Coalescence-Time Distributions in a Serial Founder Model of Human Evolutionary History

Michael DeGiorgio,* James H. Degnan,† and Noah A. Rosenberg*,‡
*Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, †Department of Mathematics and Statistics, University of Canterbury, Christchurch 8140 New Zealand, and ‡Department of Biology, Stanford University, Stanford, California 94305

ABSTRACT Simulation studies have demonstrated that a variety of patterns in worldwide genetic variation are compatible with the trends predicted by a serial founder model, in which populations expand outward from an initial source via a process in which new populations contain only subsets of the genetic diversity present in their parental populations. Here, we provide analytical results for key quantities under the serial founder model, deriving distributions of coalescence times for pairs of lineages sampled either from the same population or from different populations. We use these distributions to obtain expectations for coalescence times and for homozygosity and heterozygosity values. A predicted approximate linear decline in expected heterozygosity with increasing distance from the source population reproduces a pattern that has been observed both in human genetic data and in simulations. Our formulas predict that populations close to the source location have lower between-population gene identity than populations far from the source, also mirroring results obtained from data and simulations. We show that different models that produce similar declining patterns in heterozygosity generate quite distinct patterns in coalescence-time distributions and gene identity measures, thereby providing a basis for distinguishing these models. We interpret the theoretical results in relation to their implications for human population genetics.

EQUILIBRIUM population structure models, which assume that the rules specifying the evolution of alleles within and among populations do not change with time, have achieved much success in describing genetic variation. Although equilibrium models are convenient for obtaining analytical results that can be used to test hypotheses and predict patterns of genetic variation, nonequilibrium models often provide more realistic representations of patterns that occur in real populations. Nonequilibrium models assume that the rules specifying the evolution of alleles change as a function of time. In nonequilibrium models, however, with some exceptions (*e.g.*, Takahata *et al.* 1995; Wakeley 1996a, b,c; Jesus *et al.* 2006; Efromovich and Kubatko 2008), analytical formulas have been relatively scarce because model complexity can make them difficult to obtain.

Recently, a nonequilibrium structured population model, the “serial founder model,” has been proposed for describing the colonization of the world by modern humans (Ramachandran *et al.* 2005). The colonization process in this model starts with a single source population. The source population sends a subset of its individuals to migrate outward and found a new population. This newly founded population has a small size at its founding and subsequently expands to a larger size. After the expansion, it then sends out migrants to form the next population. The founding process is iterated until K populations have been founded. The appeal of this model is that using both forward (Ramachandran *et al.* 2005; Deshpande *et al.* 2009) and backward (coalescent) simulations (DeGiorgio *et al.* 2009; Hunley *et al.* 2009), it has been successful in describing observed patterns of human genetic variation, such as the decline in expected heterozygosity observed with increasing geographic distance from a putative African source location.

In addition to the initial serial founder model of Ramachandran *et al.* (2005), a variety of models that contain the geographic expansions and bottlenecks characteristic of the serial founder model have recently been studied (Austerlitz *et al.* 1997; Le Corre and Kremer 1998; Edmonds *et al.* 2004; Ray *et al.* 2005; Klopfstein *et al.* 2006; Liu *et al.* 2006; Excoffier and Ray 2008; Hallatschek and Nelson

doi: 10.1534/genetics.111.129296
Manuscript received April 6, 2011; accepted for publication July 7, 2011
Available freely online through the author-supported open access option.
Corresponding author: Center for Computational Medicine and Bioinformatics, University of Michigan, 2017 Palmer Commons, 100 Washtenaw Ave., Ann Arbor, MI 48109-2218. E-mail: degiormi@umich.edu

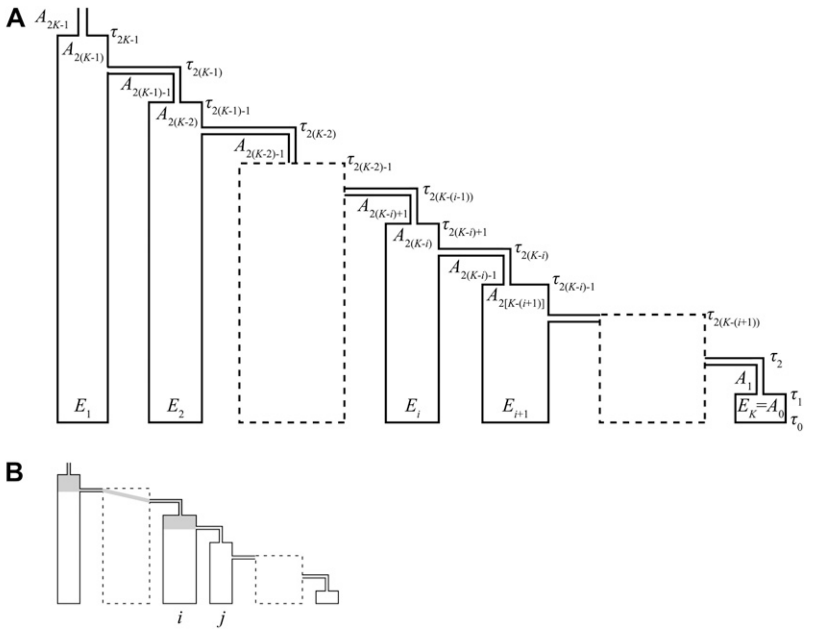

Figure 1 Serial founder model. (A) Serial founder model with K extant and $2K$ ancestral populations. At time τ_{2K-1}, ancestral population A_{2K-1} expands to a larger size to form ancestral population $A_{2(K-1)}$. Next, at time $\tau_{2(K-1)}$, ancestral population $A_{2(K-1)}$ splits to form extant population E_1 and newly founded ancestral population $A_{2(K-1)-1}$. At time $\tau_{2(K-1)-1}$, population $A_{2(K-1)-1}$ expands to a larger size to form ancestral population $A_{2(K-2)}$. In general, at time $\tau_{2(K-i)}$, ancestral population $A_{2(K-i)}$ splits into extant population E_i and newly founded ancestral population $A_{2(K-i)-1}$. At time $\tau_{2(K-i)-1}$, ancestral population $A_{2(K-i)-1}$ expands to a larger size to form ancestral population $A_{2[K-(i+1)]}$. (B) Scenario in which lineages are sampled from populations E_i and E_j, $i \leq j$ ($i < j$ is shown here). Regions in which coalescence can occur are shaded.

2008; DeGiorgio *et al.* 2009; Deshpande *et al.* 2009; Hunley *et al.* 2009). Among formulations with a one-dimensional geographic structure, some models (*e.g.*, Austerlitz *et al.* 1997; Deshpande *et al.* 2008) allow migration after the initial founding of populations and assume that once a population is founded, it logistically grows to its carrying capacity. When carrying capacity is reached, or shortly thereafter, migrants exit the population to found the next population. Other models (*e.g.*, DeGiorgio *et al.* 2009) do not permit migration after populations are founded and assume that population growth is instantaneous. In these models, after a population is founded, it experiences a small size for some length of time before instantaneously expanding to a larger size. For the former class of models, Austerlitz *et al.* (1997) presented recursions to generate the distribution of coalescence times for pairs of lineages sampled either from the same population or from different populations. These equations can then be used to calculate geographic patterns in summary statistics such as gene diversity and F_{ST}. For the latter class, DeGiorgio *et al.* (2009) and Hunley *et al.* (2009) approached similar problems using simulations. The relative simplicity of the population growth and migration assumptions in this latter group of models, however, potentially permits explicit formulas, rather than recursions or simulations, to be investigated.

Here, generalizing the coalescent-based version of the serial founder model as formulated by DeGiorgio *et al.* (2009), we provide an analytical distribution of the coalescence time for a pair of lineages at a randomly selected locus, along with corresponding expected coalescence times, expected homozygosity values, and F_{ST} values. In this nonequilibrium model, we show that the decrease in expected heterozygosity and the corresponding increase in homozygosity with increasing distance from the source population can be predicted analytically. We then provide analytical results for the expected identity for two alleles drawn randomly from a given pair of populations, and we find that the qualitative patterns produced by the formulas closely match those observed from human genetic data and the simulations of Hunley *et al.* (2009). Furthermore, we discuss how our results can be used to obtain analytical formulas for summary statistics for an archaic serial founder model, for the nested-regions model of Hunley *et al.* (2009), and for the instantaneous divergence model of DeGiorgio *et al.* (2009). Our new analytical formulas on within-population gene diversity, between-population gene identity, and pairwise F_{ST} motivate an analysis of empirical trends in these summary statistics in worldwide human genetic data. Because a serial founder process is largely consistent with worldwide patterns of human genetic variation, the analytical results presented here are useful both for generating and for testing hypotheses about human origins.

Serial Founder Model

In this section, we begin by formally defining the serial founder model. This model was used in a simulation of DeGiorgio *et al.* (2009), and here, we provide a more complete generalization. We obtain the probability density of coalescence times for two lineages sampled under the model. Utilizing this density, we obtain mth moments of coalescence times, mth moments of homozygosities, and F_{ST} values between pairs of populations.

Model

We formulate the serial founder model in a coalescent setting. A diagram of the model appears in Figure 1A. Our generic formulation contains a sequence of bottlenecks in which bottleneck sizes, population sizes, bottleneck lengths, and the times for the population founding events are allowed to vary. The model considers K extant populations, denoted $E_1, E_2, \ldots, E_K$. For $i < j$, the founding of extant population E_i took place at least as far back in time as that of extant population E_j. The model has $2K$ ancestral populations, denoted $A_0, A_1, \ldots, A_{2K-1}$. For $i < j$, the founding of ancestral population A_j took place at least as far back in time as that of ancestral population A_i. N_i denotes the size of ancestral population A_i, $i = 0, 1, \ldots, 2K - 1$. Note that for $i = 1, 2, \ldots, K$, the size of extant population E_i is equal to $N_{2(K-i)}$, which also is the size of ancestral population $A_{2(K-i)}$. Time is measured in generations, and the present has time $\tau_0 = 0$.

