Inference in two dimensions: allele frequencies versus lengths of shared sequence blocks

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

[a] *Institute of Science and Technology*
Am Campus I
A-3400 Klosterneuberg
Austria
[b] *Department of Statistics*
University of Oxford
1 South Parks Road
Oxford OX1 3TG
UK
[c] *Institute of Evolutionary Biology*
University of Edinburgh
Kings Buildings
West Mains Road
Edinburgh EH9 3JT
UK
[d] *Centre de Mathématiques Appliquées*
École Polytechnique
Route de Saclay
91128 Palaiseau Cedex
France

Abstract

We outline two approaches to inference of neighbourhood size, $\mathcal{N}$, and dispersal rate, σ^2, based on either allele frequencies or on the lengths of sequence blocks that are shared between genomes. Over intermediate timescales (10–100 generations, say), populations that live in two dimensions approach

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

[1] Work supported by European Research Council grant 250152

[2] Work supported in part by EPSRC grant EP/I01361X/1

[3] Work supported by EPSRC grant EP/I013091/1

[4] Work supported by the chaire Modélisation Mathématique et Biodiversité of Veolia Environment - École Polytechnique - Museum National d'Histoire Naturelle - Fondation X and by the ANR project MANEGE (ANR-09-BLAN-0215)

a quasi-equilibrium that is independent of both their local structure and their deeper history. Over such scales, the standardised covariance of allele frequencies (i.e. pairwise F_{ST}) falls with the logarithm of distance, and depends only on neighbourhood size, $\mathcal{N}$ and a 'local scale', κ; the rate of gene flow, σ^2, cannot be inferred. We show how spatial dependencies can be accounted for, assuming a Gaussian distribution of allele frequencies, giving maximum likelihood estimates of $\mathcal{N}$ and κ. Alternatively, inferences can be based on the distribution of the lengths of sequence that are identical between blocks of genomes: long blocks (> 0.1cM, say) tell us about intermediate timescales, over which we assume a quasi-equilibrium. For large neighbourhood size, the distribution of long blocks is given directly by the classical Wright-Malécot formula; this relationship can be used to infer both $\mathcal{N}$ and σ^2. With small neighbourhood size, there is an appreciable chance that recombinant lineages will coalesce back before escaping into the distant past. For this case, we show that if genomes are sampled from some distance apart, then the distribution of lengths of blocks that are identical in state is geometric, with a mean that depends on $\mathcal{N}$ and σ^2.

Keywords: Inference, spatial structure, gene flow, F-statistics, identity in state, recombination.

---

1. Introduction

1.1. Motivation

Over the past century, much of population genetics has been devoted to making sense of spatial patterns. Genetic data can be used to estimate rates of gene flow and to infer population history - of interest in themselves, but also important for conservation and population management. Population structure also interacts with selection, impeding adaptation and promoting divergence, and a sound null model of population structure is essential if we are to detect selection at specific loci. A wide variety of methods have been developed for analysing genetic data from spatially structured populations, but these are often ad hoc, with no clear relation to each other or to any theoretical analysis.

Here, we exploit a particular feature of large spatially structured populations: the ancestral lineages of two genes sampled adjacent to one another will either coalesce quickly, or wander away from each other, and coalesce only in the distant past (Wakeley, 2008). This highly variable distribution of coalescence times reflects a separation of timescales between local and global random drift. It is seen both in the island model, and in populations

that are spread over two dimensions. This separation will, in principle, allow robust inference of local population structure, based either on allele frequencies from multiple loci, or on the length of sequence shared between pairs of genomes.

A variety of models for evolution in two dimensions have been proposed. Wright (1943b) and Malécot (1948) proposed a simple diffusion approximation, which, though ill-defined (Felsenstein, 1975), is a close approximation to the discrete stepping stone model (Kimura and Weiss, 1964). We recently proposed an alternative scheme, in which reproduction and dispersal occur through local 'extinction and recolonisation' events, which (in order to incorporate large-scale demographic events) can occur over a range of scales (Etheridge, 2008; Barton et al., 2010a,b). In all these models, over intermediate timescales (large enough that we do not see the details of the local reproduction mechanism, but small enough that we do not see new mutations or the effects of selection or large-scale demographic events), genetic structure is determined by just two parameters: the rate of diffusion of single ancestral lineages, σ^2, and the 'neighbourhood size', $\mathcal{N}$, which is inversely proportional to the probability of 'local' coalescence. (We define $\mathcal{N}$ more precisely in the context of the models employed here in §2.) Moreover, using the fact that the generating function of the time to the most recent common ancestor of two individuals can be interpreted as their probability of identity under an infinitely many alleles mutation model, we can use the classical Wright-Malécot formula to study the coalescence time of the ancestral lineages of a sample of size two taken over such intermediate spatial scales. In all the models above the local population density is fixed. Barton et al. (2002) investigate the probability of identity of two genes sampled from a population which is subject to density dependent regulation of population size. The Wright-Malécot formula remains an excellent approximation to this probability provided that we replace the forwards in time dispersal rate by the dispersal rate of ancestral lineages as we trace backwards in time. This may be quite different from the forward rate: if local population density is regulated through competition, then an individual's ancestors may have survived by moving rapidly away from their own close relatives (Barton et al., 2002). For such populations our methods will yield an approximation for this *backwards* dispersal rate. Indeed, this is all one can hope to infer from any method of inference based on just the current population.

If we are interested in deep history, then only a small sample of individuals is needed - large samples will in any case soon coalesce down to a few ancestral branches. However, many loci must be sampled (ultimately,

whole genomes), since any one locus will have an idiosyncratic history. In contrast, if we are interested in recent history, coalescence is unlikely in the recent past and we have negligible information from mutation in any short sequence. We must therefore either take large samples at multiple loci (as in traditional surveys of F_{ST}), or large samples of a sequence long enough to have accumulated mutations (as in the study of Chinese mtDNA of Kong et al. (2011)), or consider long blocks of sequence and take information from recombination rather than mutation (Ralph and Coop, 2012). In our spatial context, we are interested in estimating the two parameters σ^2 and $\mathcal{N}$. Our aim is to illustrate how, in this setting, any of these sampling strategies can be employed.

2. Allele frequencies vs. block lengths

2.1. Our framework

In all that follows, we shall be interested in a population uniformly spread over some two-dimensional (discrete or continuous) space. In our analysis, we shall always assume that space is infinite, although the simulations presented in the next sections show that the theory applies well to populations with a large but limited range. As mentioned in §1, we shall also assume that the motion of an ancestral lineage can be described by a symmetric random walk (or Brownian motion) with variance parameter $\sigma^2 \in (0, \infty)$, and that two lineages can only coalesce locally, that is, when they are 'reasonably' close to each other. More precisely, we shall focus on models where the probability of identity in state is well-approximated by the classical Wright-Malécot formula: if T stands for the coalescence time of two lineages sampled at some distance x and if μ denotes the rate at which mutations fall on the genealogical tree, this formula reads

$$\mathbb{E}_x\left[e^{-2\mu T}\right] \approx \begin{cases} \frac{K_0(x/\ell_\mu)}{\mathcal{N}+\log(\ell_\mu/\kappa)} & \text{for } |x| > \kappa, \\ \frac{\log(\ell_\mu/\kappa)}{\mathcal{N}+\log(\ell_\mu/\kappa)} & \text{for } |x| \leq \kappa, \end{cases} \tag{1}$$

where $\ell_\mu = \sigma/\sqrt{2\mu}$, $\mathcal{N} > 0$ is the *neighbourhood size* of interest, κ is a local scale over which the probability of identity in state is approximately constant (see Appendix A for the details) and K_0 is the modified Bessel function of the second kind of degree zero.

The fact that (1) holds in the 2d stepping-stone model with deme size $2N$ and a finite-variance symmetric migration kernel is well-know, but for the ease of reference we derive it in a general setting in Appendix A. On an

infinite space and with the motion of a lineage determined by a (discretized) Gaussian kernel, we have $\mathcal{N} = 4N\pi\sigma^2$. Equation (A.4) gives a more general expression for $\mathcal{N}$. Most of the analysis and simulations presented here will be carried out under the 2d stepping-stone model.

To deal with populations distributed over a continuous space, we shall also use the *spatial* Λ*-Fleming-Viot model* mentioned in §1. We shortly describe it in §3, and show that the Wright-Malécot formula also provides an excellent approximation to the probability of identity of two individuals under this model. In this framework we have $\mathcal{N} = \nu/u$, where ν is for the number of potential parents during a single reproduction event, and u is the fraction of the local population replaced during this event. As an aside, at the end of §3, we illustrate how large scale events, deep in the history of the population, leave identity unchanged over small scales, even though their effects are clearly visible over large scales, in support of the claims underlying the approach to inference of the parameters governing local structure adopted here.

Symmetric migration in two dimensions and over large timescales implies a separation of timescales for the behaviour of two near-by lineages: because their separation behaves like a 2d random walk with finite variance, either they will coalesce quickly, or they will separate at large distance and take such a long time to come back together, that mutations are most likely to hit them before their coalescence. Here we implicitly assume that mutation is so slow that only the pairs of lineages that manage to escape from each other have the time to become differentiated before they coalesce again. This assumption is reasonable if we concentrate on intermediate sampling distances x such that $x^2/\sigma^2 \ll \mu^{-1}$ (observe that $\mathcal{O}(x^2/\sigma^2)$ is the minimal amount of time required for the lineages to meet and coalesce). Our goal is to exploit this phenomenon to infer (a subset of) the parameters σ^2, $\mathcal{N}$ and the less interesting κ.

The last key idea that we shall use derives from another separation of timescales argument. Indeed, although it is unlikely that the whole population reaches to a global equilibrium in $\mathcal{O}(\mu^{-1})$ units of time (for example, in high latitudes at least, species' ranges have changed drastically over times much shorter than those set by species-wide coalescence and by mutation), a 'quasi-equilibrium' will be reached over intermediate spatial and temporal scales. This is not a new idea (Slatkin and Barton, 1990), though it has largely been ignored. As a consequence, if we consider a large sampling area A whose diameter satisfies $\mathrm{Diam}(A) \ll \sigma\sqrt{\mu^{-1}}$, we can find a timescale t^* such that $\mathrm{Diam}(A)^2/\sigma^2 \ll t^* \ll \mu^{-1}$ and within the last t^* generations, the allelic distribution in A has remained approximately constant and no new

mutations have appeared. In particular, the alleles present in the sample are all distinguished from each other by mutation much deeper than t^* in the genealogy.

Thus, we fix some 'reference time', $t^* \ll \mu^{-1}$, a few hundred generations ago say, and consider a sampling area A such that $\mathrm{Diam}(A) \ll \sigma\sqrt{t^*}$. According to the argument of the previous paragraph, by time t^* in the past a significant amount of the 'fast' local coalescence of ancestral lineages has taken place, but those lineages that manage to escape far from one another have not come back together yet. The separation of timescales arguement for lineages then implies that most of these 'escapees' will have the time to accumulate mutations before they come back together, rendering their long excursion away from each other observable.

As a first approach to the inference of $\mathcal{N}$ and σ^2, we propose an analysis based on an analogue of F_{ST}. As in an island model, where the migration rate between discrete demes and neighbourhood size can only be inferred from F_{ST} if the population has reached an equilibrium between gene flow and random drift (Whitlock and McCauley, 1998), our proposal relies heavily on the 'quasi-equilibrium' of the allele frequencies reached over the intermediate area A. It also uses the Wright-Malécot formula (1) extensively. The advantage of our approach is then that through using only genetic structure over these intermediate scales, our estimates will be robust to the deep, and possibly complex, history of the population - and specifically to the effects of occasional major extinctions and recolonisations. We provide some justification for this claim at the end of §3.

In a second step, we perform an analysis of long blocks of genome in §5.1 which will also rest on the Wright-Malécot formula. The latter will allow us to calculate the distribution of block lengths in the case when all recombinations are 'effective'. These results are only valid for large neighbourhood size, which we illustrate through simulations of the spatial Λ-Fleming-Viot model. Finally, we use this model to investigate what happens when neighbourhood size is small and many recombinations are ineffective.

