Spread of pedigree versus genetic ancestry in spatially distributed populations

J. Kelleher[a,1], A. M. Etheridge[b,2], A. Véber[c,3], N. H. Barton[d,4]

[a] *Wellcome Trust Centre for Human Genetics*
University of Oxford
Roosevelt Drive
Oxford OX3 7BN
UK
[b] *Department of Statistics*
University of Oxford
1 South Parks Road
Oxford OX1 3TG
UK
[c] *Centre de Mathématiques Appliquées*
École Polytechnique
Route de Saclay
91128 Palaiseau Cedex
France
[d] *Institute of Science and Technology*
Am Campus I
A-3400 Klosterneuberg
Austria

Abstract

Ancestral processes are fundamental to modern population genetics and spatial structure has been the subject of intense interest for many years. Despite this interest, almost nothing is known about the distribution of the locations of pedigree or genetic ancestors. Using both spatially continuous and stepping-stone models, we show that the distribution of pedigree ancestors approaches a travelling wave, for which we develop two alternative approximations. The speed and width of the wave are sensitive to the local details of the model. After a short time, genetic ancestors spread far more slowly than pedigree ancestors, ultimately diffusing out with radius $\sim\sqrt{t}$ rather than spreading at constant speed. In contrast to the wave of pedigree ancestors, the spread of genetic ancestry is insensitive to the local details of the models.

Keywords: Coalescent simulation, ancestral wave, pedigree ancestry, Fisher-KPP wave.

1. Introduction

There has long been interest in the flow of genes through spatially structured populations, and in the ancestral relationships between genes—classically, through the concept of identity by descent (Wright, 1943; Jacquard, 1974), and more recently, through the coalescent (Kingman, 1982; Hudson, 1983b). Yet, until very recently, there has been little work on genealogies within spatially extended populations, or on biparental ancestry, in which the 'pedigree' ancestors of sexually reproducing individuals are traced. Here, we combine these issues by following the spatial locations of pedigree ancestors backwards through time.

The coalescent process describes the ancestry of single genes; the extension to recombination on a linear genome is easily stated (Hudson, 1983a), but leads to an ancestral recombination graph (Griffiths and Marjoram, 1997) which has proved intractable. Surprisingly, there has been little interest in pedigree ancestry, in which each individual necessarily has two parents. The pedigree keeps track of all ancestors of the individuals in our sample, irrespective of whether they carry relevant genetic material. Yet, we can now directly observe pedigree relatedness, over the past few generations by inferring parentage, and up to ~10 generations back, by finding long blocks of shared sequence (Browning and Browning, 2011; Huff et al., 2011). This has allowed the decline in pedigree relatedness with geographic distance to be estimated (Ralph and Coop, 2013).

Chang (1999) showed that in a single well-mixed population, pedigree ancestry mixes rapidly, so that it is almost

Email addresses: `jerome.kelleher@well.ox.ac.uk` (J. Kelleher), `etheridg@stats.ox.ac.uk` (A. M. Etheridge), `amandine.veber@cmap.polytechnique.fr` (A. Véber), `n.barton@ist.ac.at` (N. H. Barton)

[1] Work supported by EPSRC grant EP/I013091/1 and Wellcome Trust grant 100956/Z/13/Z

[2] Work supported in part by EPSRC grants EP/I01361X/1 and EP/K034316/1

[3] Work supported by the chaire Modélisation Mathématique et Biodiversité of Veolia Environment - École Polytechnique - Museum National d'Histoire Naturelle - Fondation X

[4] Work supported by ERC Advanced Grant 'Selection and information' 250152

certain that $\log_2 N$ generations before the present, an individual existed who is ancestral to the whole present-day population. At any time more than $\sim 1.77 \log_2 N$ generations in the past, all individuals that have any descendants in the present day population are almost certainly ancestors of *all* present individuals: in this sense, we all share the same ancestry in the relatively recent past. Nevertheless, the relative contribution of each ancestor (that is, its *reproductive value*) follows a broad distribution: this can be found explicitly, and (conditional on survival) is somewhat less variable than an exponential (Derrida et al., 1999, 2000; Barton and Etheridge, 2011).

This model of biparental inheritance can be seen as the limit for an infinitely long genome. However, any one path through a pedigree of depth t generations will on average transmit only a fraction 2^{-t} of genes, and so even for a very long genome (for example, the human genome, for which the effective population size times the total map length is $N_e R \sim 3 \times 10^5$), most pedigree ancestors may not be genetic ancestors (Donnelly, 1983). There are many open questions concerning the distribution of genetic ancestry, even in the simplest case of a single population.

Rohde et al. (2004) simulated a hierarchically structured population, which aimed to represent human demography over the past few millenia; they found that this structure does not drastically slow the expansion of pedigree ancestry: under their model, we are all likely to share the same ancestry a few thousand years ago. In this paper, we examine a different structure, in which dispersal is local, so that tracing backwards in time, ancestral lineages diffuse out from the location of a sampled individual. This process was first analysed by Wright (1943), for single genes. Wright argued that as we trace back, we can suppose that ancestral lineages follow a Gaussian distribution, with a variance that increases linearly with time; the probability of identity t generations back depends on the products of the densities of these ancestral distributions. This argument cannot be strictly correct: because density must be regulated, the reproduction of nearby individuals must be negatively correlated, so that ancestral lineages cannot move independently in a spatial continuum (Felsenstein, 1975; Barton et al., 2002, 2010b).

This difficulty has traditionally been avoided by assuming a discrete grid of demes, each of which has strictly constant size. We shall consider two such models, a Wright-Fisher model and a Moran model. We also model a truly continuous population, in which reproduction occurs in 'events', which may affect a few individuals, or a substantial fraction of the population. This has a well-defined backwards process, which can readily be simulated, and can include a wide range of biological processes. However, we focus on a special case in which dispersal is strictly local and the effective population density is very high. All three models are defined in Section 2.

In all that follows, we shall assume that individuals are haploid, each with two parents. We consider a finite number l of linearly arranged loci, with recombination occurring with some probability ϱ between any pair of neighbouring loci during a reproduction event (most of the time we shall take $\varrho = 1/2$, but see Section 4.2). We also assume that the population size is large but regulated in such a way that, locally, it does not fluctuate much over time. For simplicity we sample a single individual, from the origin of our space at the present time, and we trace back the spatial distribution of its ancestors as a function of time. We first consider its pedigree ancestors. Since each individual has two parents, sitting in some neighbourhood around it (see Section 2 for more details), initially the population of ancestors behaves like a branching random walk with every ancestor branching into two geographically close (but different) ancestors at some given rate as we go backwards in time. However, because local population sizes remain bounded through time, when the density of ancestors becomes large enough the chance that they all correspond to different individuals in the past population decreases. At that moment, the ancestral lineages that are close-by may either find a common parent and coalesce, or they may escape from each other by having parents that are farther and farther away from the bulk of the ancestral population. We thus expect that the range of the population of pedigree ancestors will expand at some positive speed due to the regions of low density of ancestors at the edge that still develop into the available space like a branching random walk. On the other hand, the number of ancestors in the bulk should saturate to some value due to coalescence. In other words, we expect that the population of pedigree ancestors should behave like a travelling wave and it is one of our main goals to describe this wave, how its characteristics depend on the parameters of the models and to find some appropriate approximations for it under the three models of Section 2. In particular, we shall see that the wave speeds and widths are heavily model dependent.

We then consider the subset of genetic ancestors, that is the individuals that carry some genetic material which is ultimately transmitted to our sampled individual. One of the main differences from the pedigree ancestors is of course that there can be at most l genetic ancestors by assumption. Again, at the beginning the population of genetic ancestors will grow exponentially, but each ancestor will carry less and less genetic material. Although the coalescence of nearby ancestral lineages still occurs, they are less and less likely as the ancestors become spread out in space (recall that two lineages have to be geographically close to have a chance to coalesce and that even then recombination will eventually split them apart again). In the long run, we thus expect to see about l independent lineages diffusing according to symmetric random walks, giving rise to a wave of genetic ancestors expanding at speed $\sqrt{t}$ instead of linearly in time. This is what we shall study in Section 4, mostly through simulations. In contrast to the wave of pedigree ancestors, we shall see that the spread of genetic ancestors is almost independent of the underlying model, provided that we fix certain parameters. More precisely, if we restrict attention to genetic ancestors, then the distribution of ancestral locations depends primarily on the effective population density, ρ_e, and the dispersal rate, σ^2.

In Section 2, we describe our three models and record expressions (which are derived in Appendix A) for the corresponding effective population density and dispersal rate. In Section 3 we consider the location and density of pedigree ancestors in a one-dimensional population, and show that, tracing backwards in time, the ancestors form a wave of advance. We develop two approximations: one, based on Wright's (1943) idea that ancestors diffuse out in a Gaussian distribution, and the other, based on a partial differential equation that approximates the continuum model. Both lead to a travelling wave for pedigree ancestry. Using simulations, we compare the waves that develop under our three models to one another, and to these approximations. We also compare the behaviour to the travelling wave solution to the classical Fisher-KPP equation. In Section 4 we extend this analysis by considering the location and density of the genetic ancestors of a sample. Genetic ancestry soon lags behind pedigree ancestry, and spreads much more slowly, as a diffusion rather than a travelling wave. Although we focus almost exclusively on unlinked loci, we also briefly discuss discrete loci on a linear genome. Following this, in Section 5 we examine the corresponding process in two dimensions and see the same qualitative behaviour. In Section 6 we conclude by discussing these results in the context of previous work.

