On the Choice of Genetic Distance in Spatial-Genetic Studies

Paul Fearnhead

Department of Mathematics and Statistics

Lancaster University

Summary We look at how to choose genetic distance so as to maximise the power of detecting spatial structure. We answer this question through analysing two population genetic models that allow for a spatially structured population in a continuous habitat. These models, like most that incorporate spatial structure, can be characterised by a separation of time scales: the history of the sample can be split into a scattering and collecting phase, and it is only during the scattering phase that the spatial locations of the sample affects the coalescence times. Our results suggest that the optimal choice of genetic distance is based upon splitting a DNA sequence into segments, and counting the number of segments at which two sequences differ. The size of these segments depends on the length of the scattering phase for the population genetic model.

Keywords *Isolation by Distance, MLST data, Spatial Autocorrelation*

Introduction

We consider the problem of learning about spatial structure from population genetic data. We focus on the situation where we have both genetic and spatial data from a random sample of individuals from a population in a continuous habitat. The spatial information relates to the sampling location of the individuals, and the genetic information will be the genetic type of those individuals at a series of loci. From this data we would like to answer questions such as whether there is spatial structure within the population (as opposed to the data being consistent with a panmictic population), and if so to quantify features of how this structure affects the genetic diversity of the population.

A simple, but commonly used, approach to answering whether there is spatial structure is to look for correlation between the spatial and genetic distance between two individuals from the population. This can be calculated by considering all pairs of individuals within the data set, calculating the correlation between the set of paired spatial and genetic distances, and then assessing the significance of any observed correlation through a permutation test (Sokal and Oden, 1978; Shimatani and Takahashi, 2003). This idea can be extended to look at the relationship of spatial separation on genetic difference by plotting a smoothed estimate of how genetic distance varies with spatial separation for the pairs of individuals within the data set (see e.g. Shimatani and Takahashi, 2003; French *et al.*, 2005).

However to implement these approaches requires the definition of spatial and genetic distance for a pair of individuals. Often Euclidean distance is a natural choice for spatial distance. However, there can be multiple possible choices of genetic distance, and in some situations the choice of distance can effect the results of the subsequent analysis (Shimatani and Takahashi, 2003).

As a motivating example, consider the study of *Campylobacter jejuni* in French *et al.* (2005). Here the genetic data for each *C. jejuni* isolate consisted of multi-locus sequence types (MLSTs). An MLST records the DNA sequence of the isolate at ≈500bp fragments of 7 housekeeping genes which are roughly evenly spread around the genome. If we consider the data from two isolates at a single gene, then two natural measures of genetic distance are (i) the number of polymorphic differences between the two sequences; (ii) whether or not the sequences are identical. There are also alternative measures of distance that could be considered (see METHODS). A natural and important question is which choice of distance is best in terms of detecting and learning about the effect of any spatial structure on genetic diversity.

We investigate this question via analysis of two spatial population genetic models (see METHODS). Both models assume a population that exists in a continuous habitat, and that the spatial location of an offspring is centred around the location of its parent. Both models only apply to non-recombining loci, and thus we focus on the choice of genetic distance for a single non-recombining locus. (We are unaware of appropriate spatial genetic models which incorporate recombination.)

METHODS

Spatial Genetic Models

Our results are based on two population genetic models for continuous spatial habitats, also known as Isolation by Distance (IBD) models. The first assumes complete density regulation: that is that the population density is constant through space and time. This model can be constructed as the limit of a 2-dimensional stepping stone model as the number of demes tends to infinity. This model has been analysed by Maruyama (1971) Malécot (1975), Barton and Wilson (1995), Barton and Wilson (1996) and Barton *et al.* (2002) amongst others. However here we use the simulation method and analytic approximations of Wilkins (2004), and throughout this paper we call this model the Wilkins's IBD model.

The second is based on the Isolation by Distance model of Wright (1943). We call this Wright's IBD model. This model has no density regulation, which has the disadvantage that it produces infinite clumping of the population (Felsenstein, 1975).

As we are interested in the property of estimators that use the genetic and spatial information on pairs of chromosomes, we consider samples of size 2 from these models. We consider a single non-recombining locus and assume this locus consists of L sites, with two alleles at each site. We further assume the same mutation rate at each site, and parameterise the mutation rate in terms of a scaled rate per site $\theta = 2N_e u$ where N_e is the effective (haploid) population size and u is the per generation mutation rate for the locus. The effective populations size is defined so that the mean number of mutations in the locus that separates a randomly sampled pair of haploid individuals will be $L\theta$.