Forward in time, ancestral population A_{2K-1} expands to a larger size at time τ_{2K-1} to create ancestral population $A_{2(K-1)}$, the population directly ancestral to the source population E_1. At time $\tau_{2(K-1)}$, ancestral population $A_{2(K-1)}$ splits into extant population E_1 and ancestral population $A_{2(K-1)-1}$, a newly founded population during the time in which it experiences a small size prior to expansion. At time $\tau_{2(K-1)-1}$, ancestral population $A_{2(K-1)-1}$ expands to a larger size to form ancestral population $A_{2(K-2)}$. At time $\tau_{2(K-2)}$, ancestral population $A_{2(K-2)}$ splits to form extant population E_2 and ancestral population $A_{2(K-2)-1}$, the next founded population during its bottleneck phase. This process is iterated until extant population K has been founded. In general, at time $\tau_{2(K-i)}$, $i = 1, 2, \ldots, K - 1$, ancestral population $A_{2(K-i)}$ splits into extant population E_i and a newly founded ancestral population $A_{2(K-i)-1}$. At time $\tau_{2(K-i)-1}$, $i = 0, 1, \ldots, K - 1$, ancestral population $A_{2(K-i)-1}$ expands to a larger size to form ancestral population $A_{2[K-(i+1)]}$. Note that by construction, extant population E_K and ancestral population A_0 are the same population.

We note that several past studies (*e.g.*, Austerlitz *et al.* 1997; Ramachandran *et al.* 2005; Liu *et al.* 2006; Deshpande *et al.* 2009) utilized formulations of the serial founder model that involved logistic growth of newly founded populations, migration between neighboring populations after their initial founding, or both of these model features. In contrast, for the purpose of obtaining analytical results, our model has a mathematically simpler formulation that involves an instantaneous expansion of a newly founded population to a larger size and that does not permit migration between neighbors after founding events.

Coalescence times

In this section, we derive the probability density of coalescence times for a pair of lineages sampled under the serial founder model. We begin by deriving the probability density function $f_{ij}(t)$ for the coalescence time of a pair of lineages, one randomly sampled from extant population E_i and the other from extant population E_j (where j is not necessarily distinct from i). This function is defined piecewise over the space of possible coalescence times $t \in [0, \infty)$. Using our formula for $f_{ij}(t)$, we derive mth moments of coalescence times, from which we obtain mean pairwise coalescence times. We use the result from coalescent theory that coalescence times are exponentially distributed with a rate that is inversely proportional to the population size (Kingman 1982; Hudson 1983; Tajima 1983). Also, we use the result that the number of mutations along a genealogical branch is Poisson distributed, and because we restrict our attention to neutral loci, we separate the mutation process from the genealogical process (Tavaré 1984; Hudson 1990).

Let T_{ij} be a random variable that denotes the coalescence time for a pair of lineages, one from extant population E_i and the other from extant population E_j, with $i \le j$. If $i < j$, then the two lineages cannot coalesce until they are in the same ancestral population (*i.e.*, more ancient than $\tau_{2(K-i)}$). Suppose the two lineages are in the same population during time interval $[\tau_h, \tau_{h+1})$, where $h \ge 2(K - i)$. The probability density for coalescence at time $t \in [\tau_h, \tau_{h+1})$ is the product of the probability that the lineages do not coalesce in the more recent time intervals,

$$\exp\left[-\sum\nolimits_{\ell=2(K-i)}^{h-1} \frac{\tau_{\ell+1} - \tau_\ell}{N_\ell}\right],$$

and $(1/N_h)e^{-(t-\tau_h)/N_h}$, the probability density for coalescence at time t conditional on failure to coalesce by time τ_h.

If $i = j$, then the two lineages can also coalesce in the interval $[\tau_0, \tau_{2(K-i)})$. Suppose the two lineages exist in the same population during time interval $[\tau_0, \tau_{2(K-i)})$. The probability density for coalescence at time $t \in [\tau_0, \tau_{2(K-i)})$ in extant population E_i is $(1/N_{2(K-i)})e^{-(t-\tau_0)/N_{2(K-i)}}$. The probability that the lineages do not coalesce in time interval $[\tau_0, \tau_{2(K-i)})$ is $e^{-(\tau_{2(K-i)}-\tau_0)/N_{2(K-i)}}$ (we write τ_0 for notational consistency, but recall $\tau_0 = 0$).

For $i \le j$ and $h \in \{2(K - i), 2(K - i) + 1, \ldots, 2K - 1\}$, denote the probability that a coalescence has not occurred by time τ_h for two lineages, one from E_i and one from E_j, by

$$\Lambda_{ijh} = \exp\left(-\delta_{ij}\frac{\tau_{2(K-i)} - \tau_0}{N_{2(K-i)}} - \sum_{\ell=2(K-i)}^{h-1} \frac{\tau_{\ell+1} - \tau_\ell}{N_\ell}\right),$$

where δ_{ij} is the Kronecker delta. We then arrive at the density function for the time to coalescence of a pair of lineages sampled from extant populations E_i and E_j, $i \le j$,

$$f_{ij}(t) = \begin{cases} \delta_{ij}\dfrac{e^{-(t-\tau_0)/N_{2(K-i)}}}{N_{2(K-i)}} & , \quad 0 \le \tau_0 \le t < \tau_{2(K-i)} \\ \Lambda_{ijh}\dfrac{e^{-(t-\tau_h)/N_h}}{N_h} & , \quad \begin{array}{l}\tau_h \le t < \tau_{h+1} \\ \text{and } h = 2(K-i), \ldots, 2K-1\end{array} \\ 0 & , \quad \text{otherwise,} \end{cases} \tag{1}$$

where $\tau_{2K} = \infty$. This density for the pairwise coalescence time consists of a collection of shifted exponential distributions, each defined on a different interval.

Equipped with the density in Equation 1, we next derive mth moments for the distribution of coalescence times. We are interested primarily in the mean, but the derivation for arbitrary m is no more difficult than that for $m = 1$.

$$
\begin{aligned}
\mathbb{E}[T_{ij}^m] &= \int_0^\infty t^m f_{ij}(t)dt \\
&= \int_{\tau_0}^{\tau_{2(K-i)}} t^m \delta_{ij} \frac{e^{-(t-\tau_0)/N_{2(K-i)}}}{N_{2(K-i)}} dt \\
&\quad + \sum_{h=2(K-i)}^{2K-1} \int_{\tau_h}^{\tau_{h+1}} t^m \Lambda_{ijh} \frac{e^{-(t-\tau_h)/N_h}}{N_h} dt \\
&= \delta_{ij} \frac{e^{\tau_0/N_{2(K-i)}}}{N_{2(K-i)}} \int_{\tau_0}^{\tau_{2(K-i)}} t^m e^{-t/N_{2(K-i)}} dt \\
&\quad + \sum_{h=2(K-i)}^{2K-1} \Lambda_{ijh} \frac{e^{\tau_h/N_h}}{N_h} \int_{\tau_h}^{\tau_{h+1}} t^m e^{-t/N_h} dt.
\end{aligned}
$$

Using the result (Gradshteyn and Ryzhik 2007, p. 106) that

$$
\int x^m e^{ax} dx = e^{ax} \sum_{\ell=0}^{m} \frac{(-1)^\ell \ell! \binom{m}{\ell}}{a^{\ell+1}} x^{m-\ell}, \tag{2}
$$

we obtain

$$
\begin{aligned}
\mathbb{E}[T_{ij}^m] = \sum_{\ell=0}^{m} \ell! \binom{m}{\ell} \Bigg\{ & \delta_{ij} N_{2(K-i)}^\ell \left[\tau_0^{m-\ell} - \tau_{2(K-i)}^{m-\ell} e^{-(\tau_{2(K-i)} - \tau_0)/N_{2(K-i)}} \right] \\
& + \sum_{h=2(K-i)}^{2K-1} \Lambda_{ijh} N_h^\ell \left[\tau_h^{m-\ell} - \tau_{h+1}^{m-\ell} e^{-(\tau_{h+1} - \tau_h)/N_h} \right] \Bigg\}.
\end{aligned} \tag{3}
$$

Setting $m = 1$, the expected coalescence time is

$$
\begin{aligned}
\mathbb{E}[T_{ij}] = {} & \delta_{ij} \left[\tau_0 + N_{2(K-i)} - \left(\tau_{2(K-i)} + N_{2(K-i)} \right) e^{-(\tau_{2(K-i)} - \tau_0)/N_{2(K-i)}} \right] \\
& + \sum_{h=2(K-i)}^{2K-1} \Lambda_{ijh} \left[\tau_h + N_h - (\tau_{h+1} + N_h) e^{-(\tau_{h+1} - \tau_h)/N_h} \right].
\end{aligned} \tag{4}
$$

Using the density in Equation 1, we can investigate how the initial divergence time and the severity of bottlenecks influence the distribution of coalescence times. Figure 2B displays density plots for coalescence times in the serial founder model in Figure 2A. Analytical density functions closely match the histograms generated in 10^7 coalescent simulations using MS (Hudson 2002), following the simulation method of DeGiorgio *et al.* (2009). Figure 2B shows that multiple modes appear in the distributions of pairwise coalescence times, as a result of the increased coalescence rate during bottlenecks. Coalescence-time distributions for pairs of lineages from different populations are shifted by the divergence time of the two populations, so that coalescence times for pairs of lineages from distinct populations tend to exceed those of pairs from the same population.

We can consider the effect of bottleneck size by examining the coalescence-time distribution for a pair of lineages in two scenarios that are identical except that one has a smaller bottleneck size. In Figure 2B, considering a pair of lineages from population 4, with bottleneck size 1000 individuals, most of the coalescence-time distribution accumulates early because of the strong bottleneck during the time interval $[\tau_1, \tau_2) = [5000, 10000)$. Much of the remainder of the distribution accumulates during the next strong bottleneck, in the interval $[\tau_3, \tau_4) = [15000, 20000)$.

Increasing the bottleneck size in Figure 2A, from 1000 to 5000, the coalescence rate within bottlenecks decreases. Because of this decrease, lineages are more likely to persist farther into the past without coalescing. Thus, Figure 2C shows that decreasing the severity of the bottleneck by increasing the bottleneck population size reduces the probability that the lineages coalesce during the most recent bottleneck. A fourth mode of the coalescence-time distribution then becomes visible during the bottleneck in the interval $[\tau_5, \tau_6) = [25000, 30000)$.

