

Natural Selection in Spatially Structured Populations Case for support

Part 1: Previous research track record

The main purpose of theoretical population genetics is to understand the complex patterns of genetic variation that we observe in the world around us, and to show how these lead to the evolution of adaptation and diversity. Its origins can be traced to the pioneering work of Fisher, Haldane and Wright. This formed the basis of the ‘modern evolutionary synthesis’ in which Darwin’s theory of evolution by natural selection was finally reconciled with Mendelian genetics. Mathematical modelling was crucial in the work of all three and nearly a century later it continues to play a central rôle in the quest to understand the relative importance of the different forces that shape genetic variation. Equally, this fundamental biological question has stimulated exciting mathematical developments. For example, one of the outstanding successes of mathematical population genetics is Kingman’s coalescent, originally introduced to provide a description of the genealogical trees that relate a random sample of neutral genes from a panmictic (and therefore exchangeable) population. Just under thirty years later, exchangeable coalescents are subject to intensive study for their elegant mathematical properties and unexpected connections with a myriad of other questions in probability theory. This proposal, motivated by the same fundamental biological question, is concerned with exploiting a new mathematical framework for modelling biological populations to better understand how natural selection acts and, in particular, how it interacts with the spatial structure of the population. This will further develop our understanding of a new class of infinite-dimensional stochastic processes which have the potential for wide application.

Kingman’s coalescent describes the genealogical trees relating a sample of genes from a population in which every individual is equally fit, each is equally likely to mate with every other, and all individuals experience the same conditions. But real populations are not like this: they are structured by spatial location and genetic type. It has been customary to assume that spatially structured populations are subdivided into *demes* of (large) constant size and to model the genealogical trees relating individuals in a sample from the population using the *structured coalescent*. This corresponds to a population dynamics forwards in time given by Kimura’s stepping stone model [20].

Although the stepping stone model is widely accepted as a ‘standard’ model for structured populations, many real populations are not subdivided, but instead are spread across a spatial continuum. Wright and Malécot ([30, 24]) derived expressions for the ‘probability of identity’ (the Laplace transform of the distribution of the time since the most recent common ancestor) of two individuals sampled from a two-dimensional population by assuming, on the one hand, that genes reproduce and disperse independently of one another and, on the other, that they are scattered in a stationary Poisson distribution. However, these assumptions are incompatible ([15, 27]). The assumption of independent reproduction will result in ‘clumping’ of the population and some local regulation is required to control the local population density. Felsenstein dubbed this problem ‘the pain in the torus’. Moreover, if one models the ancestral lineages of a sample from the population as a system of Brownian motions in which each pair of lineages coalesces at a rate determined by a function of their separation ([30]), the coalescent obtained does not exhibit *sampling consistency*. That is, if we construct the genealogical tree corresponding to a sample and then examine the induced genealogical tree for a randomly chosen subsample, this will not have the same distribution as the tree that we obtain directly from the subsample. (Whenever one of the lineages in the subsample is involved in a coalescence event in the full tree it will jump.) Furthermore, there is no corresponding model for the evolution of the population forwards in time.

Mathematically consistent models for the evolution of a population distributed across a spatial continuum and the (backwards in time) genealogical trees relating individuals in a sample from the population evaded theoreticians for more than half a century. However, recently we introduced a new framework that provides a whole family of such models [12, 4, 5]. In essence, forwards in time, allele frequencies are determined by a spatial version of the generalised Fleming-Viot process, while, backwards in time, genealogies are spatial versions of multiple merger coalescents. Our work on EPSRC grants E066070/E065945 is a preliminary investigation of this framework in the context of selectively neutral populations. The concern here is to extend the approach to allow for natural selection. In addition to their biological importance (relating to a range of issues such as evolution of recombination, levels of selection, analysis of sequence data etc.), the models that arise in this way are of intrinsic mathematical interest.

Part 2: Description of Proposed Research and its Context

The principal aim of this project is to investigate the action of natural selection in spatially extended populations. The proposal has five interrelated strands:

1. the action of selection at multiple levels;
2. genic selection in a spatial environment: stochastic perturbations of the Fisher wave;
3. the evolutionary advantage of recombination in a spatially extended population;
4. genealogies under selection: extending the ancestral selection graph;
5. detecting selection from sequence data: distinguishing causes of heterogeneity along the genome.