Wilkins's IBD model

We consider a haploid population inhabiting a square habitat $[0, 10] \times [0, 10]$. The model is parameterised in terms of a population density, ρ, and a dispersion parameter σ^2. A simple description of the ancestral process for this model is as described below (see Wilkins, 2004; Wilkins and Wakeley, 2002, for fuller details). Note that this model is equivalent to one for a habitat $[0, 10/c] \times [0, 10/c]$ with population density $c^2\rho$ and dispersal rate σ^2/c^2, for any $c > 0$.

We consider a sample taken from known locations. We can then trace the ancestry of our sample back in time. At any time in the past this ancestry will consist of a number of lineages, which correspond to the unique descendants of the population at that time. The position of a lineage undergoes a two-dimensional symmetric Gaussian random walk, with variance σ^2 in each direction. (We assume reflecting boundaries at the edge of the habitat.) Two lineages coalesce (share a common ancestor) if the lineages fall within an area containing a single individual (which is of size $1/\rho$).

Wilkins (2004) shows that qualitatively the genealogy from this model can be split into two phases, known as the "scattering" and "collecting" phase (this terminology was first used in Wakeley, 1999). The scattering phase is the initial phase of the genealogy, and corresponds to the period of time that the coalescence times depends on the sampling locations. This is then followed by the collecting phase, when coalescences are independent of the sampling locations and the genealogy can be closely approximate by Kingman's coalescent (Kingman, 1982).

During the collecting phase, the distribution of the genealogy is described by a single parameter: the effective population size N_e. This governs the rate of coalescence of a pair of lineages (which is $1/N_e$). Wilkins (2004) gives various approximation for N_e in terms of the parameters of the model; and this can also be estimated through simulation. Within the scattering phase, the distribution of the coalescent time for a pair of individuals sampled at x_1 and x_2 respectively depends on the scaled distance $||x_1 - x_2||/\sigma$, where $|| \cdot ||$ is the standard Euclidean distance.

To show this we plotted the hazard function of the coalescent time distribution for a range of distances between the sampled individuals. The hazard function of a random variable T is defined as $\Pr(T = t)/\Pr(T \geq t)$. Under a panmictic population model, the hazard function of the coalescent time would be constant through time and equal to $1/N_e$. Figure 1(a) shows the hazard functions we obtained, and we see that these tend to a constant value of approximately $1/N_e$ regardless of the position of the sample. In

