participate in the first (A,B) event is n_1m_1.

Among the coalescences that reduce the remaining n_1-1 lineages from population A and m_1 lineages from population B to one lineage, in order to produce k total founding lineages, $k-1$ interpopulation coalescences must occur. The number of ways of obtaining $k-1$ founding lineages when starting with n_1-1 lineages from population A and m_1 lineages from population B is the product of the number of possible sequences of coalescences for n_1-1+m_1 lineages and the probability of $k-1$ founding lineages, or $H_{n_1-1+m_1}P(K=k-1|n_1-1,m_1)$.

Dividing by the total number of sequences of coalescences for q_A and q_B lineages, $H_{q_A+q_B}$,

$$P(K=k|q_A,q_B)=\frac{1}{H_{q_A+q_B}}\sum_{n_1=k}^{q_A}\sum_{m_1=1}^{q_B}I_{q_A,n_1}I_{q_B,m_1}\times W_2(q_A-n_1,q_B-m_1)m_1n_1H_{n_1-1+m_1}\times P(K=k-1|n_1-1,m_1). \quad (32)$$

This equation can be simplified to produce the form of the equation in (31).

In situations where the exact expression (25) is cumbersome to evaluate, we can use the recursion expression to compute the probabilities for large numbers of founding lineages. Even when the maximal value of K is relatively small ($K \leqslant 3$), it can be much faster to compute the probability distribution of K using the recursion (31) than using (25).

2.10. Divergence times larger than zero

Until now, we have assumed that populations A and B have not been separated: $T = T_A = T_B = 0$. In this case, which describes the events that occur above time T (Fig. 1), the sampled numbers of A and B lineages, r_A and r_B, equal the numbers of ancestral A and B lineages, q_A and q_B. If $T_A > 0$ and $T_B > 0$, we have to consider the probabilities that the r_A and r_B sampled lineages have q_A and q_B ancestors at the time of divergence (note that T_A need not equal T_B, as an effect of different sizes of populations A and B). If we assume $T_A = T_B = T$, then the events in populations A and B *before* T (going backwards in time) are independent of each other and the events *after* T depend only on q_A and q_B. Thus, we must sum over all possible values of q_A and q_B to obtain $P(K = k|r_A, r_B, T_A, T_B)$:

$$P(K = k|r_A, r_B, T_A, T_B) = \sum_{q_A=1}^{r_A} \sum_{q_B=1}^{r_B} g_{r_A,q_A}(T_A) g_{r_B,q_B}(T_B) \times P(K = k|q_A, q_B). \quad (33)$$

The numbers of ancestral lineages of r_A and r_B decrease with T_A and T_B because $g_{r_A,1}(T_A)$ and $g_{r_B,1}(T_B)$ increases monotonically with T_A and T_B. The probability $P(K = 1|q_A, q_B)$ monotonically increases with decreasing q_A and q_B, and because q_A and q_B decrease with T_A and T_B, $P(K = 1|r_A, r_B, T_A, T_B)$ is monotonically increasing with T_A and T_B.

2.11. Expected number of founding lineages

The expected number of founding lineages for a given sample size can be computed from the probability distribution $P(K|r_A, r_B, T_A, T_B)$,

$$E(K|r_A, r_B, T_A, T_B) = \sum_{k=1}^{r_A} kP(k|r_A, r_B, T_A, T_B). \quad (34)$$

When $T = 0$, a lower bound on (34) is

$$\begin{aligned} E(K|q_A, q_B) &= \sum_{k=1}^{q_A} kP(K = k|q_A, q_B) \quad (35) \\ &\geqslant q_A P(K = q_A|q_A, q_B) \\ &\quad + 1[1 - P(K = q_A|q_A, q_B)] \\ &= (q_A - 1)P(K = q_A|q_A, q_B) + 1. \quad (36) \end{aligned}$$

Since $K \geqslant 1$ and $K \leqslant q_A$ and because we have shown that $\lim_{q_B \to \infty} P(K = q_A|q_A, q_B) = 1$ (28), we have convergence in mean when $q_B \to \infty$:

$$\lim_{q_B \to \infty} E(K|q_A, q_B) = q_A. \quad (37)$$

As $q_B \to \infty$, the sequence of random variables K_{q_A,q_B} converges in probability to q_A, where K_{q_A,q_B} denotes the random number of founding lineages when the present-day sample sizes are q_A and q_B. Since K_{q_A,q_B} is bounded above by the same constant q_A for any value of q_B, Eq. (37) follows from Serfling (1980, 1.3.6).

3. Results

3.1. The number of founding lineages when no time has passed since the divergence

We first assume that no time has passed since population divergence ($T_A = T_B = T = 0$), so that the sample sizes r_A and r_B from the present populations A and B are equal to the numbers of lineages ancestral to the samples ($r_A = q_A$ and $r_B = q_B$). The number of founding lineages (of the sample of population A) naturally cannot exceed the sample size of population A.

Fig. 3i shows an example of the probability distribution of the number of founding lineages, in which 50 individuals have been sampled from each population and the divergence time is zero. This distribution is almost symmetrical around $k = 39$.

Fig. 4 displays as functions of sample size the probabilities of k founding lineages when no time has passed since divergence, showing how the probabilities of particular values of k depend on the sample size. When r_A (or both r_A and r_B) is small, only a few values of k are likely, whereas when r_A (or r_A and r_B) becomes larger, the number of likely values of k increases. As r_A increases, the maximal probability over all values of k decreases at the same time as a wider range of values of k becomes likely. Only for small sample sizes does the probability exceed 0.5 for any particular k.

Fig. 3ii shows the cumulative probability distribution of the number of founding lineages for different sample sizes when no time has passed since divergence. It is clear from Figs. 4 and 3ii that increasing the sample size of either population makes larger values of k more likely. However, increasing the sample size of population A has a much larger effect on the probable values of k than increasing the sample size of population B (Fig. 3ii). This result is not surprising, as k is upwardly bounded by the sample size r_A from population A, while there is no corresponding upper bound based on the value of r_B.

It is noteworthy that for a sample of size r_A the number of founding lineages with the highest probability need not equal r_A even if $T = 0$ (Fig. 5). For example, for $r_A = r_B = 15$, $k = 12$ founding lineages has the highest probability, and the probability of 15 founding lineages is low, $P(K = 15|15, 15) = 0.0268$. This effect is also seen in Fig. 3i for a sample of size 50 lineages from both

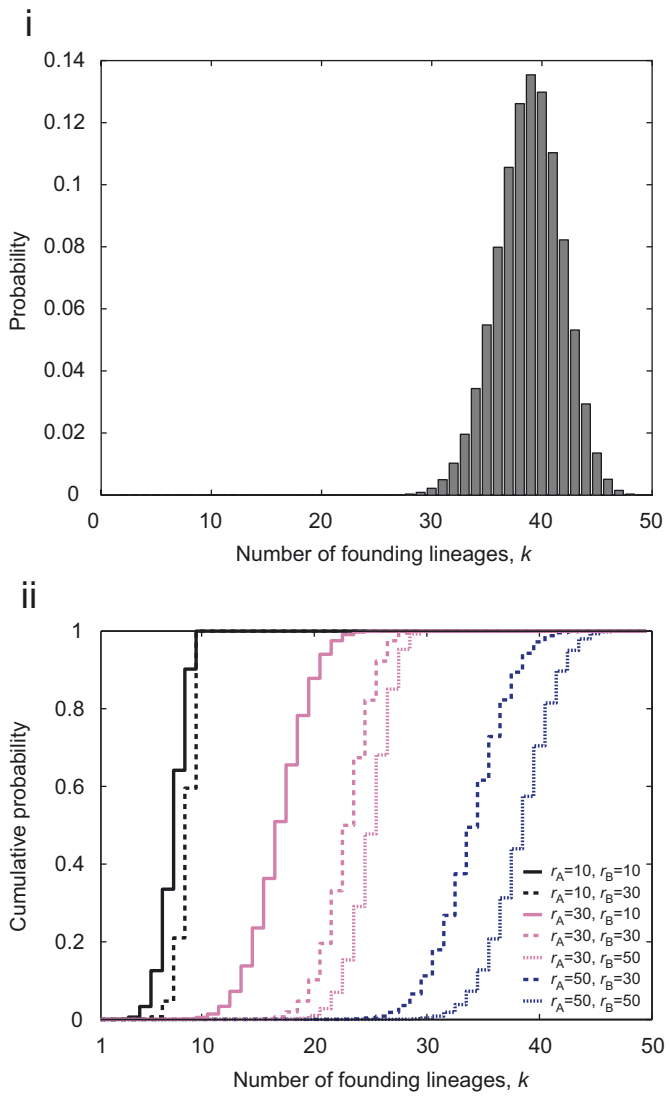

Fig. 3. Probability distribution of the number of founding lineages when no time has passed since the separation of populations A and B ($T_A = T_B = 0$). (i) The probability distribution of the number of founding lineages for a sample size of 50 from each population. (ii) The cumulative probability distribution of the number of founding lineages for varying sample sizes from populations A and B.

populations. The difference between the probability of r_A founding lineages and that of the value of k with the highest probability (arg $\max_k[P(K = k)]$) depends on r_B (see Fig. 4). When r_B is large and $T = 0$, arg $\max_k[P(K = k)]$ approaches r_A (Fig. 5). In fact, we know that $P(K = r_A|r_A, r_B) \to 1$ when $r_B \to \infty$ (see Eq. (28)). Thus, if $r_B \gg r_A$ (and $T = 0$) then there will very likely be exactly r_A founding lineages.