2. The models

In this section we describe the three different models that we shall study and identify some key summary statistics which allow us to compare them. Our main focus is one spatial dimension, but we describe the continuum model in arbitrary dimensions as it is on this model that the two-dimensional simulations of Section 5 are based.

2.1. Continuum model

In this model, individuals occupy a fixed location in a continuous habitat during their lifetime. All movement, death and reproduction occur as a consequence of replacement events, which fall randomly throughout the habitat. Events span a range of scales, from the regular process of reproduction within neighbourhoods to large scale demographic shifts, in which substantial fractions of the population are affected. A variety of different replacement mechanisms may be employed (Barton et al., 2010b, 2013b), but we concentrate on the well-studied 'ball' model (Etheridge, 2008; Barton et al., 2010a, 2013a) here.

As the basis of this model, we consider a population of individuals distributed uniformly at random with density D on a continuous range which we shall take to be the torus, $\mathbb{T}_L^d$, of side L in $\mathbb{R}^d$. This population evolves through replacement events, which occur at some fixed rate λL^d. At an event, we choose a centre z uniformly at random within the range $\mathbb{T}_L^d$. Then, we let S be the set of individuals within distance r (the event 'radius') of z, and (if S is non-empty) we choose a small number of parents ν uniformly from S. In this article we shall think of each event as a single reproduction event and, since we are interested in biparental mating systems, we shall fix $\nu = 2$. Once we have selected the parents, we then kill a fraction u (the 'impact' of an event) of the individuals in S. Finally, we repopulate the area by throwing down a Poisson number of offspring, with mean $DuV_d(r)$ (where $V_d(r)$ is the volume of the ball radius r in $\mathbb{T}_L^d$), each at a location chosen uniformly from the ball of radius r around z. (See Berestycki et al. (2009) for more details of this prelimiting model in the uniparental case.)

As D tends to infinity, the forwards-in-time process of local allele frequencies converges to the *spatial Λ-Fleming-Viot process*, for which many results have been derived (Barton et al., 2013b). In this limit, the evolution mechanism is essentially the same as in the prelimiting model, except that now density is high enough to ensure that there are always enough individuals in the set S from which to choose parents. We can then define a simple coalescent process, which can be simulated efficiently (Kelleher et al., 2013, 2014). In this coalescent, we begin with a sample of n genes and trace their ancestry backwards in time, recording the effects of events that intersect with ancestral lineages. At an event, any lineage within the affected ball has a probability u of being an offspring of the event, in which case the location of the lineage jumps to the location of a parent of the event, which is uniformly distributed within the ball. If two or more lineages are offspring of the same parent of the same event then they coalesce.

Recombination can be incorporated into this coalescent process in a natural way (Etheridge and Véber, 2013; Barton et al., 2013a; Kelleher et al., 2014). We now sample a set of n individuals with l linearly arranged loci. Recombination occurs when two adjacent loci derive from different parents at an event, which occurs with probability ϱ between pairs of adjacent loci. In the coalescent process, we proceed as before, moving backwards in time until at least one ancestral lineage is an offspring of an event. Then, we distribute the genetic material present in this individual to its parents. The first locus is assigned to a parent (chosen at random). Moving along the genome, all loci are assigned to the same parent until we reach a recombination event (after a random number of loci with a geometric distribution). After the recombination, all loci are assigned to the other parent until we reach the next recombination event when we flip back to the first parent, and so on. Importantly, if we are only interested in genetic ancestry, then if either of the parents does not carry any genetic material ancestral to the sample, we do not trace its subsequent history.

On the other hand, if we wish to trace the history of the pedigree ancestors of our sample, we always follow both parents of each event in which at least one of the lineages we are following is an offspring (even if one of them does not carry any of the genetic material in which we are interested). This leads to a rapidly growing population of ancestors, whose structure we wish to understand. This is primarily investigated using stochastic simulations of the ancestral processes just defined. These simulations use the `discsim` Python module, which is freely available at `https://pypi.python.org/pypi/discsim`. The large effective densities and numbers of loci used in this article would not be possible without recent advances in simulating these ancestral processes (Kelleher et al., 2014). The code used to run the simulations and generate the plots here is freely available under the terms of the GNU General Public License at `https://github.com/jeromekelleher/ancestral-waves`.

In the model, time proceeds by incrementing a global clock by an exponentially distributed value with parameter λL^d each time an event occurs in $\mathbb{T}_L^d$. We refer to time measured in this way as 'model time'. For comparison to other models, it is natural to measure time in terms of generations. One way in which we might define a generation is as the average lifetime of an individual (i.e., the amount of time between when they are born and when they die). Events fall at rate λL^d uniformly over a volume L^d and each event covers a volume of $V_d(r)$, where $V_d(r)$ is the volume of a ball of radius r in d-dimensions. Thus, the expected time until an event intersects with an individual is $1/(\lambda V_d(r))$. Since the individual has a probability u of dying in an event, we therefore know that one generation corresponds to $1/(\lambda u V_d(r))$ units of model time.

2.2. Wright-Fisher Model

In the Wright-Fisher model, we have a set of L demes in a (circular) one-dimensional array with deme spacing equal to 1. Each deme holds N individuals. We shall be interested in the case in which we sample a single individual at the present time, and then trace backwards in time generation-by-generation. For the pedigree simulation, each individual in each deme is represented by an integer which is 1 if the individual is ancestral to the sampled individual and 0 if it is not. In a generation, each individual, independently, chooses two parents at random. Each parent (independently) is chosen from the deme in which the individual is situated or one of the adjacent demes with probabilities $(m/2, 1-m, m/2)$. If an individual is a pedigree ancestor, then both of its parents will also be pedigree ancestors.

An ancestral individual's genetic material is distributed between its parents just as in the continuum model. However, since we shall focus on unlinked loci (corresponding to $\varrho = 1/2$), it suffices to count the number of loci at which the genetic material is ancestral to the sample; their location on the genome is unimportant. In this case, the value associated with an individual is the number of ancestral loci it is carrying. Initially, we sample an individual with l ancestral loci, and we track this ancestral material as we proceed backwards in time. Now, after an individual has chosen its parents, it must divide up its ancestral material between them. Under the assumption of free recombination, if an ancestor carries k ancestral loci, then the number of ancestral loci assigned to the first parent is a random value with distribution $\text{Binom}(k, 1/2)$ and the other parent is assigned the remainder.

2.3. The Moran model

In the Moran model, we also have L demes with N individuals in each. Now, we proceed backwards in time event-by-event. Each event corresponds to the birth of exactly one individual to replace an individual chosen uniformly at random from the whole population. As for the continuum model, we choose the rate of events in such a way that the expected lifetime of an individual, that is the expected time that it waits before it is replaced by the offspring of an event, is one unit of time. Thus events occur at rate LN. When an individual is born, its parents are chosen in the same way as in the Wright-Fisher model. Thus, in the pedigree simulation, we choose two parents independently at random. Each parent is chosen from the deme in which the offspring was born with probability $1-m$, otherwise it is picked from one of the adjacent demes (with equal probabilities). Once again, we sample a single individual from the whole present population (in deme 0, say), and trace back the locations of its ancestors. For the pedigree simulation we label individuals ancestral to the sampled individual with a 1 and all other individuals are labelled 0. At a reproduction event, parents of an individual labelled 1 will change their label to 1. On the other hand, the individual that was replaced in the event is necessarily not an ancestor and so is labelled 0.

For the genetic simulations, genetic material is distributed between parents exactly as in the Wright-Fisher model above.

2.4. Definitions of some key parameters

We define 'effective population density', ρ_e, and the 'dispersal rate', or 'rate of diffusion', σ^2. For the former, for the continuous time models, we let $h(x)$ be the instantaneous rate at which two ancestral lineages currently at separation x coalesce. Then

$$\frac{1}{2\rho_e} := \int_{\mathbb{T}_L^d} h(x)dx,$$

for the continuum model, and

$$\frac{1}{2\rho_e} := \sum_i h(i)$$

for the Moran model (where the sum runs over demes in the discrete torus). For the Wright-Fisher model, $h(i)$ is replaced by the probability that two ancestral lineages coalesce in the previous generation. The dispersal rate is the mean square displacement of an ancestral lineage after one generation. We sometimes call it the 'rate of diffusion' as over large spatial scales the motion of a single ancestral lineage can be approximated by a Brownian motion with this rate of diffusion.

We are interested in populations at high density, corresponding to large N in the discrete models, or small u in the continuum model. In this setting, the number of pedigree ancestors of a single ancestor sampled from the origin, say, will increase to high density and spread out in a 'wave'. In the 'bulk' of the wave (close to the origin), the density of pedigree ancestors will reach a stationary distribution. The mean of this equilibrium will be approximated by a quantity that we shall denote by ρ^*.

Values of the parameters ρ_e, σ^2 and ρ^* are recorded in the table below, with derivations deferred to Appendix A.

	Continuum	Wright-Fisher	Moran
σ^2	$2r^2/3$	m	m
ρ_e	$\frac{1}{2ur}$	$N/2$	$N/2$
ρ^*	$2\rho_e\omega$	$2\rho_e\omega$	$N\varphi(m,N)$,

where

$$\omega = 1 + \frac{W\big(-2e^{-2}\big)}{2},$$

with W the Lambert W function, and

$$\varphi(m,N) = \frac{1 + (1-m)^2 + \frac{2(1-m^2)}{N} - \frac{2(1-m)^2}{N^2}}{3 + (1-m)^2 - \frac{2(1-m)^2}{N}}.$$

3. Pedigree ancestors

In this section we outline some analytic approximations to the wave of pedigree ancestors of our sampled individual before comparing simulations of our three models. We concentrate on one spatial dimension.