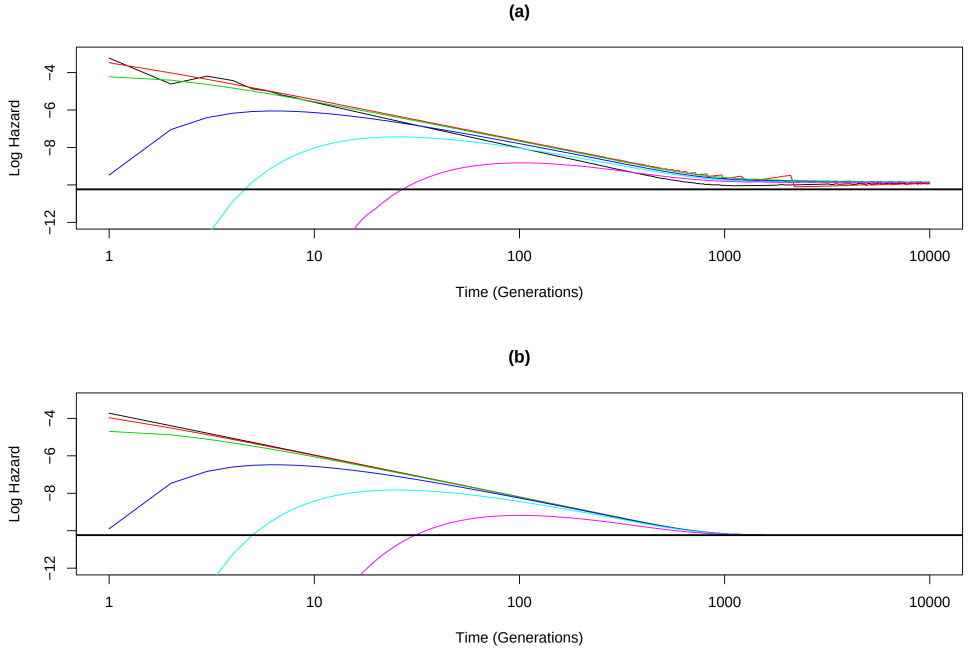

Figure 1: Plot of log hazard function for the coalescent time distribution of a sample of size 2. (a) Wilkins's IBD model with $\sigma = 0.1$, $\rho = 200$; (b) Wright's IBD model with $\sigma = 0.1$. In both cases $N_e = 28,000$ and the habitat was $[0, 10] \times [0, 10]$. The different curves in each plot correspond to different degrees of separation of the sample, and the values chosen where 0, σ, 2σ, 5σ, 10σ and 20σ. (The lines are ordered with smaller distances having larger hazard values at small numbers of generations.) The sampled locations were chosen to be in the middle of the habitat. Hazard function for (a) was calculated via the approximation's in Wilkins (2004), but simulation results gave qualitatively identical results.

this case convergence occurs at around the 1,000th generation. Prior to this time, we notice quite difference behaviour in the hazard functions.

A further important parameter of the model is the time at which the scattering phase ends and the collecting phase starts, which we call T_c. Wilkins (2004) gives ways of calculating this; though we have resorted to using visual pictures such as Figure 1 to estimate an appropriate value. (In practice this time is not clearly defined, and rough estimates, such as the value of 1,000 generations for Figure 1a are sufficient for our needs.)

We considered a range of parameter values for the results we present here. In each case we calculated the distribution of the coalescence time for a sample of two individuals. We examined this using both the analytic approximation of Wilkins (2004), and through simulation using the `tracker` program (available from `http://www.santafe.edu/∼wilkins/software.html`).
In all cases we sampled individuals from close to the centre of the habitat, to avoid any edge-effects of the model.

Wright's IBD Model

This is also a model for a haploid population. We consider a slight generalisation of the IBD model of Wright (1943).

We consider a random sample from a structured population. By random, we mean that the probability of an individual being sampled does not depend on its genetic type. We do allow the sampling to depend on the location of the individuals, and calculate the distribution of the coalescence time of a pair of individuals conditional on their sampling locations. To calculate this conditional distribution we first need to consider the unconditional distribution of the coalescence time, and the distribution of the spatial locations given the coalescence time.

Forward in time, the model assumes a fixed population size evolving over discrete generations. Ignoring the spatial information, the evolution of the population is given by the Wright-Fisher model: each descendent in the next generation "chooses" its parent independently and uniformly from the individuals in the current generation. Conditional on this, the location of the descendent is a small perturbation of the location of its parent. By considering a suitable limit as the population size tends to infinity we have a model where, marginal to the spatial information, the genealogy of a sample is given by Kingman's coalescent (Kingman, 1982). Conditional on the genealogy and the location of the MRCA of the sample, the location of the sampled individuals is obtained by simulating Brownian motion along the branches of the genealogy. Thus if we consider a branch in the genealogy of length t, if the parental node is at location y then the location of the daughter node (or tip) is drawn from a transition density $p_t(\mathbf{x}|\mathbf{y})$, where $p_t(\mathbf{x}|\mathbf{y})$ is a (2-dimensional) Gaussian distribution with mean y and variance $t\sigma^2$ in each direction.

We consider the special case of a model for a population on a closed habitat. We again choose the habitat to be $[0, 10] \times [0, 10]$. For this model the transition density, $p_t(x|y)$, is defined to be that of 2-dimensional Brownian motion constrained to the habitat. So if $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$

$$p_t(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{2} \left[\sum_{k=-\infty}^{\infty} \left(\mathcal{N}(x_i + 20k; y_i, \sigma^2 t) + \mathcal{N}(20k - x_i; y_i, \sigma^2 t) \right) \right],$$

where $\mathcal{N}(\cdot; \mu, \sigma^2)$ is the density of the Gaussian random variable with mean μ and variance σ^2. The infinite sum in this expression is to allow for the reflecting boundary of the habitat (see the Appendix of Wilkins, 2004). This model is chosen to have the same spatial dynamics as the Wilkins's IBD model. For this model, the distribution of the location of the MRCA is given by the stationary distribution of the transition density $p_t(x|y)$, which is just uniform on the habitat.

Intuitively this model behaves similarly to a neutral coalescent model, where we treat the location of the individual in the same way as a genetic locus. In this case the "type" of the individual at this locus lies in $[0, 10] \times [0, 10]$, and the "mutation" process is Brownian motion with reflecting boundaries. The type of the MRCA is drawn from the stationary distribution of this mutation process.