They will be approached with a mixture of analytic techniques and simulation.

Background

Since the seminal work of Fisher ([16]) a large literature has developed on the interaction of selection with spatial structure. Traditionally, the deterministic action of gene flow and genic selection is approximated by what we now call the Fisher-KPP equation and predictions from that equation are compared to data. For example, gradients in allele frequency have a length scale σ/\sqrt{s} (with σ^2 the diffusion constant and s the selective advantage of the allele), and when gradients at multiple loci coincide, their interaction causes a characteristic steepening; such patterns lead to estimates of gene flow, selection, and numbers of genes (e.g. [28, 17]). However, many important questions depend on how selection and gene flow interact with a third force, *random genetic drift*, and this poses significant new mathematical challenges.

Random genetic drift is the stochastic fluctuation due to reproduction in a finite population. Our theoretical understanding of drift derives from several different points of view. On the one hand, drift consists of the increase in relatedness between genes in the population as one traces backwards in time. For large populations, this is generally described through Kingman’s coalescent. Alternatively, drift causes fluctuations in allele frequencies which, in a large homogeneous population, are well-approximated by a diffusion. This route, developed by Kimura ([20]), was stimulated by the neutral theory of molecular evolution. Powerful mathematical techniques allow us to identify these approaches as two sides of the same coin [11].

Even in this classical setting genetic drift interacts with natural selection in complicated ways. With a single genetic locus, forwards in time, the frequency of a single selectively favoured allele is described by a stochastic differential equation with nonlinear drift (in the mathematical sense). The effect of selection on genealogies can be introduced in two ways. When selection is very weak, genealogies can be recovered from the ancestral selection graph ([22, 10]). However, as the strength of selection increases, one rapidly sees a proliferation of ‘potential ancestors’. With strong selection, genealogies can be simulated by first conditioning on the backwards in time allele frequencies and then viewing the population as *structured* into a finite number of genetic types [19, 2, 3]. For multiple genetic loci things are more complicated. Since genes are organised on chromosomes, different genetic loci do not evolve independently, leading to competition between favoured alleles. For non-recombining genomes, this results in theoretical bounds on the rate at which populations can accumulate beneficial or deleterious mutations (e.g. [26, 9, 25, 31] and references therein). These bounds are relaxed by recombination, which allows new highly favourable combinations of alleles to be formed. Understanding how recombination may be selected in this way is a key question, which remains poorly understood ([1]).

At the root of the difficulties of introducing spatial structure into the mix is the way in which genetic drift interacts with spatial structure. For a population evolving in a spatial continuum, one would like to replace the stochastic (ordinary) differential equation that describes the frequency of an allele in an unstructured population by a stochastic partial differential equation. However, other than in one spatial dimension, the natural candidate for that stochastic pde has no solution. As a result, even for populations evolving in a continuum, it is usual to consider the population as subdivided into *demes*, sitting at the vertices of a discrete graph (a caricature of the true geography). Allele frequencies in different demes are modelled by a system of sde’s coupled through migration along the edges of the graph. The genealogical trees relating individuals in a sample from the population are determined by random walks which can coalesce (pairwise) when they are within the same deme. If we think of each deme as a ‘neighbourhood’, we are implicitly assuming that neighbourhood size is large. An immediate concern is that in a spatial continuum, neighbourhood size could be small in which case pairwise coalescence of lineages may not dominate.

When compared to data, the classical models of the way that genetic types spread through a spatially structured population fail in (at least) three ways: they cannot explain patterns in allele frequencies observed over large spatial scales; they predict much more genetic diversity than is observed; and they assume that unlinked genetic loci evolve independently of one another. In our new framework for modelling populations evolving in a continuum, these difficulties are successfully overcome.