3.1. The Gaussian approximation

Following Wright's (1943) argument, we expect that any single lineage will diffuse out in a Gaussian distribution with variance $\sigma^2 t$, as we trace it back t generations into the past. At that time, there will be 2^t pedigree ancestors, and so the density of pedigree ancestry at a distance x from the location from which we took our sample is:

$$\psi(x) = \frac{2^t e^{-x^2/(2\sigma^2 t)}}{(2\pi\sigma^2 t)^{d/2}}, \qquad (1)$$

where d is the dimension. Because local population sizes remain bounded through time, the number of distinct ancestors in a given region can not grow indefinitely and even for modest t this density will far exceed the actual density of ancestral individuals. A crude approximation is as follows. We suppose that the population density is ρ. Consider a small neighbourhood of the point x. If we trace the pedigree of our sampled individual back t generations, then there are on average $\psi(x)$ routes through that pedigree that end at location x. Each such route must lead to an individual that was alive at the point x at time t before the present. If we assume that, for each route, the corresponding individual was picked independently and uniformly at random from those at x, then the total number of times that a given individual at x is picked is approximately Poisson($\psi(x)/\rho$)-distributed. In particular, the probability that it is picked at least once is approximately $1 - e^{-\psi(x)/\rho}$. The expected density of pedigree ancestors is then approximated by $\rho(1-e^{-\psi(x)/\rho})$. In fact, since reproductive value varies considerably, pedigree ancestry will tend to be concentrated into a smaller number of individuals, and so $\rho(1-e^{-\psi(x)/\rho})$ is probably an overestimate of the density of pedigree ancestry; this effect might be included by using some $\rho_e < \rho$.

Let us use the above approximation to describe the spatial distribution of pedigree ancestors at time t in the past. For x small enough, we have $\psi(x) \gg \rho$ and so the density of pedigree ancestors saturates at ρ. When $\psi(x)$ becomes of the order of ρ, the density of ancestors starts decreasing and we reach the edge of the wave. Solving for the x at which $\psi(x) = \rho$ will thus give us the location of the front of the wave. This yields

$$x \sim \sigma t\sqrt{2\log 2}\sqrt{1 - \frac{\log\left(\rho\left(\sigma\sqrt{2\pi t}\right)^d\right)}{t\log 2}} \qquad (2)$$

or approximately, as t becomes large,

$$\sigma t\sqrt{2\log 2} - \sigma\frac{\log\left(\rho\left(\sigma\sqrt{2\pi t}\right)^d\right)}{\sqrt{2\log 2}} + O\left(\frac{(\log t)^2}{t}\right).$$

Using the same approximation, $\rho(1-e^{-\psi/\rho})$, for the shape of the expanding wave of pedigree ancestors, we use ρ times the inverse of the maximum slope (in absolute value), which we denote by w_0, as a proxy for the width of the wave. Since the slope is $\psi' e^{-\psi/\rho}$, we have

$$w_0 := \frac{\rho}{\max_x |\psi'(x)e^{-\psi(x)/\rho}|}.$$

The maximum is attained at the point of inflection which, using our previous calculation, we expect to be at $x \approx \sigma t\sqrt{2\log 2}$. This yields

$$w_0 = \frac{\sigma}{\sqrt{2\log 2}}\frac{1}{Q\log(1/Q)},$$

where $Q = e^{-\psi/\rho}$ at the inflection point. Since we expect $Q \sim 1/2$ there, this suggests that the width will converge to a definite value. Note that the speed of a Fisher-KPP wave with the same intrinsic growth rate is $\sigma\sqrt{2\log 2}$, as in equation (2).

Suppose that there are l unlinked loci; we can imagine these as being thrown down independently onto the pedigree. That is, the ancestral lineages at each of these loci start in the same individual (our sample) and at each time in the past when a pedigree lineage splits into two ancestral lineages, the genetic material that this lineage carries is split between the two parental lineages in such a way that distinct loci 'choose' their ancestors independently. Note that when several pedigree lineages find a common ancestor and merge into a single lineage, that ancestor may thus gather the genetic material corresponding to different loci, leading to a coalescence of some genetic ancestral lineages too. When $2^t \ll l$, pedigree ancestry and genetic ancestry coincide. However, once $l \ll 2^t$, the number of genetic ancestors approaches l, and using the same approximation, is $\sim 1 - \psi^*/\rho$, where ψ^* is defined as ψ but with 2^t replaced by l in equation (1). Once $\psi^* < \rho$ at the origin, the distribution of genetic ancestry will diffuse outwards, with radius $\sim \sigma\sqrt{t}$. This is because over sufficiently long timescales, the ancestral material will spend most of the time scattered across l distinct ancestors, each with ancestral genetic material at exactly one locus, which move according to independent random walks until they are in the same deme. When in the same deme, they have a small chance of coalescing (since N is large), but with high probability they move apart again. Even if they do coalesce, we only expect to wait a short time until the ancestral material is once again split apart into two separate ancestors.

3.2. Approximating the continuum model

We continue to work in one spatial dimension; that is we take a continuous circular range of circumference L. Once again we sample a single individual from the origin and follow the pedigree ancestors as we trace backwards in time. All movement and reproduction occurs as a consequence of replacement events, and so we can reconstruct the entire history of the sample by examining the effects of events that 'hit' the sampled individual and its pedigree ancestors.

For the purposes of this analysis, we fix some of the parameters of the model. In particular, we suppose that the event radius $r = 1$ and the rate per unit area at which events fall $\lambda = 1$. These parameters can simply be seen as scaling factors for the size of the range L and time, respectively. In our simulations, we use a range size $L = 1000$ throughout, and unless stated otherwise, an effective density of $\rho_e = 100$ (corresponding to an impact of $u = 1/200$). As explained in Section 2.1, in order to view our model in units of generation time, we must rescale time by a factor $1/(2u)$.

We wish to quantify the local size of the population of pedigree ancestors at a given time, and so we let $N(t, x)$ be a random variable counting the number of individuals within distance 1 of position x at time t. Note that this is the total number of individuals in an interval of length two and, in particular, is not the same as the density of individuals. The reason for this choice becomes clear in the derivation in Appendix B; we can approximate the expected value of this quantity through an autonomous partial differential equation, whereas considering, for example, the expected density of pedigree ancestors would require an integro-differential equation. The full distribution of $N(t, x)$ is complicated, and we therefore simplify by considering its expectation, $n(t, x) = \mathbb{E}[N(t, x)]$.

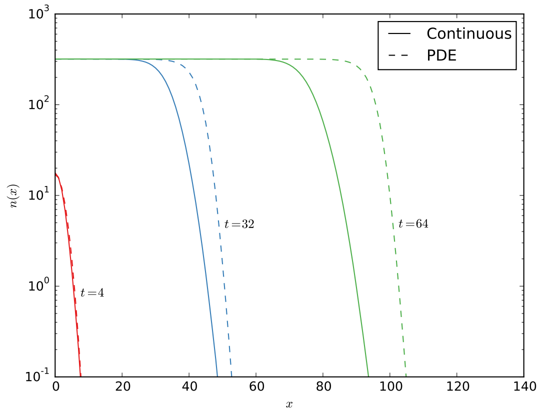

Figure 1: Comparison of numerical solutions of equation (3) with the expected wave from simulations after 4, 32 and 64 generations. Simulation values are the mean $n(t, x)$ over 10000 replicates. Equation (3) was solved numerically using FiPy (Guyer et al., 2009) version 3.1. The apparent widening of the simulated waves is an artefact of the way that replicates were averaged (see text).

As derived in Appendix B, measuring time in units of generations, we approximate n by the solution to

$$\frac{\partial n}{\partial t} = -n + \frac{2}{u}(1 - e^{-un}) + \frac{2}{3}e^{-un}\frac{\partial^2 n}{\partial x^2}. \tag{3}$$

Figure 1 compares the result of numerical solutions of equation (3) with simulation data. For small t, the approximation is very accurate, and the PDE captures the dynamics of the wave very well. However, as the local population size reaches equilibrium near the origin and the wave front becomes established, the numerical solution and observed values begin to diverge. After 64 generations, the numerical solution is well in advance of the observed wave, and becomes less and less accurate over time. The apparent widening of the simulated waves is due to the fact that the results of different replicates have simply been averaged to give an 'expected wave'. Fluctuations in the wave speed for the different replicates leads to a widening of the distribution with time. Figure 2 shows the true shape of the wave front, obtained by averaging wave shapes which are measured relative to their centres. Note that there is no widening effect in the approximation to the expected wave (3) because the fluctuations of the wave speed are lost in the approximation of the nonlinear term.

One reason why equation (3) overestimates the true wave speed is that under this approximation, instead of following a random walk, a single ancestral lineage will follow a Brownian motion. In Appendix D we shall see an analogous effect when we compare the wavespeeds for discrete and continuous space versions of the classical Fisher-KPP equation.