Pairwise homozygosity and heterozygosity

Two commonly used summary statistics are expected homozygosity (gene identity) and expected heterozygosity (gene diversity). Let J_{ij} be a random variable that denotes the homozygosity for a pair of lineages, one randomly sampled from extant population E_i and the other from extant population E_j (where j is not necessarily distinct from i). Further, let $H_{ij} = 1 - J_{ij}$ be a random variable that denotes the heterozygosity for a pair of lineages, one randomly sampled from E_i and the other from E_j. We define homozygosity as the probability that two alleles sampled at a locus are identical by descent (the definition of locus used here is flexible and can range from a single site to a haplotype). Assuming an infinite alleles mutation model and a time interval of length T generations, if mutations are Poisson distributed, then homozygosity, or the probability that no mutation occurs on an interval of length T, is $e^{-2\mu T}$, where μ is the per-generation mutation rate (Wakeley 2009, p. 107). We can therefore find mth moments of homozygosity as

$$
\begin{aligned}
\mathbb{E}[J_{ij}^m] &= \int_0^\infty e^{-2m\mu t} f_{ij}(t) dt \\
&= \int_{\tau_0}^{\tau_{2(K-i)}} e^{-2m\mu t} \delta_{ij} \frac{e^{-(t-\tau_0)/N_{2(K-i)}}}{N_{2(K-i)}} dt + \sum_{h=2(K-i)}^{2K-1} \int_{\tau_h}^{\tau_{h+1}} e^{-2m\mu t} \Lambda_{ijh} \frac{e^{-(t-\tau_h)/N_h}}{N_h} dt \\
&= \frac{\delta_{ij}}{1 + 2N_{2(K-i)} m\mu} \left[e^{-2m\mu\tau_0} - e^{-2m\mu\tau_{2(K-i)} - \frac{\tau_{2(K-i)} - \tau_0}{N_{2(K-i)}}} \right] \\
&\quad + \sum_{h=2(K-i)}^{2K-1} \frac{\Lambda_{ijh}}{1 + 2N_h m\mu} \left[e^{-2m\mu\tau_h} - e^{-2m\mu\tau_{h+1} - \frac{\tau_{h+1} - \tau_h}{N_h}} \right].
\end{aligned} \tag{5}
$$

By the binomial theorem, the mth moment of heterozygosity is $\mathbb{E}[H_{ij}^m] = \mathbb{E}[(1 - J_{ij})^m] = \sum_{\ell=0}^m \binom{m}{\ell} (-1)^\ell \mathbb{E}[J_{ij}^\ell]$. Setting $m = 1$ in Equation 5, we obtain the expected homozygosity and heterozygosity for two lineages, one sampled from population E_i and the other from E_j,

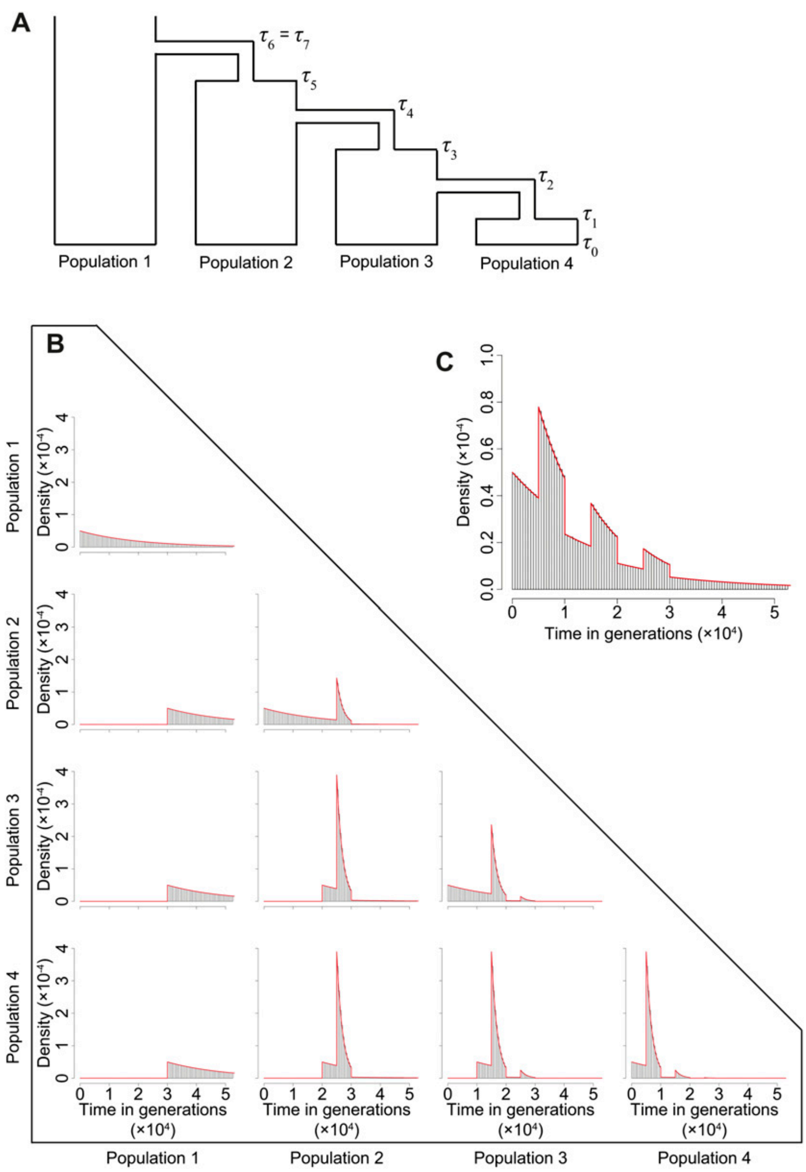

Figure 2 Distributions of coalescence times in the serial founder model. (A) Serial founder model with four extant populations. Thick population sizes represent 10000 diploid individuals and thin population sizes represent 1000 diploid individuals. The times of founding events and population expansions are $\tau_0 = 0$, $\tau_h = \tau_{h-1} + 5000$ for $h = 1, 2, \ldots, 6$, and $\tau_6 = \tau_7 = 30000$ generations. (B) Probability density of coalescence times. Each subplot is the probability density of coalescence times for a pair of lineages sampled from the pair of populations listed in the row and column. (C) Probability density of coalescence times for a pair of lineages sampled from population 4, with identical parameter values to part *A* except that the bottlenecks (thin populations) have 5000 diploid individuals instead of 1000 diploid individuals. The figure can be compared with the plot for two lineages from population 4 in part *B*. Histograms are based on 10^7 coalescent simulations using MS (Hudson 2002), and the red lines represent the analytical densities obtained from Equation 1.

$$\mathbb{E}[J_{ij}] = \frac{\delta_{ij}}{1+2N_{2(K-i)}\mu}\left[e^{-2\mu\tau_0} - e^{-2\mu\tau_{2(K-i)}-((\tau_{2(K-i)}-\tau_0)/N_{2(K-i)})}\right] + \sum_{h=2(K-i)}^{2K-1}\frac{\Lambda_{ijh}}{1+2N_h\mu}\left[e^{-2\mu\tau_h} - e^{-2\mu\tau_{h+1}-((\tau_{h+1}-\tau_h)/N_h)}\right] \tag{6}$$

$$\mathbb{E}[H_{ij}] = 1 - \frac{\delta_{ij}}{1+2N_{2(K-i)}\mu}\left[e^{-2\mu\tau_0} - e^{-2\mu\tau_{2(K-i)}-((\tau_{2(K-i)}-\tau_0)/N_{2(K-i)})}\right] - \sum_{h=2(K-i)}^{2K-1}\frac{\Lambda_{ijh}}{1+2N_h\mu}\left[e^{-2\mu\tau_h} - e^{-2\mu\tau_{h+1}-((\tau_{h+1}-\tau_h)/N_h)}\right]. \tag{7}$$

Using the model in Figure 2A, Figure 3 plots the expected heterozygosity of two lineages sampled from population 4 as a function of both bottleneck population size and bottleneck length. When the bottleneck has length zero, bottlenecks do not increase genetic drift and hence the expected heterozygosity reaches its maximum. Increasing the bottleneck length causes a monotonic decrease in expected heterozygosity. Decreasing the population size of the bottlenecks further decreases the heterozygosity. The smallest expected heterozygosity shown is reached with the combination of the smallest bottleneck population size (100 diploid individuals) and the largest bottleneck length (5000 generations).

Pairwise F_{ST}

Our computation of expected coalescence times in Equation 4 provides a basis for obtaining the commonly used measure of genetic differentiation, pairwise F_{ST} between populations. Using the results of Slatkin (1991) on F_{ST} at small mutation rates, we can write $F_{\mathrm{ST}} = (\bar{T} - \bar{T}_0)/\bar{T}$, where $\bar{T}_0$ is the mean coalescence time of two lineages randomly drawn from the same population and $\bar{T}$ is the mean coalescence time of two lineages randomly drawn from any two populations (same

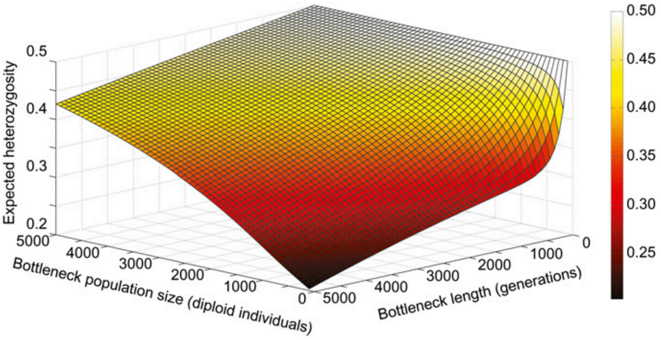

Figure 3 Expected heterozygosity for a pair of lineages sampled from population 4 of Figure 2A (Equation 7), as a function of population size for bottlenecks and bottleneck length measured in generations. A per-generation mutation rate of $\mu = 2.5 \times 10^{-5}$ is assumed.

or different). By using the expected coalescence times in our serial founder model (Equation 4), we can define these times for pairwise comparisons of populations E_i and E_j ($i < j$) as $\bar{T}_0 = (1/2)\mathbb{E}[T_{ii}] + (1/2)\mathbb{E}[T_{jj}]$, $\bar{T}_{\text{diff}} = \mathbb{E}[T_{ij}]$ (the mean pairwise coalescence time for two lineages from different populations), and $\bar{T} = (1/2)\bar{T}_0 + (1/2)\bar{T}_{\text{diff}}$. Therefore,

$$F_{\text{ST}}^{ij} = \frac{\mathbb{E}[T_{ij}] - (1/2)\mathbb{E}[T_{ii}] - (1/2)\mathbb{E}[T_{jj}]}{\mathbb{E}[T_{ij}] + (1/2)\mathbb{E}[T_{ii}] + (1/2)\mathbb{E}[T_{jj}]}, \tag{8}$$

where the quantities $\mathbb{E}[T_{ij}]$ are defined in Equation 4.