We describe the approach in its simplest form and briefly summarise the results obtained so far. Individuals in our population are assumed to have a *type* taken from $[0, 1]$ and a spatial position in a metric space E which, because these are the most biologically relevant cases, we usually take to be \mathbb{R}^2 (or the torus $\mathbb{T}(L)$ in \mathbb{R}^2). The *spatial Λ -Fleming-Viot process*, $\{\rho(t, x, \cdot), x \in \mathbb{R}^2, t \geq 0\}$ specifies a probability measure on the type space $[0, 1]$ for every $t \geq 0$ and (almost) every $x \in \mathbb{R}^2$. The dynamics are driven by a Poisson point process Π on $\mathbb{R}_+ \times \mathbb{R}^2 \times (0, \infty)$ with intensity $dt \otimes dx \otimes \mu(dr)$. If $(t, x, r) \in \Pi$, the first component represents the time of a reproduction event. The event will affect only individuals in $B(x, r)$, the closed ball of centre x and radius r . We choose $u \in [0, 1]$ independently according to the measure $\nu_r(du)$. We also select a point z at random from $B(x, r)$ and a type k at random according to $\rho(t-, z, \cdot)$. For all $y \in B(x, r)$,

$$\rho(t, y, \cdot) = (1 - u)\rho(t-, y, \cdot) + u\delta_k.$$

Sites outside $B(x, r)$ are not affected, that is $\rho(t, y, \cdot) = \rho(t-, y, \cdot)$ for every $y \notin B(x, r)$. Of course we must impose conditions on the intensity of the Poisson point process of events in order for this process to exist ([4]). The interpretation is that if we sample an individual from x at time t , then its type will be determined by sampling from $\rho(t, x, \cdot)$.

One can obtain this process as a limit of individual based models in which, in the notation above, at each reproduction event, if $B(x, r)$ is empty nothing happens. Otherwise, a parent is chosen at random from within $B(x, r)$. Each individual in $B(x, r)$ dies (independently) with probability u . Offspring (of the type of the parent) are thrown down according to a Poisson process with intensity $u m \text{Leb}|_{B(x, r)}$ (with Leb Lebesgue measure). This prelimiting process is studied in detail in [6]. Now let $m \rightarrow \infty$. Work in progress (Etheridge & Kurtz) exploits the framework of [23] to prove simultaneous convergence of the process of allele frequencies (to the spatial Λ -Fleming-Viot process) and of the genealogical trees relating individuals in a sample from the population to a *spatial Λ -coalescent*. Ancestral lineages move according to (dependent) compound Poisson processes, jumping whenever they are affected by a reproduction event. Lineages coalesce if they are affected by the same event. In particular, since a non-trivial proportion of individuals in a neighbourhood are descended from a common parent, coalescences can involve more than two lineages.

Our approach differs from older spatial models in three key ways. First, density dependent reproduction is achieved by basing reproduction events on neighbourhoods rather than on individuals. It is this that successfully circumvents the ‘pain in the torus’. Second, the offspring of a single individual can form a significant proportion of the population in a neighbourhood about the parent, capturing the essentially finite nature of the local population size. Third, and most importantly, large-scale events, in which the reproductive success of many individuals are correlated, are explicitly incorporated. This reflects the extreme extinction and recolonisation events (caused for example by climate change) that dominate the demographic history of many species. It is these events that lead to large-scale patterns in allele frequencies and correlations in genealogies of samples from unlinked genetic loci. They also provide one possible explanation for the (several orders of magnitude) discrepancy between levels of genetic variation and that predicted from census population size and genetic drift. This is quantified in [4] for a population distributed across a large two-dimensional torus. ‘Small’ reproduction events are augmented by ‘large’ extinction-recolonisation events. Unless large events affect a non-negligible proportion of the species range, as the size of the torus grows to infinity, the genealogy of a random sample from the population will converge to Kingman’s coalescent with an ‘effective population size’ (c.f.[32]), determined by the rates of both ‘large’ and ‘small’ events. If extinction-recolonisation events affect regions of the same order as the species range, then a mathematically much richer picture emerges. Depending on the relative rates of ‘large’ and ‘small’ events, we can also obtain a (non-spatial) Λ -coalescent limit or a spatial Λ -coalescent in which, in between coalescence events, ancestral lineages follow independent Brownian motions.

We emphasize that this should be viewed as a *framework* for modelling. There are many variants of the model described above, some of which are outlined in [12]. [5] investigates a version in which, instead of replacing a portion u of the population in a disc at the time of a reproduction event, the proportion of individuals affected decays (in a Gaussian distribution) with the distance from the ‘centre’ x of the event. This allows a detailed analysis of the corresponding prelimiting model which has a Poisson stationary distribution. A comparison with the classical Malécot formula ([24]), provides a graphical demonstration of the effect of large-scale extinction-recolonisation events on patterns of genetic diversity over large spatial scales.