2.2. Inference based on allele frequencies

2.2.1. An appropriate F-statistic

The basic theory of allele frequency-based inference for spatially structured populations was set out by Wright (1943b) and Malécot (1948) in the middle of the last century, yet the commonly used statistical tests are based on the island model which fails to incorporate the limited dispersal characteristic of most species (Meirmans, 2012). The most widely used statistic is Wright's F_{ST}, which is simply a standardised variance of allele frequency,

and contains no information on spatial location. Wright (1943b) did include spatial structure by calculating variance over different scales, giving a hierarchy of F statistics; this approach he applied to data on flower colour in *Linanthus* (Wright, 1943a) and later work examined lethal allelism in spatially continuous habitats (Paik and Sung, 1969; Wallace, 1966). However, all these studies were limited by the difficulty of the computations and by lack of genetic markers. We now have an abundance of computational power and of genetic data, and yet most analyses are either descriptive (e.g. spatial autocorrelation; see Epperson (2003)) or use Monte Carlo methods to fit data (Beerli and Felsenstein, 2001; Rousset and Leblois, 2007), without using the explicit theory developed by Wright and Malécot. A wide variety of other F-statistics have also been proposed (Sokal et al., 1989; Slatkin and Barton, 1990; Slatkin and Arter, 1991; Epperson, 2003; Rousset, 2003), but for the most part they have not been justified for application to a two-dimensional population.

Our proposal in §4.1 rests on a variant of Wright's F_{ST}. We first express it in terms of the probability of coalescence of two ancestral lineages before our reference time t^*. Theoretical predictions for these probabilities can thus be obtained by (numerically) inverting the generating function of the coalescence time (determined by the classical Wright-Malécot formula which, for ease of reference, we derive in Appendix A). However, for large neighbourhood size and under our separation of timescales assumption there is a simple analytic approximation to the distribution of the F-statistic, which we obtain in (8): if $r > 0$ stands for the distance between two given sampling locations in A, we show that

$$F(r) \approx \frac{\log(\bar{r}/r)}{\mathcal{N} + \log(\bar{r}/\kappa)},$$

where $\bar{r}$ denotes the geometric mean of the separation of all the individuals sampled in A in the process of estimating the global allele frequencies, and $\mathcal{N}, \kappa$ are the population parameters appearing in the Wright-Malécot formula. Note that the last parameter of the Wright-Malécot formula, σ^2, does not appear in $F(r)$.

We then show that $F(r)$ can be estimated from the statistic

$$\mathcal{F}(x,y) = \frac{1}{a-1} \sum_{i=1}^{a} \frac{(p_i(x) - \bar{p}_i)(p_i(y) - \bar{p}_i)}{\bar{p}_i}, \tag{2}$$

where a is the total number of alleles seen in the region A, $p_i(x)$ is the local frequency of allele i at site x and $\bar{p}_i$ is the frequency of allele i in the whole region A.

2.2.2. Inference in practice

Based on the results described in the previous paragraph, how should we estimate neighbourhood size from allele frequencies? The simplest method is to sample individuals from n distinct locations and regress pairwise F_{st} against the logarithm of distance (Rousset, 1997, 2003). (Actually, Rousset suggests regressing $F_{st}/(1-F_{st})$ against $\log r$, whereas our derivation suggests that regressing F_{st} on $\log r$ is more natural; this makes little difference in practice, however, unless F_{st} is unusually large). The n^2 points in this regression are not independent, but the accuracy of estimates can be found by bootstrapping (Rousset, 1997). Rousset and Leblois (2007) have implemented a Monte Carlo method which uses coalescent simulations to take account of the full distribution of allele frequencies. However this is computationally demanding (especially in two dimensions), and we argue here that it is unlikely to yield information about more than just $\mathcal{N}$. In principle, there is information in higher-order relationships - for example, in the rates of multiple mergers, which may be significant in two-dimensional populations with small $\mathcal{N}$. However, we are concerned with local patterns, for which current mutation is negligible: all that we can observe are the frequencies of alleles that are distinguished by mutations that occurred far back. Unless F_{st} is unusually high, the distribution of allele frequencies will be close to multivariate normal, and so we cannot go beyond pairwise relationships, which essentially depend only on $\mathcal{N}$.

Assuming normality, we can account for spatial dependencies by fitting a covariance matrix; moreover, if we sample over patches small enough that identity falls with $\log(1/r)$, this matrix has a unique form that depends on the known locations of the sampled genes, plus a single 'local scale', κ. This is straightforward, and more transparent than a simulation-based approach; it extends to allow for fluctuations in selected clines and across barriers to gene flow (Barton and Gale, 1993; Barton, 2008); it is implemented in the ANALYSE package (Barton, N.H. and S.J.E. Baird 1996. Software for the analysis of geographic variation and hybrid zones. University of Edinburgh, UK. Available via http://helios.bto.ed.ac.uk/evolgen/). However, this approach has hardly been applied, even theoretically (though see Barton and Wilson (1995); Tufto et al. (1996)). A thorough comparison of different statistical methods for estimating $\mathcal{N}$ is needed.

2.3. Inference based on shared blocks of genome

2.3.1. Philosophy and results

By confining our attention to the genetic structure generated over intermediate time scales, we are led to a rather 'classical' analysis based on

allele frequencies, because we have lost the ability to date coalescence events through mutations. However, if, instead of a few discrete loci, we sample sufficiently long stretches of recombining genome, then we *will* see recombination events on the genealogies generated over these time periods. Recombination can then be used in place of mutation to set a time scale. Two genomes that share an ancestor t generations in the past will share a portion 2^{-t} of their genomes, in blocks of map length $\sim 1/t$. Thus, sharing exceptionally long blocks indicates recent common ancestry and the block size gives an approximate date for that ancestry. This idea is exploited by Ralph and Coop (2012) to identify recent shared ancestry in a sample of 2,257 Europeans (the POPRES dataset). If two genomes do share a recent ancestor, then they are likely to share multiple blocks, so that ancestry at different unlinked loci is not independent (Wakeley, 2008). Of course, if the sampled genomes are close relatives, then we can estimate the pedigree that connects them by using the fraction of alleles shared and the lengths of shared blocks of genome. However, as before, we focus on scales of tens to hundreds of generations, intermediate between reconstruction of pedigree relationships, and the deep history of the population.

There is a subtlety that must be taken into account if we are to exploit recombination in this way. We are trying to extract information about the genetic structure generated in the previous t^* generations. If a recombination event occurs, then at that moment in time, the two resulting ancestral lineages are adjacent to one another. Often they will coalesce again before time t^* and, since we are assuming that there will be no mutations over that period, the recombination event will not be visible in our data. However, with some probability, they do not coalesce by time t^*. If this happens, it is typically because they have escaped far from one another. As a result they only coalesce in the distant past and we *do* expect to see mutations on their ancestral lineages before that time. It is these recombination events that we can expect to detect and we shall call them *effective* recombination events. Because an 'ineffective' recombination can change the genealogy in a way that we cannot detect in data, the resulting distribution of detectable block lengths is very complex. Here we consider two scenarios in which progress can be made: in both cases the length of shared blocks will be determined by an exponential distribution (or geometric if we consider discrete loci).

In §5.1 we shall assume that neighbourhood size is large and that we are sampling lineages from sufficiently close to one another that the time to coalesce is on the same order as the time to coalescence of two adjacent lineages. Because neighbourhood size is large, almost all recombination is effective and we shall make the approximation that *all* recombination events

can be detected. On the other hand, by sampling sufficiently close together, we ensure that there is sufficient correlation between loci that there will be some signal in the data. When all recombination is effective, we can use the Wright-Malécot formula (1) to find the full distribution of block lengths. We show that if we fix a focal locus and move along the genome in a given direction from there, the length B (measured in Morgans) of the portion of genome shared by two individuals satisfies

$$\mathbb{P}[B \geq b] \approx \frac{-\log r - \log(\sqrt{1-e^{-2b}}/\sigma)}{\mathcal{N} - \log(\sqrt{1-e^{-2b}})},$$

where r is the distance at which the two individuals were sampled and $\sigma^2, \mathcal{N}$ are the parameters appearing in the Wright-Malécot formula. The latter can thus be inferred from the tail distribution of the empirical CDF of the size of a 'half-block' of shared sequence.

In §5.2 we show by simulation that this approximation breaks down for small neighbourhood size. We then outline an approach to calculating the effective recombination rate, at least if we sample individuals from sufficiently far apart. In this case, we show that with high probability, an 'ineffective' recombination does not change the genealogy and so block lengths follow a geometric distribution (or exponential for a continuous linear genome) with parameter approximately given by

$$\gamma(r) = \frac{\rho_{\text{eff}}(r)}{\rho_{\text{eff}}(r) + r^2/(2\sigma^2)} \left(1 - \frac{K_0(\sqrt{2})}{\mathcal{N} + \log(\delta/(\kappa\sqrt{2}))}\right),$$

where r is the distance at which the two genomes are sampled, K_0, κ and $\mathcal{N}$ are as in (1), and $\rho_{\text{eff}}(r)$ is the *effective recombination rate* for two lineages sampled at distance r. The expression of $\rho_{\text{eff}}(r)$ is derived in §5.2 for the Λ-Fleming-Viot model, but similar expression can be obtained for other particular models that meet the different assumptions made in this work. By concentrating on the occurrence of long shared blocks, we can then learn about recent ancestry.

In both scenarios, crucially, we do not investigate the genetic structure before time t^*, but simply assume that there is enough variability in the population at time t^* that we have a reasonable chance of detecting effective recombination events. Thus, just as with our inference based on allele frequencies, these methods will be robust to the deep history of the population.

2.3.2. Inference in practice

??

3. The spatial Λ-Fleming-Viot process: a model for evolution in a spatial continuum

Here we present the model in continuous space that we shall use and compare to the stepping stone model. In fact, in the absence of large scale extinction-recolonisation events, our model can be thought of as a continuum analogue of the stepping stone model on $\mathbb{Z}^2$. Its novelty is that reproduction is not based on individuals in the population, but instead on a random sequence of events which prescribe the spatial region in which reproduction takes place. It is this that overcomes the difficulties with the approaches of Wright (1943b) and Malécot (1948) identified by Felsenstein (1975).

In the simple version of the spatial Λ-Fleming model we describe here, we assume that reproduction events 'fall' on $\mathbb{R}^2$ at rate λ per unit area. That is, the time to wait until the next event whose centre falls within a given region A is exponentially distributed with parameter $\lambda \mathrm{Vol}(A)$. The centre of the event is then chosen uniformly at random over A. The reproduction events share the same characteristics: all of them have radius $R > 0$ and during each we

- sample ν parental types uniformly at random within the ball of radius R around its centre,
- kill a fraction $u \in (0, 1]$ of the local population and replace it by offspring of the chosen parents in equal proportions.

If we focus on a single locus, this completely describes the evolution of the population. When we consider $L \geq 2$ loci and want to take recombination into account, we have to specify how the offspring inherit their allelic types from their parents. In this case, we assume that $\nu \geq 2$ and that each offspring chooses two potential parents at random from which it inherits its allele at each locus in such a way that the probability that a recombination occurs between two adjacent loci is equal to $\rho \in (0, 1]$, independently of the other pairs of loci. At every time $t \geq 0$, the population is described by a measure m_t defined by the property that for every $E \subset \mathbb{R}^2$ and every subset $\tilde{K}$ of the (compact) set K of possible alleles,

$$\int_{E \times \tilde{K}} m_t(x, dk)dx = \text{ mass of individuals in } E \text{ with an allele in } \tilde{K}.$$

The total mass of individuals is kept locally constant by the fact that every portion of population killed during an event is replaced by the same amount of new individuals. Some preliminary analyses of this model were carried

out in a series of recent papers (Etheridge, 2008; Barton et al., 2010a,b; Etheridge and Véber, 2012; Berestycki et al., 2012), and a survey can be found in Barton et al. (2012).

As we show in §6.1 of (Barton et al., 2010a), a single lineage follows a continuous-space symmetric random walk with variance parameter

$$\sigma^2 = \frac{\lambda u \pi R^4}{2}. \tag{3}$$

Furthermore, two lineages can coalesce only when they are at a distance less than $2R$, since this condition enables them to be overlapped by the same event and thus to have the same parent. In Appendix A, we characterize the probability of identity in state of two individuals in terms of an integral equation. This equation can be solved numerically, but as illustrated in Fig. 1, the Wright-Malécot formula provides an excellent approximation if we define the neighbourhood size $\mathcal{N}$ by

$$\mathcal{N} := \frac{\nu}{u}. \tag{4}$$

(See Appendix A for the derivation of (4).)