The expected number of founding lineages of a sample from population A is also dependent on the size of the sample from population B (Fig. 6). Unless the sample from population A is very small or the sample from population B is very large, the expected number of founding lineages is substantially lower than the number of lineages sampled from population A. In fact, we know that the expected number of founding lineages of a sample from population A approaches the sample size q_A when $q_B \to \infty$ [see Eq. (37)].

3.2. Effect of sample size

The effect of sample size on the probability distribution of the number of founding lineages of a sample from population A when no time has passed since the divergence can be seen in Figs. 3ii, 4–6. For a symmetric sample of the two populations (Fig. 4i), the number of likely values of k (say with probability >0.01) increases with sample size. For a given sample size, however, the likely values of the number of founding lineages are confined to a relatively small range. For example, given a sample size of 20 individuals from both populations, the probability that there are between 12 and 19 founding lineages is 0.9768. The probability functions for given values of k are in general wider for small samples from population B (Fig. 4iii) than for large samples (Fig. 4v). This result affects the number of founding lineages that is most likely for a given sample from population A. For a sample size of, say, 20 individuals from population A, the most likely number of founding lineages is 9 if 3 B lineages are sampled and 17 if 30 B lineages are sampled (Figs. 4iii and 4v).

How large a sample from population B is necessary for the most likely value of k to equal r_A? When r_A is small, a relatively small r_B is needed before the most likely value equals r_A. For example, in Fig. 4iv, where $r_A = 3$, for $r_B = 2$ the probability $P(K = 3)$ already is the largest (although a much larger r_B is necessary to approach the asymptotic value of 1 for $P(K = 3)$). On the other hand, if $r_A = 10$, a sample of at least 29 B lineages is necessary for $k = 10$ to have the highest probability (Fig. 4ii).

For $T = 0.2$, the probability distribution of K is still dependent on sample size (Fig. 7i). However, if $T = 2$, the probability distribution of K changes very little when the sample size exceeds 10 (Fig. 7ii). Note that $P(K = 2)$ is almost unaffected by changes in sample size when $T = 2$, which is not the case when $T = 0.2$ (Fig. 7i and 7ii). If the divergence time is even larger ($T = 4$), then regardless of sample size, almost all A lineages have coalesced to one lineage before they are able to coalesce with any B lineages. Hence, $P(K = 1|r_A, r_B, T_A = 4, T_B = 4)$ is close to 1 even for relatively large sample sizes (Fig. 7iii). The value of T clearly affects the impact of the sample size. For small T the sample size r_B has a large effect on q_A, whereas when T is large, the sample size r_B has only a weak effect on q_A. Thus, increasing the sample size r_B would improve the accuracy of an estimate of the number of founding lineages of the whole population, if the population is recently diverged. On the other hand, if population B was founded a long time ago, increasing the sample size r_B would have a small effect on the accuracy of an estimate of the number of founding lineages of population B.

Fig. 8i shows the expected number of founding lineages $E(K)$ as a function of sample size for different values of the

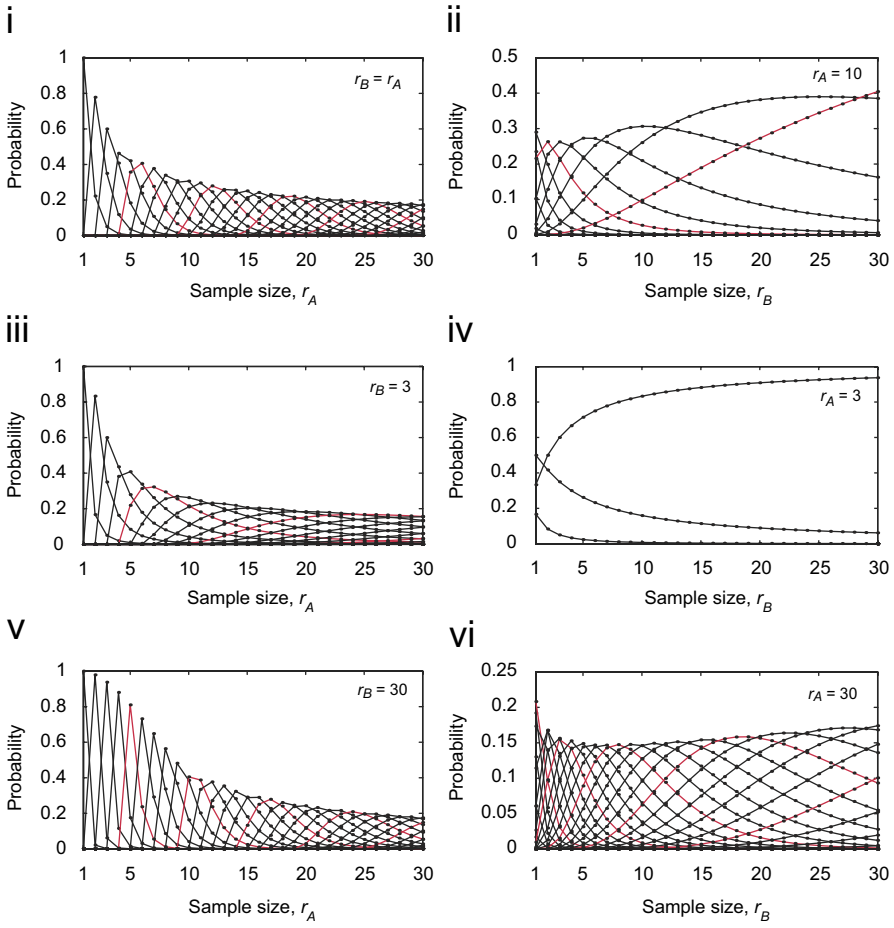

Fig. 4. The probabilities of k founding lineages computed from Eqs. (13), (16), (22), (25) and (31), plotted against sample size. The lines represent the probabilities of $k = 1, \ldots, 30$ founding lineages. The leftmost curve is $P(K = 1)$, the next curve to the right of $P(K = 1)$ is $P(K = 2)$, and so on. Every fifth curve is colored red, starting with $P(K = 5)$, to aid in telling the curves apart. At each value of the sample size on the x-axis, the probabilities sum to 1. Note that no time has passed since the separation of populations A and B ($T_A = T_B = T = 0$). Thus, $r_A = q_A$ and $r_B = q_B$. (i) $r_B = r_A$, (ii) $r_A = 10$, (iii) $r_B = 3$, (iv) $r_A = 3$, (v) $r_B = 30$, (vi) $r_A = 30$.

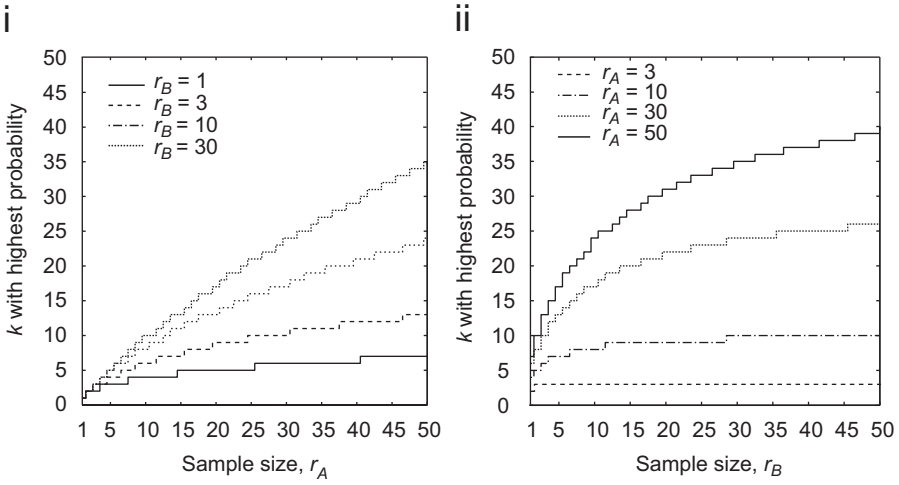

Fig. 5. The number of founding lineages k that has the highest probability as a function of (i) the sample size r_A of population A and (ii) the sample size r_B of population B, when no time has passed since the separation of the two populations ($T_A = T_B = 0$). Note that $P(K = 1|q_A = 1, q_B) = 1$ for any value of q_B.