Patterns Observed in Human Population Data

In this section we describe a worldwide human population-genetic data set and patterns in summary statistics calculated from the data set. The summary statistics we investigate are within-population gene diversity, between-population gene identity, and pairwise F_{ST}. Analytical formulas for these summary statistics under the serial founder model are obtained in Equations 6–8. We compare patterns in these summary statistics observed in data to those predicted by specific models of human evolutionary history. Through these comparisons, we discuss which models of human history are compatible with patterns of genetic variation observed in present-day human populations. Note that only one of the three summary statistics that we study (gene diversity) was discussed by DeGiorgio *et al.* (2009).

We analyzed data from the Human Genome Diversity Panel (HGDP) (Cann *et al.* 2002; Cavalli-Sforza 2005), using 783 autosomal microsatellite loci in 1048 individuals sampled from 53 worldwide populations (Ramachandran *et al.* 2005; Rosenberg *et al.* 2005). For a given population, gene diversity was calculated using DeGiorgio and Rosenberg's (2009) Equation 10, averaged across loci; the values were taken from Figure 7C of DeGiorgio and Rosenberg (2009). For distinct populations A and B, between-population gene identity was calculated as $J_{AB} = (1/L)\sum_{\ell=1}^{L}\sum_{i=1}^{I_\ell}\hat{p}_{\ell i}\hat{q}_{\ell i}$, where $\hat{p}_{\ell i}$ and $\hat{q}_{\ell i}$ are the sample frequencies of the ith distinct allele at locus ℓ in populations A and B, respectively, and I_ℓ is the number of distinct alleles in the pair of populations at locus ℓ (Nei 1987). Pairwise F_{ST} was calculated using Weir's (1996) Equation 5.3.

Figure 4 displays patterns observed for the three summary statistics in the HGDP data set. Figure 4A shows an approximate linear decline of gene diversity with increasing geographic distance from a putative East African location of modern human origins. Figure 4B shows a heat map of gene identity between all pairs of populations, illustrating that pairs closer to Africa generally have lower between-population gene identity than pairs farther from Africa. Figure 4C displays a heat map of pairwise F_{ST} between populations. F_{ST} is lower for pairs of populations that are close geographically

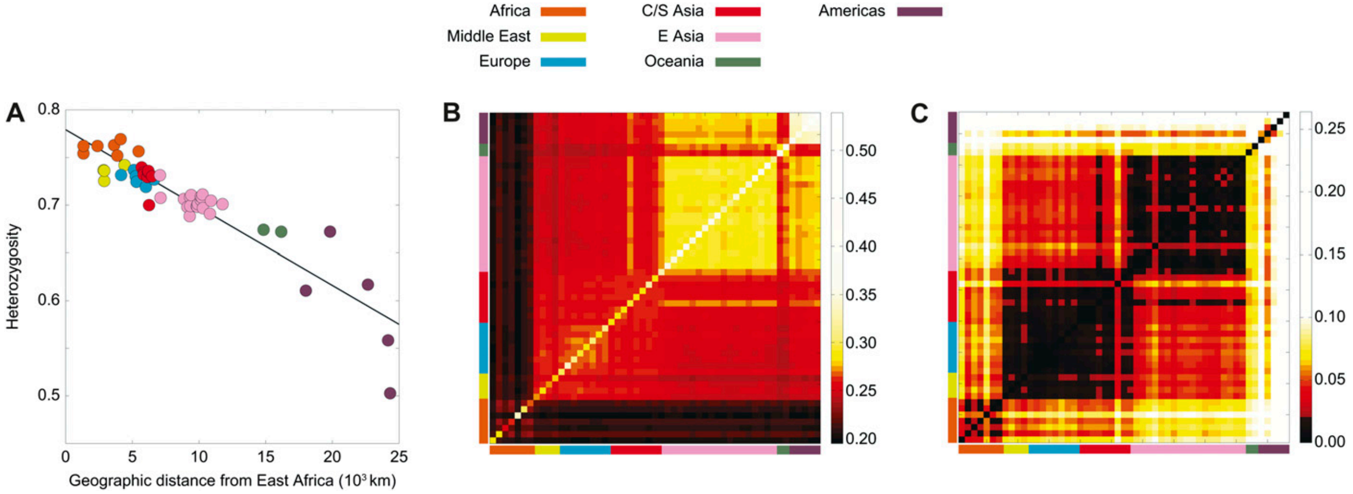

Figure 4 Patterns of within- and between-population summary statistics observed in human population-genetic data. Plots are based on 783 microsatellite loci from 53 worldwide populations in the HGDP data set (Ramachandran *et al.* 2005; Rosenberg *et al.* 2005). (A) Gene diversity as a function of distance from East Africa (redrawn from Degiorgio *et al.* 2009). Each point represents a particular population. (B) Between-population gene identity. Columns and rows each represent populations, and an entry in the matrix represents the gene identity for the population pair represented by the row and column. (C) Pairwise F_{ST} calculated from the same populations as in B.

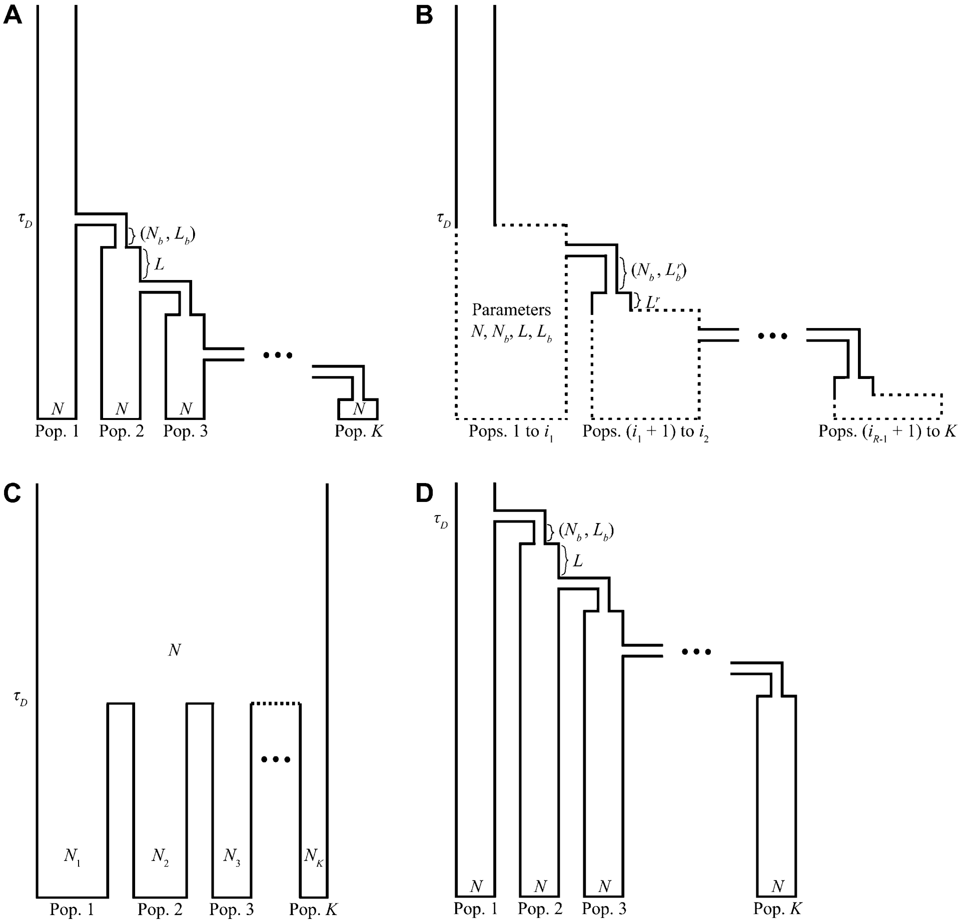

Figure 5 Models to which the general serial founder model reduces. (A) Modern serial founder model. (B) Nested regions model with R regions. (C) Instantaneous divergence model. (D) Archaic serial founder model. The models in A and C are exactly the models discussed by DeGiorgio *et al.* (2009).

than for pairs of populations that are geographically distant. Additionally, F_{ST} values between populations in the Americas are generally larger than F_{ST} values between pairs of non-American populations. In Figure 4, a slight jump in the values of summary statistics is visible at the boundaries of geographic regions. That is, separate values of gene diversity computed within populations from the same geographic region, and gene identity and F_{ST} values for pairs of populations from the same region, tend to be more similar to each other than to corresponding values involving populations from different regions.

We can now compare the three patterns in summary statistics observed from the HGDP data set with patterns predicted by models of human evolutionary history. We consider several special cases of our general serial founder model that are chosen on the basis of previous investigations of human evolution. These cases include a modern serial founder model (Ramachandran *et al.* 2005; DeGiorgio *et al.* 2009; Deshpande *et al.* 2009), a nested regions model in which bottlenecks between continental regions are more severe than those within continental regions (Hunley *et al.* 2009), an instantaneous divergence model in which all populations diverged at the same time in the past (DeGiorgio *et al.* 2009), and an archaic serial founder model in which the founding process started distantly in the past (DeGiorgio *et al.* 2009). Using Equations 6–8, we now examine the patterns in gene diversity, between-population gene identity, and pairwise F_{ST} generated by these four special cases of the general serial founder model. We consider the extent to which each model can reproduce the patterns observed in worldwide human genetic data in the three statistics.