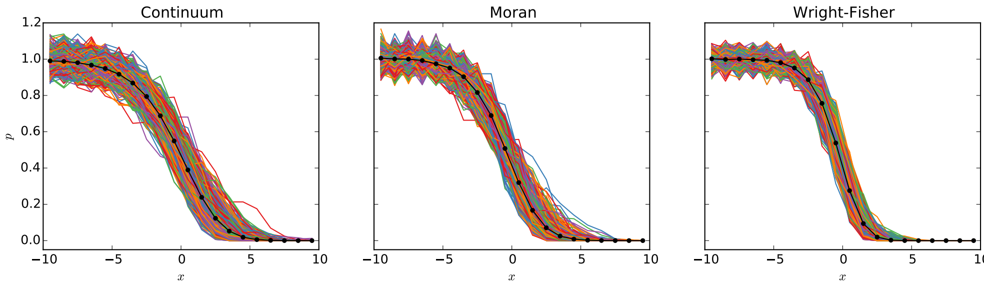

Figure 2: Profile of the pedigree waves in the continuum, Wright-Fisher and Moran models. Simulations for the three models with $\rho_e = 100$ and $\sigma = 1/\sqrt{2}$ were run for 20 generations over 1000 replicates each. For each replicate we then calculated the wave centre and plotted the wave front relative to this. Also shown in plain lines are the mean front shapes. The x-axis shows the distance from the wave centre and the y-axis shows the local population size relative to the mean population size at the origin.

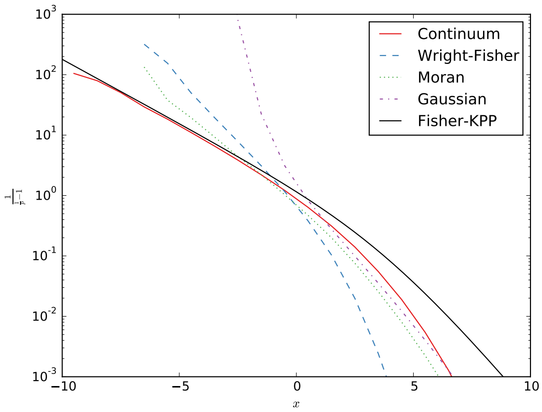

Figure 3: Mean wave shapes for the data outlined in Figure 2 along with theoretical predictions. On the x-axis is distance relative to the front centre, and on the y-axis (on a logarithmic scale) is $p(x)/(1-p(x))$, where $p(x)$ is the ancestral population size relative to the mean size at the origin. (A logistic curve would be linear.) Also shown is the predicted wave shape for the Gaussian approximation along with a solution to the Fisher-KPP equation.

3.3. Shape of the wave front

For all our models we expect the pedigree ancestors to spread out as a travelling wave. In all cases, the wave's behaviour should be at least qualitatively similar to a travelling wave solution to the classical Fisher-KPP equation with the same dispersion and intrinsic growth rate. In Appendix C we show how to calculate the shape of the wavefront in that case. In Figure 3, we compare the mean wave shapes in our three models to that predicted by the Fisher-KPP equation. We see that they are all quite different, suggesting that the shape of the front is very sensitive to the details of the local reproduction mechanism.

3.4. Estimating wave statistics

It is hard to quantify the shape of the front for the waves of pedigree ancestors, so instead we investigate some simple summary statistics. We define $p(x)$ to be the size of the population at x relative to the mean size of the population at the origin. We can then make the following definitions of wave centre z and width w:

$$z = \int p(x)\,dx \qquad w = 4\int p(x)(1-p(x))\,dx.$$

When calculating the width for pedigree waves, we use an arbitrary cutoff to prevent fluctuations in the bulk from accumulating and distorting the calculated width.

In Figure 4 we show simulations of wave centre and width for our three models for three different effective densities. The dispersal rates are chosen to match between the three models. As we would expect from travelling wave solutions, after an initial period as the wave establishes, the centre moves linearly with time and the width remains 'tight'. Note that for all three models these statistics are insensitive to the effective population density. That the speed of the wave should be independent of ρ_e is expected; the wave is a 'pulled' wave, like the travelling wave solution to the classical Fisher-KPP equation, and so its speed is determined by the behaviour in the 'tip' where the density of ancestors is very low and so we do not feel the effect of ρ_e.

We can also ask about the speed of the wave relative to that of the Fisher-KPP equation. In Appendix D we show how to estimate the speed of the travelling wave solution to a discrete deme version of the Fisher-KPP equation. It is dramatically slower than that obtained for the classical continuum version (0.78 vs. 1.177 for the parameters used in Figure 4) and provides a good fit for the speed of the wave in our simulations under the Wright-Fisher model, as shown qualitatively in Figure 4. The continuum estimate is close to the simulated speed for the Moran model, so this slowdown is a result of discretisation of time as well as space.

4. Genetic ancestors

4.1. Unlinked loci

In the standard coalescent with recombination, the number of ancestors carrying genetic material ancestral to the

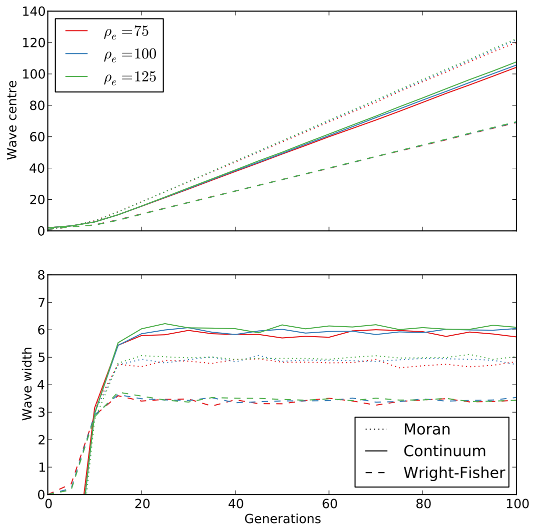

Figure 4: Estimated pedigree wave centre and width for a range of effective densities in the continuum, Wright-Fisher and Moran models. The mean wave centre and width are estimated from 1000 replicate simulations. For each replicate, we estimate the centre and width independently and then take the mean of these values over all replicates.

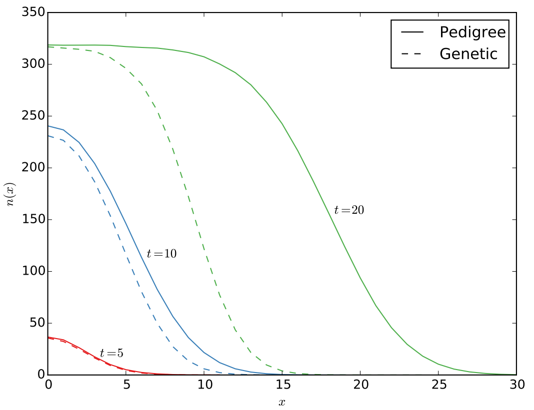

Figure 5: Comparison of the expected waves of pedigree and genetic ancestors after 5, 10 and 20 generations under the continuum model. In the case of genetic ancestors, 10^5 loci were used, with a recombination probability $\rho = 1/2$. In both cases, the mean wave was estimated from 1000 replicates.

sample grows until a steady state is reached. At this equilibrium, the increase in the number of ancestors caused by recombination is balanced by the decrease caused by coancestry. Wiuf and Hein (1997) showed that with a linear genome, the mean number of ancestors at equilibrium is approximately $R/\log(1+R)$, where R is the scaled recombination rate. Derrida and Jung-Muller (1999) obtained more precise results for the expected number of ancestors at equilibrium, although these results are only valid for small chromosome or population sizes.

In this section we are concerned with the same question. What is the number of *genetic* ancestors of an individual for a given number of loci and recombination rate over time? Clearly, we cannot have more genetic ancestors than pedigree ancestors and we cannot have more genetic ancestors than the number of discrete loci. At each event, the ancestral material from one or more individuals is distributed among the two parents. A parent is then a genetic ancestor only if it is assigned at least some material which is ancestral to the sample.

We can find an elementary bound on the probability that in the case of unlinked loci an ancestor will carry material ancestral to the sample at more than one locus: imagine that at time t there are 2^t pedigree ancestors and that for each locus, independently, the ancestral material is in an ancestor chosen uniformly at random among all pedigree ancestors. Then the probability that no pedigree ancestor carries ancestral genetic material at more than one locus is

$$\frac{(2^t-1)!}{(2^t-l)!}\frac{1}{(2^t)^{l-1}},$$

which, by Stirling's formula is approximately

$$\left(1-\frac{l}{2^t}\right)^{l-1/2}.$$

This quantity is small as long as 2^t is at most of order l^2. However for times bigger than around $2\log_2 l$ we expect the genetic material of our sampled individual to be separated into l distinct ancestors and the genetic ancestors then evolve as independent random walks. In fact this approach will somewhat overestimate the chance that all l genes are descended from different ancestors, because of variation in reproductive value.

When large numbers of loci are sampled, the waves of genetic and pedigree ancestors are qualitatively very similar, as we see in Figure 5. In this figure, we simulate the history of the pedigree ancestors as before, and also simulate the history of 10^5 freely recombining loci. In both cases, the population of ancestors around the origin grows until it reaches the equilibrium density, and then forms a wave of advance. In this example, the waves are indistinguishable after 5 generations, but are beginning to diverge after 10. After 20 generations, both waves have reached the equilibrium density, but the pedigree wave has advanced much farther than the wave of genetic ancestors. Note that $2^{10} \ll 10^5 \ll 2^{20}$, so this matches our heuristic

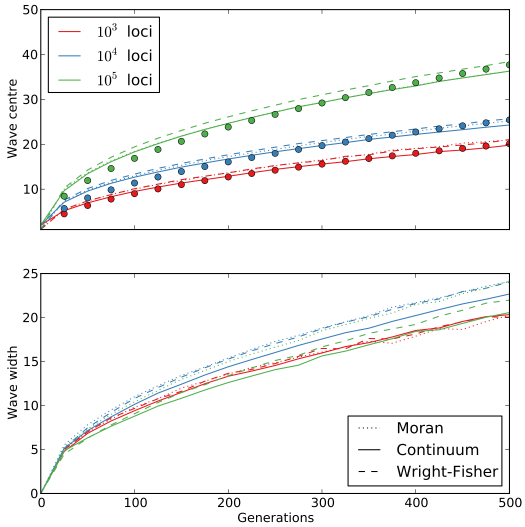

Figure 6: Estimated wave centre and width for the genetic ancestors for different numbers of loci under free recombination ($\rho_e = 100$). Also shown here in dots is $a\sqrt{t}$ fitted for each value of l.