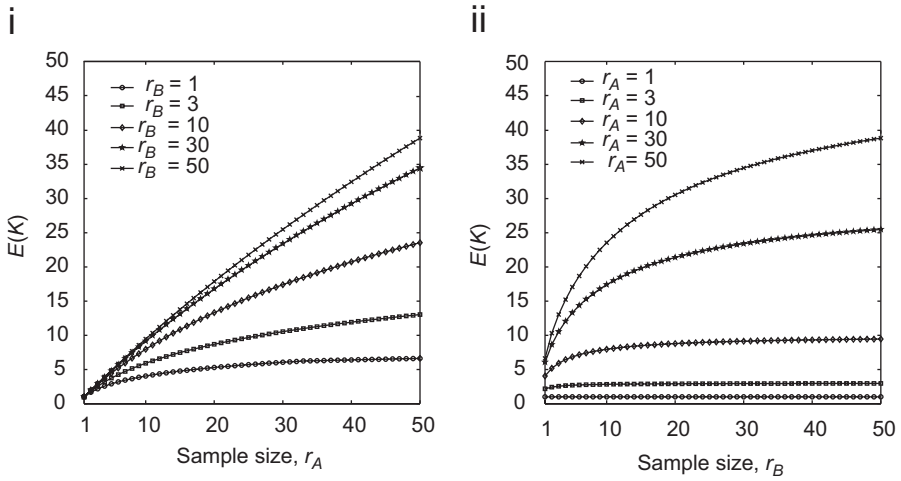

Fig. 6. The expected number of founding lineages $E(K)$ of a sample from population A when the divergence time is zero, as a function of (i) the sample size r_A from population A, and (ii) the sample size r_B from population B. The expected values were computed from Eq. (35).

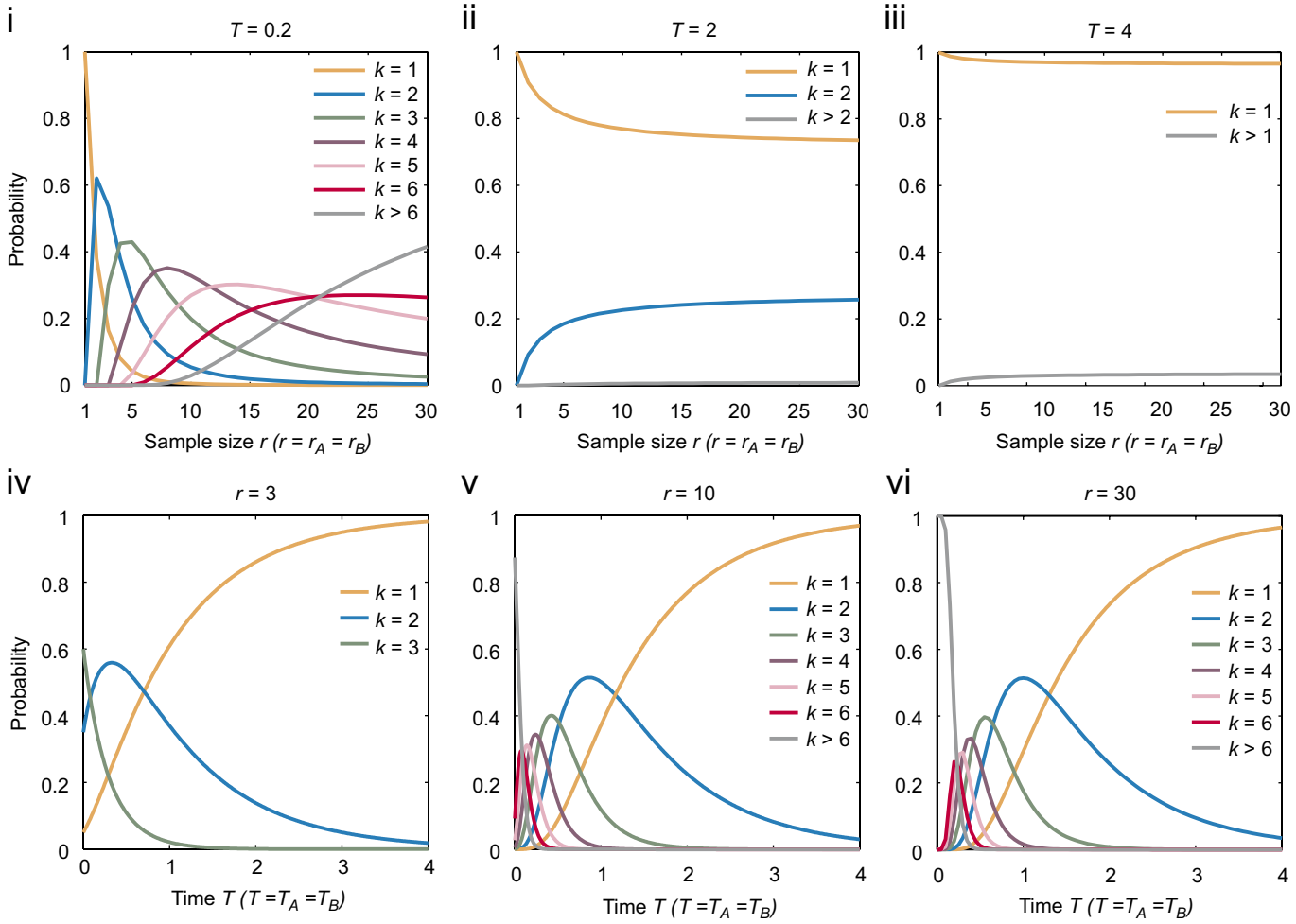

Fig. 7. The probabilities of k founding lineages as functions of sample size r_A (i)–(iii), and as functions of divergence time T (iv)–(vi), obtained from Eq. (33). Both populations have the same sample size ($r_A = r_B$) and the same amount of time has passed in the two populations ($T_A = T_B = T$). At each point on the x-axis the probabilities sum to 1. (i) $T = 0.2$, (ii) $T = 2$, (iii) $T = 4$, (iv) $r = 3$, (v) $r = 10$, (vi) $r = 30$.

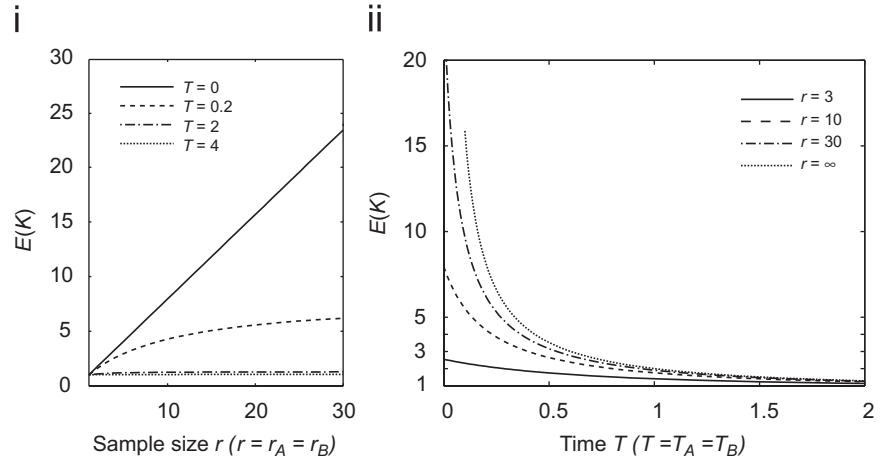

Fig. 8. The expected number of founding lineages $E(K)$ as a function of sample size r ($r = r_A = r_B$) and time T ($T = T_A = T_B$), obtained from Eqs. (35) and (34). (i) $E(K)$ as a function of r for different values of T. (ii) $E(K)$ as a function of T for r of 3, 10, 30 and ∞. Note that the dotted curve representing $E(K|r = \infty)$ in (ii) is shown only for $T \geqslant 0.1$ because we are using the approximation described after Eq. (9).

divergence time. When no time has passed since divergence and with a symmetrical sample from the two populations, $E(K|r_A = r, r_B = r, T = 0)$ is almost perfectly linear ($y = 0.7725x + 0.2302$). The residual sum-of-squares for $r = 1, \ldots, 50$ is approximately 2.9×10^{-5} and $R^2 \approx 1 - 4.7 \times 10^{-9}$. When $T = 0.2$, the expected number of founding lineages $E(K|r_A = r, r_B = r, T = 0.2)$ increases slowly with sample size and no longer fits a straight line. When $T = 4$ or even $T = 2$, the expected number of founding lineages is very close to 1.