Modern Serial Founder Model

Motivation and model

A modern serial founder model (Figure 5A) is a special case of our general formulation (Figure 1). To obtain the DeGiorgio *et al.* (2009) serial founder model with K populations, suppose that the bottleneck length is L_{b} generations and that the time between the end of a bottleneck and the founding of a new population is L generations. In other words, suppose $\tau_{2h+1} - \tau_{2h} = L$ for $h = 0, 1, \ldots, K-2$ and $\tau_{2h} - \tau_{2h-1} = L_{\mathrm{b}}$ for $h = 1, 2, \ldots, K-1$. Let $\tau_0 = 0$. Modern population 1 founds modern population 2 at time $\tau_{2(K-1)} = \tau_{2K-1} = \tau_{\mathrm{D}}$. Each bottleneck has size N_{b} diploid individuals, and all other populations have size N. For the exact serial founder model studied by DeGiorgio *et al.* (2009), we set $K = 100$, $L_{\mathrm{b}} = 2$, $L = 19$, $\tau_{\mathrm{D}} = 2079$, $N = 10000$, and $N_{\mathrm{b}} = 250$. These values were chosen to represent reasonable values for human populations: τ_{D} was chosen to lie within an estimated interval of

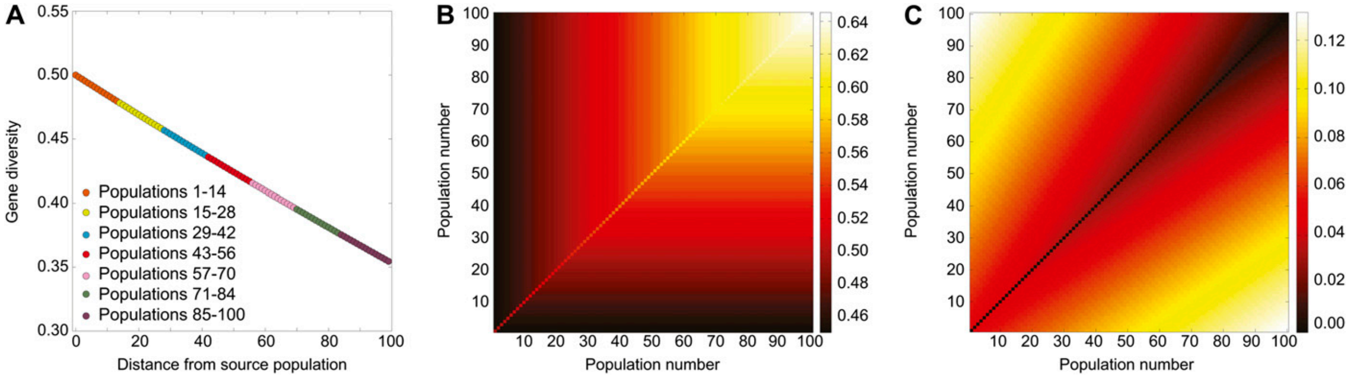

Figure 6 Patterns of genetic variation in a modern serial founder model. The values of the model parameters are indicated in the section *Modern Serial Founder Model*. (A) Gene diversity of the populations as a function of distance from the source population, where distance is measured in units of populations. (B) Between-population gene identity for pairs of populations. (C) Pairwise F_{ST} for pairs of populations.

time for the out-of-Africa migration (*e.g.*, Relethford 2008), N was chosen as a commonly used value to represent the present-day effective size of human populations (*e.g.*, Takahata *et al.* 1995), N_b was chosen to represent a size typical for small isolated hunter–gatherer populations (Cavalli-Sforza 2004), L_b was chosen to represent a process in which individuals migrate in the first generation and finalize the settlement of a population in the second generation, and L was chosen such that founding events were distributed uniformly over $\tau_D = 2079$ generations. Utilizing this parameterization and a per-generation mutation rate of $\mu = 2.5 \times 10^{-5}$, we examine whether the modern serial founder model can reproduce observed patterns of human genetic variation.

Patterns generated by the model

Figure 6 displays patterns of genetic variation generated by the modern serial founder model. As was observed previously in simulations (Ramachandran *et al.* 2005; DeGiorgio *et al.* 2009; Deshpande *et al.* 2009), the modern serial founder model reproduces the approximate linear decline in gene diversity with distance from the source population (Figure 6A). Figure 6B displays a heat map of pairwise gene identity values between pairs of modern populations. The heat map shows that populations close to the source population have smaller between-population gene identities than populations far from the source, as is observed in human population data (Figure 4B). Figure 6C displays a heat map of F_{ST} values between pairs of modern populations, demonstrating that pairs of populations that are geographically distant tend to have larger F_{ST} than pairs of populations that are geographically close. The model largely recovers the pattern observed in human data (Figure 4C); however, it also predicts small F_{ST} between pairs that are far from the source population, a pattern that is not observed for human populations distant from Africa.

The pattern of decrease in gene diversity with increasing distance from a source population is due to the decrease in pairwise coalescence time within populations caused by a cumulative increase in genetic drift with increasing distance from the source. Pairs of lineages from distinct populations distant from the source have the potential to coalesce more recently than do pairs of lineages close to the source, thereby explaining the increased gene identity for pairs of populations distant from the source. However, F_{ST} between populations that are geographically distant from the source is smaller than F_{ST} between populations that are close to the source, as the effect of reduced between-population coalescence times in decreasing F_{ST} for populations distant from the source outweighs the effect of their reduced within-population coalescence times in increasing F_{ST}.

Our results show that the modern serial founder model largely recovers the patterns observed from human genetic data (Figure 4). Two exceptions are that it does not predict either a peculiar pattern of small gene identities observed between Oceanian and non-Oceanian populations (Figure 4B) or the large F_{ST} values observed in the Americas (Figure 4C).

Nested Regions Model

Motivation and model

One aspect of the trends in genetic diversity that was not captured by our parameterization of the modern serial founder model above is the larger difference in diversity observed between populations from different continental regions than between populations from the same continental region (Figure 4A). This observation motivates the nested regions model (Figure 5B) simulated by Hunley *et al.* (2009), in which the set of populations is distributed across several “regions” separated by barriers to migration. Examples of such regions include different continents, areas separated by mountain ranges, or islands within an archipelago. Because crossing between regions is more difficult than migration within a region, significant genetic drift might occur during the expansion into a new region. The nested regions model incorporates this increase in genetic drift during the geographic expansion through increased bottleneck severity between regions relative to bottleneck severity within regions.

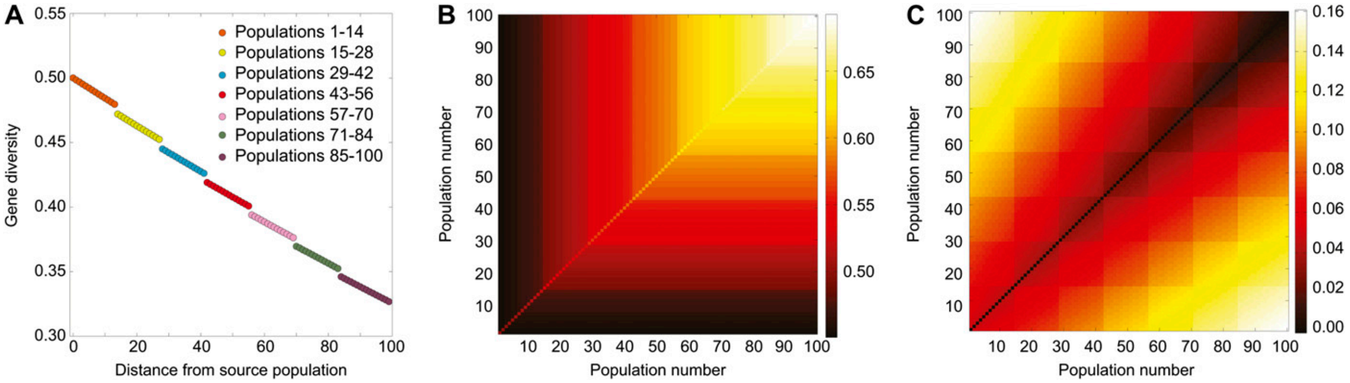

Figure 7 Patterns of genetic variation in a nested regions model. The values of the model parameters are the same as those in Figure 6, with the exception that the bottleneck lasts 16 generations instead of 2 generations during the founding of modern populations 15, 29, 43, 57, 71, and 85. (A) Gene diversity of the populations as a function of distance from the source population, where distance is measured in units of populations. (B) Between-population gene identity for pairs of populations. (C) Pairwise F_{ST} for pairs of populations.

We incorporate severe bottlenecks into the modern serial founder model (Figure 5A) by increasing the bottleneck lengths to $L_b^r = 16$ generations instead of $L_b = 2$ during the founding of modern populations 15, 29, 43, 57, 71, and 85. Hence, the length of time between the end of any of these bottlenecks and the founding of the next population is $L^r = 5$ generations instead of $L = 19$, so that the time between founding events is still $L_b + L = 21$ generations. These severe bottlenecks subdivide the set of $K = 100$ modern populations into $R = 7$ regions.

Patterns generated by the model

Figure 7 depicts patterns of genetic variation generated by the nested regions model. As was observed in simulations of Hunley *et al.* (2009), the nested regions model reproduces the approximate linear decline in gene diversity with distance from the source population, with small discontinuities in genetic diversity at region boundaries (Figure 7A). Similarly, as was observed in the simulations of Hunley *et al.* (2009), the nested regions model reproduces the patterns of between-population gene identity observed from human data, with pairs of populations far from the source displaying larger gene identity than pairs close to the source (Figure 7B). Also, in the nested regions model, pairs of populations that are geographically distant tend to have larger F_{ST} than pairs of populations that are geographically close (Figure 7C). The nested regions model predicts regional boundaries in the gene identity and F_{ST} heat maps (Figure 7, B and C) that partly reproduce the block structure in the human population data (Figure 4, B and *C*). However, as in the modern serial founder model, the nested regions model predicts small F_{ST} between pairs that are far from the source population, a pattern that is not observed for populations in the Americas (contrast Figure 4C and Figure 7C).