We have presented the spatial Λ-Fleming-Viot process in the special case in which the dynamics of the population are determined entirely by 'small-scale' (indeed fixed radius) reproduction events. In fact, part of the original motivation for the model was to provide a framework in which one can readily incorporate the large-scale extinction-recolonisation events which dominate the demographic histories of many species. For a slight variant of this model (in which instead of replacing a fixed proportion of individuals in the disc $B(x, R)$, one replaces individuals according to a Gaussian density centred on x), Barton et al. (2010b) investigate the effect of such large-scale events on the probability of identity. Here we perform a similar investigation for the 'disc model' in which we allow just two different values of R. Small reproduction events happen frequently, large extinction-recolonisation events happen rarely. The result is in Fig. 2. We see that although the probability of identity is changed over large scales, over intermediate scales - not so small that the Wright-Malécot approximation breaks down, but not so big that the large scale events start to matter - the rate of decay of identity is almost the same with and without the large scale events. In other words, by sampling over such intermediate scales we should be able to infer the parameters that govern the Wright-Malécot formula for a population driven entirely by small-scale reproduction events, namely dispersal rate and neighbourhood size. This adds credence to the approach we adopt in this paper.

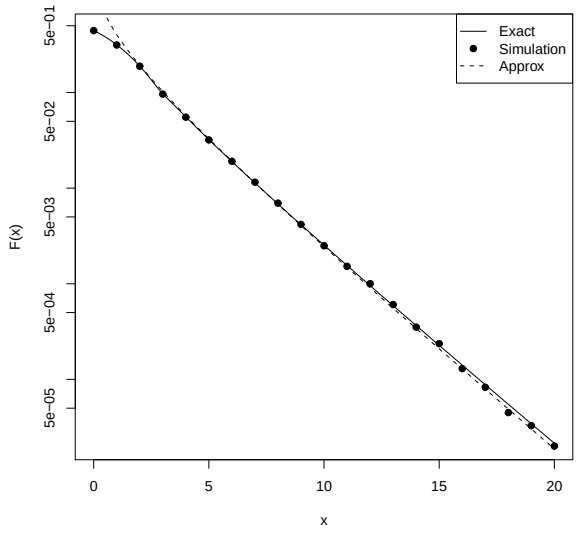


Figure 1: Probability of identity in state plotted against distance for the spatial Λ-Fleming-Viot model with parameters $\nu = 1$, $R = 1.5$, $\lambda = 1$, $u = 0.5$ with a mutation rate $\mu = 10^{-4}$ on a torus of diameter 64. The numerical solution of $\phi_\mu(x)$, simulations and the Wright-Malécot solution (with $\kappa \approx 1.34$), are shown. Simulation results report the mean identity over 10^5 replicates.

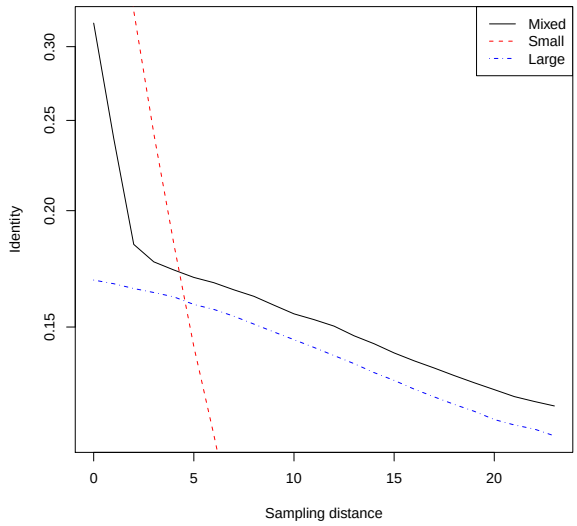


Figure 2: The probability of identity under the spatial Λ-Fleming-Viot process when we allow a mixture of frequent small-scale reproduction events and rare large scale events. Here, small events with $R = 1$ fall at rate 1, with $\nu = 1$ and $u = 0.5$. Large events with $R = 10$ fall at rate 0.1, with $\nu = 1$ and $u = 0.05$.

4. Inference based on allele frequencies

In this section, we present an approach to inferring neighbourhood size from allele frequencies in the population through an appropriate F-statistic. As we already mentioned in Section 2, for simplicity we shall present our results and some simulations in the context of the stepping stone model. However, they will remain valid whenever the separation of two ancestral lineages can be described by a 2-dimensional symmetric random walk (or its continuous analogue, Brownian motion) with variance parameter $2\sigma^2$, until they merge through a local coalescence mechanism that can be summarised by a single parameter $\mathcal{N}$ (the 'local' coalescence rate).

In the following section, we use the duality between allele frequencies (at a single locus) and genealogies to express the correlations between local genetic diversities in terms of the probability of coalescence of two ancestral lineages by time t^*. We then define a statistic $\mathcal{F}$ and in §4.2 we use this as the basis for a maximum likelihood approach to estimating $\mathcal{N}$. A separate estimate for σ^2 cannot be found.

4.1. Identity in state and measures of F_{ST}

Suppose we are able to sample perfectly from the local population diversity at a given point (or rather in a small area around a given point); in an analysis of real data, sampling variance can readily be incorporated in the inference method we derive in §4.2. We assume that we observe exactly a distinct alleles in our sample and we write $p_i(x)$ for the frequency of the ith allele in deme x. In keeping with the classical F-statistics, we shall compare correlations in allele frequencies between individuals sampled at a specific separation to those observed over a larger 'patch'. Recall that this patch, which we denote by A, is assumed to be such that its diameter satisfies $\mathrm{Diam}(A) \ll \sigma\sqrt{t^*}$. Recall also that mutation is assumed to be slow enough that no mutations have appeared recently and the genetic diversity in A is due to mutations in the remote past (i.e., more than t^* units of time ago) only.

Let us write $\mathbb{P}_{x,y}$ for the distribution of the lineages ancestral to two individuals sampled from locations $x, y \in A$ and T for the random time at which they coalesce. In accordance with our assumption of local equilibrium in A, we assume that if the two lineages have *not* coalesced by time t^*, then the chance that they are of different types is independent of their initial separation and is given by $H(t^*)$, the heterozygosity at time t^*. If $|x-y| = r$,

we can thus write

$$\begin{aligned} 1 - H_r &:= \mathbb{E}\left[\sum_{i=1}^{a} p_0(x,i)p_0(y,i)\right] \\ &= \mathbb{P}_{x,y}[T \le t^*] + \mathbb{P}_{x,y}[T > t^*](1 - H(t^*)), \end{aligned}$$

from which, in an obvious notation,

$$H_r = \mathbb{P}_r[T > t^*]H(t^*). \tag{5}$$

We shall also write H_A for the heterozygosity in A at the present time, that is

$$H_A = 1 - \frac{1}{|A|^2}\int_A \int_A \sum_{i=1}^{a} p(x,i)p(y,i)dxdy.$$

Writing $\mathbb{P}_A$ for the distribution of the two lineages ancestral to two individuals sampled independently and uniformly at random from A, we also have $H_A = \mathbb{P}_A[T > t^*]H(t^*)$.

Our analogue of Wright's F_{ST} is defined by

$$F(r) = \frac{H_A - H_r}{H_A} = \frac{\mathbb{P}_A[T > t^*] - \mathbb{P}_r[T > t^*]}{\mathbb{P}_A[T > t^*]}. \tag{6}$$

Notice, in particular, that this is independent of $H(t^*)$.

Remark 1. *Our approach differs from the standard one in which one does not fix t^*, but instead works with an infinitely many alleles mutation model with mutation rate μ and compares heterozygosity at different separations. This corresponds to replacing t^* by an exponentially distributed random variable with parameter 2μ, and following Slatkin (1991), the F-statistic becomes*

$$\frac{\left(1 - \mathbb{E}_A[e^{-2\mu T}]\right) - \left(1 - \mathbb{E}_r[e^{-2\mu T}]\right)}{1 - \mathbb{E}_A[e^{-2\mu T}]} \approx \frac{\mathbb{E}_A[T] - \mathbb{E}_r[T]}{\mathbb{E}_A[T]},$$

which (analogous to the independence of $H(t^*)$ in (6) above) is independent of μ. However, we are working on an infinite range and so $\mathbb{E}_r[T] = \infty$. One can try to circumvent this by working instead on a large (but finite) range. However, the terms in both the numerator and denominator will grow very rapidly.

If neighbourhood size is large, since we are sampling from a region A with $\mathrm{Diam}(A) \ll \sigma\sqrt{t^*}$, the distribution of $\mathcal{F}(x,y)$ takes a particularly simple form. Indeed, we first have that

$$1 - F(0) = \frac{\mathbb{P}_0[T > t^*]}{\mathbb{P}_A[T > t^*]}.$$

Using estimates on continuous-space random walks to compare $\mathbb{P}_0[T > t^*]$ to $\mathbb{P}_r[T > t^*]$, and observing next that $\mathbb{P}_A[T > t^*]$ is an average over all the pairs of individuals sampled in A of quantities of the form $\mathbb{P}_r[T > t^*]$, as in the Wright-Malécot approximation we obtain that

$$F(r) \approx \frac{1 - F(0)}{\mathcal{N}} \log\left(\frac{c}{r}\right), \tag{7}$$

where c is determined by the geometric mean of the separation of individuals sampled from A since (see Appendix A for a complete argument). This will be our basis for inference.

To understand the form of the corresponding statistic, let us write $\bar{p}_i$ for the expected value of the frequency of allele i in the region A. Note that $\bar{p}_i$ is also the probability that an individual sampled at random within A is of type i. Now consider the quantity $\mathcal{F}(x, y)$ introduced in (2), namely

$$\mathcal{F}(x, y) = \frac{1}{a - 1} \sum_{i=1}^{a} \frac{(p_i(x) - \bar{p}_i)\,(p_i(y) - \bar{p}_i)}{\bar{p}_i}$$

and write $r = |x - y|$. Expanding the brackets and using Relation (5) together with the fact that the allele frequencies over the intermediate region A are at 'quasi-equilibrium' since time t^* before the present, we obtain that

$$\mathcal{F}(x, y) = 1 - \frac{\mathbb{P}_r[T > t^*]}{\mathbb{P}_A[T > t^*]} = F(r).$$

Thus $\mathcal{F}(x, y)$ provides a statistic on which to perform maximum likelihood. In practice we do not know $\bar{p}_i$ and so we must also estimate that from the data. The distribution of $\mathcal{F}(x, y)$ can be obtained numerically from the generating function approach used to derive the Wright-Malécot formula, see Appendix A.

4.2. A maximum likelihood approach to inference

In general the method of the last section leads us to

$$F(r) \approx \frac{\log(\bar{r}/r)}{\mathcal{N} + \log(\bar{r}/\kappa)}, \tag{8}$$

where the 'local scale' κ is chosen so that

$$F(0) \approx \frac{\log(\bar{r}/\kappa)}{\mathcal{N} + \log(\bar{r}/\kappa)},$$

and $\bar{r}$ is the geometric mean of the separation of individuals sampled uniformly at random from A except that we replace all separations of less than κ by κ.