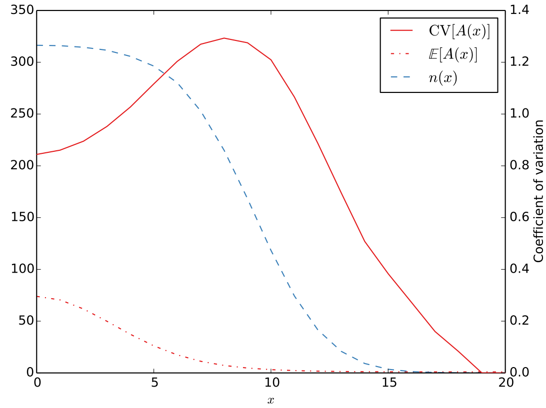

Figure 7: The amount of ancestral material present around a given location as a function of distance and the variation in this value for free recombination. We simulate the genetic ancestry of two individuals with 10^5 loci with a recombination probability of $\varrho = 1/2$ under the continuum model for 20 generations. We show $n(x)$ and the mean amount of ancestral material per ancestor in the interval $[x-1, x+1]$ on the left hand axis. The right-hand axis shows the coefficient of variation of the amount of ancestral material in the interval $[x-1, x+1]$. The mean and variance of the amount of ancestral material in each interval was calculated by accumulating the amount of ancestral material falling in each interval over 1000 replicates.

argument. This is only a single example of a genetic wave, however, since we have used a particular number of loci. In general, the parameters describing the evolution of the centre and of the width of a genetic wave depend logarithmically on the number of loci (see Figure 6). Indeed, the time at which genetic ancestry falls behind pedigree ancestry should be of order at most $\log_2 l$, since the ancestral material can be spread over at most l pedigree ancestors.

The summary statistics for genetic waves are shown in Figure 6, where we plot the wave centre and width against time for different numbers of freely recombining loci. As we would expect from a diffusion, the wave spreads as $a\sqrt{t}$ for a constant a. In contrast to the pedigree wave (see Figure 4), for the genetic wave the three different models match one another very closely.

On a finite range, the wave will continue to collapse and flatten until ancestors are distributed uniformly throughout the range. However, given the slow rate of advance of the genetic wave, the time scale over which ancestors are uniformly distributed is quite large (order L^2). On an infinite range, ancestors spread out indefinitely and the wave becomes more and more diffuse over time.

The lack of ancestral material at the tip of the wave helps to explain the differences in wave speed between the pedigree and genetic waves. In the pedigree wave, if a single individual is born in an event, then we are guaranteed that there will be an increase in the local population as there are exactly two parents in each event. For the genetic wave, however, there must be sufficient ancestral material present to share between two parents. Clearly, if the individual only carries one piece of ancestral material, then recombination cannot occur and the population does not grow.

The expected amount of ancestral material is only part of the picture however; the variation around this value is also very important. Figure 7 also shows the coefficient of variation (i.e., the standard deviation divided by the mean) for the amount of ancestral material per ancestor. This plot shows that, even with free recombination, there is a great deal of variation in the amount of ancestral material the ancestors carry, and that this variation changes with respect to the position of the wave front. At the very tip of the wave there is almost no variation in the amount of ancestral material individuals carry, and variation quickly increases with the local density of ancestors.

It is this variation that makes extending the analysis of Section 3.2 problematic. It is not difficult to write down a system of differential equations in which we track both $n(x)$ and $a(x)$, defined to be the expected number of genetic ancestors within distance 1 of x. We can then write down the increase and decrease in population size using the probability that a given event results in one or two parents, and the change in $n(x)$ and $a(x)$ that results. As we saw, however, tracking the expectation of these values leads to substantial errors due to the nonlinear nature of the system. The variation in the amount of ancestral material per ancestor through space makes this approach unfruitful.

4.2. Linked loci

Free recombination is an interesting limit, but we must also consider the effects on the genetic wave of lower levels

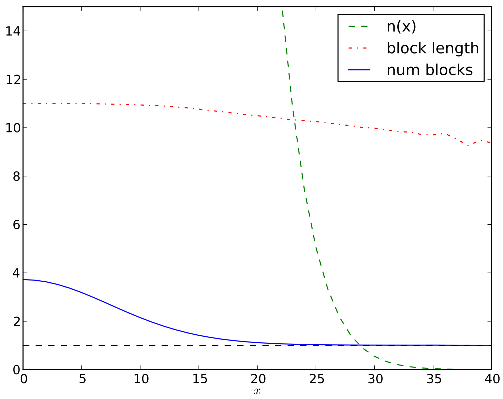

Figure 8: The number and size of blocks of shared ancestry after 100 generations, with 10^5 loci and a recombination probability of $\rho = 10^{-3}$. A block of shared ancestry is defined as a set of contiguous loci sharing the same genealogy. The expected number and length of blocks within the interval $[x-1, x+1]$ were calculated by accumulating blocks within the interval over 10^5 replicates. A large number of replicates is required here to obtain an accurate estimate of the mean length of blocks at the tip of the wave, where there are necessarily only a few individuals in each replicate.

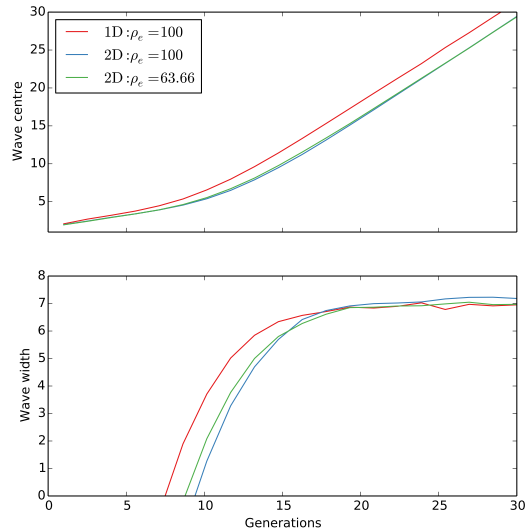

Figure 9: Comparison of the summary statistics for the waves of pedigree ancestors in one and two dimensions under the continuum model, based on simulations. Here we plot the wave centre and width for the wave in 1D with effective density equal to 100, and the waves in 2D for this effective density and also the value of u corresponding to an effective density of 100 in 1D.

of recombination. In Figure 8 we see the wave of genetic ancestors for a linear genome, along with the expected number and size of blocks of ancestry. The number of blocks follows a similar pattern to the total amount of ancestral material seen in Figure 7. The size of blocks is approximately constant until we reach the tip of the wave. This suggests that methods developed to understand the wave of genetic ancestors under free recombination may extend to cover more general recombination rates.

5. Two dimensions

The previous sections have discussed the ancestral wave in the context of a one dimensional continuous habitat, which is of limited interest biologically. Many more species occupy a two dimensional continuum, and so we briefly extend our analysis to illustrate that very similar patterns arise in this case. We concentrate on the wave of pedigree ancestors, and show that the methods derived for a one-dimensional wave generalise quite simply to two dimensions.

The population evolves on a two-dimensional torus of side L, and we sample an individual at the origin at the present time, as before. We then consider the location of its ancestors as we trace backwards in time. The population of ancestors forms a wave of advance spreading radially from the origin. To quantify the size of the local population, we let $N(t, \mathbf{x})$ be a random variable counting the number of individuals within a disc of radius 1 centred at point $\mathbf{x}$ on the torus. (For convenience, we do not distinguish notationally between this and the one dimensional versions of this function from the preceding sections.)

Since the wave is radially symmetric about the origin, it is convenient to consider the expected population size at a given radial distance. Therefore, we let $n(t, x)$ be the expectation of $N(t, \mathbf{x})$, where $\|\mathbf{x}\| = x$. Using this definition, we can then directly compare the waves of ancestors in one and two dimensions.

The results of estimating wave centre and width over time from simulation data for a range of effective densities is shown in Figure 9. In this figure we also show the corresponding estimated summary statistics for the 1D wave for comparison. The 2D wave starts somewhat more slowly that the 1D wave, but it does appear to move at approximately the same constant speed as the 1D wave. At least asymptotically, once again one expects this to be true by analogy with the Fisher-KPP equation. We also see that, over the time scale considered, there is very little difference in the wave speeds for the different effective densities.

The behaviour of the genetic wave is captured in Figure 10. In two dimensions, once ancestral material is separated into distinct pedigree ancestors, it is very unlikely that two individual carrying material ancestral to the sampled individual will become geographically close again and coalesce, so we expect that the genetic wave will rapidly look like l independent random walkers.