Now we can calculate the conditional distribution of the coalescence time of a sample of size 2 given their sampling locations. Consider a sample taken from specified locations, $\mathbf{x}$ and $\mathbf{y}$ say. For a population of effective population size N_e the conditional distribution of the number of generations until the sample has a common ancestor is given, using Bayes' formula, by

$$p(t|\mathbf{x}, \mathbf{y}) \propto \exp\{-t/N_e\} p_{2t}(\mathbf{x}|\mathbf{y}).$$

Here the first time comes from the exponential prior distribution of the coalescence time, and the second term is the conditional distribution of the location of the sample given the coalescence time. (This simplifies due to the reversibility of the dispersal process.) Note that here we have defined time in terms of generations, so σ^2 for our model is defined in terms of the variance of the dispersal over one generation.

Plots of the hazard function of the coalescence times for this model again show a separation of time-scales effect (Figure 1b). Again it will be appropriate to summarise the model by two parameters: the effective sample-size, N_e, and the time at which the collecting phase starts (the hazard rate is approximately $1/N_e$), T_c. For this model the time depends only on σ and the habitat size. For our habitat $T_c \approx 10/\sigma^2$.

Genetic Distances

The two most natural measures of genetic distance between a pair of sequences at a locus are (i) the number of segregating sites, which we call d_L; and (ii) whether or not the sequences at the locus are the same, which we call d_1. Note the $d_1 = \mathrm{I}(d_L > 0)$.

We can obtain a range of measurements in between these extremes by considering segments made up of subsets of the L sites at the locus. Consider such a set of l such segments, each of which consists of L/l sites. (The natural definition would be to consider the first segment to consist of the first L/l sites; the second the next L/l sites, and so on.) Now define the genetic distance to be the number of these segments at which the two sequences differ. We define this to be d_l.

Power to detect Spatial Structure

Consider a choice of genetic distance. Let $\tilde{\mu}$ be the expectation of this distance under a panmictic population model; and $\mu(x)$ and $V(x)$ be the expectation and variance of this under a spatial model for samples chosen at distance x from each other. Then we (indirectly) measure the power to detect spatial structure through

$$D(x) = (\mu(x) - \tilde{\mu})^2/V(x). \tag{1}$$

This is a natural measure, as it directly relates to the non-centrality parameter of a chi-squared statistic to detect whether $\mu(x) \neq \tilde{\mu}$.

RESULTS

We first examined the distribution of the number of segregating sites in a sample of size 2. We simulated data assuming a bi-allelic mutation model at each site, with mutation rate $\theta = 2N_e u = 0.01$. Thus for a locus consisting of 500 bases we would expect around 5 segregating sites in a sample of 2 chromosomes, which is consistent with *Camplyobacter jejuni* MLST data. (Fearnhead *et al.*, 2005).