3.3. Effect of time

If we assume that $N_A = N_B$ we only need to consider the case where $T = T_A = T_B$. Fig. 7iv–vi shows probabilities of the number of founding lineages as functions of time for various sample sizes. In Fig. 7iv the probabilities of various values of k are shown as functions of T for a sample of size 3 for both populations. The probability of one founding lineage increases with T, whereas the probability of two founding lineages has a maximum at $T \approx 0.3307$. $P(K = 3)$ decreases with T. Thus, when no time has passed, $k = 3$ has the highest probability, but it is only slightly higher than $P(K = 2)$. As time passes, the probability increases that one pair among the three A lineages coalesce with each other before they coalesce with any B lineage, causing $P(K = 2)$ to be the highest until $T \approx 0.705$, when it is overtaken by $P(K = 1)$. The probability $P(K = 1)$ is monotonically increasing in the whole interval of T. To obtain $P(K = 1) = 0.95$ the time needed is $T \approx 3.008$. In Fig. 7v, when the sample size is 10 for both populations, the probabilities $P(K = k)$ for $k = 1, 2$ and 3 are shifted to the right, that is, towards higher values of T. The time required to obtain $P(K = 1) = 0.95$ is slightly higher in Fig. 7vi, in which the sample size is 30 from each population, equaling $T \approx 3.634$.

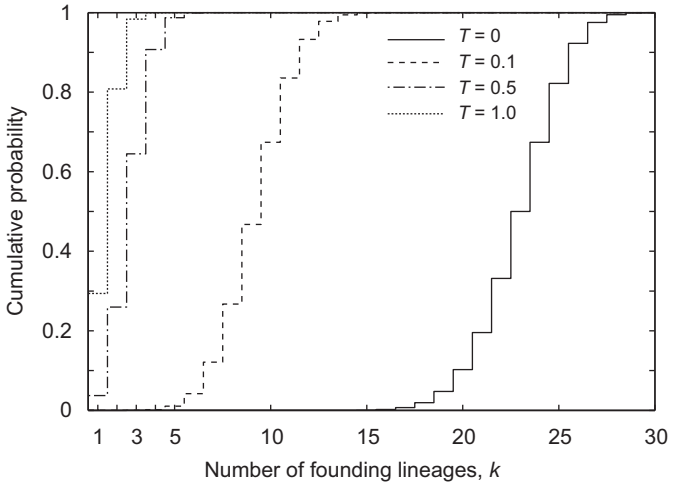

Fig. 9. The cumulative probability distribution of the number of founding lineages for samples of size 30 each from populations A and B, for divergence times of $T = 0, 0.1, 0.5$ and 1. The population sizes were assumed to be the same for populations A and B ($T_A = T_B$). The probabilities were obtained from Eq. (33).

Fig. 9 shows the cumulative probabilities of the number of founding lineages for four different values of the divergence time T, given sample sizes of 30 individuals from each population. Even though the likely values of K decrease relatively quickly with T, for moderate divergence times the likely numbers of founding lineages include values that are larger than one (Fig. 9). The probability that the present population A has one founding lineage increases slowly, and substantial divergence times are required for that probability to approach one.

From Fig. 7 we see that when $T > 2$, the probability distribution of K is quite similar for different sample sizes. When $T < 2$ the probability distribution is very different for

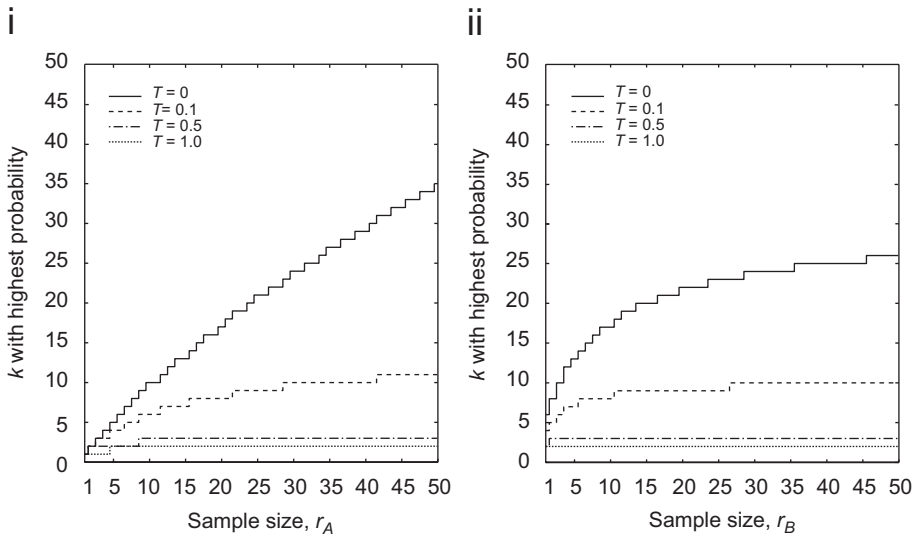

Fig. 10. The number of founding lineages k that has the highest probability as a function of (i) the sample size r_A of population A when $r_B = 30$, and (ii) the sample size r_B of population B when $r_A = 30$, for different values of the divergence time T ($T_A = T_B = T$).

differing sample sizes. This result is also visible in Fig. 8ii, which shows the expected number of founding lineages as a function of time, and in Fig. 10, which shows the most likely number of founding lineages for different values of the divergence time.

3.4. Effect of different population sizes

By varying the divergence time T, the impact of varying population sizes on the probability distribution of K can be explored. To investigate the properties of different sizes of populations A and B, we set $T_A = \beta T_B$, where β is a constant. Time then elapses faster in population A if $\beta > 1$ and slower if $\beta < 1$. Using (33), the probabilities of various values of k can be computed as functions of time for different values of β (Fig. 11). From Figs.11i, 11iii and 11v, we see that small sizes of population A lead to high probabilities of one founding lineage when the sample size is 30 (the results are similar for larger sample sizes, results not shown). If $N_A = N_B/5$ and $r = 30$, then $P(K = 1)$ almost equals 1 after one unit of coalescent time has passed in population B (5 units of coalescent time in population A). If N_A is large in comparison with N_B, then $P(K = 1)$ increases slowly with time (Fig. 11ii, iv and vi). However, $P(K > 6)$ decreases relatively fast, and the possible values of K are much smaller than q_A. Suppose, for example, that population A was founded by population B $3T_B$ coalescent units ago. Then $P(K = 1) > P(K > 1)$ if $N_A < 2N_B$. On the other hand, if $N_A = 5N_B$, the most likely number of founding lineages is 2, and if $N_A = 20N_B$, then $k = 4$ has the highest probability.

3.5. Implications for the whole population

Given a finite sample, what is known about the number of founding lineages of the whole A population? Since the expected number of founding lineages grows almost linearly with sample size when $T = 0$ (solid line in Fig. 8i), we would need a sample size that is close to the total population size to infer the number of founding lineages correctly when no time has passed since the divergence of the two populations (or one could use the linear relationship in Fig. 8i if the total population size was known). However, if $T > 0$, the expected number of founding lineages grows slowly with the sample size (see Fig. 8i). Assuming that N_A and N_B are large, we can use (9) to compute the expected number of ancestral lineages T units of time in the past. The expected number of founding lineages L of the whole population A can now be obtained from (34)

$$E(L|N_A = \infty, N_B = \infty, T_A, T_B) = \sum_{k=1}^{\infty} k \sum_{s_A=1}^{\infty} \sum_{s_B=1}^{\infty} g_{\infty,s_A}(T_A) g_{\infty,s_B}(T_B) P(k|s_A, s_B). \quad (38)$$

It is clear that a larger sample from both populations will estimate the number of founding lineages of the whole population more accurately than will a small sample (Fig. 12i). With a sample size of 50 individuals and $T = 0.1$, the ratio of the expected number of founding lineages of the sample and the expected number of founding lineages of the whole population, $E(K)/E(L)$, equals ≈ 0.72. When $T \geqslant 0.62$, for the same sample size, $E(K)/E(L) \geqslant 0.95$. For a fixed sample size from population A, little information about L is gained by increasing the sample size from population B (Fig. 12ii). Conversely, for a fixed sample size from population B, the information about L is greatly increased just by increasing a small sample size from population A to a moderate sample size (Fig. 12iii).