Of course there are many possible explanations for observed levels of genetic diversity. How could one detect large-scale extinction-recolonisation events in data? One possibility is through the decay in identity of [5]. An alternative is to exploit the fact that, whereas in the classical models loci on different chromosomes evolve independently, in our framework we see correlations in allele frequencies across unlinked loci.

To develop a statistic based on this, one must model multiple genetic loci. In our preliminary investigations, two approaches have emerged. In the first, reproduction is modelled explicitly by ‘small’ events. In the second, we assume a deterministic decay in correlations in between large scale events. Work to develop a statistical test based on either approach continues.

Research Hypothesis and Objectives

So far we have discussed the *pairwise* interactions of selection, gene flow and drift. The challenge before us is to find tractable approximations to the *three way* interaction between these evolutionary processes. Only then can we answer key questions: How fast do advantageous alleles spread? How does their spread influence linked loci, via ‘hitchhiking’? Will recombination facilitate their spread, and can this maintain high rates of recombination? Does random genetic drift lead to sufficient relatedness to allow selection to act on groups, as well as on individuals?

Our starting point will be the framework for modelling drift in a continuum described above. The first step is to introduce selection into the models. There are many different ways to do this, depending on the biological scenario that one wishes to investigate. In principle one can make all parameters of the model depend on the current configuration of genotypes in the population, but in practice we would be unlikely to learn very much by doing so. Thus, at least in the first instance, we focus on some specific forms of selection, at each stage relating the results stemming from our approach to those already available in the literature.

To mirror the selection on juveniles of [13] (see Part I) one takes the proportion, u , of the population to be replaced in an event to depend on the type of the parent. To mimic classical approaches to genic selection in the Wright-Fisher or Moran models, we allow the choice of parent to depend on type. For example, suppose that within the region affected by a reproduction event, the proportion of the population carrying a favoured allele is w , then the chance that the offspring are of this favoured type is $(1 + s)w/(1 + sw)$.

In this second model, if reproduction events affect only regions of bounded radius, then preliminary calculations show that there is a space-time rescaling (in which the parameters u and s are also suitably rescaled) under which the spatial Λ -Fleming-Viot process converges to a classical Fisher wave (in *any* spatial dimension). This is an appealing feature of the framework as it allows us to relate our work to classical theory, but more importantly it provides a way of considering stochastic perturbations of the Fisher wave in the previously problematic two-dimensional case. Indeed, this approach to adding stochastic perturbations to partial differential equations is of intrinsic mathematical interest.

Before turning to some precise questions, we mention Sewall Wright’s ingenious theory of adaptation, which allows selection to act on alternative *adaptive peaks*, corresponding to different deterministic equilibria (such as alternative chromosome arrangements). Individual selection pushes populations towards one or other peak, which may then spread by various kinds of group selection ([8]). This theory has been influential as a motivation for studying population structure and epistasis, but the component of group selection has hardly been modelled. Our framework is a natural setting in which to do so.

Programme and methodology

We now describe a selection of problems to be addressed. These are not exhaustive and evidently may change as our understanding increases. Those labelled ‘analytic’ will mostly be addressed by the PIs and the Oxford PDRA (who will be expected to have a background in mathematics). Those labelled ‘simulation’ will initially be attacked by the Edinburgh PDRA who is expected to be experienced in large scale simulations. However, it is, of course, envisaged that simulations will inform analysis and vice versa.

1. Genic selection in a spatial environment: spatial hitchhiking and stochastic perturbations of the Fisher wave.

1. (*Analysis*) First we must prove rigorously the convergence to the Fisher wave described above. We will then study the ‘perturbed’ wave, obtained by including the first order correction term as we pass to the limit. We emphasize that this makes sense in *arbitrary* spatial dimension. To a first approximation the dual system of ancestral lineages will be a branching Brownian motion with an ‘effective’ branching rate, which determines the speed of the perturbed Fisher wave (a little less than that of its deterministic counterpart). Many of the calculations required for this project already appear in [4]. In one dimension, this will be compared to the wave speed with a small Wright-Fisher noise term ([7]).