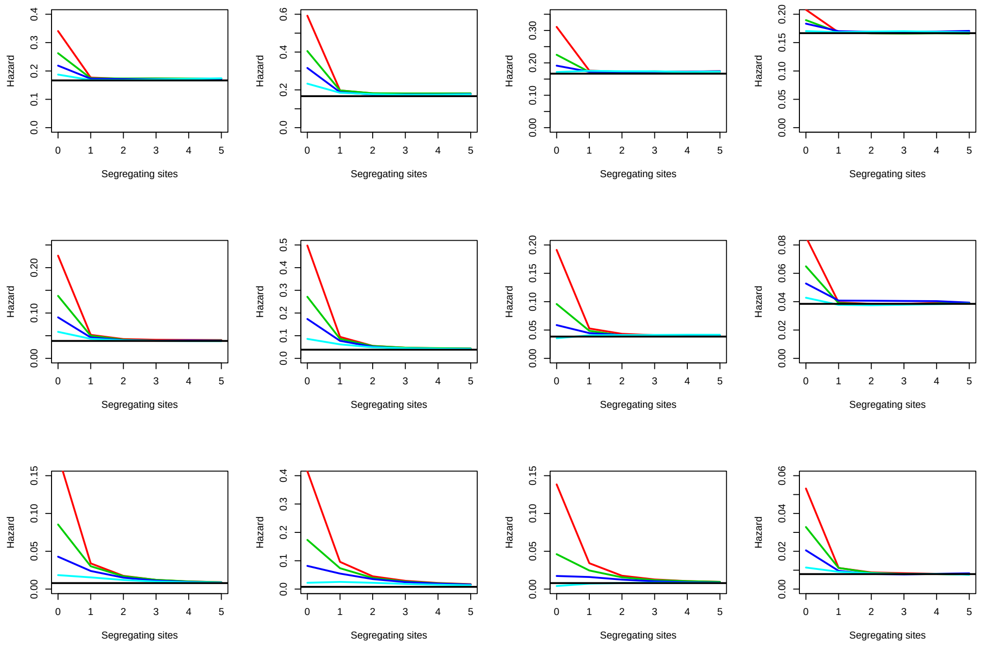

Figure 2: Plot of the hazard function for the number of segregating sites for Wilkins's IBD model. The top row corresponds to a locus with 500bp; the middle row one with 2500bp and the bottom row one with 12500bp. The columns correspond to different demographic parameters (from left): $\sigma = 0.1$, $\rho = 200$; $\sigma = 0.1$, $\rho = 50$; $\sigma = 0.2$, $\rho = 50$; $\sigma = 0.1$, $\rho = 1000$. (These correspond to N_e values of 28,000; 13,000; 6,700; and 108,000.) The different colour lines represent different distances between the sampled chromosomes: these are 0 (red), 5σ (green), 10σ (blue) and 20σ (light blue). (Note that the curves are ordered, with smaller distances producing higher hazard values.) Results are based on 10,000 draws from the distribution of coalescence times, and 100 simulated data sets for each draw. The horizontal line shows the hazard under a panmictic model.

In Figure 2 we plot the hazard function for the number of segregating sites for the Wilkins's IBD model. We consider a range of demographic parameters, and three sizes of locus: 500bp, 2500bp and 12500bp. If S is the the number of segregating sites then the hazard function evaluated at the value s is defined as $\Pr(S = s)/\Pr(S \geq s)$. We plot this as under a panmictic model the hazard function is constant; and thus it highlights deviation from the panmictic model.

Each plot shows the hazard for four different degrees of separation of the sample chromosomes (for full details see figure caption). The common feature of the plots is that for the 500bp locus, the only noticeable difference in the hazard is for 0 segregating sites, with greater probability for closely sampled pairs of chromosomes. As the size of the locus increases (and with it the mutation rate of the locus) we observe differences in the hazard for other numbers of segregating sites.

The importance of this result is that for small loci, the only information about the spatial model will be found in the proportion of pairs of identical chromosomes at different distances. (Conditional on $S > 0$ the hazards are almost identical for different spatial separations, so there is almost no information in the conditional distribution of the number of S given $S > 0$.) Thus measuring genetic distance via whether two chromosomes are identical at that locus will be optimal. However, for larger loci there is likely to be information over and above whether two chromosomes are identical: and in these situations other genetic distances may perform better at detecting spatial structure.

The reason for this dependence on locus size is related to the separation of time-scales property of spatial models. It is only during the scattering phase (up to time T_c see METHODS) that the hazard of coalescence times depends on the spatial separation of the chromosomes. This difference will manifest itself in the hazard only for small numbers of segregating sites. For two lineages that coalesce prior to time T_c, the expected number of mutations will be bounded by $2T_cLu = L\theta T_c/N_e$, and only when this is of the order of 1 or more will you observe a noticeable difference in the hazard function at 1 or more segregating sites. For the columns of Figure 2, $T_c/N_e \approx 0.04$, 0.08, 0.04 and 0.01 respectively. Thus $L\theta T_c/N_e < 1$ for $L = 500$ in all cases, and also for $L = 2500$ for the rightmost column.

Note that the results depend primarily on the value of $L\theta$, and changing L and θ whilst keeping this product the same has little affect on the results (at least while $\theta << 1$).

