Effects of the population pedigree on genetic signatures of historical demographic events

John Wakeley[a,1], Léandra King[a], and Peter R. Wilton[a]

[a]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

Edited by John C. Avise, University of California, Irvine, CA, and approved April 19, 2016 (received for review February 13, 2016)

Genetic variation among loci in the genomes of diploid biparental organisms is the result of mutation and genetic transmission through the genealogy, or population pedigree, of the species. We explore the consequences of this for patterns of variation at unlinked loci for two kinds of demographic events: the occurrence of a very large family or a strong selective sweep that occurred in the recent past. The results indicate that only rather extreme versions of such events can be expected to structure population pedigrees in such a way that unlinked loci will show deviations from the standard predictions of population genetics, which average over population pedigrees. The results also suggest that large samples of individuals and loci increase the chance of picking up signatures of these events, and that very large families may have a unique signature in terms of sample distributions of mutant alleles.

coalescence | population pedigree | genealogy | population genetics

The degree to which a sample may be considered representative of a population is a fundamental question in any application of statistics. In the complicated world of evolutionary and population genetics, where it is sometimes not even clear which aspects of ancestry or data should be modeled as random processes, questions of this sort assume greater significance still, and simple mistakes can have drastic effects on inference. These issues are brought to the fore in the field of phylogeography, which was first developed by Avise and colleagues in the 1980s after the introduction of genotyping technologies into evolutionary biology and which takes as its starting point the fact that hierarchical patterns of genetic variation contain information about the locations of populations and species in the past, as well as their relative population sizes and other factors of biological interest (1).

The core debate about randomness in the subsequent development of phylogeography was about whether individual gene genealogies should be treated as outcomes of highly variable random processes, which need to be modeled, or as simple observations from which conclusions about the past may be drawn more or less directly (2–6). There will be cases in which the size and shape of a single gene genealogy contain substantial information about population-level or intraspecific ancestry but, as noted in a recent review (7), this debate has come down on the side of modeling. The reasons for this are that gene genealogies are in fact the results of random processes, likely at the population level but certainly at the level of Mendelian genetic transmission, and that it is not known a priori whether a given set of data comes from one of those cases in which gene genealogies are individually informative (8–10). Although this particular issue may be considered settled, debates about the proper application of random models in phylogeography continue to arise (11, 12).

We consider an additional question about the application of random models that has received comparatively little attention either in phylogeography or population genetics. Namely, what is the extent to which genealogies in the family sense—also known as organismal pedigrees (13) or population pedigrees (14)—constrain gene genealogies and thus genetic variation? Two points distinguish this question from the initial core debate about randomness in phylogeography.

First, whereas in phylogeography the focus has been on the undesirable effects of making inferences conditional on a single gene genealogy estimated from data, here it is on the validity of inferences based on standard population-genetic models that average over population pedigrees when in fact there is only one. It turns out that in relatively large well-mixed populations with constant demography over time, the predictions of standard models are generally quite accurate even though they involve this conceptual error (13, 14). The second point is that the variation we are interested in here is variation among loci for a set of sampled individuals. Even though the population pedigree may itself be the outcome of a random process, all loci in the genome share the same pedigree. The population pedigree should thus be considered a given, fixed quantity because peculiarities of genetic variation among loci in the genome may be due to peculiarities in the pedigree.

Work on the effects of population pedigrees began in 1990 with Ball et al. (13), who made the fundamental observation that standard-model predictions for a single well-mixed population fit the distributions of pairwise measures of diversity among independent loci on a given pedigree surprisingly well. Follow-up work on subdivided populations came to similar conclusions but also illustrated that sampling small numbers of transmission pathways through a pedigree can give results quite different from corresponding standard-model predictions (15) and that pedigrees can substantially affect the probabilities of gene-tree topologies in isolation-by-distance migration models (16). These works used simulations to generate pedigrees and to model genetic transmission within each pedigree.

Chang (17) explored two key aspects of ancestry within population pedigrees analytically, proving for a population of N individuals that (i) the most recent common ancestor of all present-day individuals in the pedigree sense (i.e., an individual through which all present-day individuals are cousins) will typically be observed at $\log_2(N)$ generations in the past, and (ii) by about $1.77\log_2(N)$ generations in the past, the ancestries of all present-day individuals overlap completely. Underlying these results is the key fact that the number of pedigree ancestors of an individual grows by a factor of two each generation. Rohde et al. (18) used simulations and analysis of human population structure and history to suggest that our ancestries overlap in these same ways only a few thousand years ago. The $\log_2(N)$-generation time scale for pedigree ancestry is dramatically shorter than the N-generation time scale for common ancestry in the genetic sense (9), which for humans corresponds to hundreds of thousands of years (e.g., ref. 19).

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "In the Light of Evolution X: Comparative Phylogeography," held January 8–9, 2016, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/ILE_X_Comparative_Phylogeography.

Author contributions: J.W., L.K., and P.R.W. designed research; J.W., L.K., and P.R.W. performed research; J.W., L.K., and P.R.W. analyzed data; and J.W. and P.R.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: wakeley@fas.harvard.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1601080113

Subsequent work using both analysis and simulations has emphasized the rapid approach to equilibrium of shared ancestry in pedigrees. Reproductive values of individuals across the population (20), which are proportional to the probabilities that a genetic lineage sampled randomly today traces back to each individual in a given past generation, reach a stationary distribution on this same $\log_2(N)$ time scale (21, 22). Correspondingly, deviations from the predictions of standard population-genetic models, in particular those of coalescent theory, are evident for distributions of times to common ancestry within the recent $\log_2(N)$ generations but disappear as lineages are traced into the more distant past (14).

Pedigrees are, of course, a mainstay of medical genetics, where they allow powerful inferences about the genetics of human disease (23). These are not population pedigrees, which cover entire populations or species for all times, but partial recent pedigrees of sampled individuals. Pedigree analyses of this sort are being applied to a growing number of natural populations, ones for which patterns of reproductive relationship are known, to disentangle the genetics of complex traits and understand patterns and consequences of inbreeding (24). Observed partial pedigrees have also been used to make inferences about recent historical demography—for example, the French settlement of Quebec (25)—directly from pedigree shape without genetics.

Population pedigrees have less frequently made their way into the models of population genetics. Beyond the examples above (13–16, 21, 22), they have been invoked to study the length distribution of admixture tracts in a descendant population (26) as well as to describe the ways in which ancestors in the pedigree sense are numerous, whereas the genetic ancestors among them are comparatively few (27, 28).

Here, we use simulations to assess the potential for two kinds of demographic events to alter the shape of population pedigrees so dramatically that they have marked signatures on genetic variation across the genome, specifically among independently segregating loci without intralocus recombination. We begin by emphasizing the assumptions of standard population-genetic models, which determine how they should be applied, and the resulting conceptual error involved in using standard models to explain variation across the genome in diploid biparental organisms. The first kind of demographic event we consider is the case of a very large family at some generation in the past. The second is the introduction and sweep through the population of a strongly advantageous mutant allele. In both cases, we ask whether data from unlinked loci will deviate from standard predictions for the same demographies without these special events. We restrict our attention to well-mixed populations. This provides a baseline set of results against which subsequent work (e.g., on geographically structured populations) may be compared.