2. (*Simulation*) The small noise perturbation described above corresponds to considering populations that can live at very high density, such as laboratory populations of bacteria. Krone (e.g. [21]) has proposed alternative (lattice based) mathematical models that he investigates via simulations and compares to data from his lab. We will compare our approach to his. The Fisher wave, and an alternative strong selection limit, can also be thought of as models of range expansion and so this work also promises analytic understanding of the beautiful pictures of [18].

3. (*Analysis informed by simulation*) Because different genetic loci do not evolve independently, a neutral allele, that happens to be linked to a favourable mutation when it first arises, can receive a boost in frequency. This *genetic hitchhiking* is reasonably well understood in the non-spatial setting (see [14]). In the spatial setting the pattern of hitchhiking is much less clear.

2. Selection acting at multiple levels. (*Simulations informed by analysis*) The two forms of selection described above capture a simple advantage that one genetic type has over another. However, a key feature of adaptation is that different genotypes confer advantages in different settings. In our framework, selection can act in a multitude of different ways which might depend on both the size of the event and the configuration of types within it. There is no reason, for example, to suppose that an allele which confers a selective advantage in small-scale reproduction events should not be disadvantageous in large-scale extinction-recolonisation events. Just as for recombination, one can expect two different approaches to develop, one in which reproduction events are explicitly modelled by frequent ‘small’ events and one in which allele frequencies evolve deterministically between ‘large’ events.

We will approach the full two-dimensional model in stages. First we will deal (analytically) with the non-spatial version and relate it to existing published results. This will then be extended to an analogue of the island model in which some appreciable fraction of demes can go extinct (c.f. [29]). These results will inform a numerical investigation of the full spatial model and that in turn will inform an analytic investigation of ‘homogenisation’ in this setting. Specific issues to address include:

1. *Kin selection.* The change in allele frequencies in a population in which related individuals interact with one another depends on *Hamilton’s inclusive fitness*. This accounts for the increase in frequency of an individual’s genes both through its own reproduction and through its effects on the reproduction of its relatives. There is a longstanding debate in evolutionary biology about whether such *kin selection*, or more generally selection acting on groups, population, or species, is ever significant. The key argument against higher-level selection is that selection on individuals acts more rapidly, and will dominate. However, is that the case for our model, where a small number of parents can generate a large fraction of the population, and where such large-scale events dominate random sampling drift?

2. *Price’s equation.* Kin selection and other forms of interaction can be elegantly described using *Price’s equation*. This partitions the change in mean value of a trait into a component equal to the covariance between the fitness and the trait and a component due to ‘transmission’. The levels-of-selection debate is often quantified in terms of the Price equation, which gives the change due to selection as a sum of covariances between trait, and fitness at different levels (in our case, within and between large events). We will also develop more general interpretations of the Price equation, applying it to the propagation of genealogies rather than genes as suggested by the intriguing result in [2].

3. The evolutionary advantage of recombination in a spatially extended population. (*Analysis informed by simulation*) If chromosomes were passed down from parent to offspring as indivisible units, then two selectively advantageous mutations on the same chromosome could only become fixed in the population if the second mutation arose in an individual that already carried the first. This provides an evolutionary advantage to recombination, which allows new favourable combinations of alleles to be formed. Recurrent sweeps of favourable mutations through the whole species are much too rare to maintain costly recombination ([1]), but recurrent *local* sweeps provide a plausible source for the level of selection required. Random associations between loci, generated by population structure, may impede selection much more than in a panmictic population. We will investigate the strength of this effect, and the consequent selection for recombination.

4. Genealogies under selection: extending the ancestral selection graph. (*Analysis*) In [13], the genealogy of a sample from a non-spatial Λ -Fleming-Viot process with viability selection is recovered from a system of branching and coalescing lineages with rates determined by the probabilities of sampling different configurations of types from the stationary distribution of genetic types for the population. This distribution is not known explicitly even in the non-spatial setting and so rather than extend this approach here, we shall take a different route.