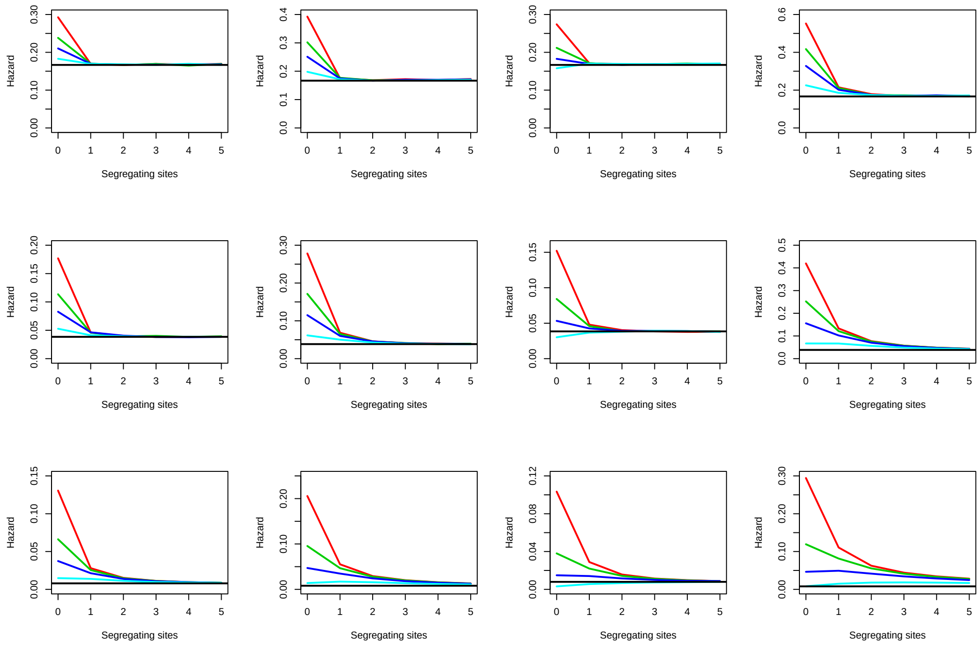

Figure 3: Plot of the hazard function for the number of segregating sites for Wright's IBD model. The top row corresponds to a locus with 500bp; the middle row one with 2500bp and the bottom row one with 12500bp. The columns correspond to different demographic parameters (from left): $\sigma = 0.1$, $N_e = 28,000$; $\sigma = 0.1$, $N_e = 13000$; $\sigma = 0.2$, $N_e = 6700$; $\sigma = 0.1$, $N_e = 5000$. (The first three have the same σ and N_e values as the model in the corresponding column of Figure 2). The different colour lines represent different distances between the sampled chromosomes: these are 0 (red), 5σ (green), 10σ (blue) and 20σ (light blue).(Note that the curves are ordered, with smaller distances producing higher hazard values.) Results are calculated analytically. The horizontal line shows the hazard under a panmictic model.

Furthermore the patterns we observe in Figure 2 are representative of a range of choices of the demographic parameters (results not shown). Note that for Wilkins's IBD model there is a limit on the range of values of T_c/N_e that can be obtained (due to their co-dependence on σ); and the choices given in Figure 2 show models with a reasonable spread over the possible values of T_c/N_e.

Similar qualitative results are obtained for Wright's IBD model (see first three columns of Figure 3). The first three columns of Figure 3 have identical T_c and N_e values to the first three columns of Figure 2, and we obtain qualitatively very similar results. For the rightmost column we have chosen N_e to be sufficiently small so that $T_c/N_e = 0.2$ and in this case we observe differences in the hazard for $S = 1$ even when $L = 500$.

Secondly, we looked at the effect of different measures of genetic distance on the power to detect spatial structure. We did this through fixing a spatial separation x and calculating the normalised distance $D(x)$, (see Eqn 1 in METHODS). Larger values of $D(x)$ correspond to greater power. The top two rows of Figure 4 correspond to four different spatial models, for $x = 0$ and $x = 5\sigma$. There are similar patterns for the different values of x for a given model - the main difference is that the power to detect variation from a panmictic model increases as x decreases. This pattern is observed over a greater range of x values (results not shown).

For each plot in the top two rows of Figure 4 we plot the power for four different choices of genetic distance, as a function of the size of the locus being analysed. These genetic distances include the two extremes, namely the number of segregating sites (denoted d_L, see METHODS) and whether or not the two sequences are identical (d_1). The two further distances are based on splitting the locus into 2 or 3 segments, and counting the number of segments at which the two sequences differ (d_2 and d_3 respectively). The optimal choice of distance varies with both size of locus and with the spatial model.

To investigate this further the bottom row of Figure 4 shows equivalent results in the $x = 0$ case, but now each curve is based on splitting the locus into segments of different size. We see that for each scenario there appears to be an optimal size of segment; and now increasing the size of the locus has little effect, with the $D(x)$ values converging to a fixed value for each choice of size of segment.

The optimal size of segment varies between spatial models; and the reason for this is the different T_c/N_e values for these models. Choosing different sized segments is equivalent to looking at differences in the distribution of coalescent events over different time-scales. Very large segments correspond to focusing on differences over very short time-scales; whereas small segment focus on differences over much large time-scales. Thus models with small T_c/N_e values should use large segments; and vice versa. As an approximate rule, the optimal segment size has population-scaled mutation rate θ_s where $\theta_s T_c/N_e \approx 1$. This can be seen from the bottom row of plots in Figure 4. The N_e/T_c values are approximately 28, 100, 28 and 5 for the four models. Over the values of θ_s considered, the optimum appears to be 100, 20 and 5 for the last three models; with $\theta_s = 20$ and $\theta_s = 50$ giving very similar results for the first model.