Two Conceptually Different Random Experiments

One of the most familiar results of population genetics is the probability there will be j copies of an allele in the next generation given there are currently i copies of it in a population of N individuals,

$$P(j|i) = \binom{2N}{j}\left(\frac{i}{2N}\right)^{j}\left(\frac{2N-i}{2N}\right)^{2N-j}, \qquad [1]$$

which is derived from the diploid monecious Wright–Fisher model (20, 29) with the possibility of selfing as a result of random mating or random union of gametes (30). There is no reference to the specific outcome of reproduction among the N individuals (i.e., to what could be called the single-generation pedigree) because Eq. **1** is an average over all possible outcomes of reproduction. Using the theory of Markov processes or diffusion approximations for large N, predictions over longer periods of time can be derived from Eq. **1** (31). Such predictions about the probabilities of outcomes of evolution from a given starting point can be compared directly to the results of laboratory experiments, in which allele frequencies are measured but pedigrees typically are not.

The classic experiments of Buri (32), in which the entire evolutionary process was repeated independently a large number of times, provide the appropriate sort of data. In one experiment, Buri recorded allele frequencies of a selectively neutral mutation (bw^{75}) at the *brown* (eye-color) locus in *Drosophila melanogaster* over 19 generations in 107 replicate laboratory populations. Populations were founded each generation by a random sample of eight male and eight female offspring of the adult flies of the previous generation. Every population began with a relative frequency of 0.5, or 16 copies of the mutant allele out of a total of $2N = 32$. The results are displayed in Fig. 1*A*, with corresponding predictions providing a fit to the data shown in Fig. 1*B*. Over the course of the 19 generations, each population's allele frequency drifted randomly. Some populations became fixed for and others lost the bw^{75} allele. By the end of the experiment, roughly 54% of the populations were monomorphic and the remainder were distributed more or less evenly among the polymorphic allele frequencies.

Now consider another standard population-genetic prediction, in this case for the distribution of the number (K_2) of SNP differences between a pair of sequences at a locus,

$$\begin{aligned} P(K_2 = k) &= \int_0^\infty f_{T_2}(t)P(K_2 = k|T_2 = t)dt \\ &= \frac{1}{\theta+1}\left(\frac{\theta}{\theta+1}\right)^{k} \qquad k = 0,1,2,\ldots. \end{aligned} \qquad [2]$$

Eq. **2** holds under the infinitely many sites mutation model with parameter $\theta = 4N_e u$, in which N_e is the coalescent effective population size (33), and without intralocus recombination (34). The first line of Eq. **2** shows how a typical derivation of this result proceeds by conditioning on the underlying, unknown coalescence time (T_2) between the pair of sequences, that is, with $T_2 \sim$ Exponential(1) and $K_2|T_2 = t \sim$ Poisson(θt). Because the distribution of T_2 is obtained starting from the single-generation probability of coalescence which, like Eq. **1**, is an average over the process of reproduction, the exponential distribution of T_2 is an average over the long-term process of reproduction, or over the population pedigree.

Thus, Eq. **2** is an equilibrium result that captures the balance between genetic drift and mutation. It predicts what would be observed if two sequences at a locus were sampled at random from such a population. For most organisms, it is not feasible to perform long-term experiments analogous to those of Buri (32) to create multiple replicate populations for comparison with Eq. **2** or other similar predictions. Instead, these predictions are applied to datasets of multiple loci genotyped in the same set of individuals sampled from a single population (or species). Although this type of application is conceptually wrong because the loci share the pedigree, standard-model predictions match simulated pedigree-coalescent data surprisingly well for large, well-mixed populations (13, 14).

An example of this standard type of application is given in table 3 of ref. 35, which gives the numbers of loci showing zero, one, two, three, or four SNP differences between pairs of sequences at 12,027 loci ranging in length between 400–700 bp in one of the first major SNP-typing studies in humans. Fig. 2 plots these data alongside the corresponding predictions from Eq. **2**. The coalescent model in Fig. 2 and the more sophisticated one in table 3 of ref. 35, which takes variation in the lengths of loci and the mutational opportunity among loci into

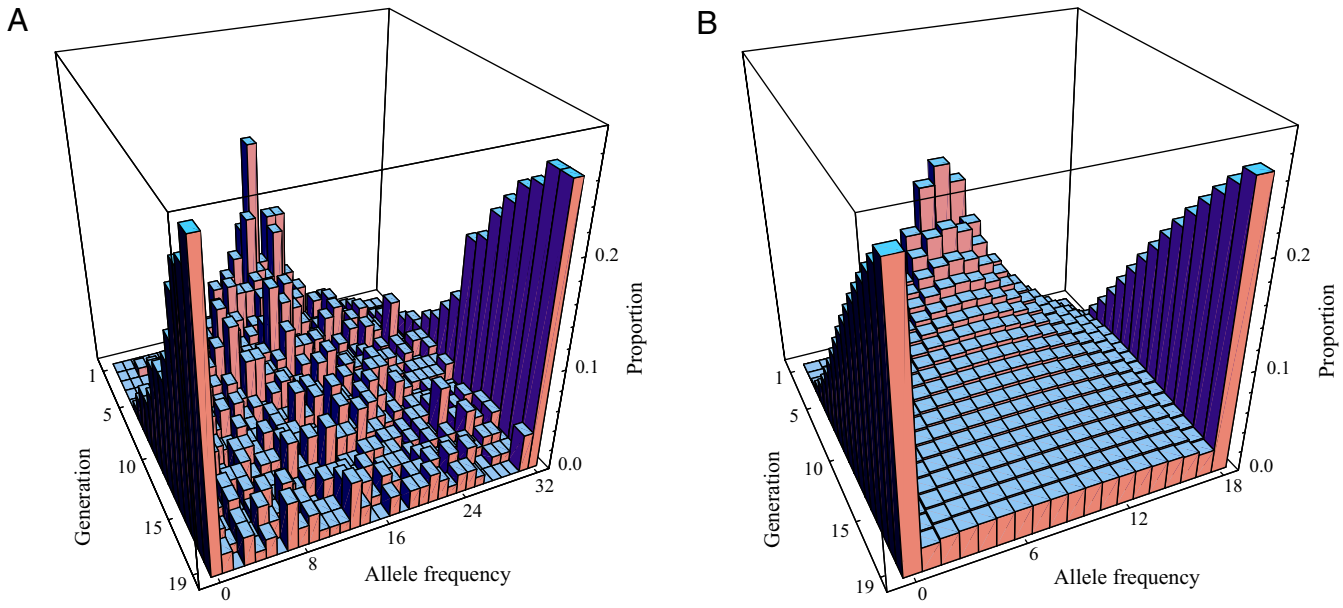

Fig. 1. (*A*) Data from series I (table 13 of ref. 32) for generations 1–19. In each generation, the proportion for each allele frequency is the fraction of the total 107 populations that showed that particular frequency. Generation 0 is not depicted but would have allele frequency equal to 16 and proportion equal to 1. (*B*) Corresponding theoretical prediction using Eq. **1** iteratively, but with the effective population size $N_e = 9$ estimated by Buri (32) rather than $N = 16$ as in the experiment. With $N = 16$, only about 23% of the populations would have been monomorphic by generation 19 instead of the ~54% observed in the experiment.

account, can both be rejected using a χ^2 test. However, it is not clear that this is due to the pedigree, because humans deviate from the assumptions of standard models in other ways (e.g., growth and population structure).