As was seen with the modern serial founder model above, the nested regions model recovers most of the patterns observed in human population-genetic data (Figure 4). Because of the increased bottleneck severity between regions, unlike the modern serial founder model, the nested regions model also reproduces the larger differences in values of the three summary statistics observed between regions compared to values observed within regions (Figure 4).

Instantaneous Divergence Model

Motivation and model

DeGiorgio *et al.* (2009) found that another model, the instantaneous divergence model, was capable of generating patterns that were compatible with observed patterns of within-population gene diversity, linkage disequilibrium, and the ancestral allele frequency spectrum. Because we investigated only within-population summary statistics, however, it was not examined whether the gene identity and F_{ST} patterns observed in Figure 4, B and C, could also be generated by the instantaneous divergence model.

The instantaneous divergence model (Figure 5C) is a model in which all populations diverge at the same time in the past and populations that are farther from the source population have a smaller population size than those that are closer to the source. The motivation for this model is that populations that have traveled a greater distance from a source population will likely have lost alleles through genetic drift. The instantaneous divergence model allows for this increased drift for populations that are located far from the source population by assigning such populations a smaller size. An increase in genetic drift causes a decrease in gene diversity due to the random loss of alleles, as also occurs in bottlenecks. DeGiorgio *et al.* (2009) found that when the size of population *i* in the instantaneous divergence model was set so that the elapsed coalescent time was the same as in modern population *i* in the modern serial founder model, the approximate linear trend in gene diversity with distance from the source population was virtually indistinguishable from that of the modern serial founder model.

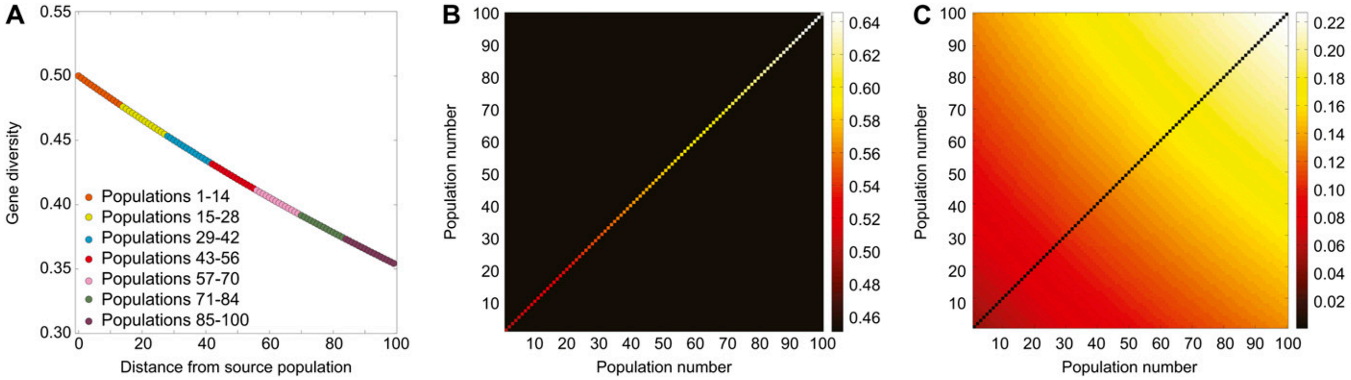

Figure 8 Patterns of genetic variation in the instantaneous divergence model. The values of the model parameters are indicated in the section *Instantaneous Divergence Model*. (A) Gene diversity of the populations as a function of distance from the source population, where distance is measured in units of populations. (B) Between-population gene identity for pairs of populations. (C) Pairwise F_{ST} for pairs of populations.

Suppose a modern serial founder model is parameterized as in Figure 6A. We obtain the instantaneous divergence model of DeGiorgio *et al.* (2009) by setting the divergence time of all K populations to τ_D, the ancestral diploid population size to N, and the diploid size of population i to

$$N_i = \frac{\tau_D}{[\tau_D - (i-1)L_b]/N + (i-1)L_b/N_b}, \tag{9}$$

for $i = 1, 2, \ldots, K$, where τ_D, N, N_b, L, and L_b are the parameters in the modern serial founder model in the section *Modern Serial Founder Model* (DeGiorgio *et al.* 2009). The value of N_i is chosen so that τ_D/N_i is the total duration in coalescent units of population i. To obtain the exact instantaneous divergence model described by DeGiorgio *et al.* (2009), we set $\tau_D = 2079$, $N = 10000$, $N_b = 250$, $L = 19$, and $L_b = 2$. These values are the same values used for the modern serial founder model in Figure 6A. Using Equation 9 for the size of population i allows population i to experience the same level of genetic drift as modern population i in the modern serial founder model.

Patterns generated by the model

Figure 8 depicts patterns of genetic variation generated by the instantaneous divergence model. As was observed in the simulations of DeGiorgio *et al.* (2009), this model reproduces the approximate linear decline in gene diversity with increasing distance from the source population (Figure 8A). In contrast, between-population gene identity and pairwise F_{ST} yield patterns that are quite different from those observed in human data (contrast Figure 8, B and C, with Figure 4, B and C). All off-diagonal entries of Figure 8B have identical small gene identities. Also, pairs of populations that are close to the source population have smaller F_{ST} than pairs that are far from the source (Figure 8C).

The approximate linear decline in gene diversity produced by the instantaneous divergence model (Figure 8A) is caused by the loss of alleles and consequent decrease in heterozygosity due to increased genetic drift within populations that are far from the source population (DeGiorgio *et al.* 2009). However, the fact that all off-diagonal entries of Figure 8B are identical indicates that no correlation exists with geography for between-population gene identity under the instantaneous divergence model. This lack of correlation causes the pattern of pairwise F_{ST} values to be driven completely by the sizes of population pairs. Hence, population pairs far from the source location, which have smaller population sizes, and therefore smaller within-population coalescence times, have higher F_{ST} values.

Because the approximate linear decline in gene diversity (Figure 8A) generated by the instantaneous divergence model matches the pattern observed from human genetic data (Figure 4A), we can conclude that the pattern of within-population gene diversity observed from human data reflects the cumulative increase in genetic drift with increasing distance from Africa (DeGiorgio *et al.* 2009). However, the patterns of between-population summary statistics generated by the instantaneous divergence model (Figure 8, B and C) do not match the patterns observed from data (Figure 4, B and C). Thus, a model that incorporates only a cumulative increase in genetic drift with increasing distance from a source is not sufficient to predict observed patterns of between-population genetic diversity.

Archaic Serial Founder Model

Motivation and model

The serial founder model was motivated as a model to explain how modern humans expanded out of Africa and colonized the world. Our general serial founder model, however, does not place restrictions on the time of the first founding event. Therefore, our general model reduces to an archaic serial founder model (Figure 5D) when the time to the first founding event occurs distantly in the past. The archaic serial founder model, although it has an identical mathematical form to the modern serial founder model, is conceptually different in the sense that it is motivated by hypotheses regarding expansions of ancient hominids out of Africa, whereas the modern serial founder model is motivated by hypotheses of recent expansion of anatomically

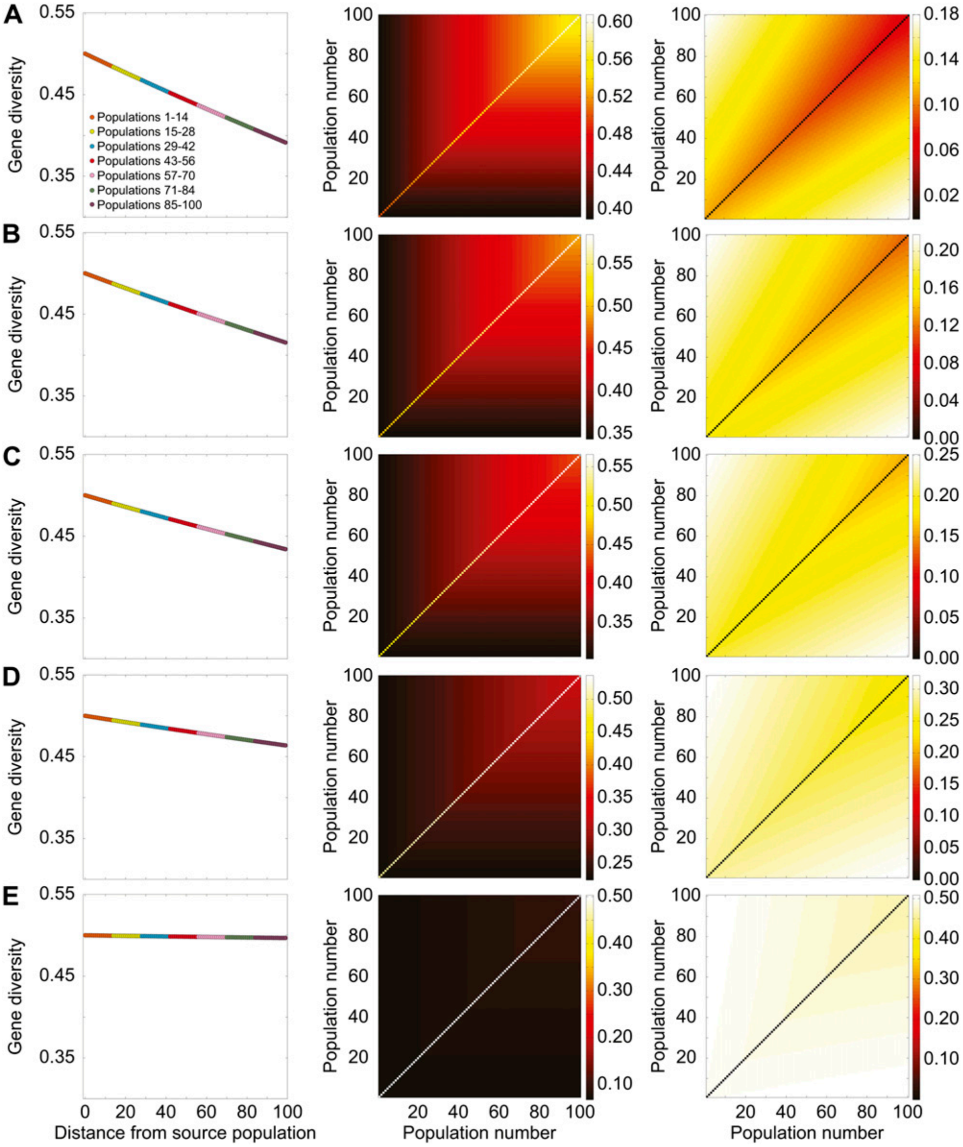

Figure 9 Patterns of genetic variation in an archaic serial founder model, as a function of varying divergence time τ_D. The values of the model parameters for A–E are the same as in Figure 6A, with the exception that the divergence time τ_D, measured in generations, varies across the plots. The first column is gene diversity of the populations as a function of distance from the source population, where distance is measured in units of populations. The second column is between-population gene identity for pairs of populations. The third column is pairwise F_{ST} for pairs of populations. (A) $\tau_D = 5000$. (B) $\tau_D = 7500$. (C) $\tau_D = 10000$. (D) $\tau_D = 16000$. (E) $\tau_D = 40000$.

modern humans out of Africa. The effect of increasing the time of the first founding event can be investigated in the serial founder model while holding all other parameters in the model constant.