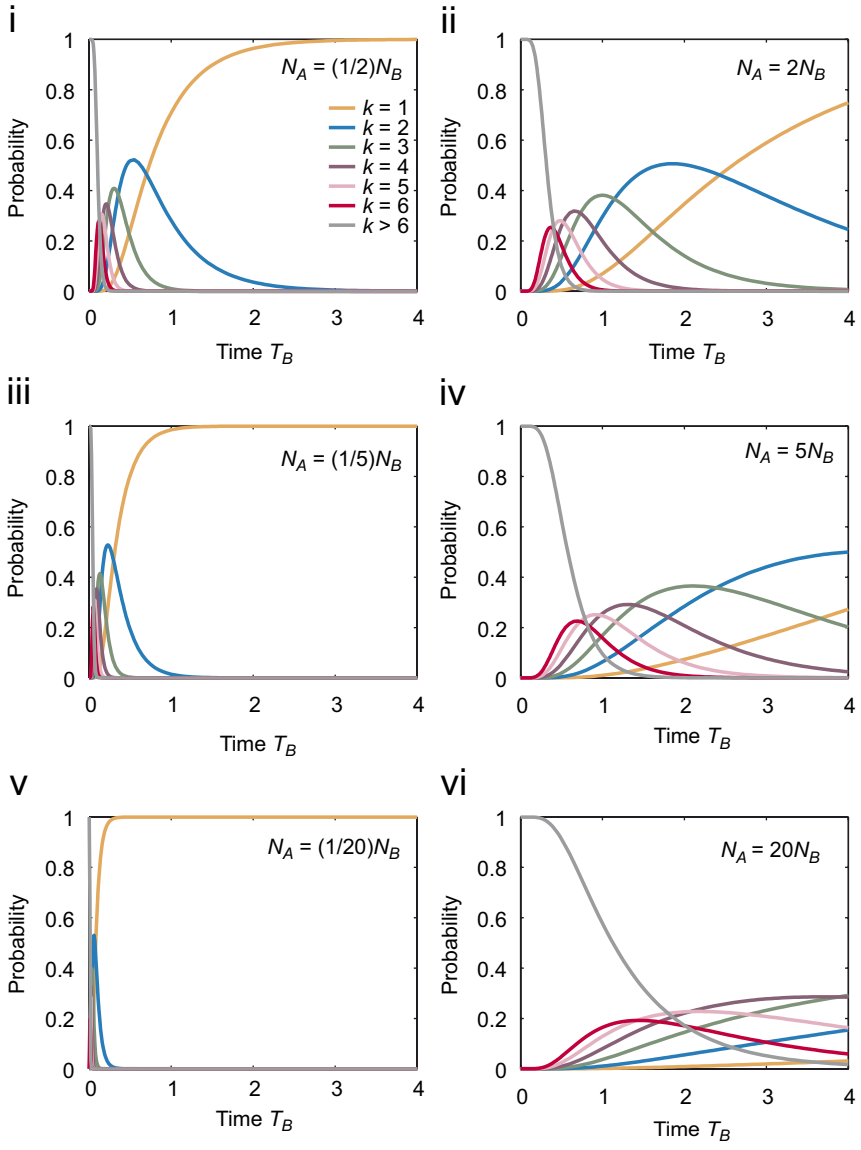

Fig. 11. The probabilities of small values of the number of founding lineages k as functions of T_B when $r = r_A = r_B = 30$. The functions were obtained from Eq. (33). The probabilities of k when $N_A = N_B$ ($T_A = T_B$) can be seen in Fig. 7vi. (i) $N_A = (1/2)N_B$, (ii) $N_A = 2N_B$, (iii) $N_A = (1/5)N_B$, (iv) $N_A = 5N_B$, (v) $N_A = (1/20)N_B$ and (vi) $N_A = 20N_B$.

3.6. Implications for the whole genome

If the sample sizes r_A and r_B are large enough, the probability of a certain k can be viewed as the fraction of the genome of population A that has exactly k founding lineages. In the same way, the probability $P(K \leqslant \gamma)$ can be thought of as the fraction of the whole genome of population A that has at most γ founding lineages. Let $M_i(T_A)$ be Eq. (33) for $i = k$ when $r_A = r_B = \infty$, $T_A = \beta T_B$ and β is a constant. The probability that there were at most γ founding lineages is

$$M_\gamma(T_A) = \sum_{i=1}^{\gamma} M_i(T_A). \tag{39}$$

M_γ is an increasing function of T_A, and let M_γ^{-1} be its inverse function. The waiting time T_α until a fraction $1 - \alpha$ of the genome of population A has at most γ founding lineages is then

$$T_\alpha = M_\gamma^{-1}(1 - \alpha). \tag{40}$$

Table 1 shows the waiting times for different values of α and γ. The waiting time until a large fraction of the genome of population A has one founding lineage is much longer than the time until the population has at most two founding lineages. Increasing the allowed number of founding lineages decreases the waiting time considerably. For example, the waiting time is ≈9.904 units of coalescent time until one founding lineage contributed to 99.99% of the genome of population A (the corresponding waiting time for three founding lineages is ≈1.860). This does not mean that the genome of population A was founded by the same founding individual; it just means that for 99.99% of

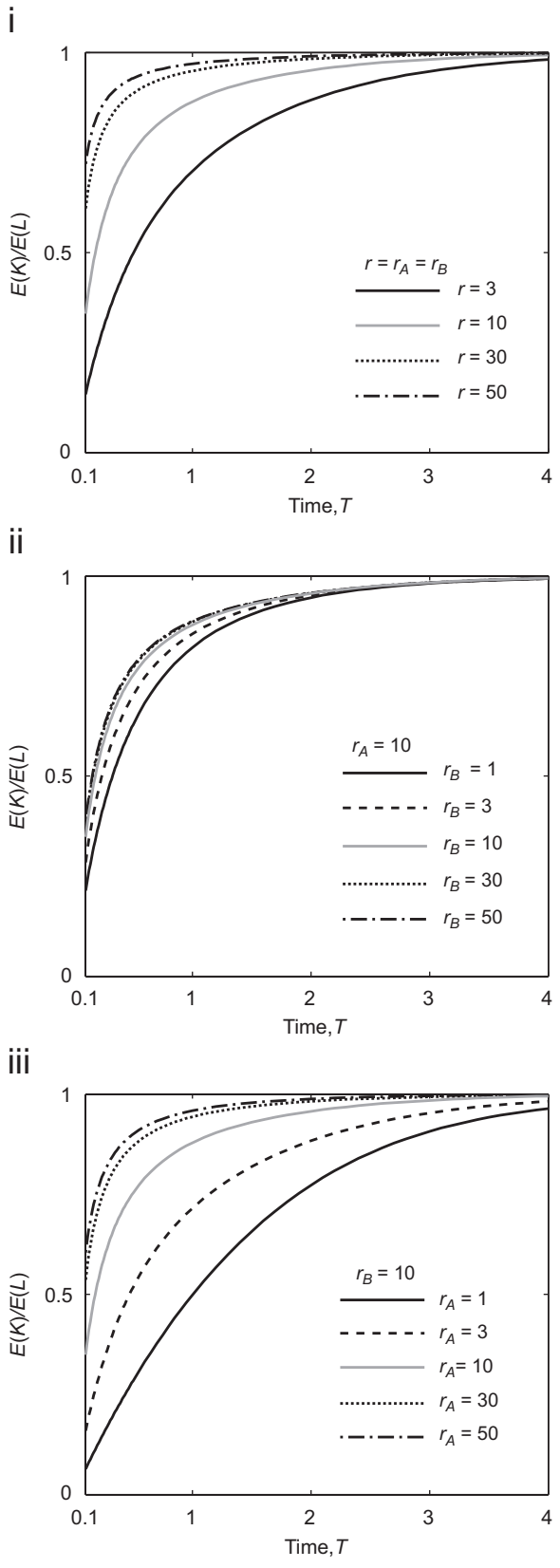

Fig. 12. The ratio of the expected number of founding lineages for a particular sample size, $E(K)$, and the expected number of founding lineages in the whole population, $E(L)$, as functions of T ($T = T_A = T_B$) when $N_A = \infty$ and $N_B = \infty$. The expected number of founding lineages was obtained from Eqs. (34) and (38). (i) $r = r_A = r_B$, (ii) $r_A = 10$, (iii) $r_B = 10$.

Table 1
Waiting times T_α until a fraction $1 - \alpha$ of the genome of population A has at most γ founding lineages

α	Maximal no. founding lineages, γ				
	1	2	3	5	10
0.5	1.440	0.737	0.498	0.302	0.153
0.05	3.701	1.491	0.894	0.479	0.212
0.01	5.301	1.997	1.144	0.583	0.244
0.001	7.601	2.735	1.500	0.725	0.286
10^{-4}	9.904	3.487	1.860	0.866	0.325
10^{-5}	12.206	4.246	2.225	1.007	0.364
10^{-6}	14.509	5.010	2.595	1.149	0.403
10^{-7}	16.811	5.776	2.969	1.292	0.441

The population sizes of both populations were assumed to be equal ($\beta = 1$). The waiting times are measured in coalescent units and were computed using Eq. (40).

the genome, all individuals of the current population have one common ancestor before they coalesce with any lineage from population B. It is a well-known property of the coalescent model that the number of lineages decreases rapidly with time when the number of lineages is large (e.g. Nordborg, 2001). Thus, the number of A lineages decreases rapidly down to a small number in a relatively short amount of time, and this is the reason for short waiting times when $\gamma \gg 1$.