This standard type of application is assumed to be appropriate for loci that are far enough apart in the genome (on different chromosomes in the extreme case) that they assort essentially independently into gametes. Whether or not they assort independently, Eq. **2** is not the correct prediction because Eq. **2** involves the implicit assumption that the loci do not share the same pedigree. Loci on different chromosomes are independent, but only conditional on the population pedigree. They might collectively show patterns of times to common ancestry or genetic variation that depend on the specific features of the pedigree.

In fact, the population pedigree completely determines the probabilities of coalescence in any given generation. Fig. 3 shows a four-generation piece of the Spanish Hapsburg royal family from a study of inbreeding in the demise of this ruling family line (36). Two alleles, one sampled from Mary of Portugal and one sampled from Philip II, would have zero chance of coalescing in the previous two generations, then a substantial probability of coalescing in past generation 3. Thus, the probability of coalescence is not constant over time, as assumed in standard models, and it may not be clear whether it should ever be equal to familiar result $P(\text{coal}) \approx 1/(2N_e)$ even under the idealized diploid, monecious Wright–Fisher model.

Simulations for a variety of models of reproduction show that standard predictions, such as the exponential distribution of T_2, are robust to the presences of the shared population pedigree (13, 14). One exception is when the sample being analyzed has recent common pedigree ancestors, in which case predictions such as Eq. **2** are drastically wrong. However, it is unlikely to sample related individuals from a large population, so the main effect of the shared population pedigree is to make coalescence impossible (as in Fig. 3) until the ancestries of the sampled individuals overlap (14).

In what follows, we consider the effects of extreme pedigrees on distributions of time to coalescence, pairwise SNP differences, and frequencies of mutations in a sample. The results are from simulations of population pedigrees and coalescence of alleles from sampled individuals within pedigrees. In large part, our findings provide further support of the robustness of standard models that average over pedigrees but also suggest that some demographic events might leave signatures detectable in large samples of loci and individuals.

Pedigree Effects of a Large Family

An extensive recent study of human Y-chromosome variation (37) identified a number of descent clusters present at unusually high frequencies in Asia and inferred that these represent the genetic heritages of a corresponding number of highly reproductively successful men. It was surmised that one of these men was Genghis Khan, who had previously been suggested as the

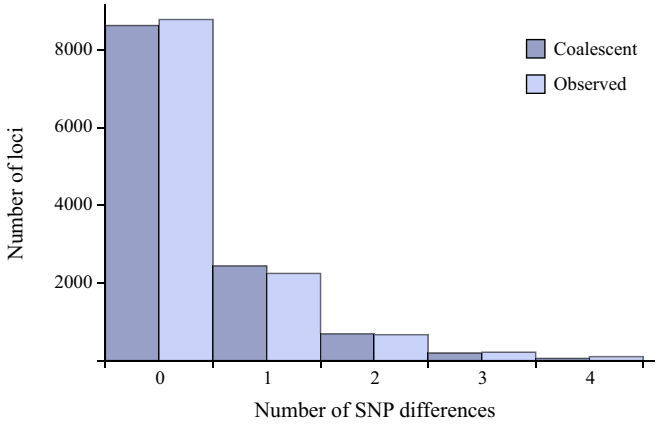

Fig. 2. Observed data from table 3 in ref. 35 and coalescent expectations fitted to have the same average number of SNP differences (0.394), with the distribution truncated at four SNP differences for technical reasons described in ref. 35.

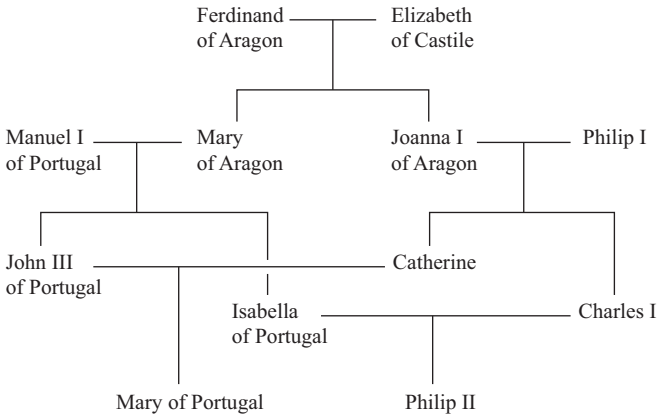

Fig. 3. A small portion of the human population pedigree, from Alvarez et al. (36). Spanish Habsburg King Charles II, who is not shown but would be three generations below, is inferred to have had an inbreeding coefficient of $F = 0.254$ and could neither rule effectively nor continue the family line.

source of a particular Y-chromosome haplotype found at ~8% frequency across a large region of Asia (38). The larger sample and finer-scale geographical sampling seemed to uphold this finding and further revealed substantially higher frequencies of this haplotype in some local populations in central Asia, with one from Middle Kyrgyzstan, for example, showing a sample frequency of ~68% (37).

We consider a hypothetical, extreme scenario based on these inferences about Genghis Khan, in which a single man has a very large number of children at generation 28 in the past. Details of our simulations are given in *Materials and Methods*. We present results for distributions of pairwise coalescence times among autosomal loci in a pair of individuals, assuming independent assortment but conditional on a single shared population pedigree. We also present results for pairwise SNP differences and site frequencies, for which we use $\theta = 0.5$ per locus and assume the infinitely many sites mutation model without intralocus recombination (34). Considering the observed population heterozygosity of about 7×10^{-4} in humans (e.g., see ref. 39), $\theta = 0.5$ corresponds to loci of length ~700 bp, and an average number of SNP differences between pairs of sequences equal to 0.5. Note that the per-site recombination rate is of a similar order of magnitude as the per-site mutation rate in humans (40) and in many other organisms (see table 4.1 in ref. 41). Modeling these relatively short loci, which should have on average only about 0.5 recombination events between a pair of sequences, is one way to minimize the consequences of assuming no intralocus recombination.

Fig. 4*A* shows the probabilities of pairwise coalescence, or the proportion of loci expected to coalesce, in each of the past 40 generations assuming "Genghis Khan's" children comprise 8% of the population. There is very little coalescence in the most recent generations, 1–20, due the strong population growth assumed, but there would still be little coalescence during this time in a population of constant size (here $N = 10{,}000$). In generation 28, there is a spike in the chance of coalescence. Its height is small, though, because coalescence occurs only when both lineages are among that 8% of the population, both trace back to the father, and they descend from the same allele. Thus, the increase in probability is only $0.08^2 \times (1/2)^2 \times 1/2 = 0.0008$.