This immediately suggests that we can obtain an estimate for $\mathcal{N}$ and κ through regression of $F(r)$ on $\log r$. First if we sample n distinct locations from A, then the geometric mean sampling distance that we need to calculate F corresponds to $n(n-1)$ ordered non-zero pairs and n on-diagonal terms, each of which we replace by κ. Thus

$$n^2 \log \bar{r} = n(n-1) \log(\bar{r}^*) + n \log \kappa,$$

where $\bar{r}^*$ is the geometric mean based on the $n(n-1)$ non-zero pairs. Substituting in (8),

$$\log\left(\frac{\bar{r}}{\kappa}\right) = \left(1 - \frac{1}{n}\right) \log\left(\frac{\bar{r}^*}{\kappa}\right) = \frac{F(0)}{1 - F(0)} \mathcal{N}.$$

We also have that the regression of $F(r)$ on $\log r$ has slope

$$m = \frac{1}{\mathcal{N} + \log\left(\bar{r}/\kappa\right)} = \frac{1 - F(0)}{\mathcal{N}}.$$

This allows us to estimate

$$\mathcal{N} \sim \frac{1 - F(0)}{m}, \qquad \kappa \sim \bar{r}^* \exp\left(-\frac{F(0)}{m} \frac{n}{n-1}\right).$$

Since the relationship is only logarithmic, there may be little power in a method based on regression. We expect to obtain a better estimate by using maximum likelihood based on the $F(r)$ to estimate the parameter $\mathcal{N}$. We make the approximation that fluctuations in allele frequencies are small, so that we can approximate them by a multivariate Gaussian distribution. Suppose that we sample from a given set of n locations. We write F_{ST} for the resulting matrix of standardised covariances. We write $\tilde{F}$ for the observed covariances and F^* for the expected covariances, then the log-likelihood function takes the form

$$\log(L) = -\frac{1}{2} \left(\log(\det(F^*)) + \sum_{j,k} \tilde{F}_{j,k} {F^*_{j,k}}^{-1} \right).$$

We approximate F_{ST} as in (8), replacing r by κ on the diagonal. Since we have estimated the mean allele frequency from the data, the covariance of deviations from the mean is

$$F^*_{j,k} = F_{ST\,j,k} - F_{ST\,j,*} - F_{ST\,*,k} + F_{ST\,**},$$

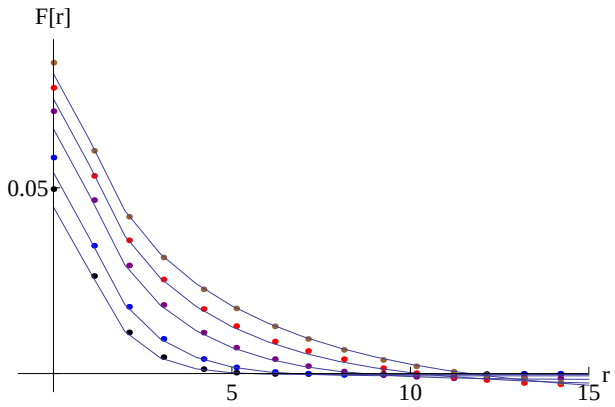


Figure 3: Standardised covariance of allele frequencies, F, against distance under a stepping stone model. A 40×40 toroidal grid of demes, each with $2N = 20$ haploid individuals, was simulated for 200 generations; there was migration between each of the four nearest neighbours at rate $m/2 = 0.125$. There were three alleles, with initial frequencies $\{0.1, 0.4, 0.5\}$, and no mutation. Dots show the average of 10 independent replicates at times 10, 20, 50, 100, 200 (bottom to top); lines join the theoretical prediction for this discrete-space model, allowing for estimation of the population mean from the realised values. Agreement is close, apart from a small underestimation of $F(0)$. This simulation represents sampling of ten independent loci from a population of 40×40 demes.

where $*$ represents an average over an index. This matrix is singular, having one zero eigenvalue, and so $\det(F^*)$ is calculated as the product of the $n-1$ positive eigenvalues. For a given set of locations, F^* depends only on κ and the maximum likelihood estimator for $\mathcal{N}$ can be found explicitly:

$$\hat{\mathcal{N}} = \frac{n-1}{\sum_{j,k} \tilde{F}_{j,k} F^*_{j,k}} - \log(\bar{r}/\kappa). \tag{9}$$

4.3. Results based on allele frequencies

Figure 3 shows the standardised covariance of allele frequencies as a function of separation for a stepping stone model on $\mathbb{Z}^2$ with nearest neighbour migration. After 200 generations, for deme spacings of less than about 7, it is very close to the logarithmic approximation (8), as demonstrated in Fig. 4.

Figure 5 shows the result of using regression of $F(r)$ on $\log r$ to estimate $\mathcal{N}$ for simulations of the stepping stone model. Finally, Fig. 6 shows the likelihood surface for the parameters $\mathcal{N}$ and κ obtained by implementing the scheme of §4.2 for the population simulated in Fig. 3.

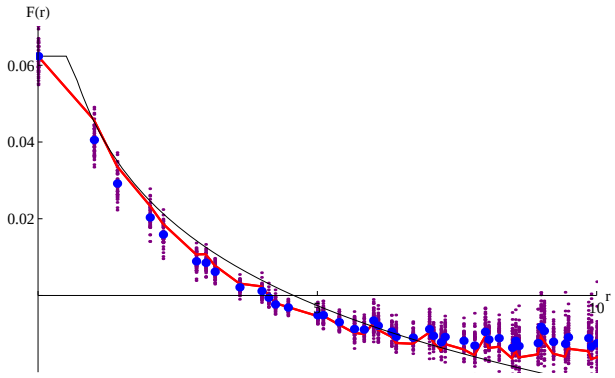


Figure 4: The covariance against distance for a 10×10 subsample of demes, taken from ten realisations of the 40×40 population of Fig. 3; this can be thought of as sampling from ten independent loci, since the allele frequencies at each locus evolve independently. Each dot is the mean over ten replicates; different dots are different locations of the 10×10 grid. The red line is the theoretical prediction, and the black line, the simple prediction assuming $F(r) = (1 - F(0))/\mathcal{N} \log(c/r)$, $F(0)$ being estimated from the sample. This is independent of the constant c. It can be rewritten as $\log(\bar{r}/r)/(\mathcal{N} + \log(\bar{r}/\kappa))$ where $\kappa = 0.57$ is estimated from $\bar{r} \exp(-\mathcal{N} F(0)/(1 - F(0))$. The theoretical prediction fits well: it is jagged because $F(\{1, 1\})$ is slightly lower than $F(\{0, 2\})$ even though the points are nominally closer. The black curve is the naive logarithmic prediction, fitted to the observed $F(0) = 0.062$. It declines slightly less steeply than expected: $F(0) = 0.106$ estimated from the regression of F on $\log r$ which should have slope $(1 - F(0))/\mathcal{N}$.

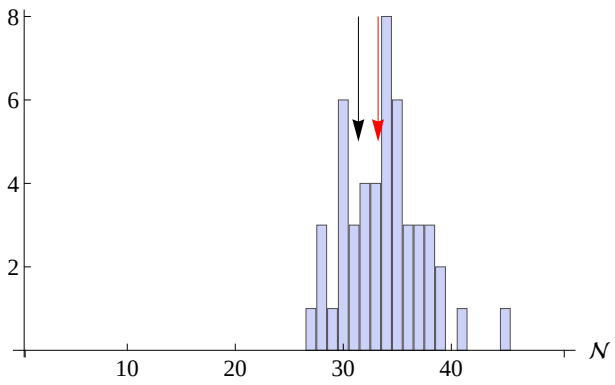


Figure 5: Estimates of $\mathcal{N}$, each based on ten independent realisations (representing ten independent loci). These estimates are derived from the slope of the regression of F_{ST} on $\log r$. The distribution shows 49 replicate 10×10 patches, taken from the same 40×40 population, but starting at $\{\{1,1\},\{6,1\},\ldots,\{6,1\},\ldots,\{31,31\}\}$; this represents the variation in estimates that would be obtained by sampling from different places within a stationary distribution; since the replicates are correlated, this underestimates the variation between independent realisations. The black arrow shows the true $\mathcal{N}=4\pi\times10\times0.25=31.4$ and the red arrow the mean across replicates which is a slight overestimate, 33.2; this bias is smaller than the scatter between different patches from the same population.

5. Estimation based on patterns of recombination

Our approach based on allele frequencies only allowed us to estimate neighbourhood size and not the other key parameter, dispersal rate. We now turn to our second approach, based on a small sample (two in what follows) of long genomes. The idea is that, although we have confined our attention to sufficiently small scales that genealogies cannot be reconstructed from patterns of mutation at a single locus, nonetheless, if we are considering sufficiently long blocks of genome, we can use recombination to determine the timescale of coalescence. We shall thus consider pairs of long genomes sampled from our population and investigate the lengths of shared blocks of sequence.

Suppose that a pair of genes coalesce t generations in the past (so that the genealogy relating them has length $2t$). Moving out from the focal locus in either direction, the distance along the genome (measured in Morgans, where $1M$ is the block length over which we see one recombination per time unit on average) before we encounter a recombination event is exponentially distributed with density $2te^{-2tb}$. The total length of the block around the focal locus which is bordered by these two recombination events is therefore

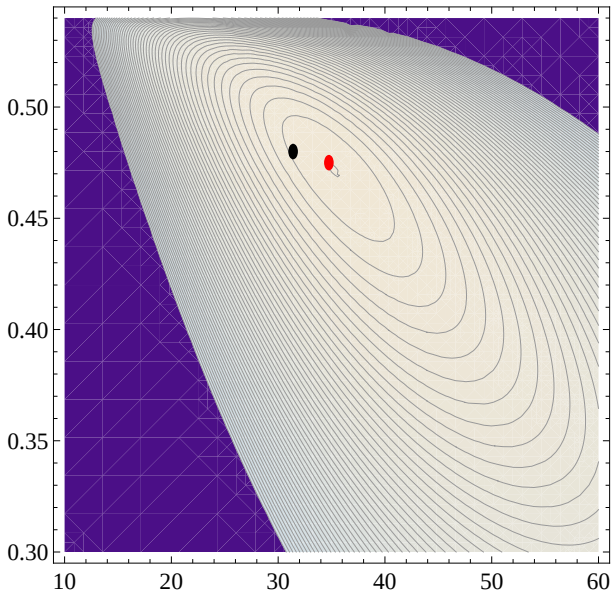


Figure 6: Likelihood surface based on ten 'loci' sampled from a 10×10 patch within the 40×40 population of Fig. 3; log likelihood is plotted against $\mathcal{N}$ (x-axis) and κ (y-axis). The MLE is $\mathcal{N} = 34.75$, $\kappa = 0.475$; the true $\mathcal{N} = 31.4$, $\kappa = 0.48$, with log-likelihood lower by 1.5. Contours are spaced at 2 units of log-likelihood, so that he inner circle indicates the support limits for each parameter.

Gamma-distributed with density $4t^2be^{-2bt}$. This represents a size-biased pick from the distribution of block lengths, because we effectively condition on the focal locus being within the block. The distribution of the length of a randomly chosen block (as opposed to the block around a randomly chosen point) is exponential with density $2te^{-2tb}$. Since each can be derived in essentially the same way as what follows, to keep things simple we are going to work with the joint distribution of B and T where B is the length of a 'half-block', that is the distance we move in just one direction before encountering the next recombination, and T is the coalescence time. The distribution of the coalescence time of lineages ancestral to two individuals sampled at separation x, $\psi_t(x)$, was derived in Appendix A and the joint density of B and T becomes $2te^{-2bt}\psi_t(x)$.

As outlined in §2.3, analyses of shared blocks are complicated by the fact that we may not be able to detect all recombination events in our data. Such recombinations were called 'ineffective'. In order to make progress, we shall assume in §5.1 that neighbourhood size is large. Under this assumption, local coalescence is difficult to achieve and we expect two lineages that have just been created through a recombination event to coalesce again after a very large amount of time, even though they sit at nearby locations initially. We should thus be able to detect essentially all recombination events that affect either of the genomes in our sample before time t^* and the Wright-Malécot formula can be used to write down the distribution of the lengths of shared blocks. In §5.2 we show (by simulation) that this approximation breaks down for small neighbourhood size and then outline a preliminary analysis in that setting. However, the results will only apply if we sample the genomes from sufficiently far apart that if the ancestral lineages at a locus coalesce before time t^*, then they do so at a time close to t^*, and if a recombinant lineage coalesces back into the genealogy before time t^*, then with overwhelming probability it coalesces with the lineage from which it split off and therefore the genealogy of the recombinant block is the same as that of the focal locus. The difficulty is that one would like to base a method of inference on long shared blocks, but sampling at these separations, such blocks are very rare.

5.1. Large neighbourhood size

In this section we shall suppose that neighbourhood size is large so that essentially all recombination is 'effective'. That is, if a recombination occurs before our reference time t^*, then we can ignore the probability that the resulting lineages coalesce (either with each other or with the lineage ancestral to the other individual in the sample) by time t^*. In this setting,

the probability density function corresponding to the distribution of lengths of blocks of genomes that are shared between the two individuals in our sample can be determined from the Wright-Malécot formula. If we sample two adjacent genomes, then using equation (A.5) in Appendix A, we obtain that the length of a half-block has density

$$\mathbb{E}\big[2Te^{-2bT}\big] = -\frac{\partial}{\partial b}\left(\frac{1}{1-2\mathcal{N}/\log(1-e^{-2b})}\right) = \frac{2\mathcal{N}}{(e^{2b}-1)\big(2\mathcal{N}-\log(1-e^{-2b})\big)^2}. \quad (10)$$

In Fig. 7(a), we show the resulting distribution of block sizes for two genomes sampled adjacent to one another. We also show the cumulative contribution due to coalescence events in successive generations as we trace back in time. This can be calculated from (A.2). In our numerical examples we have taken dispersal to be governed by a discretised Gaussian. We also implemented nearest neighbour dispersal and the results were indistinguishable from those displayed here. We see that very many short blocks can be attributed to coalescence events in the distant past, but such events make a negligible contribution to long blocks. A similar approach yields the distribution of block lengths shared between individuals sampled at different separations. This is illustrated in Fig. 7(b). In Fig. 8 we show the density of block sizes for different separations. Since the chance of a very recent coalescent event declines rapidly with separation, we see a deficit of large blocks between well-separated genomes.