6. Discussion

The number of pedigree ancestors of an individual doubles with each generation, so that it takes surprisingly few

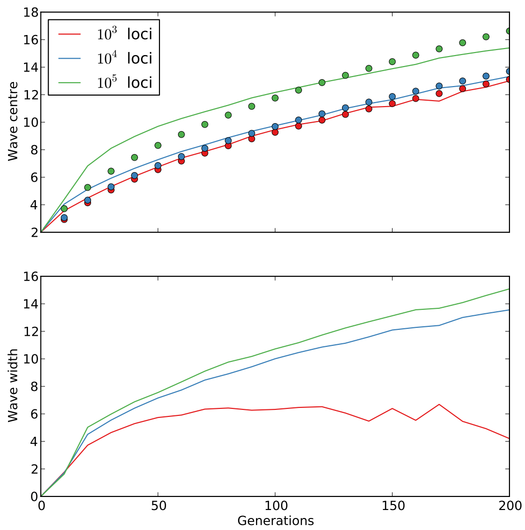

Figure 10: Summary statistics for genetic waves in 2D for the continuum model with $\rho_e = 100$ and disc radius $r = 1$. Solid lines show the mean values estimated from 1000 replicate simulations. Dots show the result of fitting $a\sqrt{t}$ to the data in each case.

generations for every individual to share the same ancestry. Chang (1999) showed that in a single population, all individuals share the same ancestry by $\sim 1.77 \log_2 N$ generations into the past; Rohde et al. (2004) simulated a hierarchical model of migration to show that population subdivision need not substantially slow the sharing of ancestry. Here, we use a different model, in which migration is local, and can be approximated by diffusion. The distribution of pedigree ancestors then spreads as a travelling wave, with a speed that depends on the local details of the model, but even in the fastest case (the Moran model) the speed is approximately one dispersal range per generation ($\sigma\sqrt{2 \log 2}$). This is far slower than the rate of mixing simulated by Rohde et al. (2004), whose model allowed long-range migration from one region to another, so that the whole range could be crossed in a few steps.

We simulated three different models, and found that wave speed and shape differ somewhat between them: this is because the speed and shape of such a travelling wave is determined by details of the reproduction of individuals right at the tip. In such a wave, around the original location from which our present day individual was sampled, there are many pedigree ancestors, each represented by very many routes through the pedigree; but at the tip of the wave, individuals are typically ancestors only via a single route. Because pedigree ancestry spreads as a wave with constant speed, approximately equal to the dispersal distance, ancestry will be shared much deeper into the past than would be the case in a single population. For our own species, assuming a dispersal distance of a few kilometres, it would take some thousands of generations for ancestors to spread world-wide - much longer than the hundreds of generations taken in the model of Rohde et al. (2004). However, occasional long-range movements can greatly accelerate spread, and so the actual time may lie between these extremes. Moreover, all these times are far more recent than the timescale for coalescence of lineages in shared ancestors at individual loci.

Because the number of pedigree ancestors increases very rapidly, as 2^t, it soon becomes much larger than the number of discrete genetic loci, or the number of unrecombined ancestral blocks on a linear genome. Thus, the location of ancestors that actually contribute genetic material to their descendants spreads more slowly than the pedigree ancestors, who become far more numerous: crucially, while the distance of pedigree ancestors from the origin increases linearly with time, genetic ancestry ultimately spreads as $\sqrt{t}$. Comparing Figures 4 and 6 we see that with 10^5 discrete loci, genetic ancestry falls behind pedigree ancestry after ~ 15 generations; this is illustrated in Figure 5. With linkage, the qualitative pattern is the same; for example, Figure 8 shows that with a linear genome of 100 Morgans (about three times the human genome), genetic ancestors occupy a much smaller region than pedigree ancestors after 100 generations (referring back to Figure 4). As one would expect for a simple diffusion, the distribution of genetic ancestors is almost independent of the precise model of reproduction (Figure 6).

Once the pedigree and genetic ancestral waves corresponding to a single individual are understood, it is not difficult to extrapolate to describe the ancestry of a sample of individuals. Indeed, the pedigree ancestry will also behave like a travelling wave: in the bulk, the density of ancestors will remain bounded due to the regulation of local population sizes (resulting from a trade-off between the potentially exponential growth of pedigree ancestors and local coalescence), while the families at the edges will expand thanks to the available space around them from which some of their ancestors will come. A question that may deserve some attention would be whether the ancestors at the edges would always be ancestors of the left-most and right-most individuals of the sample (thinking in one dimension), or whether the ancestry of other individuals may also reach the tip of the pedigree wave. This is left to future enquiries. As concerns the genetic ancestors of the sample, their number is bounded by l times the sample size and so the same reasoning as for a sample of size one applies: the ancestral wave will expand at a speed of the order of $\sqrt{t}$ and its characteristics will depend only on ρ_e and σ^2.

Matsen and Evans (2008) showed that, in an unstructured population, while the pedigree follows a branching process with rate 2 (just as in our spatial extension to their results), genetic ancestry is much more restricted, and the amount of genetic material passed on is only loosely related to the number of genealogical descendants. Indeed, Gravel and Steel (2015) show that a fraction of the ancestral population may be ‘ghosts’, who are pedigree ancestors of all present-day individuals, and yet pass on no genetic material whatever. In our spatial model, pedigree ancestry extends over a much wider area than genetic ancestry, and so most individuals in the pedigree wave that are ahead of the genetic wave are likely to be 'ghosts' for our sampled individual (although they may of course contribute some genetic material to other individuals in the present population).

One might argue that pedigree ancestry is irrelevant, since it cannot be determined for more than a few generations, whether through historical records or by genetic inference. Indeed, this was the view expressed in the discussion that followed Chang (1999)'s first results on pedigree ancestry (Donnelly et al., 1999). However, it is important to understand that all genetic relationships are constrained by the pedigree. Wakeley et al. (2012) show that this constraint makes the standard coalescent seriously misleading over recent generations. Usually, a sample either contains no close relatives, in which case recent coalescence is impossible; however, if it does contain close relatives, then there is an appreciable chance of recent coalescence. Thus, even if an extremely large number of unlinked loci are observed, the distribution of coalescence times will not converge to the standard exponential distribution in any particular sample. This illustrates the more general point that even though pedigree ancestry is essentially unobservable, it is helpful to think of the evolution of a sexual population as consisting in first, the random generation of a pedigree, and second, the random percolation of genes down through this pedigree.

We cannot directly observe pedigree ancestry (except through historical records, which rarely go back for more than ~ 10 generations; (Gillespie et al., 2013). However, whole genome sequencing allows relatives to be identified: individuals are unlikely to share long blocks of identical sequence unless these are derived from a recent common ancestor. This approach allows ancestry shared up to ~ 7 generations ago to be identified from human data (Browning and Browning, 2010; Huff et al., 2011)—and considerably further if phased data were available. It is now possible to sequence ancient DNA, from remains up to ~ 1000 generations old. One might imagine that relationships might then be identified across approximately double the timespan, if a long shared block indicated that the present-day individual descends directly from the ancient individual. However, the timespan will not quite double, because the common ancestor may be some generations earlier than the older sample. Nevertheless, the methods outlined here and in Barton et al. (2013a) allow the distribution of shared blocks to be predicted in such a situation, and might make inference of ancestry feasible over perhaps up to 15 generations. This is just the timescale over which genetic and pedigree ancestry separate: though most pedigree ancestors will not contribute any genetic material, the few that do may contribute blocks of detectable size.

Appendix A. Expressions for the key parameters

We derive the values in the table in Section 2.4 for our key parameters.

Appendix A.1. Effective density

We begin with the continuum model. Recall that effective density is defined via

$$\frac{1}{2\rho_e} = \int h(x)\,dx$$

where $h(x)$ is the instantaneous rate at which genes separated by distance x become identical. Thus, for the ball model with 2 parents in 1D we have

$$h(x) = \frac{1}{2r\lambda u}\frac{\lambda u^2(2r-x)}{2}$$

(recalling that one generation is $1/(2r\lambda u)$ units of model time). Thus, in 1D we have

$$\frac{1}{2\rho_e} = \int_0^{2r} u\frac{(2r-x)}{2r}\,dx$$

and so

$$\rho_e = \frac{1}{2ru}.$$

A similar calculation for 2D gives $\rho_e = 1/(\pi r^2 u)$.

For the Wright-Fisher model, ancestors can only coalesce if they are at most two demes apart and

$$\begin{aligned} h(0) &= \frac{1}{N}\left((1-m)^2 + \frac{m^2}{2}\right), \\ h(-1) &= h(1) = \frac{1}{N}m(1-m), \\ h(-2) &= h(2) = \frac{1}{N}\frac{m^2}{4}. \end{aligned}$$

Summing we obtain that $1/2\rho_e = 1/N$ and so $\rho_e = N/2$.

A similar calculation for the Moran model yields the same expressions as in the Wright-Fisher model, so $\rho_e = N/2$.

Appendix A.2. Rate of dispersal

Evidently $\sigma^2 = m$ for both the Moran and Wright-Fisher models.

For the continuum model, we must calculate the mean square displacement of an individual when it is affected by an event. Since the individual gene's location and the location of the parent from which it was inherited are independently and uniformly distributed on the region affected by the event, the mean square displacement will be $\mathbb{E}[V^2]$, where V is the distance between two independent uniform random variables on $[0, 2r]$. A simple calculation yields $\sigma^2 = 2r^2/3$.

Appendix A.3. Equilibrium density

Finally we compute ρ^*, our approximation for the equilibrium density of pedigree ancestors.