Looking at the same scenario over the much longer time frame relevant to coalescence, in Fig. 4*B*, this extra mass of coalescence probability has no discernible effect on recent coalescence (leftmost bin in Fig. 4*B*) now corresponding to coalescence within the recent $0.1N$, or 1,000, generations. In sum, we cannot expect to observe the effects of even this fairly dramatic demographic event in a large sample of loci from a pair of individuals, which would amount to taking many random draws from the distribution in Fig. 4. Fig. 4*B* is indistinguishable from the simple coalescent predictions from an exponential distribution with mean 1 corresponding to $2N$ generations.

The situation changes when the children make up 68% of the population. Fig. 5*A* shows a dramatic effect even on the overall distribution of coalescence times. In this case the increase in the chance of coalescence is $0.68^2/2^3 = 0.0578$, which roughly doubles the proportion of loci expected to coalesce within the first $0.1N$, or 1,000, generations. We might expect this increase to be observable in data, for example in pairwise SNP differences. However, for the relatively short (~700 bp or $\theta = 0.5$) loci we model here, a fairly large proportion of loci should be monomorphic even if their coalescence times are greater than $0.1N$ generations. Fig. 5*B* compares a simulated distribution of pairwise SNP differences among loci on a single population pedigree for this case to a simulated distribution for a pedigree with the same demography but without any special demographic event. The distributions differ, but it would take more than 8,300 loci to distinguish between them at the 1% level using a χ^2 homogeneity test.

We also investigated the possibility there would be greater power to detect the pedigree effects of a large family using site-frequency data. We again simulated ancestries of very many loci starting from the same set of individuals sampled without replacement from the current generation, only now we sampled 1,000 individuals and followed 1,000 genetic lineages, creating pseudodata for each locus then counting the number of copies of each mutant in the sample. Fig. 6 shows these "unfolded" site-frequency

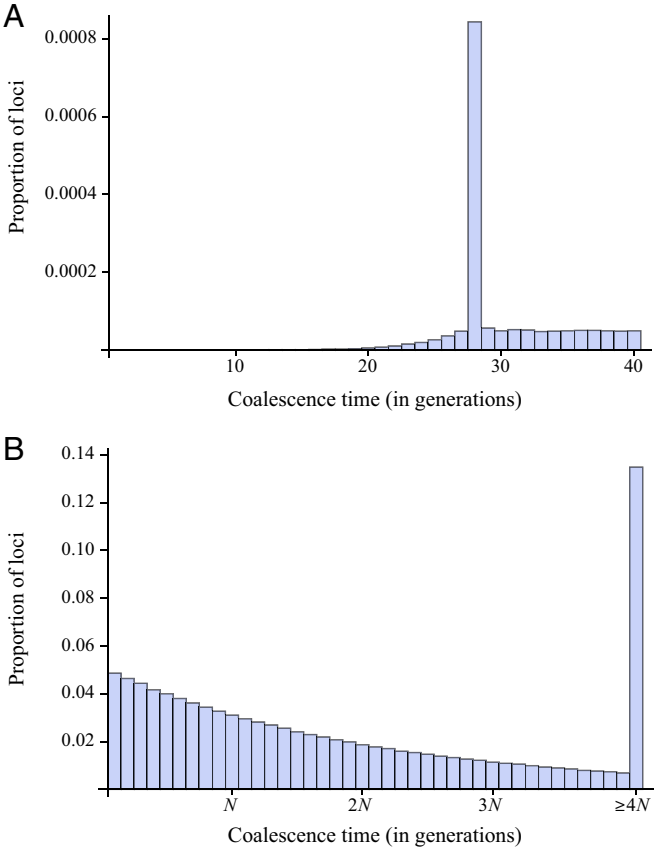

Fig. 4. Simulated distributions of coalescence times conditional on a population pedigree for the case of a large family described in the text, in which the children comprise 8% of the population in generation 27. Each panel is based on a single population pedigree and single pair of sampled individuals. (*A*) Only the most recent generations. (*B*) The whole range of coalescence times on the coalescent time scale of the ancestral population ($N = 10{,}000$). Proportions in *A* are estimated based on 10^8 replicates and in *B* from 10^6 replicates.

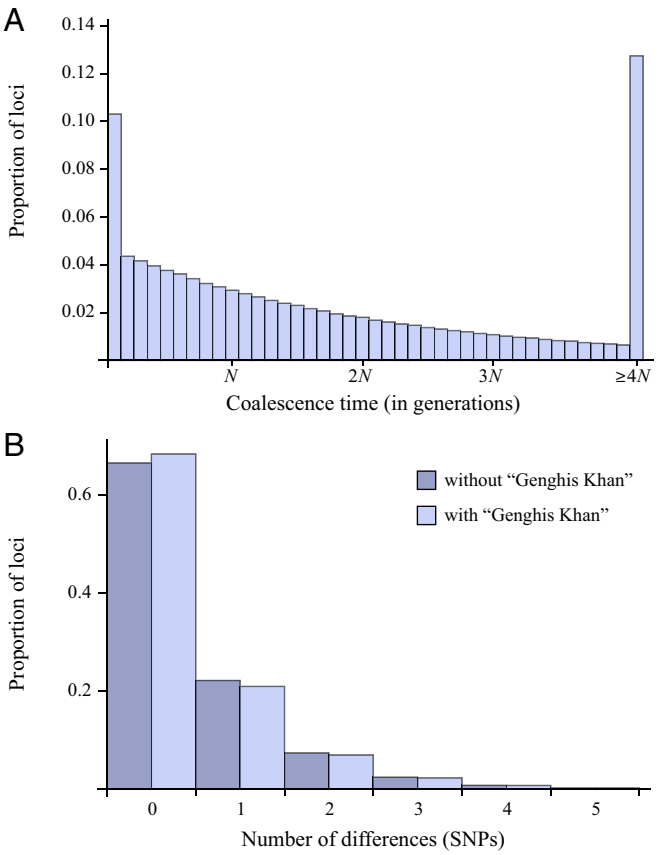

Fig. 5. Simulated distributions of pairwise coalescence conditional on a population pedigree in which the children in generation 27 comprise 68% of the population. Each panel is based on a single population pedigree and single pair of individuals sampled. (*A*) A plot of coalescence times, analogous to Fig. 4*B*. (*B*) The distribution of genetic variation among loci with or without the demographic event of such a very large family. Proportions are estimated based on 10^6 replicates.

distributions (42) for the case in which the children comprise 8% of the population (Fig. 6*A*) and in which they comprise 68% of the population (Fig. 6*B*).