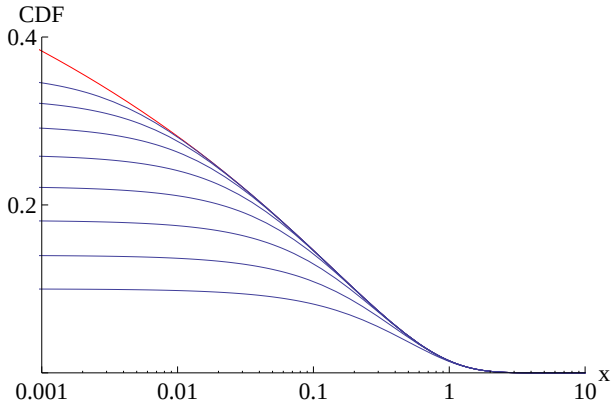
Equation (A.6) with $z = e^{-2b}$ provides an expression for the probability that two genomes share a half-block of length at least b. As we have seen, substantial blocks are determined by recent ancestry and we only see this if we sample our two genomes from relatively close to one another. In this setting we can approximate $K_0(r\sqrt{1-z}/\sigma)$ by $-\log(r\sqrt{1-z}/\sigma)$. Then the probability that a randomly chosen block has length at least b is

$$\mathbb{E}_r[e^{-2bT}] \approx \frac{-\log r - \log(\sqrt{1-e^{-2b}}/\sigma)}{\mathcal{N} - \log(\sqrt{1-e^{-2b}})}.$$

Thus plotting this probability against $\log r$ we obtain a graph which, over these local scales, is approximately a straight line with slope

$$\frac{1}{\mathcal{N} - \log(\sqrt{1-e^{-2b}})}$$

(a)



(b)

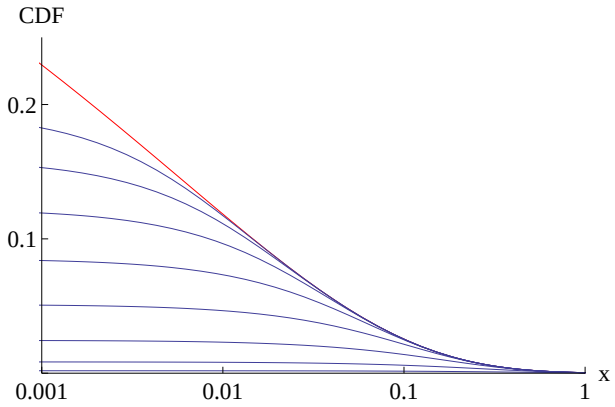


Figure 7: (a) The CDF of block size for two genes sampled at the same place and with $\mathcal{N} = 5$. The top line shows the total distribution, and the lower lines the contribution up to times $1, 2, 4, \ldots, 64, 128$ generations. Note that distant generations make a negligible contribution to the distribution of large blocks. For example, blocks greater than 0.1 in length are almost all contributed by coalescence within ~ 10 generations. However, there are very many small blocks contributed by very distant coalescence. (b) The same, but for genes separated by $r = 4\sigma$. Now, the first few generations make hardly any contribution, but later generations contribute in essentially the same way.

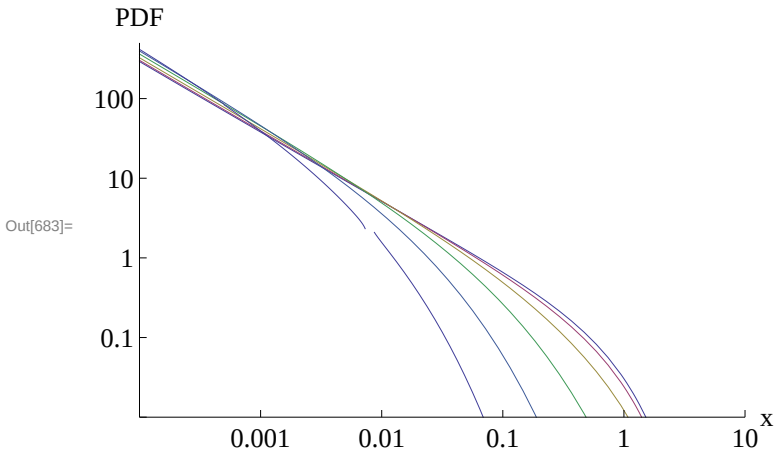


Figure 8: This shows the density of block size, for separations $r = 0, 1, 2, 4, 8, 16$ (top to bottom). The density of small blocks is independent of separation, but there is a deficit of large blocks between well-separated genes.

and intercept

$$\frac{-\log(\sqrt{1-e^{-2b}}/\sigma)}{\mathcal{N}-\log(\sqrt{1-e^{-2b}})}.$$

These can then be used to infer both $\mathcal{N}$ and σ. In Fig. 9 we plot the proportion of blocks of length at least $0.1M$ against the logarithm of the sampling distance for the parameters of Fig. 8. The result is indeed approximately a straight line, at least provided the sampling distance is not too large. To test the feasibility of performing inference in this way, we truncated the graphs in Fig. 9 and fitted the parameters. This resulted in an estimate for σ of 1.17 (true value $\sigma = 1$) and for $\mathcal{N}$ of 5.05, 10.09, 20.17 (true values 5, 10, 20 respectively). Of course, since the relationship between the cumulative distribution of block length and sampling distance is only logarithmic, this method may have little power and certainly, for real data it would be better to fit the actual formula. Nonetheless, this suggests that one can use the prevalance of long shared blocks to infer our two key parameters of local evolution.

5.2. Small neighbourhood size

The approach of §5.1 will break down for small or moderate neighbourhood size since then we expect that there will be recombinant lineages that coalesce back into the genealogy before time t^*. As a result, not only will there be recombination events which we do not detect in our data, but we

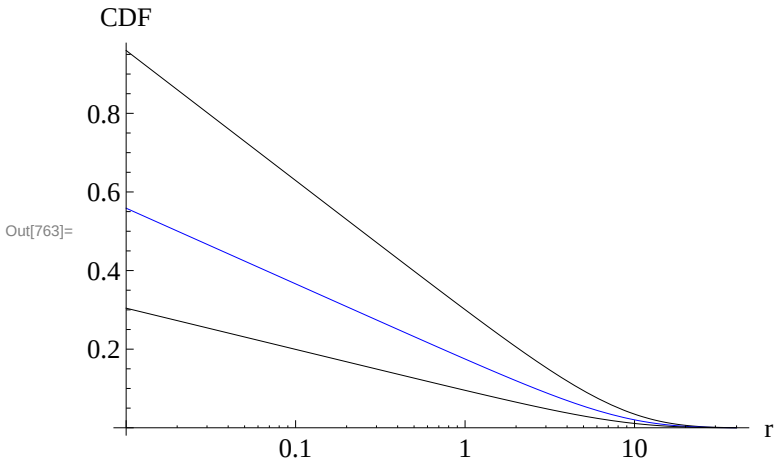


Figure 9: This shows the logarithmic regime, where the CDF is proportional to $\log r$, with slope $1/(\mathcal{N} - \log(\sqrt{1 - e^{-2x}}))$. The three lines are the probability of a half-block being longer than 0.01 for $\mathcal{N} = 5, 10, 20$ (top to bottom).

also see correlations in block lengths as we scan along the genome. In Fig. 10 we illustrate this through a simulation of the spatial Λ-Fleming-Viot process. It is this model which will form the basis of both our simulations and analysis in this section.

Since our simulations of long genomes assume discrete loci arranged along a linear genome (instead of the continuous genome of the previous section), our analysis in this section will assume discrete loci with 'links' between. The intuitively clear, but notationally challenging, incorporation of recombination into our spatial Λ-Fleming-Viot process is spelled out in detail in **??**. The mechanism is simple. From the point of view of genealogies, when a lineage experiences a reproduction event, there is a fixed probability, which we denote by ρ, that there is a crossover event between any two adjacent loci. We sample the individual with which the lineage recombines uniformly at random from the region affected by the reproduction event. Starting at a focal locus, we select one of the two recombinant lineages as parent. Scanning along the genome, whenever we encounter a crossover event, we switch to the other parent.

The philosophy is as before. We suppose that we sample two genomes at separation δ, assumed to be large compared to the radius R of a reproduction event. Whenever two recombinant lineages are created, either they coalesce very quickly, or they manage to 'escape' from one another and only coalesce in the distant past, by which time they have accumulated many mutations.

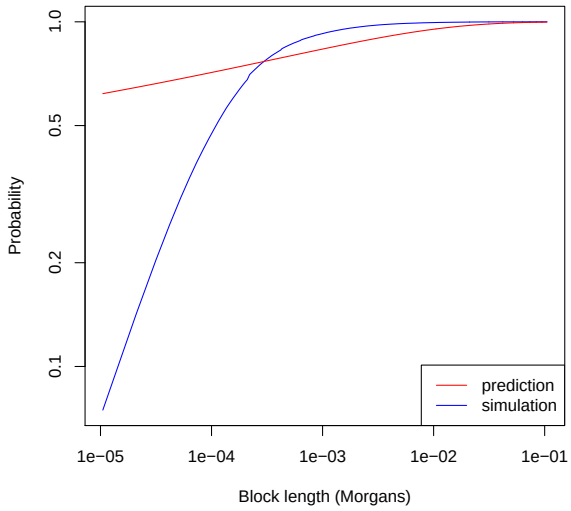


Figure 10: Breakdown of the Wright-Malécot approximation to block length distribution when neighbourhood size is small. To find the empirical distribution of block lengths we simulate the ancestry of two individuals with 5×10^4 loci sampled at distance $\delta = 10$ from each other on a torus of diameter 1000, with $R = 1$, $u = 0.75$, $\nu = 2$ and $\rho = 10^{-5}$. This model is simulated backwards in time for 10^{10} events, and we then calculate the ECDF of the length of blocks across 200 independent replicates. Loci that have not coalesced are discarded. Note that $1M$ corresponds to $\rho^{-1} = 10^5$ loci here and that a distance δ in the simulations of the spatial Λ-Fleming-Viot process corresponds to $r = (\delta\sqrt{2})/R$.

In order for two loci to be identical in state, they must have coalesced in the recent past, and we are interested in the length of blocks of consecutive loci that are identical in state. Our first task is to define what we mean by quick, or 'early' coalescence. Since the spatial Λ-Fleming-Viot process has overlapping generations, it is natural to replace the fixed reference time t^* by an exponentially distributed time. If we are sampling two genomes at separation δ, then the time that the lineages ancestral to any given locus require to come together is $\mathcal{O}(\delta^2/(2\sigma^2))$. We shall define a coalescence to be *early* if it takes place before an exponentially distributed time, T_ζ, with mean $\zeta \equiv \delta^2/(2\sigma^2)$. We are interested in the lengths of blocks of genome which are identical by descent, which, under the assumption that we can detect all 'ineffective recombination events' in our data, in this terminology correspond to blocks of consecutive loci, all of which experience an early coalescence. The following result is proved in Appendix C.

Theorem 1. *Suppose that we sample two individuals at separation δ and let X be the length of a block of consecutive loci at which the two individuals are identical in state. Then X follows (approximately) a geometric distribution with parameter $\gamma(\delta)$ given by*

$$\gamma(\delta) = \frac{\rho_{\text{eff}}(\delta)}{\rho_{\text{eff}}(\delta) + \zeta(\delta)} \left(1 - \frac{K_0(\sqrt{2})}{\mathcal{N} + \log(\delta/(\kappa\sqrt{2}))} \right), \tag{11}$$

where K_0 and κ are as in (1), $\mathcal{N} = \nu/u$ is again the neighbourhood size, and

$$\rho_{\text{eff}}(\delta) = (\Lambda u \pi R^2 \rho)\, \alpha(\delta),$$

the quantity $\alpha(\delta)$ being the 'escape' probability of two recombinant lineages for which we find a characterisation in Appendix C.2.