In this section, we discuss how the patterns for within-population gene diversity, between-population gene identity, and pairwise F_{ST} change as the serial founding process is pushed farther into the past. To obtain an archaic serial founder model, we assume that except for divergence time τ_D, all parameters are the same as in the modern serial founder model considered in Figure 6. We consider divergence times of $\tau_D = 5000$, 7500, 10000, 16000, and 40000 generations ago. Divergence times $\tau_D = 16000$ and $\tau_D = 40000$ are of particular interest because, assuming a generation time of 25 years, they approximate estimates of the divergence of modern humans with Neanderthal (400 KYA; Green *et al.* 2006; Noonan *et al.* 2006) and *Homo erectus* (1 MYA; Takahata 1993) populations, respectively.

Patterns generated by the model

For $\tau_D = 5000$, relative to the modern serial founder model in which $\tau_D = 2079$, a decrease occurs in the magnitude of the slope of the decline of gene diversity with increasing distance from the source population (Figure 9A). The increased gene identity and decreased F_{ST} between populations that are far from the source population relative to between populations that are close to the source, although still observable, are less distinct with the increased divergence time. Further increasing the divergence time to $\tau_D = 7500$ (Figure 9B) and $\tau_D = 10000$ (Figure 9C) leads to a progressive decrease in the differences among populations in values of the three summary statistics. For a serial founder model with a divergence time of $\tau_D = 16000$, at a putative time of the Neanderthal divergence, differences in values among populations for each of the three summary statistics are small (Figure 9D). For the *H. erectus* serial founder model with $\tau_D = 40000$, differences in values among populations for each of the three summary statistics are nearly negligible, displaying almost no trend (Figure 9E).

As τ_D increases, the differences among populations in values of gene diversity, between-population gene identity, and F_{ST} decrease. These smaller differences result from the smaller degree of influence that ancient bottlenecks have on genetic diversity in comparison with recent bottlenecks of identical severity. This lack of influence of ancient bottlenecks on present-day gene diversity is reflected most strongly in the small difference in gene diversity between population 1 and population 100 in the *H. erectus* serial founder model (Figure 9E). Furthermore, with greater τ_D, the difference between the divergence time for two populations sampled close to the source and for two populations sampled far from the source is small relative to τ_D. This small difference in divergence times causes between-population summary statistics such as gene identity and F_{ST} to have little correlation with geography (*i.e.*, most off-diagonal entries have similar values) at large divergence times (Figure 9E).

These results imply that the patterns in gene diversity, gene identity, and F_{ST} observed from empirical data cannot be predicted solely by an archaic serial founder process using our parameterization; specifically, the observed patterns are not consistent with a serial founder process that occurs too far back in the past. Pushing back the time of the first founding event while holding all other parameters constant decreases the ability of the serial founder model to generate the patterns observed in Figure 4.

Discussion

In this article, we have derived pairwise coalescence-time distributions for a serial founder model. Under the model, we have provided analytical formulas for expected coalescence times, expected homozygosity, and pairwise F_{ST}. In addition, we have analytically described the trend of decreasing gene diversity with increasing distance from the source population, and the patterns observed in between-population gene identity and pairwise F_{ST}. Using coalescence-time densities in various special cases, we have found that the modern serial founder model and the nested regions model are consistent with geographic patterns of within- and between-population genetic diversity observed in human data. Our work demonstrates the utility of using theoretical computations on between-population summary statistics in conjunction with similar computations on within-population statistics to predict geographic patterns in genetic data.

One pattern that was not predicted by any of our models was the large F_{ST} observed in the Americas. Whereas the modern serial founder and the nested regions models predict small F_{ST} between populations far from the source, F_{ST} values in the Americas are large. It is possible that the models provide a poor fit to the pattern of evolution in the Americas after the initial founding of the Native American population, as they also are inconsistent with the large differences in gene diversity among populations in the Americas. During the initial migration into the Americas, small individual populations may have experienced highly variable levels of genetic drift as they spread over a large unoccupied region (*e.g.*, Wang *et al.* 2007; Goebel *et al.* 2008; Meltzer 2009). Such a migration process could have given rise to highly variable levels of genetic diversity across the region, as well as a somewhat irregular pattern in F_{ST}. If we were to modify our model to incorporate this variability along with stronger bottlenecks or smaller population sizes within the Americas relative to those in non-American populations, then we might be able to produce patterns that agree with the observed data. Indeed, Hunley *et al.* (2009) found that model parameters can be chosen to enable patterns of within- and between-population genetic diversity to closely match those empirically observed in the Americas.

Another pattern that was not predicted by any of our models is the small between-population gene identity observed between pairs of populations, one from Oceania and the other not from Oceania (Figure 4B). This pattern could potentially be explained either by an ancient divergence of the Oceanian populations from the non-Oceanian populations through a separate migration out of Africa to Oceania (*e.g.*, Derricourt 2005; Bulbeck 2007; Field *et al.* 2007; Szpiech *et al.* 2008; Kayser 2010) or by admixture of the populations in Oceania with an archaic human population (*e.g.*, Reich *et al.* 2010). A separate founding process could have generated low levels of within-population gene diversity for the Oceanian populations while simultaneously producing the low levels of between-population gene identity between Oceanian and non-Oceanian populations. Alternatively, because the increase in between-population coalescence times that would be caused by ancient admixture would result in a decrease in between-population gene identity, such admixture could potentially explain the disagreement of the data with our model predictions. Separate migrations or ancient admixture could potentially be incorporated into a more general version of our model to investigate the plausibility of these scenarios.

By increasing the time of the first founding event, we have determined that the archaic serial founder model is not able to reproduce patterns of gene diversity, between-population gene identity, and pairwise F_{ST} observed in human genetic data. However, limited archaic admixture coupled with a modern serial founder model might not be incompatible with the patterns we have examined. Signatures of archaic admixture might exist in modern human population-genetic data (*e.g.*, Garrigan and Hammer 2006; Plagnol and Wall 2006; Green *et al.* 2010; Reich *et al.* 2010) and as discussed above, such admixture could potentially explain anomalous observations in Oceania. However, this admixture, if it indeed occurred, must have been insufficient to generate a large signature in most of the patterns that we have studied.

Although the patterns of gene diversity produced by the serial founder and the instantaneous divergence models are virtually indistinguishable (DeGiorgio *et al.* 2009), we have shown that these models can be differentiated using between-population gene identity and pairwise F_{ST}. Ultimately, this potential for differentiation traces to distinctive distributions of pairwise coalescence times. In the instantaneous divergence model, each population has a constant size up until time τ_D and consequently, the coalescent process simply follows an exponential distribution until time τ_D and then another exponential distribution with a different rate after time τ_D. In contrast, in the serial founder model, the rate of coalescence inside a bottleneck is elevated compared to outside the bottleneck. This increased rate of coalescence causes lineages to merge within a narrow time interval. Because the serial founder model incorporates multiple bottlenecks, the distribution of coalescence times is multimodal.

Recently, many studies have found that two-dimensional spatial maps generated from principal components analysis (PCA) applied to human genetic data closely match maps of geographic sampling locations of populations (*e.g.*, Lao *et al.* 2008; Novembre *et al.* 2008; Price *et al.* 2009; Bryc *et al.* 2010; Wang *et al.* 2010; Xing *et al.* 2010). McVean (2010) demonstrated a close link between pairwise coalescence times and PCA, in which sampled lineages can be projected onto principal components through expected coalescence times for pairs of lineages. The coalescence-time distributions provided in this article can potentially be used to interpret PCA maps, so that PCA maps themselves might be used as summary statistics for testing evolutionary models.

Estimated coalescence-time distributions might also be utilized more formally for maximum-likelihood estimation of parameters such as bottleneck lengths, bottleneck sizes, and divergence times (*e.g.*, Thomson *et al.* 2000; Takahata *et al.* 2001; Tang *et al.* 2002; Rannala and Yang 2003; Tishkoff and Verrelli 2003; Garrigan and Hammer 2006; Fagundes *et al.* 2007; Blum and Jakobsson 2011). Further, these distributions might also be useful for hypothesis testing; because many of the models in this article are nested, likelihood-ratio tests can be performed. For extending our work to perform maximum-likelihood inference, it will be desirable to extend the computations to permit the sampling of multiple lineages in each population. Such an extension could potentially build upon the work of Marth *et al.* (2004), who derived the coalescence-time distribution for a sample of n lineages in a single population with multiple bottlenecks.

An additional feature of structured population models that would be desirable to incorporate is migration between populations after their initial founding. In the archaic serial founder model, some level of migration between neighboring populations might enable the model to make predictions that more closely match observations from human genetic data. For the modern serial founder model, simulations have shown that small to moderate levels of migration have relatively little impact on observed patterns of genetic diversity (DeGiorgio *et al.* 2009). In any case, inclusion of migration would enable us to examine considerably more complex versions of the models that we have investigated.