When the children make up 8% of the population, there seems to be no discernible effect on site frequencies, but a striking pattern is observed when the children make up 68% of the population. Differences appear in two parts of the distribution. First, there is a deficit of polymorphic sites at which the mutant is found in about 50–200 copies in the sample. The explanation for this is that many potential branches in the gene genealogy that would have had between roughly 50 and 200 descendants in the sample will be collapsed to zero when bunches of lineages coalesce in "Genghis Khan." Without a large family, these branches would have positive lengths and mutations on them would produce polymorphisms in these site-frequency classes. In the simulations for Fig. 6*B*, an average of 934 lineages remained by generation 27 in the past, so each of the two clusters of coalescent events in "Genghis Khan" involve an average of $943 \times 0.68/4 \approx 159$ ancestral lineages. Thus, there is a deficit of branches with roughly $1{,}000 - 943 = 66$ descendants up to the size of these two clusters (159 lineages each). These calculations are based on average numbers of lineages, whereas the simulations in Fig. 6 include a great deal of variation in each of these numbers and in patterns of coalescence.

The second effect on the site-frequency distribution is an increase in the number of high-frequency derived mutations. Similar patterns have been ascribed to positive selection (43), but U-shaped distributions of allele frequencies are observed within local populations subject to migration (29) and are not unexpected when multiple-merger coalescent events can occur (44). We do not have a quantitative explanation of this pattern in Fig. 6*B*, but, roughly speaking, it is due to the fact that both of the large clusters may be on one side of the root of the gene genealogy. As described in *Materials and Methods*, we verified the overall pattern of site frequencies for this case using a modified set of standard coalescent simulations.

Pedigree Effects of a Selective Sweep

We also investigated the potential of a strong selective sweep to structure the population pedigree in such a way that a genome-wide deviation from the predictions of the standard neutral

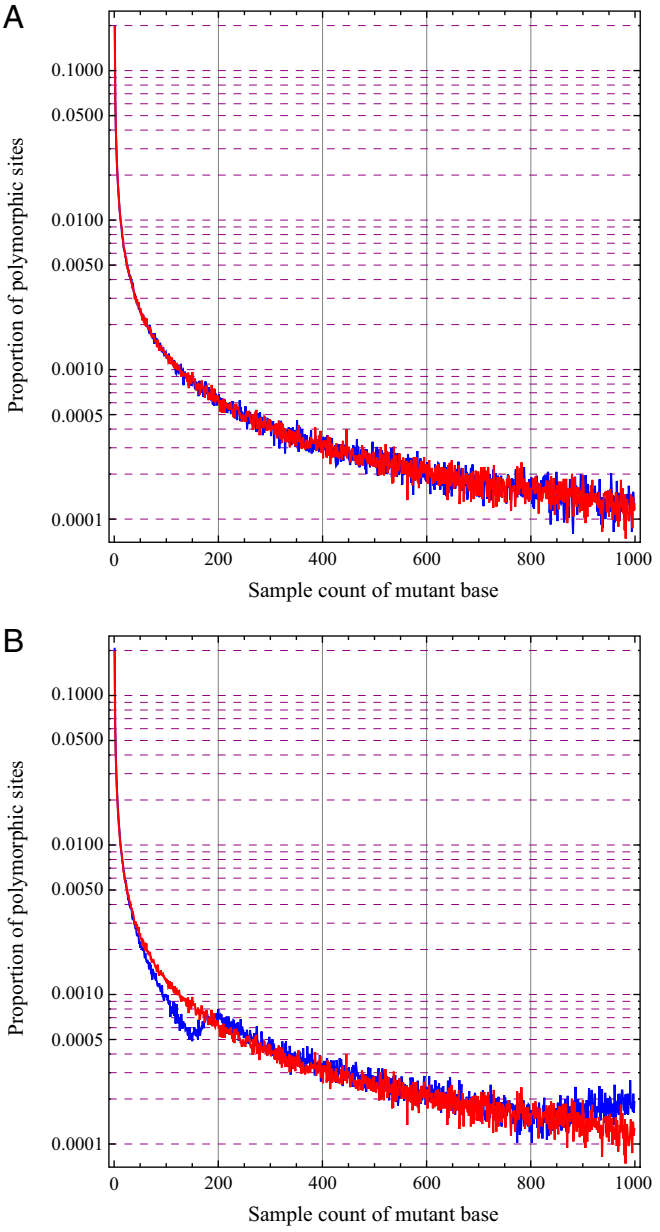

Fig. 6. Unfolded site-frequency distributions when the children of the large family comprise 8% (*A*) versus 68% (*B*) of the population. In both panels, the lines in red display results for the assumed background demography with growth but no large family and are identical in both panels, and the lines in blue show results when there is a large family. The lines in blue in *B* are based on 100,000 replicate loci; the others are based on 10,000 loci.

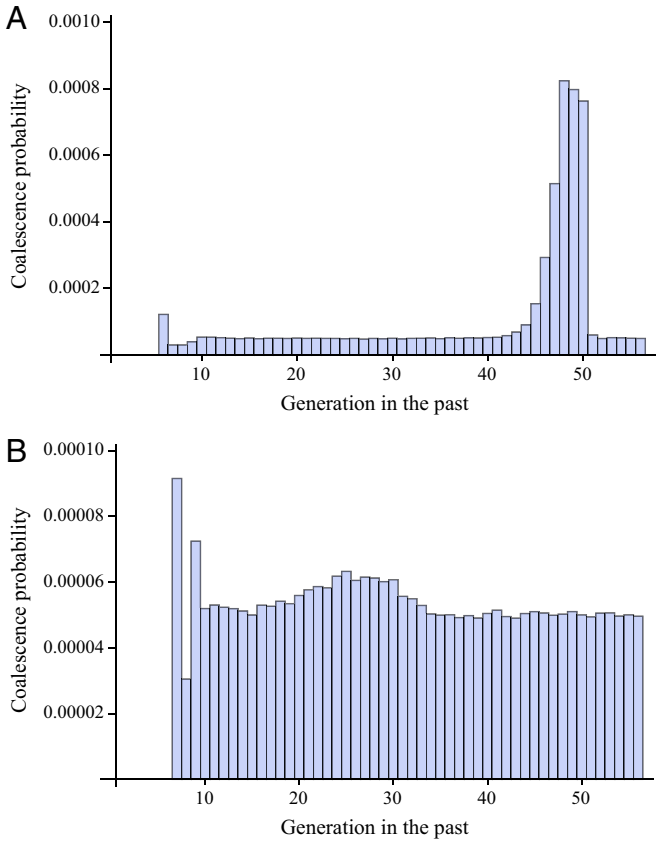

Fig. 7. Distributions of pairwise coalescence times conditional on the population pedigree for the case of a selective sweep, with either $s = 10$ (*A*) or $s = 1$ (*B*). Each panel is based on a single population pedigree and single pair of sampled individuals. In both panels, coalescence probabilities have been computed numerically given each pedigree and pair of sampled individuals.

model would be observed. Whereas the genetic effects of selective sweeps are known to be dramatic for loci linked to a locus under selection (45–47), it is generally understood that unlinked loci are not affected by sweeps. In fact, there is some small effect of a selective sweep even on unlinked loci, which may be attributed to a transient increase of the variance of offspring numbers during a selective sweep (48, 49). To investigate this effect of a sweep as mediated by the population pedigree, we simulated very strong selective sweeps beginning at generation 50 in the past in a population of constant size $N = 10{,}000$, with additive fitness effects of two alleles (*Materials and Methods*).