Although the parameters R and u appear in the formulation of this result, at least for sufficiently large δ, as we argue below, the quantity $\gamma(\delta)$ depends only σ and $\mathcal{N}$. Obviously, for this result to hold we need δ to be large enough that $K_0(\sqrt{2})/\{\mathcal{N} + \log(\delta/(\kappa\sqrt{2}))\}$ is less than one. The quantity $\rho_{\text{eff}}(\delta)$ is proportional to, but not equal to, the local recombination rate $\lambda u \pi R^2 \rho$. It is the *effective recombination rate*, described in §2.3.

The proof of Theorem 1 is given in Appendix C, but let us give here an outline of it. We are sampling individuals from so far apart that if the ancestral lineages at a locus coalesce early, they do so close to the time T_ζ. If a recombinant lineage is created which coalesces back into the focal genealogy before time T_ζ, then with high probability it coalesces with the

lineage from which it split off. Thus for every locus at which we see an early coalescence, the chance of an effective recombination event between that locus and the adjacent locus (thus bringing an end to the block) is approximately the same. We denote it by $\rho_{\text{eff}}(\delta)$. For a given pair of adjacent loci, the ratio

$$\frac{\rho_{\text{eff}}(\delta)}{\rho_{\text{eff}}(\delta) + \zeta(\delta)}$$

simply gives the probability that an effective recombination occurs before the timescale of early coalescence. The second term in the definition of $\gamma(\delta)$ is the probability that the ancestral lineages at locus $j+1$ do not coalesce early, bringing the block X to an end.

In Fig. 11 we show that the geometric block length distribution predicted by Theorem 1 is reasonably accurate if we sample from far enough apart. Somewhat surprisingly, we see that the geometric distribution obtained in Theorem 1 fits the empirical distribution of the length of a block of loci having *exactly* the same coalescence time better. Although we have no explanation for this fact, one should notice that the discrepancy between *early* blocks and *equal* blocks vanishes as the sampling distance grows, and so the meaning of Theorem 1 remains clear for reasonably large sampling distances.

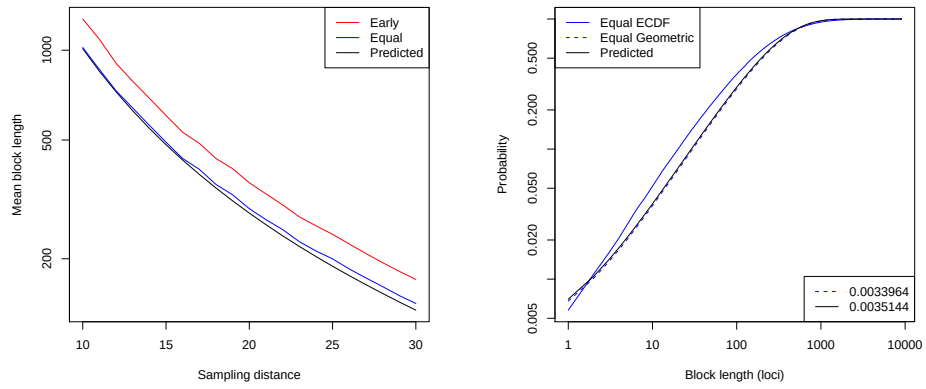
5.3. Inference for small neighbourhood sizes

Because of the similarity between Equation (A.8) defining the Laplace transform of the coalescence time T and Equation (C.2) defining $\alpha(\delta)$, we expect the function ρ_{eff} to depend only on σ, $\mathcal{N}$ and κ, at least for large δ's. Indeed, as explained in Appendix C, $\alpha(\delta)$ is the probability that two lineages starting at distance $\mathcal{O}(R)$ separate at distance δ before they coalesce again. If δ is large compared to R, this probability is essentially the same as the probability that two lineages starting at distance κ do not coalesce before the time $\mathcal{O}(\delta^2/\sigma^2)$ that they need to travel a distance δ. Hence, using the Wright-Malécot formula (1) with $2\tilde{\mu} = (\delta^2/\sigma^2)^{-1}$ (and so $\ell_{\tilde{\mu}} = \delta/\sqrt{2}$), we arrive at

$$\alpha(\delta) \approx 1 - \frac{\log(\ell_{\tilde{\mu}}/\kappa)}{\mathcal{N} + \log(\ell_{\tilde{\mu}}/\kappa)},$$

which in fact depends on the precise evolution mechanism only through the two parameters $\mathcal{N}$ and κ.

Similarly, since ρ is the *'per hit'* recombination probability, the product $\lambda u \pi R^2 \rho$ is the total rate at which two neighbouring loci recombine. When dealing with real data, this compound term should be replaced by



(a) (b)

Figure 11: Block lengths due to early coalescence; (a) plots mean block length against sampling distance δ and (b) shows the distribution of block lengths for $\delta = 20$. Simulations trace the ancestry of two individuals sampled at distance δ until time T_ζ in the past, where T_ζ is an exponentially distributed value with rate $2\sigma^2/\delta^2$ chosen independently for each replicate. Model parameters are otherwise identical to the simulations of Figure 10. The length of blocks of loci are then calculated in two different ways: we have *early* and *equal* blocks. An early block is defined as a set of contiguous loci that have coalesced by time T_ζ. An equal block is a set of contiguous loci that have coalesced by T_ζ and have equal coalescence times. In (a) we have $\sim 10^6$ early blocks and $\sim 1.2 \times 10^6$ equal blocks from 7911 independent simulations (many simulations have no early coalescences); there is an excellent correspondence between the predicted block length $1/\gamma(\delta)$ and the length of equal blocks. Panel (b) shows the CDF of the length of equal blocks for $\delta = 20$, and compares the ECDF of block lengths from simulations with a geometric distribution with parameter $\gamma(20)$. Also shown is the geometric distribution with parameter estimated from simulation data; this agrees very closely with the predicted value.

an estimate of the recombination rate. Overall, the function $\gamma(\delta)$ is really a function of σ, κ and $\mathcal{N}$ alone, and so can, in principle, be used as a basis for the inference of these parameters.

However, several problems must be overcome if we are to devise a robust inference scheme based on the results of §5.2 (or even §5.1). First, because we want to sample only a few individuals, it is natural to try to use the empirical distribution of lengths of blocks which are identical by descent between pairs to estimate the distribution of the parameter $\gamma(\delta)$ of Theorem 1. But recall that the genealogical trees at different loci are all embedded in a single pedigree. Very long blocks of identical sequence are due to very recent coalescence, and so many portions of the genome are still carried by the same ancestor at that time. Hence, the formation of a very long sequence is rare, but when it occurs, several long sequences are produced at the same time. Due to these correlations, the empirical distribution obtained in a single run of simulations either overestimates or underestimates the probability of a very long sequence (depending on whether such a sequence was created or not), while the empirical distribution obtained by merging several runs of simulations will in general overestimate the probability of very long sequences.

Second, in theory, the dichotomy between recent and distant coalescence that underlies our approach should be sufficient to detect precisely the parts of the genome in which coalescence was early. Of course it is not entirely true that such a dichotomy holds in our model, or even in practice, since we cannot find a precise timescale t^* such that lineages that have not coalesced before t^* need more that $100t^*$ (say) to find a common ancestor. Hence, it is not yet clear which portion of the tail of the distribution of block length should be used to infer γ.

In conclusion, our results represent first steps towards the design of an alternative inference method based on recombination patterns, but a non-negligible amount of work will be required before arriving at a satisfactory scheme.

6. Discussion

Interpretation of spatial genetic data is dominated by two distinct, and indeed, incompatible approaches. Pairwise relationships, measured by F_{ST} and spatial correlation, are interpreted on the assumption that populations are at equilibrium between random drift and gene flow; patterns at different loci are then independent, and in aggregate yield estimates of neighbourhood size. In contrast, phylogeography attempts to infer the specific history of

the whole population from genealogies sampled at one or more loci. This approach requires that loci share a common signal that reflects the history of the whole population. The recent flood of genetic data has motivated other approaches, which can be seen as intermediate between these two, in that they use explicit population genetic models to reconstruct the history of population subdivision and mixing (e.g. Beerli and Felsenstein (2001); Pinho and Hey (2010); Patterson et al. (2006)). However, these model discrete demes or species, whereas here we focus on populations that are spread across two dimensions.

Our central argument is that these different approaches are appropriate over different scales. We cannot hope to infer all the details of local population history, but neither can we assume that populations have reached equilibrium over large spatial and temporal scales. However, by focusing on samples taken over modest patches (a few tens of dispersal ranges across) and on long blocks of shared genome (longer than $\sim$ 0.1cM, say), we can make robust estimates by assuming a local quasi-equilibrium.

Inferences based on fluctuations in allele frequency, and on lengths of shared sequence, both depend on the underlying distribution of times when genes sampled from some distance r apart shared common ancestry. Wright (1943b) argues that this distribution of coalescence times can be found simply by imagining the ancestors of each lineage, t generations back, as being distributed in a Gaussian with variance $\sigma^2 t$; lineages coalesce at a rate proportional to the overlap of their ancestral distributions. Thus, neighbouring genes have probability $\sim \frac{1}{\mathcal{N}t}$ of coalescing at time t, and this probability falls away with $\log(1/r)$. We show that if we sample over local patches (of area $\sim \sigma^2 T$, where the population has been diffusing steadily for time $\sim T$), then we can *only* estimate neighbourhood size, $\mathcal{N}$: information about the rate of gene flow, σ^2, is lost. The logarithmic relation with distance must break down over sufficiently small scale, $\sim \kappa$, since the identity has an upper bound, $F(0)$. However, this local scale κ may be substantially different from the long-term rate of diffusion of ancestral lineages, σ; it depends on the idiosyncrasies of local population regulation, and seems to us to be of little general interest. In contrast, neighbourhood size gives the relative rate of local drift and gene flow, and determines the rate of shifts between alternative adaptive peaks in Wright's 'shifting balance' theory (Rouhani and Barton, 1987; Coyne et al., 1997). It is important to appreciate, however, that neighbourhood size has only a weak influence on the rate of spread of favourable alleles, and on the long-term rate of drift of the population as a whole.

Recombination between linear genomes occurs, in effect, over a wide

range of scales, ranging from rates between adjacent bases that are similar to mutation, up to multiple crossovers per generation over the whole genome. This extra clock allows us to use lengths of shared blocks to estimate σ^2 as well as $\mathcal{N}$. However, we cannot simply regress squared distance, r^2, on coalescence time, as has occasionally been proposed (Neigel and Avise, 1993; Lemmon and Lemmon, 2008). Estimation is complicated by the fact that the distribution of blocks shared between any two genomes reflects their particular ancestry, and so we need to sample many genomes in order to estimate $\mathcal{N}$ and σ^2 reliably. These parameters essentially measure the fraction of close relatives, and their spatial dispersion, and so the accuracy of estimates depends on the number of related pairs that we can identify. Of course, for the past few generations, we can directly reconstruct the pedigree. However, it is not clear how far back we could do this, even given full genomes. Our aim here is to use the distribution of moderately large blocks, reflecting shared ancestry $10 - 100$ generations back, to estimate $\mathcal{N}$, σ^2; it is unclear whether we could ever estimate more than these two parameters.

Li and Durbin (2011) have recently proposed an ingenious method for finding the size of an ancestral population through time, $N(t)$. They approximate the ancestral recombination graph by a hidden Markov model, in which the coalescence time between a pair of genomes jumps whenever there is an (effective) recombination event; this is reflected by jumps in the rate of heterozygous SNPs along the genome. We expect that this method will be robust to two-dimensional population structure: that would reduce the rate of coalescence over very recent times, but not by much if $F_{st} \ll 1$. Since ancestral lineages spend most of their time wandering across the whole species' range, a model that assumes a single rate of coalescence, $1/(2N(t))$, should be accurate. This method is based on the distribution of very short blocks of sequence identity, which we discard when estimating $\mathcal{N}$, σ^2.