Finally, one important quantity that we did not explore is linkage disequilibrium (LD). In simulations, we previously studied whether the spatial distribution of LD observed in worldwide human populations is consistent with a serial founder model (DeGiorgio *et al.* 2009). We found that the serial founder model can indeed predict the observed spatial distribution of LD. Moreover, we found that LD patterns can be useful in distinguishing the patterns predicted by different evolutionary models. Therefore, incorporation of LD into our theoretical models would provide a distinct type of statistic that would further enhance model identifiability. For example, because excess long-range LD is a signature of ancient admixture (*e.g.*, Plagnol and Wall 2006), incorporation of LD statistics would be useful for assessing whether models that include archaic admixture provide a better fit to observed human genetic variation than models that do not consider admixture. Because LD is such a valuable quantity, it would be informative to examine patterns of LD produced by the various models by incorporating recombination into the theory.

Acknowledgments

We thank Raquel Assis, Laurent Excoffier, Zach Szpiech, and two anonymous reviewers for their comments. Support was provided by National Institutes of Health (NIH) grant R01 GM081441, NIH training grants T32 GM070449 and T32 HG000040, a grant from the Burroughs Wellcome Fund, a University of Michigan Rackham Merit Fellowship, and the New Zealand Marsden Fund.

Literature Cited

Austerlitz, F., B. Jung-Muller, B. Godelle, and P.-H. Gouyon, 1997 Evolution of coalescence times, genetic diversity and structure during colonization. Theor. Popul. Biol. 51: 148–164.

Blum, M. G. B., and M. Jakobsson, 2011 Deep divergences of human gene trees and models of human origins. Mol. Biol. Evol. 28: 889–898.

Bryc, K., A. Auton, M. R. Nelson, J. R. Oksenberg, S. L. Hauser *et al.*, 2010 Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. USA 107: 786–791.

Bulbeck, D., 2007 Where river meets sea: a parsimonious model for *Homo sapiens* colonization of the Indian Ocean rim and Sahul. Curr. Anthropol. 48: 315–321.

Cann, H. M., C. de Toma, L. Cazes, M.-F. Legrand, V. Morel *et al.*, 2002 A human genome diversity cell line panel. Science 296: 261–262.

Cavalli-Sforza, L. L., 2004 *Examining the Farming/Language Dispersal Hypothesis*, edited by P. Bellwood, and C. Renfrew. Mcdonald Institute Monographs, Cambridge, UK.

Cavalli-Sforza, L., 2005 The Human Genome Diversity Project: past, present and future. Nat. Rev. Genet. 6: 333–340.

DeGiorgio, M., and N. A. Rosenberg, 2009 An unbiased estimator of gene diversity in samples containing related individuals. Mol. Biol. Evol. 26: 501–512.

DeGiorgio, M., M. Jakobsson, and N. A. Rosenberg, 2009 Explaining worldwide patterns of human genetic variation using a coalescent-based sesrial founder model of migration outward from Africa. Proc. Natl. Acad. Sci. USA 106: 16057–16062.

Derricourt, R., 2005 Getting "Out of Africa": sea crossings, land crossings and culture in the Hominin migrations. J. World Prehist. 19: 119–132.

Deshpande, O., S. Batzoglou, M. W. Feldman, and L. L. Cavalli-Sforza, 2009 A serial founder effect model for human settlement out of Africa. Proc. Biol. Sci. 276: 291–300.

Edmonds, C. A., A. S. Lillie, and L. L. Cavalli-Sforza, 2004 Mutations arising in the wave front of an expanding population. Proc. Natl. Acad. Sci. USA 101: 975–979.

Efromovich, S., and L. S. Kubatko, 2008 Coalescent time distributions in trees of arbitrary size. Stat. Appl. Genet. Mol. Biol. 7: 2.

Excoffier, L., and N. Ray, 2008 Surfing during population expansions promotes genetic revolutions and structuration. Trends Ecol. Evol. 23: 347–351.

Fagundes, N. J. R., N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano *et al.*, 2007 Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. USA 104: 17614–17619.

Field, J. S., M. D. Petraglia, and M. M. Lahr, 2007 The southern dispersal hypothesis and the South Asian archaeological record: examination of dispersal routes through GIS analysis. J. Anthropol. Archaeol. 26: 88–108.

Garrigan, D., and M. F. Hammer, 2006 Reconstructing human origins in the genomic era. Nat. Rev. Genet. 7: 669–680.

Goebel, T., M. R. Waters, and D. H. O'Rourke, 2008 The late Pleistocene dispersal of modern humans in the Americas. Science 319: 1497–1502.

Gradshteyn, I. S., and I. M. Ryzhik, 2007 *Table of Integrals, Series, and Products*, edited by Jeffrey, A., and D. Zwillinger. Academic Press, Burlington, MA.

Green, R. E., J. Krause, S. E. Ptak, A. W. Briggs, M. T. Ronan *et al.*, 2006 Analysis of one million base pairs of Neanderthal DNA. Nature 444: 330–336.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel *et al.*, 2010 A draft sequence of the Neandertal genome. Science 328: 710–722.

Hallatschek, O., and D. R. Nelson, 2008 Gene surfing in expanding populations. Theor. Popul. Biol. 73: 158–170.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23: 183–201.

Hudson, R. R., 1990 Gene genealogies and the coalescent, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by Futuyma, D. , and J. Antonovics. Oxford University Press, New York.

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

Hunley, K. L., M. E. Healy, and J. C. Long, 2009 The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: implications for biological race. Am. J. Phys. Anthropol. 139: 35–46.

Jesus, F. F., J. F. Wilkins, V. N. Solferini, and J. Wakeley, 2006 Expected coalescence times and segregating sites in a model of glacial cycles. Genet. Mol. Res. 5: 466–474.

Kayser, M., 2010 The human genetic history of Oceania: near and remote views of dispersal. Curr. Biol. 20: R194–R201.

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. 13: 235–248.

Klopfstein, S., M. Currat, and L. Excoffier, 2006 The fate of mutations surfing on the wave of a range expansion. Mol. Biol. Evol. 23: 482–490.

Lao, O., T. T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf *et al.*, 2008 Correlation between genetic and geographic structure in Europe. Curr. Biol. 18: 1241–1248.

Le Corre, V., and A. Kremer, 1998 Cumulative effects of founding events during colonisation on genetic diversity and differentiation in an island and stepping-stone model. J. Evol. Biol. 11: 495–512.

Liu, H., F. Prugnolle, A. Manica, and F. Balloux, 2006 A geographically explicit genetic model of worldwide human-settlement history. Am. J. Hum. Genet. 79: 230–237.

Marth, G. T., E. Czabarka, J. Murvai, and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166: 351–372.

McVean, G., 2010 A genealogical interpretation of principal components analysis. PLoS Genet. 5: e1000686.

Meltzer, D. J., 2009 *First Peoples in a New World: Colonizing Ice Age America*. University of California Press, Berkeley, CA.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York, NY.

Noonan, J. P., G. Coop, S. Kudaravalli, D. Smith, J. Krause *et al.*, 2006 Sequencing and analysis of Neanderthal genomic DNA. Science 314: 1113–1118.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko *et al.*, 2008 Genes mirror geography within Europe. Nature 456: 98–101.

Plagnol, V., and J. D. Wall, 2006 Possible ancestral structure in human populations. PLoS Genet. 2: 972–979.

Price, A. L., A. Helgason, S. Palsson, H. Stefansson, D. St. Clair *et al.*, 2009 The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genet. 5: e1000505.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman *et al.*, 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. USA 102: 15942–15947.

Rannala, B., and Z. Yang, 2003 Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645–1656.

Ray, N., M. Currat, P. Berthier, and L. Excoffier, 2005 Recovering the geographic origins of early modern humans by realistic and spatially explicit simulations. Genome Res. 15: 1161–1167.

Reich, D., R. E. Green, M. Kircher, J. Krause, N. Patterson *et al.*, 2010 Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053–1060.

Relethford, J. H., 2008 Genetic evidence and the modern human origins debate. Heredity 100: 555–563.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard *et al.*, 2005 Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 1: 660–671.

Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. 58: 167–175.

Szpiech, Z. A., M. Jakobsson, and N. A. Rosenberg, 2008 ADZE: a rarefaction approach for counting alleles private to combinations of populations. Bioinformatics 24: 2498–2504.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.

Takahata, N., 1993 Allelic genealogy and human evolution. Mol. Biol. Evol. 10: 2–22.

Takahata, N., Y. Satta, and J. Klein, 1995 Divergence time and population size in the lineage leading to modern humans. Theor. Popul. Biol. 48: 198–221.

Takahata, N., S.-H. Lee, and Y. Satta, 2001 Testing multiregionality of modern human origins. Mol. Biol. Evol. 18: 172–183.

Tang, H., D. O. Siegmund, P. Shen, P. J. Oefner, and M. W. Feldman, 2002 Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. Genetics 161: 447–459.

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

Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner, and M. W. Feldman, 2000 Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc. Natl. Acad. Sci. USA 97: 7360–7365.

Tishkoff, S. A., and B. C. Verrelli, 2003 Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu. Rev. Genomics Hum. Genet. 4: 293–340.

Wakeley, J., 1996a Distinguishing migration from isolation using the variance of pairwise differences. Theor. Popul. Biol. 49: 369–386.

Wakeley, J., 1996b Pairwise differences under a general model of population subdivision. J. Genet. 75: 81–89.

Wakeley, J., 1996c The variance of pairwise nucleotide differences in two populations with migration. Theor. Popul. Biol. 49: 39–57.

Wakeley, J., 2009 *Coalescent Theory: An Introduction*. Roberts and Co. Publishers, Greenwood Village, CO.

Wang, C., Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton *et al.*, 2010 Comparing spatial maps of human population-genetic variation using Procrustes analysis. Stat. Appl. Genet. Mol. Biol. 9: 13.

Wang, S., C. M. Lewis Jr, M. Jakobsson, S. Ramachandran, N. Ray *et al.*, 2007 Genetic variation and population structure in Native Americans. PLoS Genet. 3: 2049–2067.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer, Sunderland, MA.

Xing, J., W. S. Watkins, A. Shlien, E. Walker, C. D. Huff *et al.*, 2010 Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. Genomics 96: 199–210.

Communicating editor: L. Excoffier