The pedigree effects of a sweep may be likened to those of a large family, with the family now defined in genetic terms and where the event unfolds over a larger number of generations. Another conceptually similar phenomenon is cultural inheritance of fertility, or correlation in offspring numbers, across generations, evidence for which has been inferred from the shapes of human mitochondrial gene genealogies (50).

Fig. 7 shows probabilities of pairwise coalescence, or proportion of loci expected to coalesce, in each of the past 56 generations assuming a selection coefficient of $s = 10$ (Fig. 7*A*) or $s = 1$ (Fig. 7*B*). When selection is extremely strong, such that individuals homozygous for the advantageous mutant allele have an average of 11 offspring for every one offspring of a wild-type homozygote (Fig. 7*A*), there is a sharp peak in the distribution of coalescence times around the time of the sweep. However, the overall effect on the proportion of loci expected to coalesce during the event is only about four times greater than for our "Genghis Khan" whose children comprise 8% of the population (Fig. 4*A*), and analogously we may infer that even this exceedingly strong selective sweep should have little impact on patterns of genetic variation.

Not surprisingly, the effects of lesser sweeps are very subtle. Fig. 7*B* shows the effect of a sweep with $s = 1$ on probabilities of pairwise coalescence, plotted over the same number of generations as Fig. 7*A* but with a notably different scale on the vertical axis. In this case, where homozygotes for the advantageous mutant allele have an average of two offspring for every one offspring of a wild-type homozygote, there is just a small bump in the proportion of loci expected to coalesce during the sweep, here centered around generation 26.

In contrast to the large-family simulations that included population growth, and therefore showed little coalescence in the first ~ 20 generations, both panels in Fig. 7 illustrate the effect of recent pedigree structure on probabilities of coalescence. In the most recent $\sim \log_2(N)$ generations, here about 13 generations, probabilities of coalescence depend strongly on the ancestries of the two sampled individuals. In the case of Fig. 7*A*, these ancestries did not overlap until generation 6 in the past and in the case of Fig. 7*B* they did not overlap until generation 7 in the past. Tracing farther back, in both cases, the probability then equilibrates and stays near $1/(2N)$, which here is equal to 0.00005 because $N = 10{,}000$.

Fig. 8 provides a more detailed view of the pedigree effects of strong selective sweeps. Ten replicate populations, each with a sweep beginning in generation 50 in the past, were simulated. The probabilities of both coalescence for a pair of lineages and the frequency of the advantageous allele were computed for every generation in the pedigree. These two quantities are shown in Fig. 8 with thicker and thinner lines, respectively, and using different colors for each of the 10 replicates. Fig. 8*A* shows that sweeps with $s = 10$ occur very quickly, in about 15 generations, whereas the sweeps in Fig. 8*B* for $s = 1$ take longer, about 50 generations. Coalescence probabilities for the sweeps in both panels display the relatively great variation over time and among pedigrees in the recent $\sim \log_2(N) \approx 13$ generations as well as the characteristic settling near $1/(2N) = 0.00005$ in the more distant past.

A greater level of variation in the timing of the ten sweeps is visible in Fig. 8*B*, with $s = 1$, than in Fig. 8*A*, with $s = 10$. Fig. 8*B* also shows that differences in the timing of the increase in coalescence probability track differences in the timing of sweeps (distinguished by color). Variation in the timing of a sweep is attributable to the time it takes the favored allele to escape the effects of genetic drift when it is in low copy number in the population. Especially in Fig. 8*A*, it can be seen that coalescence tends to happen earlier in the sweep, when the favored allele is in low frequency (51). Finally, there is greater variation in the additional density of coalescence events among sweeps in Fig. 8*A* ($s = 10$) than in Fig. 8*B* ($s = 1$). We interpret this as a consequence of sweeps happening so quickly when $s = 10$ that coalescences depend strongly on the details of the initial increase of the favored allele.

Conclusions

We have explored two ways in which demographic events within populations may alter the structure of organismal genealogies, or population pedigrees, so as to produce unexpected patterns of variation across genomes. Our simulations of the effects of recent very large families and strong selective sweeps on variation among unlinked loci have primarily yielded negative results. Standard population-genetic predictions that average over pedigrees, such as $P(K_2 = k)$ in Eq. **2**, seem quite robust even to fairly extreme versions of these events. However, we have also shown that frequencies of mutant alleles across the genome in very large samples of individuals provide more sensitive indicators of extreme demographic events, compared with simpler measures such a pairwise sequence differences. Following Keinan and Clark (52), large samples have been of particular interest in human population genetics. For example, the recent update of the 1000 Genomes Project presented site frequencies in a sample of 2,504