Our results depend on the decrease in identity with $\log r$, which is peculiar to isolation by distance in two dimensions: this leads to significant fluctuations over a wide range of scales. We have deliberately discarded information from large scales, by sampling over small patches and by discarding small blocks. Barton et al. (2010b) develop a model that can account for large-scale patterns generated by extinction and recolonisation. However, in practice it will not be possible to estimate the parameters of this model, since the unique history of the population affects all loci. Estimation of a few parameters that describe local structure is possible because loci fluctuate independently under isolation by distance, whereas the deeper history is reflected in patterns common across loci.

The theory of isolation by distance originated by Wright and Malécot gives a robust framework for understanding spatial patterns in two dimensions, but the exact relationship between the biparental pedigree, distributions of allele frequency, and shared sequence identity along the genome, remains to be explored. The Wright-Malécot theory provides a clear null model against which to detect the effects of selection at specific loci (Lewontin and Krakauer, 1973; Beaumont and Balding, 2004): signals of selection may be more reliably detected from local signals than from large-scale patterns. The ideas sketched here leave open many questions for the future.

References

Alagar, V., 1976. The distribution of the distance between random points. J. Applied Probab. 13, 558–566.

Barton, N., 2008. The effect of a barrier to gene flow on patterns of genetic variation. Genetics Research 90, 139–149.

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

Barton, N., Etheridge, A., Véber, A., 2010a. A new model for evolution in a spatial continuum. Electron. J. Probab. 15, 162–216.

Barton, N., Etheridge, A., Véber, A., 2012. Modelling evolution in a spatial continuum. J. Stat. Mech.: theory and experiment To appear.

Barton, N., Gale, K., 1993. Genetic analysis of hybrid zones., in: Harrison, R. (Ed.), Hybrid zones and the evolutionary process, Oxford University Press. pp. 13–45.

Barton, N., Kelleher, J., Etheridge, A., 2010b. A new model for extinction and recolonization in two dimensions: quantifying phylogeography. Evolution 64, 2701–2715.

Barton, N., Wilson, I., 1995. Genealogies and geography. Phil. Trans. R. Soc. London Ser. B 349, 49–59.

Beaumont, M., Balding, D., 2004. Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology 13, 969–980.

Beerli, P., Felsenstein, J., 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Nat. Acad. Sci. (USA) 98, 4563–4568.

Berestycki, N., Etheridge, A., Véber, A., 2012. Large scale behaviour of the spatial Λ-Fleming-Viot process. Ann. Instit. H. Poincaré Probab. Statist. To appear.

Coyne, J., Barton, N., Turelli, M., 1997. A critique of wright's shifting balance theory of evolution. Evolution 51, 643–671.

Durrett, R.T., 2008. Probability models for DNA sequence evolution, 2nd Edition. Springer.

Epperson, B., 2003. Geographical genetics. Princeton University Press.

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

Etheridge, A., Véber, A., 2012. The spatial Λ-Fleming-Viot process on a large torus: genealogies in the presence of recombination. Ann. Applied Probab. To appear.

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

Kimura, M., Weiss, G., 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49, 561–576.

Kong, Q.P., Sun, C., Wang, H.W., Zhao, M., Wang, W.Z., Zhong, L., Hao, X.D., Pan, H., Wang, S.Y., Cheng, Y.T., Zhu, C.L., Wu, S.F., Liu, L.N., Jin, J.Q., Yao, Y.G., Zhang, Y.P., 2011. Large-scale mtdna screening reveals a surprising matrilineal complexity in East Asia and its implications to the peopling of the region. Molecular Biology and Evolution 28, 513–522.

Lawler, G., Limic, V., 2010. Random walk: A modern introduction. Cambridge University Press.

Lemmon, A., Lemmon, E., 2008. Likelihood framework for estimating phylogeographic history on a continuous landscape. Syst. Biol. 57, 544–561.

Lewontin, R., Krakauer, J., 1973. Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. Genetics 74, 175–195.

Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496.

Malécot, G., 1948. Les mathématiques de l'hérédité. Masson et Cie, Paris.

Meirmans, P., 2012. The trouble with isolation by distance. Molecular Ecology Online First.

Neigel, J., Avise, J., 1993. Application of a random walk model to geographic distributions of animal mtDNA variation. Genetics 135, 1209–1220.

Paik, Y., Sung, K., 1969. Behaviour of lethals in *Drosophila melanogaster* populations. Jap. J. Genet. 44 Suppl.1, 180–192.

Patterson, N., Richter, D., Gnerre, S., Lander, E., Reich, D., 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature doi:10.1038/nature04789.

Pinho, C., Hey, J., 2010. Divergence with gene flow: models and data. Annual Review of Ecology, Evolution and Systematics 41, 215–230.

Ralph, P., Coop, G., 2012. The geography of recent ancestry across Europe. Preprint .

Rouhani, S., Barton, N., 1987. Speciation and the 'shifting balance' in a continuous population. Theor. Pop. Biol. 31, 465–492.

Rousset, F., 1997. Genetic differentiation and estimation of gene flow from f-statistics under isolation by distance. Genetics 145, 1219–1228.

Rousset, F., 2003. Genetic structure and selection in subdivided populations. Princeton University Press.

Rousset, F., Leblois, R., 2007. Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification. Molecular Biology and Evolution 24, 2730–2745.

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

Slatkin, M., Arter, H., 1991. Spatial autocorrelation methods in population genetics. American Naturalist 138, 499–517.

Slatkin, M., Barton, N., 1990. A comparison of three methods for estimating average levels of gene flow. Evolution 43, 1349–1368.

Sokal, R., Jacquez, G., Wooten, M., 1989. Spatial autocorrelation analysis of migration and selection. Genetics 121, 845–855.

Tufto, J., Engen, S., Hindar, K., 1996. Inferring patterns of migration from gene frequencies under equilibrium conditions. Genetics 144, 1911–1921.

Wakeley, J., 2008. Coalescent theory: an introduction. Roberts and Company, Englewood, Colorado.

Wallace, B., 1966. Distance and allelism of lethals in a tropical population of *Drosophila melanogaster*. Amer. Nat. 100, 565–578.

Whitlock, M., McCauley, D., 1998. Indirect measures of gene flow and migration: F_{ST} is not equal to $1/(4Nm + 1)$. Heredity 82, 117–125.

Wright, S., 1943a. An analysis of local variability in flower color in *Linanthus parryae*. Genetics 28, 139–156.

Wright, S., 1943b. Isolation by distance. Genetics 28, 114–138.

Appendix A. The Wright-Malécot formula

Appendix A.1. Derivation under the stepping stone model

In this section we work with a stepping stone model on $\mathbb{Z}^2$. There are $2N$ genes in each deme. We suppose that the population evolves in discrete generations. In each generation, first offspring are generated by Wright-Fisher sampling within each deme. Next a proportion $g_1(x - y)$ of the offspring in deme x migrate to deme y. Rather than introducing a mutation mechanism, we think of the Wright-Malécot formula as prescribing the generating function of the number of generations back to the most recent common ancestor of two individuals sampled at separation x (a two-dimensional vector) from the population. Our derivation parallels the approach of Wright (1943b).

Let $\psi_t(x)$ be the probability that two genes sampled at separation x had their most recent common ancestor exactly t generations in the past. For $t > 1$, we decompose this quantity according to the separation of the immediate ancestors of the two genes. If the two genes arose as migrants from the same deme, then with probability $1/2N$ they have a common ancestor in the previous generation. Thus

$$\psi_1(x) = \frac{1}{2N} G_1(x),$$

where $G_1(x)$ is the convolution of two copies of g_1 (corresponding to modelling the *separation* of two lineages). If, on the other hand, they have distinct parents, at separation y, then the chance that their most recent

common ancestor was t generations in the past is $\psi_{t-1}(y)$. For $t > 1$, we arrive at the recursion

$$\psi_t(x) = \sum_y \left\{ G_1(x-y)\psi_{t-1}(y) - \frac{\mathbf{1}_{\{y=0\}}}{2N} G_1(x-y)\psi_{t-1}(0) \right\}. \qquad \text{(A.1)}$$

This can be rewritten as

$$\psi_t(x) = \frac{1}{2N}\left(G_t(x) - \sum_{\tau=1}^{t-1} G_{t-\tau}(x)\psi_\tau(0) \right), \qquad \text{(A.2)}$$

where G_t is the t-fold convolution of G_1. Writing T for the (random) time at which the two genes share their most recent common ancestor, the generating function of T, which of course depends upon the sampling distance between the two genes, is defined by $\phi(z,x) = \mathbb{E}_x[z^T]$. The subscript x in the expectation is used to indicate that the sampling distance is x. Multiplying (A.2) by z^t and summing over t yields

$$\phi(z,x) = \frac{\tilde{G}(z,x)}{2N}\left(1 - \phi(z,0)\right),$$

where $\tilde{G}$ denotes the Z-transform (discrete Laplace transform) of G,

$$\tilde{G}(z,x) = \sum_{t=1}^{\infty} G_t(x) z^t.$$

Setting $x = 0$ to find an expression for $\phi(z,0)$ and substituting gives

$$\phi(z,x) = \frac{\tilde{G}(z,x)}{2N + \tilde{G}(z,0)}. \qquad \text{(A.3)}$$

This takes a particularly simple form if g_1 is a discretised Gaussian kernel which we can then approximate by a strictly Gaussian dispersal kernel. On an infinite range,

$$\frac{1}{2N} G_t(x) = \frac{1}{2\mathcal{N}t} \exp\left(-\frac{|x|^2}{4\sigma^2 t} \right),$$

where $\mathcal{N} = 4N\pi\sigma^2$ is the *neighbourhood size*. (This corresponds to dispersal of individual lineages at rate $\sigma^2/2$, the extra factor of two arising because G_1 governs the separation between two lineages.) If we write $\eta(x)$ for the

probability that two lineages at separation x will coalesce in the previous generation, we can write

$$\mathcal{N} = \frac{2\pi\sigma^2}{\int_{\mathbb{R}^2} \eta(x)dx}. \tag{A.4}$$

With this continuous approximation to G_t,

$$\frac{1}{2N}\tilde{G}(z,0) = \frac{1}{2\mathcal{N}}\sum_{t=1}^{\infty}\frac{z^t}{t} = \frac{1}{\mathcal{N}}\log\left(\frac{1}{\sqrt{1-z}}\right)$$

and

$$\frac{1}{2N}\tilde{G}(z,x) = \frac{1}{\mathcal{N}}\sum_{t=1}^{\infty}\frac{z^t}{2t}\exp\left(-\frac{|x|^2}{4\sigma^2 t}\right).$$

Provided that $|x|\sqrt{1-z}/\sigma$ is not too small, $|x|/\sigma > 2$, say, and $z > 0.5$, this latter quantity is approximately

$$\frac{1}{\mathcal{N}}K_0\left(\frac{|x|}{\sigma}\sqrt{1-z}\right),$$

where K_0 is the modified Bessel function of the second kind of degree zero. However, as $|x| \downarrow 0$, $\mathcal{N}\tilde{G}(z,x)/(2N)$ tends to $\log(1/\sqrt{1-z})$ whereas $K_0(|x|\sqrt{1-z}/\sigma)$ diverges.

We now have the ingredients for the generating function of the coalescence times:

$$\phi(z,0) = \mathbb{E}_0[z^T] = \frac{1}{1-\frac{2\mathcal{N}}{\log(1-z)}}, \tag{A.5}$$

and, at least for sufficiently large $|x|$ and z sufficiently close to 1,

$$\phi(z,x) = \mathbb{E}_x[z^T] = \frac{\tilde{G}(z,x)}{2N+\tilde{G}(z,0)} \approx \frac{K_0\left(\frac{|x|}{\sigma}\sqrt{1-z}\right)}{\mathcal{N}-\log(\sqrt{1-z})}. \tag{A.6}$$

Essentially the same derivation can be applied to any dispersal distribution, including nearest neighbour random walk. Of course, the expression (A.6) cannot apply for very small $|x|$ as it has the problem, inherited from K_0, of divergence at $x = 0$. The exact solution for these very small sampling distances will depend upon the details of the dispersal mechanism. For Gaussian dispersal, Malécot (1948) finds the exact expression as an integral with respect to a Bessel function. Durrett (2008) (Theorem 5.7) considers the case where migration of ancestral lineages is governed by nearest neighbour random walk. One can establish a similar recursion to (A.1) for populations that are distributed across a spatial continuum (Barton et al., 2002).