DISCUSSION

We have looked at two spatial population genetic models to give us insight into the

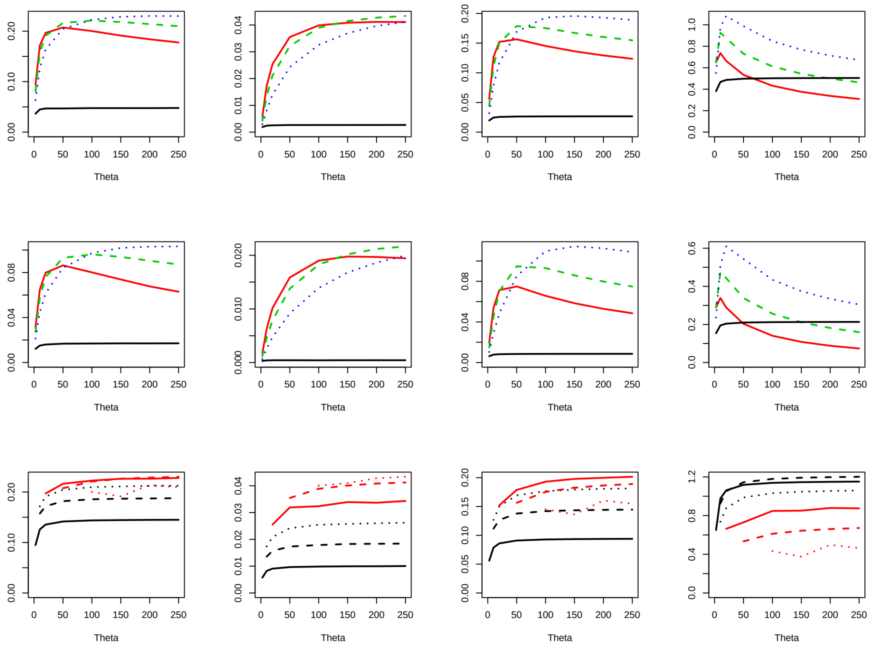

Figure 4: Plot of $D(x)$ values against locus size (measured by locus mutation rate θ) for different spatial models. Columns from right: Wilkins's IBD model $\rho = 200$, $N_e = 28,000$; Wilkins's IBD model $\rho = 1000$, $N_e = 108,000$; Wright's IBD model $N_e = 28,000$; and Wright's IBD model $N_e = 5,000$; all models use $\sigma = 0.1$, and $T_c \approx 1,000$. (Thus N_e/T_c is approximately 28, 100, 28 and 5 for these 4 models.) First two rows plot $D(x)$ against choice of genetic distance: d_1 (red, full line), d_2 (green, dashed line), d_3 (blue, dotted line) and d_L (black, full line). Top row is for $x = 0$, and middle row for $x = 5\sigma$. The bottom row shows plots for $x = 0$ but with each curve relating to segments of different lengths (see text), measured in terms of the segment's mutation rate θ_s: $\theta_s = 2$ (black, full line), $\theta_s = 5$ (black, dashed line), $\theta_s = 10$ (black, dotted line), $\theta_s = 20$ (red, full line), $\theta_s = 50$ (red, dashed line) and $\theta_s = 100$ (red, dotted line).

choice of genetic distance for detecting spatial structure in population genetic data. The models we considered differed in their assumptions of density regulation, and they each take one of the two extreme possibilities. Wilkins's IBD model assumes complete density regulation, whereas Wright's IBD model assumes no density regulation. Both models are unrealistic in real-life (in particular the Wright's IBD model leads to infinite population density) but the similarity in the results we obtain for both cases suggest that our results are informative about more realistic scenarios that lie between these two extremes.

The qualitative features of our results can be traced to the separation of time scales properties of these models: namely that there is an, often short, scattering period where the spatial location of the samples affects the genealogy, and that this is then followed by a collecting phase where the initial locations have no effect on the genealogy. The choice of distance should be based upon looking for differences across segments of DNA, where the size of the segment is chosen so that there will be of the order of 1 mutation expected between two haploid individuals that coalesce within the scattering phase. The fact that many spatial models (Wilkins and Wakeley, 2002; Wilkins, 2004; Slade and Wakeley, 2005) can be described via a separation of time scales suggests that this guideline will apply quite generally.