Consider a genome (of population A) of finite size F base pairs approximated by u unlinked subunits of length l so that $F = ul$. The probability that each subunit u of the genome of population A has $\leqslant \gamma$ founding lineages is given by Eq. (39). The probability $1 - \alpha$ that the whole genome has at most γ founding lineages then equals $[M_\gamma(T_\alpha)]^u$, and the waiting time is

$$T_\alpha = M_\gamma^{-1}[(1-\alpha)^{1/u}] \approx M_\gamma^{-1}(e^{-\alpha/u}). \tag{41}$$

Suppose that the individuals in population A have a genome size of 3000 megabases (Mb). If we assume that sites separated by 0.2 Mb have independent genealogies, then the genome has $u = 15,000$ independent units. The maximum number of founding lineages, genome wide, in a finite genome can then be computed using Eq. (41). The waiting time is 13.305 until the probability is 0.95 that only one founding lineage (though not necessarily from the same ancestral individual) has contributed to each unit of the genome of population A. The waiting times (until the probability is 0.95) for larger values of γ are much shorter: for example when $\gamma = 2$, $T_\alpha = 4.610$, and when $\gamma = 10$, $T_\alpha = 0.383$.

The waiting time until the probability is 0.95 that only one founding lineage has contributed to each unit of the genome of population A is the same as the time until population A is monophyletic across 95% of the genome (but population B does not have to be monophyletic).

Rosenberg (2003) computed the waiting time until two species are reciprocally monophyletic with probability 0.95 using the same assumptions about genome size and independent units of the genome. The waiting time until reciprocal monophyly, $T_\alpha = 13.972$ (Rosenberg, 2003), is, as expected, longer than the time to monophyly of only one population, $T_\alpha = 13.305$. However, the difference is very small. On the other hand, the waiting time until a population has two founding lineages, $T_\alpha = 4.610$, is only about a third of the waiting time until the population has one founding lineage.

4. Discussion

Using a neutral coalescent model of two populations, where an offshoot population is assumed to be founded from an ancestral population, we have derived an expression for the probability distribution of the number K of genetic founding lineages of the offshoot population. We have also obtained a recursion expression that enables rapid computation of the probabilities for more than one founding lineage. The sample size heavily affects the probability distribution of K when the divergence time is small, whereas if the divergence time is large, the sample size has a relatively small effect on the distribution. We have also found that for estimating the number of founding lineages of an offshoot population when the sizes of the two populations are similar, if one is forced to choose between increasing the sample size either in population A or in population B, one should choose to increase the sample size of the former. For a large fraction of the loci in the whole offshoot population to each have one founding lineage, the two populations have to be separated for a long time. A much shorter divergence time is required for the same fraction of the loci in the whole offshoot population to have some small number (>1) of founding lineages.

In many biological scenarios, new populations have been founded by relatively few individuals. The contributions of most of the founding individuals will quickly be lost over time due to drift unless the population is growing rapidly. The hope of estimating the number of founders in a population with a small present-day population size is often not great, simply because the contributions of many founding individuals may have been lost (see for example Fig. 11v). In many cases, however, we may be more interested in the number of founding individuals that *contributed* to the population. If that is the case, then we could instead ask: how many founding individuals contributed to the present-day population, that is, what is the number of founding *lineages* of the present-day population? Attempting to answer this question is potentially more straightforward using methods based on the results of this article.

4.1. Comparison with the number of ancestral lineages

It is of interest to compare the probability distribution of the number of founding lineages with Eq. (8), the distribution of the number of ancestors of a sample at a given point in the past. Both the number of founding lineages and the number of ancestral lineages decrease as T increases. When T is large, the distribution of the number of founding lineages is driven by events that occur within population A, that is, by coalescence within the population

Table 2
The probability $P(k|10,10,T)$ of k founding lineages from a sample of 10 lineages in each population after T units of time, and the probability $g_{10,k}(T)$ that 10 lineages have k ancestors T units of time in the past

T	Probability	k									
		1	2	3	4	5	6	7	8	9	10
0.0	$P(k\|10,10,T)$	–	–	0.001	0.005	0.027	0.092	0.210	0.307	0.261	0.098
	$g_{10,k}(T)$	–	–	–	–	–	–	–	–	–	1.000
0.1	$P(k\|10,10,T)$	–	0.008	0.049	0.158	0.280	0.282	0.163	0.052	0.008	0.001
	$g_{10,k}(T)$	–	–	0.002	0.022	0.099	0.239	0.317	0.227	0.081	0.011
0.5	$P(k\|10,10,T)$	0.088	0.363	0.385	0.142	0.021	0.001	–	–	–	–
	$g_{10,k}(T)$	0.025	0.205	0.405	0.279	0.077	0.009	–	–	–	–
1.0	$P(k\|10,10,T)$	0.372	0.502	0.119	0.007	–	–	–	–	–	–
	$g_{10,k}(T)$	0.228	0.526	0.222	0.024	0.001	–	–	–	–	–
2.0	$P(k\|10,10,T)$	0.769	0.226	0.005	–	–	–	–	–	–	–
	$g_{10,k}(T)$	0.675	0.312	0.013	–	–	–	–	–	–	–
4.0	$P(k\|10,10,T)$	0.970	0.030	–	–	–	–	–	–	–	-
	$g_{10,k}(T)$	0.955	0.045	–	–	–	–	–	–	–	–

$P(k|10,10,T)$ was computed from (33) and $g_{10,k}(T)$ was computed from (8). Probabilities below 0.0005 are denoted by –.

as time passes. However, as can be seen from the difference between the probabilities $P(k|10,10,T)$ and $g_{10,k}(T)$ in Table 2, for small and moderate values of T, the coalescent process within population A has only a minor effect on the distribution of the number of founding lineages of population A. Instead, the combinatorial nature of interspecific coalescence events in the ancestral population, as described in this article, tends to dominate. Thus, in many cases, the behavior of the number of founding lineages is quite different from the behavior of the number of ancestors.

Because many of the events affecting the properties of the former distribution occur within the ancestral population, that is, above T, events within the ancestral population are important to take into account in order to avoid misleading conclusions. For example, as was pointed out by Hudson and Turelli (2003), Palumbi et al. (2001) in defining the "three-times rule" used the number of ancestral lineages of a population in a scenario where accounting for interpopulation coalescence in the ancestral population would have been more appropriate.

4.2. Sample size and experimental design

The expected number of founding lineages of population A increases approximately linearly with a symmetric sample size when no time has passed since the divergence of populations A and B (Fig. 8i). However, when either sample size is fixed, the linear relationship of the number of founding lineages and sample size is lost (Fig. 6). Note that for any value of T, an estimate of the expected number of founding lineages of the whole population based on a subsample of the two populations may underestimate the true value, regardless of sampling strategy.

If the divergence time is larger than zero, assuming that the population sizes are the same for the two populations, the expected number of founding lineages of a sample from population A decreases quickly as the divergence time increases (Fig. 8ii). Given a certain divergence time, what sampling strategy would then be best able to capture as much information as possible about the number of founding lineages of population A? From Fig. 8ii, when $T \gtrsim 1.5$, the expected number of founding lineages is close to 1, and to estimate the expected number of founding lineages of the whole population, there is no need for large sample sizes. Note, however, that the divergence time has to be much longer for $P(K=1)$ to exceed 0.95 (Fig. 7).

For moderate divergence times ($0.1 \leqslant T \leqslant 1$), sample sizes of 50, or even 30 individuals from each of the two populations, give good estimates of the expected number of founding lineages of the whole A population (Fig. 12i). Suppose that one were restricted to a fixed sample size—say 60 individuals from both populations—for estimating the number of founding lineages of population A. Suppose also that the two populations have similar population sizes and that divergence time is somewhere between 0.1 and 1 units of time. A sample of equal numbers of individuals (30 from each population, dotted line in Fig. 12i) is only marginally preferable to a sample of 50 individuals from population A and 10 from population B (dotted-dashed line in Fig. 12iii). On the other hand, a sample of 10 individuals from population A and 50 from population B (dotted-dashed line in Fig. 12ii) is less informative than either of the two previous sampling strategies. Thus, if one were to choose between increasing the sample size in either population A or population B, one should choose the former.

4.3. Comparing nonrecombining haploid DNA and autosomal DNA

Nonrecombining haploid DNA, such as mitochondrial DNA, chloroplast DNA and Y-chromosomal DNA, has been used extensively in demographic studies of numerous species. However, inferences from nonrecombining haploid DNA are in many cases not expected to agree with those from autosomal DNA. For instance, under neutrality, in a diploid species with two sexes and with equal distributions of the number of offspring for males and females, coalescent time elapses four times faster in uniparentally inherited haploid DNA than in autosomal DNA. As described above, the number of founding individuals that contributed to a set of present-day individuals is heavily dependent on time since divergence (see e.g. Figs. 7 and 8).