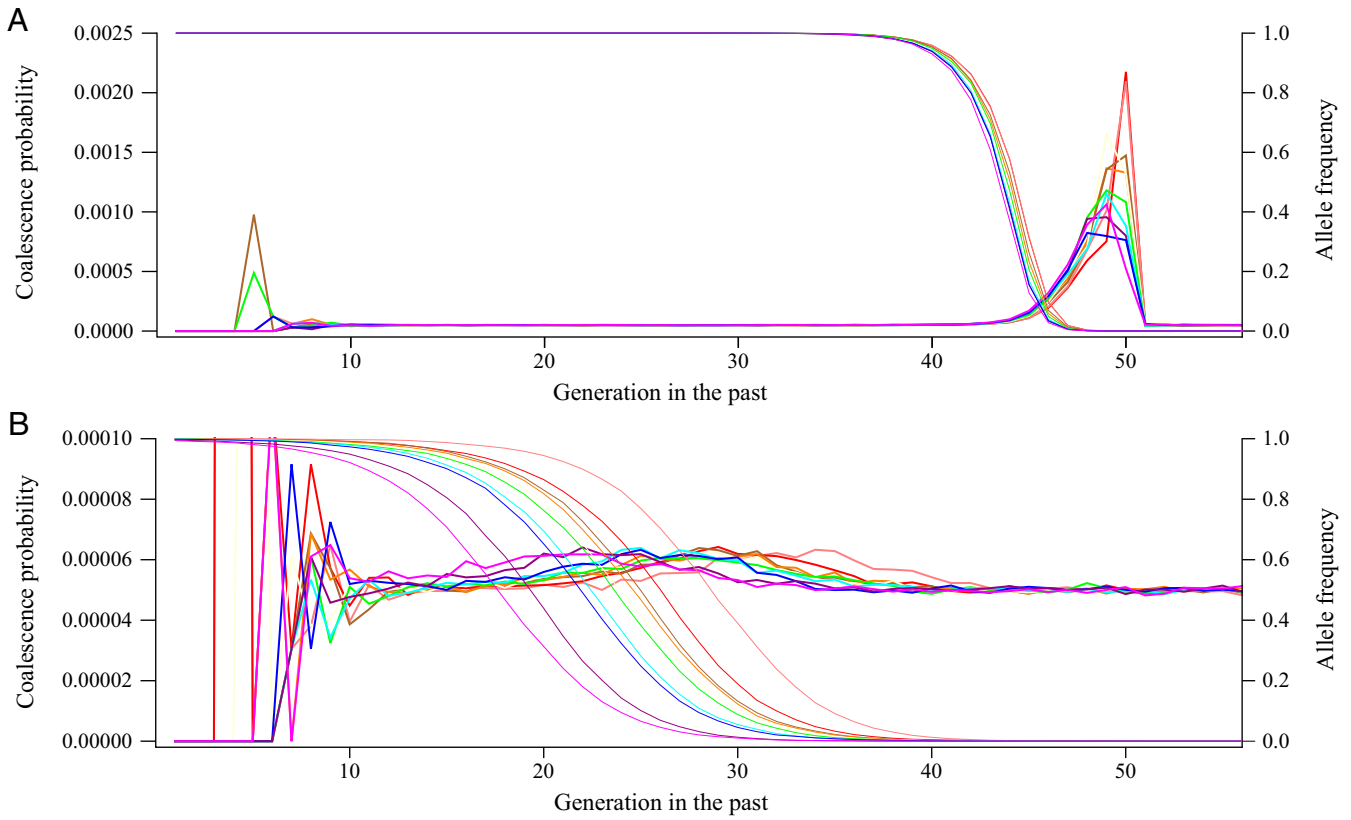

Fig. 8. Distributions of pairwise coalescence times and trajectories of selective sweeps for 10 different replicate populations. As in Fig. 7, $s = 10$ in *A* and $s = 1$ in *B*. The left vertical axes and thicker colored lines plot probabilities of coalescence and the right vertical axes and thinner colored lines plot frequencies of the favored allele as the sweeps progress, in each case beginning with a single copy in generation 50 in the past. In each panel, lines with the same color apply to the same replicate population.

people at more than 80 million SNPs (53), so the potential is there to generate similar data for more geographically localized populations and for species other than humans to investigate the detailed effects of population pedigrees.

Finally, the genetic signatures of recent demographic events that we have uncovered apply marginally to single sites—they do not take linkage and recombination into account—and we note that the pedigree effects of such events might be relatively strong for multilocus measures such as the length distribution of blocks of identity by descent (54, 55).

Materials and Methods

Simulations of Population Pedigrees and Coalescence. We simulated pedigrees according to the diploid, two-sex version of the Wright–Fisher model of random mating. That is, each individual in the next generation (forward in time) has a mother and a father chosen uniformly at random from the female and male adults of the current generation. Given a pedigree, neutral genetic loci are transmitted according to Mendel's laws. Importantly, multiple loci are independent conditional on the pedigree. For each simulated population pedigree, a sample of individuals is taken at random without replacement from the current population, which is generation 0 in the model. A single genetic lineage is followed backward in time from each sampled individual according to Mendel's law of independent segregation (i.e., going with 50% chance to the mother or the father in each generation). When two lineages trace back to the same individual, they coalesce with probability 1/2 and remain distinct in that individual with probability 1/2. For each pedigree and sample, we simulated large numbers of loci that were assumed also to follow Mendel's law of independent assortment. The programs used in this research may be downloaded from wakeleylab.oeb.harvard.edu/resources.

Pedigree Simulations Coalescent with a Large Family. We set the generation in which there was a large family to be generation 28 in the past using the fact that Genghis Khan lived about 800 y ago and a current estimate of 29 y as the average length of one human generation (56). We assume that the children of our "Genghis Khan" comprised either 8% or 68% of the population, and that for the next 27 generations the population grew at rate 0.3 per generation (52), which is similar to estimates of growth for descent clusters in Balaresque et al. (37). The results we present do not depend strongly on this growth because, either way, generation 28 in the past is very recent compared with average coalescence time. We assume an ancestral population size of $N = 10{,}000$, and in every generation there are equal numbers of males and females in the population. Pedigree simulations were as above, except in generation 27 in the past. Depending on the case, in this generation, 27, either 8% or 68% of individuals have our hypothetical "Genghis Khan" as their father. The mothers of these individuals are chosen uniformly at random as in every other generation. When considering sample frequencies of mutations, or site frequencies, we sampled 1,000 individuals. With multiple lineages, multiple coalescent events can occur in single generations, either in different individuals or within single individuals. These multiple mergers are especially important in generation 28 in the past, when large numbers of lineages may trace back to "Genghis Khan." If k lineages trace back to a single individual, each of them has chance 1/2 of descending from each of the two alleles in that individual. Therefore, a binomially distributed number of lineages, with parameters k and 1/2, will trace back to one allele and the remainder will trace back to the other allele in that parent, creating two clusters of coalescence.

In simulating genetic data, we assumed that all mutations are selectively neutral and that each mutation produces a unique polymorphic site (34). For each gene genealogy we placed a Poisson number of mutations randomly on the branches in the standard way to create pseudodata (57), with the modification that our gene genealogies are not necessarily simple bifurcating trees. For a gene genealogy with total length t generations, the number of mutations would be Poisson($\theta t/(4N)$), where N is the ancestral population size, which we set to 10,000. We assumed the mutant state could be distinguished from the ancestral state at each polymorphic site when compiling the site-frequency distribution and simply counted the number of copies of the mutant in the sample of size 1,000. To verify that

the site-frequency distribution shown in Fig. 6*B*, with a deficit of mutant counts around 150 and an increase above about 850, we performed simulations in which a sample of size 1,000 was subject to two rounds of binomial sampling. First, the number of lineages that trace back to "Genghis Khan" was given by a random draw from a binomial distribution with parameters 1,000 lineages and 0.68×0.5 for the probability of being among the children and tracing back to the father. All of these (k) lineages then coalesce into two groups, of sizes k_1 and k_2, one for each of the two chromosomes in the father. We modeled this with a second random draw as described above, that is, $k_1 \sim \text{binomial}(k, 1/2)$ and $k_2 = k - k_1$. We then generated a standard coalescent tree (57) with $1{,}000 - k + 2$ tips, where two tips had k_1 and k_2 descendants in the sample instead of the usual 1 descendant. These simulations assumed that the large-family event occurred instantaneously at time 0 and did not account for population growth, but the results were extremely similar to those in Fig. 6*B*.

Pedigree-Coalescent Simulations of Strong Selective Sweeps. We assumed that a selectively favored allele A was introduced as a mutant in a single copy in the population in generation 50 in the past, into a background of wild-type alleles a. The relative fitnesses of the three diploid genotypes were $AA : 1 + s$, $Aa : 1 + s/2$, and $aa : 1$. We conditioned on allele A not going extinct, by restarting the simulation if A was lost. Pedigrees were simulated from generation 50 onward by sampling parents nonuniformly, in proportion to their relative fitnesses. The population size was held constant at $N = 10{,}000$. Here, starting with a single pair of sampled individuals, we computed probabilities of coalescence in each generation in the past numerically (for example, see equation 3 in ref. 14) rather than by simulating large numbers of independent loci. In addition, for each simulated pedigree with a sweep, we recorded the numbers of copies of allele A present in each generation.