We first work with the continuum model. Let $\mathcal{K}$ be number of ancestors within distance $r = 1$ of a point within the 'bulk' of the wave, so that $n(t, x) = \mathcal{K}$ at equilibrium. In a reproduction event, the number of ancestors that are offspring, and therefore coalesce, is $\mathrm{Binom}(\mathcal{K}, u)$ and so the expected number of unaffected lineages is $\mathcal{K}(1 - u)$. If at least one ancestor was an offspring, which happens with probability

$$1 - (1 - u)^{\mathcal{K}} \approx 1 - e^{-u\mathcal{K}},$$

then we must add two parents to the pedigree. Thus, at stationarity, we have $\mathcal{K}u = 2(1 - e^{-u\mathcal{K}})$, and solving for $\mathcal{K}$ then yields

$$\mathcal{K} = \frac{2 + W\big(-2e^{-2}\big)}{u}, \tag{A.1}$$

where W is the Lambert W function. The density of ancestors at equilibrium is clearly given by $\rho^* = \mathcal{K}/V_d(1)$, and so we have

$$\rho^* = 2\rho_e \left(1 + \frac{W\big(-2e^{-2}\big)}{2}\right) \approx 1.59\rho_e.$$

This result is consistent with the simulations of (Barton et al., 2002), where the effective density was substantially lower than census density.

As we see from Figure 1, this equilibrium density for pedigree ancestors is one aspect of the pedigree wave that is very well captured by equation (3).

As concerns the Wright-Fisher model, suppose that each island contains $\mathcal{K}$ ancestors at equilibrium. Let us focus on island 0. The probability that a given individual on this island in the previous generation is not a parent of one of the $3\mathcal{K}$ ancestors present on islands -1, 0 or 1 is equal to

$$\left(1 - \frac{m}{2N}\right)^{4\mathcal{K}} \left(1 - \frac{1-m}{N}\right)^{2\mathcal{K}},$$

so that the mean number of such ancestors is

$$N\left(1 - \left(1 - \frac{m}{2N}\right)^{4\mathcal{K}} \left(1 - \frac{1-m}{N}\right)^{2\mathcal{K}}\right) \approx N\big(1 - e^{-2\mathcal{K}/N}\big).$$

Solving $N(1 - e^{-2\mathcal{K}/N}) = \mathcal{K}$ gives here again

$$\mathcal{K} = 2\rho_e \left(1 + \frac{W\big(-2e^{-2}\big)}{2}\right).$$

Finally, assuming $\mathcal{K}$ ancestors on each island in the Moran model, the rate $b(i)$ at which the number of ancestors changes by i is given by

$$b(-1) = \mathcal{K}\left\{m^2 + 2(1-m)m\,\frac{\mathcal{K}-1}{N} + (1-m)^2\left(\frac{\mathcal{K}-1}{N}\right)^2\right\}$$

$$\begin{aligned} b(1) &= \mathcal{K}\left\{(1-m)^2\left[\left(1 - \frac{\mathcal{K}}{N}\right)^2 + \frac{1}{N}\left(1 - \frac{\mathcal{K}}{N}\right)\right]\right. \\ &\quad \left. + m^2\left(1 - \frac{\mathcal{K}}{N}\right)\frac{\mathcal{K}}{N} + m\left(1 - \frac{m}{2}\right)\left(1 - \frac{\mathcal{K}}{N}\right)\right\} \\ b(2) &= \frac{m^2}{2}\left(1 - \frac{\mathcal{K}}{N}\right)^2, \end{aligned}$$

$b(0) > 0$ and $b(i) = 0$ for any $i \notin \{-1, 0, 1, 2\}$. Solving for $\mathcal{K}$ such that the mean change in number of ancestors is 0, we obtain

$$\mathcal{K} = N\,\frac{1 + (1-m)^2 + \frac{2(1-m^2)}{N} - \frac{2(1-m)^2}{N^2}}{3 + (1-m)^2 - \frac{2(1-m)^2}{N}}.$$

Note that when $m = 0$ and N is large, we have $\mathcal{K} \approx N/2$.

Appendix B. Deriving the continuum PDE

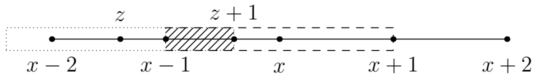

Figure B.11: Illustration of the effects of an event centred at z on the local ancestral population size $N(t, x)$. The dashed outline depicts the region within which individuals are counted by $N(t, x)$, and the dotted outline shows the region in which individuals are affected by the event centred on z. The hatched region then shows the segment within which parents of the event at z must fall in order to increase $N(t, x)$.

Consider the effects of an event centred at z on $N(x)$ as illustrated in Figure B.11 (we suppress the dependence on t for brevity). We are interested in deriving the change in $\mathbb{E}[N(x)]$ as a result of events, and we therefore consider the positive and negative change in turn. There can only be an increase in $N(x)$ if, (a) at least one individual was born in the event (since we are going backwards in time); and (b) if at least one of the resulting parents falls in the line segment $[x - 1, x + 1]$. By definition, there are $N(z)$ individuals within distance 1 of z. The probability that at least one of these individuals was born in this event is therefore $1 - (1 - u)^{N(z)}$, since the probability that none were born is $(1 - u)^{N(z)}$. We write

$$\bar{u}(k) = 1 - (1 - u)^k.$$

Then, given that there was at least one individual born in the event, we know that there are exactly two parents (as this is an assumption of our backwards in time model), which are located uniformly in the segment $[z - 1, z + 1]$. We can then only see an increase in $N(x)$ if one or both of these parents fall in the segment $[x - 1, x + 1] \cap [z - 1, z + 1]$, as shown in the hatched region in Figure B.11. The

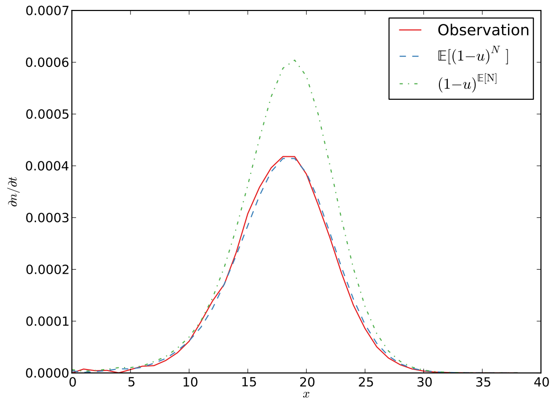

Figure B.12: Evaluation of equation (B.1) from simulation data. To calculate the derivative of $n(t, x) = \mathbb{E}[N(t, x)]$, we estimate $n(t, x)$ after 2×10^6 events, corresponding to 20 generations and after $2 \times 10^6 + 10^4$ events, and then report the difference divided by 10^4. Also shown are the results of evaluating equation (B.1) using $(1 - u)^{\mathbb{E}[N(t,x)]}$ and $\mathbb{E}[(1-u)^{N(t,x)}]$ as described in the text. Expectations were calculated by taking the mean over 10000 replicates.

probability that a given parent falls in this region is $p_z = 1 - |x - z|/2$. Thus, conditional on at least one individual being born, the expected number of parents falling in the interval is $2p_z = 2 - |x - z|$. The mean increase in $N(x)$ due to an event at $z \in [x - 2, x + 2]$ is therefore

$$\frac{1}{4}\int_{x-2}^{x+2} \mathbb{E}[\bar{u}(N(z))]\,(2 - |x - z|)\; dz.$$

To derive the mean decrease in $n(x)$ due to events, we require some new notation. Let $N^*(I)$ denote the number of individuals in the interval I, so that $N^*([x - 1, x + 1]) = N(x)$. We know that only the individuals in $[x - 1, x + 1] \cap [z - 1, z + 1]$ can potentially reduce $N(x)$ for an event centred on z. Since each individual has a probability u of being born (thereby reducing $N(x)$, as we are looking backwards in time), the mean decrease is $uN^*([x-1, x+1] \cap [z-1, z+1])$. Now, during a reproduction event the locations of the parents are always uniformly distributed over the area of the event, and so we make the approximation that individuals are uniformly distributed in the interval considered. We thus have

$$\mathbb{E}[N^*([x-1, x+1] \cap [z-1, z+1])] \approx \mathbb{E}[N(x)]\left(1 - \frac{|x - z|}{2}\right).$$

Recall from Section 3.2 that reproduction events occur homogeneously in space and time at rate 1 and that the region affected by any of them has radius 1 (more general rates and radii can be obtained by a simple change of time and space scales). Since only events with centre $z \in [x - 2, x + 2]$ may affect the individuals present in $[x - 1, x + 1]$, we obtain that the mean reduction in $n(x)$ is given by

$$\frac{un(x)}{4}\int_{x-2}^{x+2}\left(1 - \frac{|x - z|}{2}\right) dz = \frac{un(x)}{2}.$$

In units of generation time, events fall at rate $1/(2u)$ per unit time per unit area. Combining the positive and negative part of the mean change in $n(t, x)$ we have

$$n(t + dt, x) - n(t, x) = \frac{2dt}{u}\left(-\frac{un(t, x)}{2}\right.$$
$$\left.\frac{1}{4}\int_{x-2}^{x+2} \mathbb{E}[\bar{u}(N(z))]\,(2 - |x - z|)\; dz\right) + \mathcal{O}(dt^2).$$

Then, approximating again by assuming that $\mathbb{E}[\bar{u}(N(z))] \approx \bar{u}(n(z))$, we arrive at

$$\frac{\partial n(t, x)}{\partial t} = -n(t, x) + \int_{x-2}^{x+2} \frac{\bar{u}(n(t, z))}{2u}(2 - |x - z|)\, dz. \tag{B.1}$$

Equation (B.1) is a useful description of the dynamics of $n(t, x)$, but it is, unfortunately, only an approximation. The reason for the approximate nature of this equation is that $n(t, x)$ models the *expected value* of the number of ancestors in the region around x, not the full distribution. This leads to a rather poor approximation in this case because the equation contains the nonlinear term $\bar{u}(n(z, t))$. Figure B.12 plots the observed value of $\partial n/\partial t$ from simulations and compares this with predictions made from (B.1). In one prediction we evaluate the equation in the straightforward way, evaluating $\bar{u}(n(t, x))$ directly as $1-(1-u)^{\mathbb{E}[N(t,x)]}$. This method leads to substantial errors, as we would expect from Jensen's inequality. However, if we evaluate $\bar{u}(n(t, x))$ as $1 - \mathbb{E}[(1 - u)^{N(t,x)}]$ (using simulated values) we see a very close agreement between the observed and predicted values of $\partial n/\partial t$. Thus, Figure B.12 shows that equation (B.1) is correct in a limited sense, but is also fundamentally flawed due to the incorrect assumption that $\mathbb{E}[(1 - u)^{N(t,x)}]$ is equal to $(1 - u)^{\mathbb{E}[N(t,x)]}$. In fact, the main problem comes from the large fluctuations in $N(t, x)$ at the tip of the wave and we expect that a better understanding of the actual value of $\mathbb{E}[(1-u)^{N(t,x)}]$ in this region of space should yield a much better approximation of the wave.