For example, if two units of time have passed for the Y-chromosome, only 1/2 unit has passed for autosomes. In this case we expect 1.29 founding lineages for the Y-chromosome and, on average, 3.53 founding lineages for autosomal loci (assuming infinite samples of both populations in Eq. (34); see also Fig. 8ii). The probability is 0.716 that one founding individual contributed to the Y-chromosome, 0.274 that two founding individuals contributed, and 0.010 that more than two founding individuals contributed. For an autosomal locus, the probability distribution is quite different: the probabilities of 1, 2, 3, 4, 5, 6 and >6 founding lineages are 0.020, 0.148, 0.337, 0.317, 0.142, 0.032 and 0.004, respectively. Thus, this type of discrepancy should be considered when drawing conclusions about the demographic history of a species or population based on different types of DNA.

Acknowledgments

We thank Michael Blum and two anonymous reviewers for helpful comments on a previous version of this manuscript. This work was supported by a Burroughs Wellcome Career Award in the Biomedical Sciences.

Appendix A

Identity 1: For positive integers m and n (Rosenberg, 2003)

$$\sum_{k=0}^{n} \frac{k\binom{n}{k}}{\binom{m+n-1}{k}} = \frac{n(m+n)}{m(m+1)}.$$

Identity 2: For positive integers n, a, b with $a \geqslant 2$ and $b \geqslant a$

$$\sum_{k=0}^{n} k^2(a-1+k)\binom{a-2+k}{k}\binom{b-a+n-k}{n-k} = \frac{a(a-1)(an-a+n+b+1)n}{(b+2)(b+1)}\binom{b+n}{n}.$$

Proof. By noting that $(a-1+k)({}_{a-2+k}C_k) = (a-1)({}_{a-1+k}C_k)$ we see that both sides of this identity have a factor of $a-1$, so that the identity is equivalent to

$$S_2 = \sum_{k=0}^{n} k^2\binom{a-1+k}{k}\binom{b-a+n-k}{n-k} = \frac{a(an-a+n+b+1)n}{(b+2)(b+1)}\binom{b+n}{n}. \tag{42}$$

We now provide a proof for Eq. (42). The proof utilizes two very useful combinatorial identities (see e.g. Graham et al., 1994), an absorption identity,

$$r\binom{r-1}{s-1} = s\binom{r}{s} \quad \text{integers } r, s > 0 \tag{43}$$

and a version of the Vandermonde convolution (Graham et al., 1994, Eq. (5.26)),

$$\sum_{k=0}^{l}\binom{l-k}{m}\binom{q+k}{n} = \binom{l+q+1}{m+n+1}, \quad \text{integers } l, m \geqslant 0, \text{ integers } n \geqslant q \geqslant 0. \tag{44}$$

Replacing ${}_{b-a+n-k}C_{b-a}$ with $Y(k)$,

$$S_2 = \sum_{k=0}^{n} k^2\binom{a-1+k}{k}\binom{b-a+n-k}{n-k} = \sum_{k=0}^{n} k^2\binom{a-1+k}{a-1}Y(k). \tag{45}$$

We use a simple trick of rewriting k as $-a+(a+k)$ and splitting the sum S_2 into two sums:

$$\begin{aligned} S_2 &= \sum_{k=0}^{n} k[-a+(a+k)]\binom{a-1+k}{a-1}Y(k) \\ &= -a\sum_{k=0}^{n} k\binom{a-1+k}{a-1}Y(k) + \sum_{k=0}^{n} k(a+k)\binom{a-1+k}{a-1}Y(k) \\ &= -a\sum_{k=0}^{n} k\binom{a-1+k}{a-1}Y(k) + \sum_{k=0}^{n} ka\binom{a+k}{a}Y(k), \end{aligned}$$

where the last equality follows from Eq. (43). We now use the same trick as above on these two sums, obtaining

$$\begin{aligned} S_2 &= -a\sum_{k=0}^{n}[-a+(a+k)]\binom{a-1+k}{a-1}Y(k) + a\sum_{k=0}^{n}[-a-1+(a+k+1)]\binom{a+k}{a}Y(k) \\ &= a^2\sum_{k=0}^{n}\binom{a-1+k}{a-1}Y(k) - a^2\sum_{k=0}^{n}\binom{a+k}{a}Y(k) - a(a+1)\sum_{k=0}^{n}\binom{a+k}{a}Y(k) \\ &\quad + a(a+1)\sum_{k=0}^{n}\binom{a+k+1}{a+1}Y(k). \end{aligned}$$

We then use Eq. (44) on each of the four sums and obtain

$$\begin{aligned}S_2 &= a^2\sum_{k=0}^{n}\binom{a-1+k}{a-1}\binom{b-a+n-k}{b-a} - a^2\sum_{k=0}^{n}\binom{a+k}{a}\binom{b-a+n-k}{b-a}\\ &\quad - a(a+1)\sum_{k=0}^{n}\binom{a+k}{a}\binom{b-a+n-k}{b-a} + a(a+1)\sum_{k=0}^{n}\binom{a+k+1}{a+1}\binom{b-a+n-k}{b-a}\\ &= a^2\binom{b+n}{b} - a^2\binom{b+n+1}{b+1} - a(a+1)\binom{b+n+1}{b+1} + a(a+1)\binom{b+n+2}{b+2}\\ &= \frac{an\binom{b+n}{b}}{(b+1)(b+2)}(an-a+n+b+1). \qquad \square\end{aligned}$$

Identity 3: For real c and positive integers n, a, b with $b \geqslant a$

$$\begin{aligned}&\sum_{k=0}^{n}k^2(a+k+ck-c)\binom{a-1+k}{k}\binom{b-a+n-k}{n-k}\\ &\quad = \frac{(a+1)(b+n)!}{(b+3)!(n-1)!}(abn-ab+n^2a-3anc+2ac+an^2c-a+1+2bnc+2b-2c+3n+b^2+2n^2-2bc+3bn+2n^2c).\end{aligned} \tag{46}$$

Proof. The desired sum can be written as $(c+1)S_3+(a-c)S_2$ where S_2 was obtained in Identity 2.

$$S_3 = \sum_{k=0}^{n}k^3\binom{a-1+k}{k}\binom{b-a+n-k}{n-k} = \sum_{k=0}^{n}k^3\binom{a-1+k}{a-1}\binom{b-a+n-k}{b-a}. \tag{47}$$

We derive an expression for S_3 using the same approach as was used for S_2. Then Identity 3 is obtained by combining sums of the form S_2 and S_3. As before, let ${}_{b-a+n-k}C_{b-a} = Y(k)$. Then

$$\begin{aligned}S_3 &= \sum_{k=0}^{n}k^3\binom{a-1+k}{a-1}Y(k) = a^2\sum_{k=0}^{n}k\binom{a-1+k}{a-1}Y(k) - a^2\sum_{k=0}^{n}k\binom{a+k}{a}Y(k)\\ &\quad - a(a+1)\sum_{k=0}^{n}k\binom{a+k}{a}Y(k) + a(a+1)\sum_{k=0}^{n}k\binom{a+k+1}{a+1}Y(k)\\ &= -a^3\sum_{k=0}^{n}\binom{a-1+k}{a-1}Y(k) + a^3\sum_{k=0}^{n}\binom{a+k}{a}Y(k) + a^2(a+1)\sum_{k=0}^{n}\binom{a+k}{a}Y(k)\\ &\quad - a^2(a+1)\sum_{k=0}^{n}\binom{a+1+k}{a+1}Y(k) + a(a+1)^2\sum_{k=0}^{n}\binom{a+k}{a}Y(k)\\ &\quad - a(a+1)^2\sum_{k=0}^{n}\binom{a+1+k}{a+1}Y(k) - a(a+1)(a+2)\sum_{k=0}^{n}\binom{a+1+k}{a+1}Y(k)\\ &\quad + a(a+1)(a+2)\sum_{k=0}^{n}\binom{a+2+k}{a+2}Y(k).\end{aligned} \tag{48}$$

By using Eq. (44) on each sum in Eq. (48) we then obtain

$$\begin{aligned}S_3 &= -a^3\binom{b+n}{b} + a^3\binom{b+n+1}{b+1} + a^2(a+1)\binom{b+n+1}{b+1}\\ &\quad - a^2(a+1)\binom{b+n+2}{b+2} + a(a+1)^2\binom{b+n+1}{b+1} - a(a+1)^2\binom{b+n+2}{b+2}\\ &\quad - a(a+1)(a+2)\binom{b+n+2}{b+2} + a(a+1)(a+2)\binom{b+n+3}{b+3}\\ &= \frac{an\binom{b+n}{b}[a^2(n-2)(n-1)+3a(n-1)(b+n+1)+(b+2n+1)(b+n+1)]}{(b+1)(b+2)(b+3)}. \qquad \square\end{aligned} \tag{49}$$

Appendix B. Proof of Eq. (28)

Eq. (27) gives

$$P(K=q_A|q_A,q_B)=\frac{2^{q_A}(q_B-1)!}{\binom{q_A+q_B}{q_B}(q_A+q_B-1)!}\sum_{m_1=1}^{q_B} m_1\cdots\sum_{m_i=1}^{m_{j-1}} m_i\cdots\sum_{m_k=1}^{m_{k-1}} m_k.$$