One difficulty with applying this result is knowing or inferring the length of the scattering phase, T_c (or the ratio T_c/N_e). All we can do is draw some general conclusions. For models with strong density regulation, like the Wilkins's IBD model, there are constraints on the values that T_c/N_e can take, and the results that we presented are representative of the results obtained for a range of different parameter values for the model. These suggest the segment you choose should have a high mutation rate; of the order of 20–50. (It is possible to choose parameter values that require larger segments, but not smaller ones.) So for example, for the *C. jejuni* MLST data the mutation rate is of the order of 5-10 for a gene fragment. Thus the optimal choice of genetic distance will be to look at whether or not sequences for a gene fragment are identical. Note also that some spatial models are equivalent to $T_c \approx 0$ (Slade and Wakeley, 2005), in which case measuring genetic distance by whether or not the complete DNA sequence at a locus is identical would be best. Non-equilibrium changes in population structure, such as range expansion or contraction, will affect the values that T_c/N_e can take. Historic contraction would increase the value T_c/N_e which would result in segments with lower mutation rates being preferred.

In our study we have ignored recombination, this is due to the difficulty with analysing or simulating from spatial genetic models which include recombination. However, the results we present may be robust to situations where there is some recombination. The reason for this is that in looking for spatial structure, we are looking for differences in the distribution of coalescence times within the scattering phase, up to T_c. Thus it is only recombination events that occur before T_c that will affect our conclusions. As T_c/N_e is often small, the probability of such recombination events will be small except for large or highly recombinant loci. In particular, for genetic loci that do not include a recombination hotspot (McVean *et al.*, 2004), the effect of recombination may be small.

One final conclusion from our work is that, as spatial structure affects the genealogy of a sample only during the, generally short, scattering phase, large amounts of genetic data may be needed to make strong conclusions about the presence and effect of spatial structure on genetic variation. In particular, it is likely to be beneficial to type fewer individuals over more of the genome.

Acknowledgements This work was funded by Engineering and Physical Research Council grant number EP/C531558. I would like to thank Jon Wilkins for help with using `tracker`.

References

Barton, N. H. and Wilson, I. (1995). Genealogies and geography. *Philosophical Transactions of the Royal Society of London, series B* **349**, 49–59.

Barton, N. H. and Wilson, I. (1996). Genealogies and geography. In: *New uses for new phylogenies* (eds. P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith and S. Nee), Oxford University Press, Oxford, 23–56.

Barton, N. H., Depaulis, F. and Etheridge, A. M. (2002). Neutral evolution in spatially continuous populations. *Theoretical Population Biology* **61**, 31–48.

Fearnhead, P., Smith, N. G. C., Barrigas, M., Fox, A. and French, N. (2005). Analysis of Recombination in Campylobacter jejuni from MLST Population Data. *Journal of Molecular Evolution* **61**, 333–340.

Felsenstein, J. (1975). A pain in the torus: some difficulties with the model of isolation by distance. *American Naturalist* **109**, 359–368.

French, N. P., Barrigas, M., Brown, P., Ribiero, P., Williams, N. J., Leatherbarrow, H., Birtles, R., Bolton, E., Fearnhead, P. and Fox, A. (2005). Spatial epidemiology and natural population structure of campylobacter jejuni colonising a farmland ecosystem. *Environmental Microbiology* **7**, 1116–1126.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications* **13**, 235–248.

Malécot, G. (1975). Heterozygosity and relationship in regularly subdivided populations. *Theoretical Population Biology* **8**, 212–241.

Maruyama, T. (1971). Analysis of population structure. II. Two dimensional stepping stone models of finite length. *Annals of Human Genetics* **34**, 201–219.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584.

Shimatani, K. and Takahashi, M. (2003). On methods of spatial analysis for genotyped individuals. *Heredity* **91**, 173–180.

Slade, P. F. and Wakeley, J. (2005). The ancestral selection graph and the many-demes limit. *Genetics* **169**, 1117–1131.

Sokal, R. R. and Oden, N. L. (1978). Spatial autocorrelation in biology. 1. methodology. *Biol J Linn Soc* **10**, 199–228.

Wakeley, J. (1999). Non-equilibrium migration in human history. *Genetics* **153**, 1863–1871.

Wilkins, J. F. (2004). A separation-of-timescales approach to the coalescent in a continuous population. *Genetics* **168**, 2227–2244.

Wilkins, J. F. and Wakeley, J. (2002). The coalescent in a continuous, finite, linear population. *Genetics* **161**, 873–888.

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