ACKNOWLEDGMENTS. We thank Noah Rosenberg for helpful comments and discussion.

1. Avise JC (2000) *Phylogeography: The History and Formation of Species* (Harvard Univ Press, Cambridge, MA).
2. Knowles LL, Maddison WP (2002) Statistical phylogeography. *Mol Ecol* 11(12): 2623–2635.
3. Hey J, Machado CA (2003) The study of structured populations–New hope for a difficult and divided science. *Nat Rev Genet* 4(7):535–543.
4. Templeton AR (2008) Nested clade analysis: An extensively validated method for strong phylogeographic inference. *Mol Ecol* 17(8):1877–1880.
5. Templeton AR (2009) Statistical hypothesis testing in intraspecific phylogeography: Nested clade phylogeographical analysis vs. approximate Bayesian computation. *Mol Ecol* 18(2):319–331.
6. Nielsen R, Beaumont MA (2009) Statistical inferences in phylogeography. *Mol Ecol* 18(6):1034–1047.
7. Bloomquist EW, Lemey P, Suchard MA (2010) Three roads diverged? Routes to phylogeographic inference. *Trends Ecol Evol* 25(11):626–632.
8. Ewens WJ (1990) Population genetics theory – the past and the future. *Mathematical and Statistical Developments of Evolutionary Theory*, ed Lessard S (Kluwer, Amsterdam), pp 177–227.
9. Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, eds Futuyma DJ, Antonovics J (Oxford Univ Press, Oxford), Vol 7, pp 1–44.
10. Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* 3(5):380–390.
11. Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. *PLOS Comput Biol* 5(9):e1000520.
12. De Maio N, Wu C-H, OReilly KM, Wilson D (2015) New routes to phylogeography: A Bayesian structured coalescent approximation. *PLoS Genet* 11(8):e1005421.
13. Ball M, Neigel JE, Avise JC (1990) Gene genealogies within organismal pedigrees of random-mating populations. *Evolution* 44:360–370.
14. Wakeley J, King L, Low BS, Ramachandran S (2012) Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics* 190(4):1433–1445.
15. Wollenberg K, Avise JC (1998) Sampling properties of genealogical pathways underlying population pedigrees. *Evolution* 52:957–966.
16. Kuo C-H, Avise JC (2008) Does organismal pedigree impact the magnitude of topological congruence among gene trees for unlinked loci? *Genetica* 132(3):219–225.
17. Chang JT (1999) Recent common ancestors of all present-day individuals. *Adv Appl Probab* 31:1002–1026.
18. Rohde DL, Olson S, Chang JT (2004) Modelling the recent common ancestry of all living humans. *Nature* 431(7008):562–566.
19. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nat Genet* 46(8):919–925.
20. Fisher RA (1930) *The Genetical Theory of Natural Selection* (Clarendon, Oxford).
21. Derrida B, Manrubia SC, Zanette DH (2000) On the genealogy of a population of biparental individuals. *J Theor Biol* 203(3):303–315.
22. Barton NH, Etheridge AM (2011) The relation between reproductive value and genetic contribution. *Genetics* 188(4):953–973.
23. Thompson EA (1975) *Human Evolutionary Trees* (Cambridge Univ Press, Cambridge, UK).
24. Pemberton JM (2008) Wild pedigrees: The way forward. *Proc Biol Sci* 275(1635): 613–621.
25. Moreau C, et al. (2011) Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science* 334(6059):1148–1150.
26. Liang M, Nielsen R (2014) The lengths of admixture tracts. *Genetics* 197(3):953–967.
27. Matsen FA, Evans SN (2008) To what extent does genealogical ancestry imply genetic ancestry? *Theor Popul Biol* 74(2):182–190.
28. Gravel S, Steel M (2015) The existence and abundance of ghost ancestors in biparental populations. *Theor Popul Biol* 101:47–53.
29. Wright S (1931) Evolution in Mendelian populations. *Genetics* 16(2):97–159.
30. Watterson GA (1970) On the equivalence of random mating and random union of gametes models in finite, monoecious populations. *Theor Popul Biol* 1(2):233–250.
31. Ewens WJ (2004) *Mathematical Population Genetics, Volume I: Theoretical Foundations* (Springer, Berlin).
32. Buri P (1956) Gene frequency in small populations of mutant Drosophila. *Evolution* 10:367–402.
33. Sjödin P, Kaj I, Krone S, Lascoux M, Nordborg M (2005) On the meaning and existence of an effective population size. *Genetics* 169(2):1061–1070.
34. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7(2):256–276.
35. Sachidanandam R, et al.; International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409(6822):928–933.
36. Alvarez G, Ceballos FC, Quinteiro C (2009) The role of inbreeding in the extinction of a European royal dynasty. *PLoS One* 4(4):e5174.
37. Balaresque P, et al. (2015) Y-chromosome descent clusters and male differential reproductive success: Young lineage expansions dominate Asian pastoral nomadic populations. *Eur J Hum Genet* 23(10):1413–1422.
38. Zerjal T, et al. (2003) The genetic legacy of the Mongols. *Am J Hum Genet* 72(3): 717–721.
39. Palamara PF, et al.; Genome of the Netherlands Consortium (2015) Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *Am J Hum Genet* 97(6):775–789.
40. Reich DE, et al. (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 32(1):135–142.
41. Lynch M (2007) *The Origins of Genome Architectures* (Sinauer, Sunderland, MA).
42. Akashi H (1999) Inferring the fitness effects of DNA mutations from polymorphism and divergence data: Statistical power to detect directional selection under stationarity and free recombination. *Genetics* 151(1):221–238.
43. Fay JC, Wu C-I (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155(3): 1405–1413.
44. Sargsyan O, Wakeley J (2008) A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor Popul Biol* 74(1):104–114.
45. Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genet Res* 23(1):23–35.
46. Kaplan NL, Hudson RR, Langley CH (1989) The "hitchhiking effect" revisited. *Genetics* 123(4):887–899.
47. Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160(2):765–777.
48. Robertson A (1961) Inbreeding in artificial selection programmes. *Genet Res* 2:189–194.
49. Barton NH (1995) Linkage and the limits to natural selection. *Genetics* 140(2):821–841.
50. Blum MGB, Heyer E, François O, Austerlitz F (2006) Matrilineal fertility inheritance detected in hunter-gatherer populations using the imbalance of gene genealogies. *PLoS Genet* 2(8):e122.
51. Pennings PS, Hermisson J (2006) Soft sweeps III: The signature of positive selection from recurrent mutation. *PLoS Genet* 2(12):e186.
52. Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336(6082):740–743.
53. The 1000 Genome Project Consortium (2015) A global reference for human genetic variation. *Nature* 526:68–74.
54. Chapman NH, Thompson EA (2003) A model for the length of tracts of identity by descent in finite random mating populations. *Theor Popul Biol* 64(2):141–150.
55. Palamara PF, Lencz T, Darvasi A, Pe'er I (2012) Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet* 91(5):809–822.
56. Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* 128(2): 415–423.
57. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.