Some care is needed to write down models that do not suffer from Felsenstein's 'pain in the torus' (Felsenstein, 1975), but the spatial Λ-Fleming-Viot process of §3 provides one such continuum model. When working in a spatial continuum, to circumvent the divergence of the Bessel function in (A.6) it is often convenient to proceed as in Barton et al. (2002) and declare there to be a local scale κ over which the generating function is approximately constant and equal to $\tilde{\phi}(z,0)$. Writing equation (A.6) as

$$\phi(z,x) = \frac{1-\tilde{\phi}(z,0)}{\mathcal{N}} K_0\left(\frac{|x|}{\sigma}\sqrt{1-z}\right),$$

equating $\phi(z,\kappa)$ to $\tilde{\phi}(z,0)$ and rearranging (using that $K_0(y) \approx -\log y$ as $y \downarrow 0$) we obtain

$$\phi(z,x) \approx \frac{K_0\left(\frac{|x|}{\sigma}\sqrt{1-z}\right)}{\mathcal{N} - \log\left(\frac{\kappa}{\sigma}\sqrt{1-z}\right)}. \tag{A.7}$$

It is more usual to set $z = e^{-2\mu}$ with μ representing a mutation rate (per individual per generation) under an infinitely many alleles mutation model. The quantity $\phi(z,x)$ then tells us the probability that two alleles sampled at separation x are identical in state. Substituting in (A.7) we obtain the more familiar version of the Wright-Malécot formula given in (1):

$$\phi(e^{-2\mu},x) = \mathbb{E}_x[e^{-2\mu T}] \approx \frac{K_0(x/\ell_\mu)}{\mathcal{N} + \log(\ell_\mu/\kappa)}, \quad \text{for } |x| > \kappa,$$

where $\ell_\mu = \sigma/\sqrt{2\mu}$ and

$$\phi(e^{-2\mu},0) = \frac{\log(\ell_\mu/\kappa)}{\mathcal{N} + \log(\ell_\mu/\kappa)}.$$

It is to this version of the Wright-Malécot formula that we compare the probability of identity under the spatial Λ-Fleming-Viot process in §3, with a value of κ numerically evaluated as $\kappa \approx 1.34113$.

Appendix A.2. Probability of identity in the spatial Λ-Fleming-Viot model

Let us now assume that we work under the spatial Λ-Fleming-Viot model described in §3. In (3), we have seen that a single lineage moves around in space according to a continuous-time random walk with variance parameter $\sigma^2 = \frac{\lambda u \pi R^4}{2}$.

Suppose now that we have two lineages at separation x with $|x| < 2R$. Without loss of generality, we suppose that one of them is at the origin. They

are both in the region affected by any events whose centres lie in $B(0,R) \cap B(x,R)$ and during such an event they coalesce with probability u^2/ν (the factor of $1/\nu$ is because they must both choose the same parent). Thus when two lineages are at separation x, their instantaneous rate of coalescence is $\eta(x) = \lambda u^2 \mathrm{Vol}\left(B(0,R) \cap B(x,R)\right)/\nu$. We now define neighbourhood size, $\mathcal{N}$, as

$$\mathcal{N} := \frac{2\pi\sigma^2}{\int_{\mathbb{R}^2} \eta(x)\, dx} = \frac{\nu}{u}.$$

Notice that this corresponds to the definition of neighbourhood size for the stepping stone model in (A.4).

In much the same way as we arrived at the recursion (A.1) for the stepping stone model, by considering the behaviour of the ancestral lineages over an infinitesimal time interval, it is elementary to derive an equation for the probability of identity, $\phi_\mu(x) \equiv \phi(e^{-2\mu}, x)$ for two individuals sampled at separation x:

$$\begin{aligned} -\frac{2\mu}{\lambda}\phi_\mu(x) + (1-\phi_\mu(x))\,\frac{u^2}{\nu}\,L_R(x) \\ + \int_{\mathbb{R}^2} \frac{2u}{\pi R^2}\big(L_R(y) - uL_R(x,y)\big)\big(\phi_\mu(x-y) - \phi_\mu(x)\big)\,dy \\ + u^2\Big(1-\frac{1}{\nu}\Big)L_R(x)\int_0^2 f(z)\big(\phi_\mu(Rz) - \phi_\mu(x)\big)\,dz = 0, \end{aligned} \tag{A.8}$$

where $L_R(y)$ denotes the volume of the intersection $B(0,R) \cap B(y,R)$, $L_R(x,y)$ for that of the intersection $B(0,R) \cap B(x,R) \cap B(y,R)$ and

$$f(x) = \frac{x}{\pi}\left(4\arccos(x/2) - x\sqrt{4-x^2}\right), \qquad x \in [0,2], \tag{A.9}$$

is the density function of the distance between two points sampled independently and uniformly at random within the unit disc (cf. Alagar (1976)). See e.g. Equation (2) in (Barton et al., 2010b) for the derivation of the evolution equation of the probability of identity in state in a very similar context. This equation can be solved numerically, but in §3 we show that the solution is very well-approximated by the Wright-Malécot formula with the parameters σ^2 and $\mathcal{N}$ given above.

Appendix B. Proof of (7)

Here we detail the comparison between $\mathbb{P}_0[T > t^*]$ and $\mathbb{P}_r[T > t^*]$ which justifies the approximation (7).

Let us write $L_{t^*}(r)$ for the total time that a random walk with variance $2\sigma^2$, started at distance r from the origin, spends at the origin up until time t^*. Then $\mathbb{P}_r[T > t^*] = \mathbb{E}[\exp(-L_{t^*}(r)/N)]$ and so

$$\begin{aligned}\mathbb{P}_r[T > t^*] - \mathbb{P}_0[T > t^*] &= \mathbb{E}\left[e^{-(L_{t^*}(0)/N)}\left(e^{(L_{t^*}(0)-L_{t^*}(r))/N} - 1\right)\right] \\ &\approx \mathbb{E}\left[e^{-(L_{t^*}(0)/N)}\left(\frac{L_{t^*}(0) - L_{t^*}(r)}{N}\right)\right] \\ &\approx \frac{1}{N}\mathbb{P}_0[T > t^*]\mathbb{E}[L_{t^*}(0) - L_{t^*}(r)]. \qquad \text{(B.1)}\end{aligned}$$

As $t^* \to \infty$, $\mathbb{E}[L_{t^*}(0) - L_{t^*}(r)] \to a(r)$, the *potential kernel* of the random walk (see e.g. Lawler and Limic (2010)). This function takes the form

$$a(r) = 2C \log r + \mathcal{O}(1),$$

where $C = 1/(4\pi\sigma^2)$ (there is an extra factor of two in the denominator here since we are interested in the random walk governing the separation of two lineages, not the motion of a single lineage). To see where this comes from, since we are assuming that $t^* \gg r^2$, by the time of order r^2 when the random walk starting at separation r hits zero for the first time, the random walker started from the origin has spent $\sim \sum_{n<r^2} 1/n \sim 2C \log r$ units of time at 0, where $C = 1/(4\pi\sigma^2)$. From that time onwards, the walk started from r behaves like a walk started from zero. We refer to Lawler and Limic (2010) for more details. Combining the above we see that

$$F(r) \approx \frac{1 - F(0)}{\mathcal{N}} \log\left(\frac{c}{r}\right),$$

where c is determined by the geometric mean of the separation of individuals sampled from A.

Remark 2. *Note that letting t^* tend to infinity as above does not contradict our use of the comparison result when t^* is fixed to an intermediate timescale of a couple of hundred generations. Indeed, some estimates of the speed of convergence of $\mathbb{E}[L_{t^*}(0) - L_{t^*}(r)]$ to the potential kernel of the random walk can be obtained and show that when $r \ll \sqrt{t^*}/\sigma$, $t^* \approx 100$ generations is sufficiently large for the approximation to be reasonable.*

Appendix C. Proof of Theorem 1

We fix δ, and write $\xi^{j,1}$ and $\xi^{j,2}$ for the ancestral lineages of our two individuals at locus j, and T^j for their coalescence time. Let us suppose

that at a given locus $j-1$, the two individuals are identical in state due to an early coalescence of $\xi^{j-1,1}$ and $\xi^{j-1,2}$. For the lineages $\xi^{j,1}$ and $\xi^{j,2}$ corresponding to locus j not to coalesce early, one needs that:

1. Either $\xi^{j,1}$ or $\xi^{j,2}$ recombines away and become well-separated from $\xi^{j-1,1}$ or $\xi^{j-1,2}$ respectively, in less time than $\xi^{j-1,1}$ and $\xi^{j-1,2}$ need to coalesce. We shall call this step an *effective recombination.*
2. Once Step 1 has been completed, $\xi^{j,1}$ and $\xi^{j,2}$ do not coalesce early.

Let us explain why we want the lineages to become well separated during the first step. Recalling how recombination works from **??**, we see that during each event affecting a given individual, recombination breaks the link between loci $j-1$ and j with probability ρ. In this case, the common ancestral lineage of the two loci splits into two lineages, with locations independently and uniformly distributed over a ball of radius $R \ll \delta$. Hence, even though the lineages recombine away from each other, they remain close enough for their behaviours to be highly correlated for a while (they may coalesce again quickly, for instance). If neither $\xi^{j,1}$ nor $\xi^{j,2}$ managed to become well-separated from its former adjacent locus before the early coalescence of $\xi^{j-1,1}$ and $\xi^{j-1,2}$, then with high probability the event causing that coalescence would also, with high probability, result in the coalescence of $\xi^{j,1}$ and $\xi^{j,2}$. Consequently, we make the following definition.

Definition 1. *We call a recombination* effective *if the two recombinants become separated by distance δ before coalescing again.*

Remark 3. *Here we make the approximation that if $\xi^{j,1}$ manages to become decorrelated from $\xi^{j-1,1}$, it is also decorrelated from every $\xi^{j-i,1}$, $i \leq j-1$. Indeed, all the lineages corresponding to a locus $k \leq j-1$ that are still correlated remain close together, whereas decorrelation implies separation at large distances. In addition, for reasonably small $\mathcal{N}$, the number of nearby lineages is limited since most of them would coalesce quickly. Hence, $\xi^{j-1,1}$ is in fact trying to become decorrelated from a small cloud of very nearby lineages, which is essentially the same as escaping from a single one. Of course, this approximation becomes worse and worse as neighbourhood size increases.*

Step 1 focuses mainly on the decorrelation of the ancestral lines of two adjacent loci while Step 2 deals with the coalescence of two lineages which are initially far from each other. We examine each step separately.

Appendix C.1. Probability of an early coalescence

Recall that 'well-separated' in Step 1 means 'at distance δ'. The definition of an early coalescence yields directly

$$\mathbb{P}_\delta[\,\text{early coal.}\,] = \mathbb{E}_\delta\big[e^{-2\zeta T}\big],$$

and using (1) with $\zeta = (\sigma/\delta)^2$ and thus $\ell_\mu = \delta/\sqrt{2}$, we obtain

$$\mathbb{P}_\delta[\text{not an early coal.}] = 1 - \frac{K_0(\sqrt{2})}{\mathcal{N} + \log(\delta/(\kappa\sqrt{2}))}.$$

This gives us the second term in the expression for $\gamma(\delta)$.

Appendix C.2. Escape probability

Following Definition 1, we need to compute the probability $\alpha(\delta)$ that two recombinants separate to distance δ before coalescing again. Let us write $g_D(x)$ for the probability that two lineages starting at distance $x > 0$ separate to a distance at least D before coalescing. Of course $g_D(x) = 1$ for any $x \geq D$, and writing $f(y)dy$ for the distribution of the distance between two points sampled independently and uniformly at random in the ball $B(0,1)$, we have

$$\alpha(\delta) = \int_0^2 f(y) g_\delta(Ry)\, dy. \tag{C.1}$$

Now, using the description of the evolution of two lineages (c.f. §3), we obtain an equation similar to (A.8):

$$\begin{aligned} -\, g_D(x)\, \frac{u^2}{\nu}\, L_R(x) + \int_{\mathbb{R}^2} \frac{2u}{\pi R^2} \big(L_R(y) - u L_R(\vec{x}, y)\big)\big(g_D(|\vec{x} - y|) - g_D(x)\big)\, dy \\ + u^2\Big(1 - \frac{1}{\nu}\Big) L_R(x) \int_0^2 f(z)\big(g_D(Rz) - g_D(x)\big)\, dz = 0, \end{aligned} \tag{C.2}$$

with boundary condition $g_D(x) = 1$ for every $x \geq D$. This equation can be written as a Fredholm equation of the second kind and can thus be solved numerically.