Nonetheless, equation (B.1) provides us with a useful approach to characterising the wave of pedigree ancestors, and does predict some aspects of the wave to high accuracy. To simplify, we approximate (B.1) by taking the Taylor series of $\bar{u}(n(t, z))$ to order three about x, obtaining

$$\frac{\partial n}{\partial t} = -n + 2\frac{\bar{u}(n)}{u} - \frac{2}{3u}(1 - u)^n \log(1 - u)^2 \left(\frac{\partial n}{\partial x}\right)^2$$
$$- \frac{2}{3u}(1 - u)^n \log(1 - u)\frac{\partial^2 n}{\partial x^2}.$$

We are most interested in the case of large effective density and we can therefore approximate further in the case of small u to obtain the partial differential equation

$$\frac{\partial n}{\partial t} = -n + \frac{2}{u}(1 - e^{-un}) + \frac{2}{3}e^{-un}\frac{\partial^2 n}{\partial x^2}. \tag{B.2}$$

Appendix C. Wave shape for the Fisher-KPP equation

The shape of the wave front for such a wave is found (after a suitable scaling of space and time) by solving

$$0 = \frac{\partial^2 p}{\partial x^2} + c\frac{\partial p}{\partial x} + p(1-p) = 0. \tag{C.1}$$

Writing $z = \partial p/\partial x$, we see that this is equivalent to solving

$$\frac{\partial z}{\partial p} = -2 - \frac{p(1-p)}{p} \qquad \frac{\partial x}{\partial p} = \frac{1}{z(p)}$$

for $0 < p < 1$. We know that the wavefront decays exponentially (with parameter 1 when $c = 2$) close to the front and so we take initial conditions $z(0.999) = -0.009$ and $x(0.999) = 0$. This gives us a function $x(p)$ mapping the relative population size p to position in the front. We then invert this function to map position x to $p(x)$, and use this function to determine the wave centre z. We then plot $p(x)/(1-p(x))$ (on a logarithmic scale) against $(x-z)/\sqrt{2\log 2}$ to obtain the curve in Figure 2.

Appendix D. Speed of a discrete Fisher wave

Suppose that we replace the classical Fisher-KPP equation by a discrete space version, in which the Laplacian is replaced by the discrete Laplacian. The simplest way to identify the wavespeed for the classical equation is to look for solutions of the form $e^{-\gamma x}$ for equation (C.1). This gives a relationship between γ and c and the speed of the wave is the smallest c for which γ is real-valued. To mimic this in the discrete case, we seek a travelling wave solution of the form z^{i-ct}. This leads to

$$z^{-c} = \lambda\left(1 + \frac{m}{2}\left(z - 2 + \frac{1}{z}\right)\right).$$

For example, setting $m = 1/2$ and solving numerically for z and c, there is a minimum speed at $c = 0.78$ (and $z = 0.124$). This is appreciably slower than the speed of the corresponding wave in a continuum which is $\sqrt{2\log 2} \approx 1.177$.

Acknowledgment

We thank the two anonymous referees for their very careful reading and their comments, which led to many improvements in the presentation of the manuscript.

References

Barton, N. H., Depaulis, F., Etheridge, A. M., 2002. Neutral evolution in spatially continuous populations. Theor. Pop. Biol. 61 (1), 31–48.

Barton, N. H., Etheridge, A. M., 2011. The relation between reproductive value and genetic contribution. Genetics 188, 953–973.

Barton, N. H., Etheridge, A. M., Kelleher, J., Véber, A., 2013a. Inference in two dimensions: allele frequencies versus lengths of shared blocks. Theor. Pop. Biol. 87, 105–119.

Barton, N. H., Etheridge, A. M., Véber, A., 2010a. A new model for evolution in a spatial continuum. Electron. J. of Probab. 15, 7.

Barton, N. H., Etheridge, A. M., Véber, A., 2013b. Modelling evolution in a spatial continuum. J. Stat. Mech. P01002.

Barton, N. H., Kelleher, J., Etheridge, A. M., 2010b. A new model for extinction and recolonisation in two dimensions: quantifying phylogeography. Evolution 64 (9), 2701–2715.

Berestycki, N., Etheridge, A., Hutzenthaler, M., 2009. Survival, extinction and ergodicity in a spatially continuous population model. Markov Process. Relat. Fields 15 (3), 265–288.

Browning, B. L., Browning, S. R., 2010. High resolution detection of identity by descent in unrelated individuals. Am. J. Hum. Genet. 86, 526–539.

Browning, B. L., Browning, S. R., 2011. A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. 88 (2), 173–182.

Chang, J. T., 1999. Recent common ancestors of all present-day individuals. Adv. in Appl. Probab. 31 (4), 1002–1026.

Derrida, B., Jung-Muller, B., 1999. The genealogical tree of a chromosome. J. Stat. Phys. 94, 277–298.

Derrida, B., Manrubia, S. C., Zanette, D. H., 1999. Statistical properties of genealogical trees. Phys. Rev. Lett. 82 (9), 1987–1990.

Derrida, B., Manrubia, S. C., Zanette, D. H., 2000. On the genealogy of a population of biparental individuals. J. Theor. Biol. 203 (3), 303–315.

Donnelly, K. P., 1983. The probability that related individuals share some section of genome identical by descent. Theor. Pop. Biol. 23, 34–63.

Donnelly, P., Wiuf, C., Hein, J., Slatkin, M., Ewens, W. J., Kingman, J. F. C., 1999. Discussion: Recent common ancestors of all present-day individuals. Adv. in Appl. Probab. 31 (4), 1027–1035.

Etheridge, A. M., 2008. Drift, draft and structure: some mathematical models of evolution. Banach Center Publications 80, 121–144.

Etheridge, A. M., Véber, A., 2013. The spatial Λ-Fleming-Viot process on a large torus: genealogies in the presence of recombination. Ann. Appl. Probab. 22 (6), 2165–2209.

Felsenstein, J., 1975. A pain in the torus: some difficulties with the model of isolation by distance. Amer. Nat. 109, 359–368.

Gillespie, D., Russell, A., Lummaa, V., 2013. The effect of maternal age and reproductive history on offspring survival and lifetime reproduction in pre-industrial humans. Evolution 67, 1964–1974.

Gravel, S., Steel, M., 2015. The existence and abundance of 'ghost' ancestors in biparental populations. arXiv:1401.3668v2 [q-bio.PE].

Griffiths, R. C., Marjoram, P., 1997. An ancestral recombination graph. In: Donnelly, P., Tavaré, S. (Eds.), Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications. Vol. 87. Springer-Verlag, Berlin, pp. 257–270.

Guyer, J. E., Wheeler, D., Warren, J. A., 2009. FiPy: Partial differential equations with Python. Computing in Science and Engineering 11 (3), 6–15.

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

Hudson, R. R., 1983b. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37 (1), 203–217.

Huff, C. D., Witherspoon, D. J., Simonson, T. S., Xing, J., Watkins, W. S., Zhang, Y., Tuohy, T. M., Neklason, D. W., Burt, R. W., Guthery, S. L., Woodward, S. R., Jorde, L. B., 2011. Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Res. 21, 768–774.

Jacquard, A., 1974. The Genetic Structure of Populations. Springer-Verlag, New York.

Kelleher, J., Barton, N. H., Etheridge, A. M., 2013. Coalescent simulation in continuous space. Bioinformatics 29 (7), 955–956.

Kelleher, J., Etheridge, A. M., Barton, N. H., 2014. Coalecent simulation in continuous space: algorithms for large neighbourhood size. Theor. Pop. Biol. 95, 13–23.

Kingman, J., 1982. The coalescent. Stoch. Proc. Appl. 13 (3), 235–248.

Matsen, F. A., Evans, S. N., 2008. To what extent does genealogical ancestry imply genetic ancestry? Theor. Pop. Biol. 74 (2), 182–190.

Ralph, P., Coop, G., 2013. The geography of recent genetic ancestry across Europe. PLoS Biology 11 (5), e1001555.

Rohde, D. L. T., Olson, S., Chang, J. T., 2004. Modelling the recent common ancestry of all living humans. Nature 431, 562–566.

Wakeley, J., King, L., Low, B. S., Ramachandran, S., 2012. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190 (4), 1433–1445.

Wiuf, C., Hein, J., 1997. On the number of ancestors to a DNA sequence. Genetics 147, 1459–1468.

Wright, S., 1943. Isolation by distance. Genetics 28 (2), 114–138.