Suppose that we augment our reproduction events by ‘potential selective events’. At the time of such an event two ‘potential parents’ are chosen. If they are both the same genetic type, then we choose one at random. If they are different types then we choose the genetically favoured type to be the parent. Then, heuristically at least, it seems that backwards in time, genealogies will be determined by a spatial analogue of the ancestral selection graph (with multiple mergers). Many variants on this theme are possible. The above mechanism does not seem particularly natural, but does it provide a good approximation? In general, we will need to allow multiple branches as well as multiple mergers. Will that result in a proliferation of lineages or are the corresponding multiple mergers enough to keep the process tractable?

5. Detecting selection from sequence data: distinguishing causes of heterogeneity along the genome. (*Simulations*) Enormous effort is devoted to genome-wide scans for evidence of selection - typically, reduced diversity, excess linkage disequilibrium, or high geographic differentiation. The main problem is in distinguishing the effects of selection on particular genes from the background effect of demography, which can generate high heterogeneity across the genome. Mostly, this problem is examined using a simple model of population growth. Our model provides a more realistic framework for finding the effect of population structure in genome-wide diversity. Specifically, extinction events that replace only part of the population are expected to produce highly heterogeneous patterns of diversity along the genome.

This is just a sample of the sort of question we'd like to address. The overarching aim is to find ways to disentangle from genetic data the signals of the various demographic and genetic forces that have shaped the population.

Relevance to Academic Beneficiaries

We hope that our discussion illustrates how these models of selection in spatially extended populations could have important implications for both mathematics and biology. Our framework for modelling genetic drift is already attracting attention in the mathematical community. It provides a tractable, dimension independent, alternative to the established theories of spde driven by white noise (in one spatial dimension) and superprocesses in higher dimensions. (In fact superBrownian motion can be recovered as a rescaling limit of our model in $d \geq 2$.) Incorporating selection is the next big challenge for this rich class of measure-valued models. From a biological perspective, this project will help bring together research communities that have remained largely separate: genomic analysis, phylogeography, and social evolution. The flood of genomic data has stimulated a rapidly expanding field which fits complex stochastic models to sampled DNA sequences, with the aim of detecting genes that are under selection - in particular, those responsible for human disease. However, such models virtually all assume a well-mixed population, even though most actual examples of selection involve spatial heterogeneity. At the other extreme, a large community of biologists use genetic data to infer complex 'phylogeographic' histories - yet such analysis is almost entirely qualitative, and it is clear that such inferences are at best extremely noisy. Finally, there is intense interest in evolutionary biology in the evolution of cooperation, and of social interactions in general. The models that we propose set out a much broader framework for thinking about 'levels of selection', and our application of Price's equation to genealogies links the central theory of social evolution with the analytical framework used to analyse sequence data sampled from spatially structured populations.

As well as providing results of interest to both probabilists and evolutionists, and bringing together different theoretical traditions within evolutionary biology, our project addresses topics of wide interest: how a population's history can be discovered using genetic data, whether natural selection can be effective in realistically turbulent populations, and whether selection can act at the level of the whole population, as well as on individuals. The project will generate software for simulating selection on populations that are distributed continuously in space. These will necessarily run forwards in time, but we will also develop software for simulating sampled genealogies backwards in time, conditional on the forwards model. These will be directly applicable to data analysis. Software will be made freely available on the web. Further dissemination of results will, of course, be through publication in academic journals (biological and mathematical) and presentations at international conferences. Our complementary backgrounds give us access to very different communities which helps us enormously in disseminating our joint work.

References

- [1] N H Barton. Why sex and recombination? *Cold Spring Harbor Symposia Quant. Biol.*, 74, 2009.
- [2] N H Barton and A M Etheridge. The effect of selection on genealogies. *Genetics*, 166:1115–1131, 2004.
- [3] N H Barton, A M Etheridge, and A K Sturm. Coalescence in a random background. *Ann Appl Probab*, 14:754–785, 2004.
- [4] N H Barton, A M Etheridge, and A Véber. A new model for evolution in a spatial continuum. *Elect J Probab*, 15:162–216, 2010.
- [5] N H Barton, J Kelleher, and A M Etheridge. A new model for large-scale population dynamics: quantifying phylogeography. *Evolution*, To appear, 2010.
- [6] N Berestycki, A M Etheridge, and M Hutzenthaler. Survival, extinction and ergodicity in a spatially continuous population model. *Markov Processes and Related Fields*, 15:265–288, 2009.
- [7] E Brunet, B Derrida, A Mueller, and S Munier. Noisy traveling waves: Effect of selection on genealogies. *Physics Letters*, 76:1–7, 2006.