$P(K=q_A|q_A,q_B)$ can be written as the quotient $f(q_B)/g(q_B)$ of two polynomials in q_B. The numerator $f(q_B)$, based on the sums on the right side of (27), has leading term $q_B^{2q_A}/(2^{q_A}q_A!)$ since

$$\begin{aligned}
f(q_B)&=\sum_{m_1=1}^{q_B} m_1\cdots\sum_{m_i=1}^{m_{j-1}} m_i\cdots\sum_{m_k=1}^{m_{k-1}} m_k=\sum_{m_1=1}^{q_B} m_1\cdots\sum_{m_i=1}^{m_{j-1}} m_i\cdots\sum_{m_{k-1}=1}^{m_{k-2}} m_{k-1}\times\frac{1}{2}(m_{k-1}+1)m_{k-1}\\
&=\sum_{m_1=1}^{q_B} m_1\cdots\sum_{m_i=1}^{m_{j-1}} m_i\cdots\sum_{m_{k-1}=1}^{m_{k-2}}\frac{1}{2}(m_{k-1}^3+m_{k-1}^2)\\
&=\sum_{m_1=1}^{q_B} m_1\cdots\sum_{m_i=1}^{m_{j-1}} m_i\cdots\sum_{m_{k-1}=1}^{m_{k-2}}\frac{1}{2}m_{k-1}^3+Q(m_{k-1}^2)=\sum_{m_1=1}^{q_B} m_1\cdots\sum_{m_i=1}^{m_{j-1}} m_i\cdots\sum_{m_{k-2}=1}^{m_{k-3}}\frac{1}{2}\frac{1}{4}m_{k-2}^4+Q(m_{k-2}^3)\\
&=\frac{1}{2}\frac{1}{4}\cdots\frac{1}{2k}q_B^{2k}+Q(q_B^{2k-1})=\frac{q_B^{2k}}{2^k k!}+Q(q_B^{2k-1})=\frac{q_B^{2q_A}}{2^{q_A}q_A!}+Q(q_B^{2q_A-1}),
\end{aligned}$$

where $Q(n^r)$ is a polynomial in n of degree r or less. Here we have used the fact that as a polynomial in n, $\sum_{i=1}^n i^r$ has leading term $n^{r+1}/(r+1)$ (see e.g. Conway and Guy, 1996, p. 106). The denominator $g(q_B)$, based on the first part of Eq. (27), has leading term $q_B^{2q_A}$ since

$$\begin{aligned}
\frac{1}{g(q_B)}&=\frac{2^{q_A}(q_B-1)!}{\binom{q_A+q_B}{q_B}(q_A+q_B-1)!}=\frac{2^{q_A}(q_B-1)!q_A!q_B!}{(q_A+q_B)!(q_A+q_B-1)!}\\
&=2^{q_A}q_A!\frac{1}{\frac{(q_A+q_B-1)!}{(q_B-1)!}\frac{(q_A+q_B)!}{q_B!}}=2^{q_A}q_A!\frac{1}{q_B^{2q_A}+R(q_B^{2q_A-1})},
\end{aligned}$$

where $R(n^r)$ is a polynomial in n of degree r or less. We can now use L'Hôpital's rule for evaluating $f(q_B)/g(q_B)$. As $q_B\to\infty$, we obtain

$$P(K=q_A|q_A,q_B)=\frac{f(q_B)}{g(q_B)}=\left[\frac{q_B^{2q_A}}{2^{q_A}q_A!}+Q(q_B^{2q_A-1})\right]\left[\frac{2^{q_A}q_A!}{q_B^{2q_A}+R(q_B^{2q_A-1})}\right]\to 1.$$

References

Ainouche, M.L., Baumel, A., Salmon, A., 2004. *Spartina anglica* C. E. Hubbard: a natural model system for analysing early evolutionary changes that affect allopolyploid genomes. Biol. J. Linn. Soc. 82, 475–484.

Bonatto, S.L., Salzano, F.M., 1997. A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data. Proc. Natl. Acad. Sci. USA 94, 1866–1871.

Brown, M.D., Hosseini, S.H., Torroni, A., Bandelt, H.J., Allen, J.C., Schurr, T.G., Scozzari, R., Cruciani, F., Wallace, D.C., 1998. mtDNA haplogroup X: an ancient link between Europe/Western Asia and North America? Am. J. Hum. Genet. 63, 1852–1861.

Conway, J.H., Guy, R.K., 1996. The Book of Numbers. Springer, New York.

Crawford, M.H., 1998. The Origins of Native Americans. Cambridge University Press, Cambridge, UK.

Edwards, A.W.F., 1970. Estimation of the branch points of a branching diffusion process. J. R. Stat. Soc. Ser. B 32, 155–174.

Evans, B.J., Kelley, D.B., Melnick, D.J., Cannatella, D.C., 2005. Evolution of RAG-1 in polyploid clawed frogs. Mol. Biol. Evol. 22, 1193–1207.

Graham, R.L., Knuth, D.E., Patashnik, O., 1994. Concrete Mathematics: A Foundation for Computer Science, second ed. Addison Wesley, Boston, USA.

Grant, P.R., Grant, B.R., Petren, K., 2001. A population founded by a single pair of individuals: establishment, expansion, and evolution. Genetica 112–113, 359–382.

Harter, A.V., Gardner, K.A., Falush, D., Lentz, D.L., Bye, R.A., Rieseberg, L.H., 2004. Origin of extant domesticated sunflowers in eastern North America. Nature 430, 201–205.

Hedrick, P.W., 2000. Genetics of Populations, second ed. Jones and Bartlett, London, UK.

Hey, J., 2005. On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biol. 3, 965–975.

Hudson, R.R., Turelli, M., 2003. Stochasticity overrules the three-times rule: genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. Evolution 57, 182–190.

Jakobsson, M., Hagenblad, J., Tavaré, S., Säll, T., Halldén, C., Lind-Halldén, C., Nordborg, M., 2006. A unique recent origin of the allotetraploid species *Arabidopsis suecica*: evidence from nuclear DNA markers. Mol. Biol. Evol. 23, 1217–1231.

Luikart, G., Gielly, L., Excoffier, L., Vigne, J.D., Bouvet, J., Taberlet, P., 2001. Multiple maternal origins and weak phylogeographic structure in domestic goats. Proc. Natl. Acad. Sci. USA 98, 5927–5932.

Matsuoka, Y., Vigouroux, Y., Goodman, M.M., Sanchez, G.J., Buckler, E., Doebley, J., 2002. A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. USA 99, 6080–6084.

Mulligan, C.J., Hunley, K., Cole, S., Long, J.C., 2004. Population genetics, history, and health patterns in native Americans. Annu. Rev. Genomics Hum. Genet. 5, 295–315.

Nei, M., Maruyama, T., Chakraborty, R., 1975. The bottleneck effect and genetic variability in populations. Evolution 29, 1–10.

Nordborg, M., 2001. Coalescent theory. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 179–212 (Chapter 7).

Palumbi, S.R., Cipriano, F., Hare, M.P., 2001. Predicting nuclear gene coalescence from mitochondrial data: the three-times rule. Evolution 55, 859–868.

Rosenberg, N.A., 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61, 225–247.

Rosenberg, N.A., 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57, 1465–1477.

Rosenberg, N.A., 2006. The mean and variance of the numbers of *r*-pronged nodes and *r*-caterpillars in Yule-generated genealogical trees. Ann. Comb. 10, 129–146.

Schurr, T.G., Sherry, S.T., 2004. Mitochondrial DNA and Y chromosome diversity and the peopling of the Americas: evolutionary and demographic evidence. Am. J. Hum. Biol. 16, 420–439.

Schurr, T.G., Ballinger, S.W., Gan, Y.Y., Hodge, J.A., Merriwether, D.A., Lawrence, D.N., Knowler, W.C., Weiss, K.M., Wallace, D.C., 1990. Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages. Am. J. Hum. Genet. 46, 613–623.

Segraves, K.A., Thompson, J.N., Soltis, P.S., Soltis, D.E., 1999. Multiple origins of polyploidy of the geographic structure of *Heuchera grossularifolia*. Mol. Ecol. 8, 253–262.

Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.

Smith, D.G., Malhi, R.S., Eshleman, J., Lorenz, J.G., Kaestle, F.A., 1999. Distribution of mtDNA haplogroup X among Native North Americans. Am. J. Phys. Anthropol. 110, 271–284.

Soodyall, H., Jenkins, T., Mukherjee, A., du Toit, E., Roberts, D.F., Stoneking, M., 1997. The founding mitochondrial DNA lineages of Tristan da Cunha Islanders. Am. J. Phys. Anthropol. 104, 157–166.

Takahata, N., 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122, 957–966.

Tavaré, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26, 119–164.

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

Wright, S.I., Bi, I.V., Schroeder, S.G., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., 2005. The effects of artificial selection on the maize genome. Science 308, 1310–1314.