- [8] J A Coyne, N H Barton, and M Turelli. A critique of Wright's shifting balance theory of evolution. *Evolution*, 51:643–671, 1997.
- [9] M. M. Desai and D. S. Fisher. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics*, 176:1759–1798, 2007.
- [10] P J Donnelly and T G Kurtz. Genealogical processes for Fleming-Viot models with selection and recombination. *Ann Appl Probab*, 9:1091–1148, 1999.
- [11] P J Donnelly and T G Kurtz. Particle representations for measure-valued population models. *Ann Probab*, 27:166–205, 1999.
- [12] A M Etheridge. Drift, draft and structure: some mathematical models of evolution. *Banach Center Publ.*, 80:121–144, 2008.
- [13] A M Etheridge, R C Griffiths, and J E Taylor. A coalescent dual process in a Cannings model with genic selection and the lambda coalescent limit. *Theor Pop Biol*, To Appear, 2010.
- [14] A M Etheridge, P Pfaffelhuber, and A Wakolbinger. An approximate sampling formula under genetic hitchhiking. *Ann Appl Probab*, 16:685–729, 2006.
- [15] J Felsenstein. A pain in the torus: some difficulties with the model of isolation by distance. *Amer Nat*, 109:359–368, 1975.
- [16] R A Fisher. The wave of advance of advantageous genes. *Ann. Eugenics*, 7:353–369, 1937.
- [17] S J Goodman, N H Barton, G Swanson, K Abernethy, and J M Pemberton. Introgression through rare hybridisation: a genetic study of a hybrid zone between red and sika deer (genus *Cervus*), in Argyll, Scotland. *Genetics*, 152:355–371, 1999.
- [18] O Hallatschek, P Hersen, S Ramanathan, and D Nelson. Genetic drift at expanding frontiers promotes gene segregation. *Proc Nat Acad Sci USA*, 104(50):19926–19930, 2007.
- [19] N L Kaplan, T Darden, and R B Hudson. The coalescent process in models with selection. *Genetics*, 120:819–829, 1988.
- [20] M Kimura. Stepping stone model of population. *Ann Rep Nat Inst Genetics Japan*, 3:62–63, 1953.
- [21] S Krone, R Lu, R Fox, H Suzuki, and E Top. Modelling the spatial dynamics of plasmid transfer and persistence. *Microbiology*, 153:2803–2816, 2007.
- [22] S M Krone and C Neuhauser. Ancestral processes with selection. *Theor Pop Biol*, 51:210–237, 1997.
- [23] T G Kurtz and E Rodrigues. Poisson representations of branching Markov and measure-valued branching processes. *Ann Probab*, To appear, 2010.
- [24] G Malécot. *Les Mathématiques de l'hérédité*. Masson et Cie, Paris, 1948.
- [25] I Rouzine, E Brunet, and C O Wilke. The traveling wave approach to asexual evolution: Muller's ratchet and speed of adaptation. *Theor. Pop. Biol.*, 73:24–46, 2008.
- [26] I Rouzine, J Wakeley, and J M Coffin. The solitary wave of asexual evolution. *Proc Nat Acad Sci U.S.A.*, 100(2):587–592, 2003.
- [27] S Sawyer and J Fleischmann. The maximal geographical range of a mutant allele considered as a subtype of a Brownian branching random field. *Proc Nat Acad Sci U.S.A.*, 76(2):872–875, 1978.
- [28] J M Szymura and N H Barton. The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*: comparisons between transects and between loci. *Evolution*, 45:237–261, 1991.
- [29] J E Taylor and A Véber. Coalescent processes in subdivided populations subject to recurrent mass extinctions. *Elect J Probab*, 14:242–288, 1998.
- [30] S Wright. Isolation by distance. *Genetics*, 28:114–138, 1943.
- [31] F Yu, A Etheridge, and C Cuthbertson. Asymptotic behaviour of the rate of adaptation. *Ann Appl Probab*, To appear, 2010.
- [32] I Zähle, J T Cox, and R Durrett. The stepping stone model ii: genealogies and the infinite sites model. *Ann. Appl. Probab.*, 15:671–699, 2005.