How does geographical distance translate into genetic distance?

Verónica Miró Pina, Emmanuel Schertzer

How does geographical distance translate into genetic distance?

Verónica Miró Pina, Emmanuel Schertzer

Abstract

Geographic structure can affect patterns of genetic differentiation and speciation rates. In this article, we investigate the dynamics of genetic distances in a geographically structured metapopulation. We model the metapopulation as a weighted directed graph, with d vertices corresponding to d subpopulations that evolve according to an individual based model. The dynamics of the genetic distances is then controlled by two types of transitions -mutation and migration events. We show that, under a rare mutation - rare migration regime, intra subpopulation diversity can be neglected and our model can be approximated by a population based model. We show that under a large population - large number of loci limit, the genetic distance between two subpopulations converges to a deterministic quantity that can asymptotically be expressed in terms of the hitting time between two random walks in the metapopulation graph. Our result shows that the genetic distance between two subpopulations does not only depend on the direct migration rates between them but on the whole metapopulation structure.

1 Introduction

1.1 Genetic distances in structured populations. Speciation

In most species, the geographical range is much larger than the typical dispersal distance of its individuals. A species is usually structured into several local subpopulations with limited genetic contact. Because migration only connects neighbouring populations, more often than not, populations can only exchange genes indirectly, by reproducing with one or several intermediary populations. As a consequence, the geographical structure tends to buffer the homogenising effect of migration, and as such, it is considered to be one of the main drivers for the persistence of genetic variability within species (see **?** or **?**).

The aim of this article is to present some analytical results on the genetic composition of a species emerging from a given geographical structure. The main motivation behind this work is to study speciation. When two populations accumulate enough genetic differences, they may become reproductively isolated, and therefore considered as different species. As the geographic structure of a species is one of the main drivers for the genetic differentiation between subpopulations, this work should shed light on which are the geographic conditions under which new species can emerge.

Several authors have studied parapatric speciation, i.e. speciation in the presence of gene flow between subpopulations, for example **???** and **??**. In their models, some loci on the chromosome are responsible for reproductive isolation. These loci may be involved in incompatibilities at any level of biological organisation (molecular, physiological, behavioural etc) and either prevent mating (pre-zygotic incompatibilities) or prevent the development of hybrids (post-zygotic incompatibilities). The number of segregating loci increases through the accumulation of mutations, and decreases after each migration event (creating the opportunity for some gene exchange between the migrants and the host population). When the number of segregating loci between two individuals reaches a certain threshold, they become reproductively incompatible. For example, **??** studied the case of a metapopulation containing two homogeneous subpopulations.

The authors studied how the genetic distance, defined as the number of loci differing between the two subpopulations, evolves through time, using a continuous-time model. When considering metapopulations with more than two subpopulations, this kind of dynamics may translate into complex patterns of speciation. One particularly intriguing example is the case of ring species (**??**), where two neighbouring subpopulations are too different to be able to reproduce with one another but can exchange genes indirectly, by reproducing with a series of intermediate subpopulations that form a geographic 'ring'. How these patterns emerge and are maintained is still poorly understood, and we hope that our analytical result might shed some new light on the subject.

1.2 Population divergence and fitness landscapes

To study speciation by accumulation of genetic differences, we model the evolution of some loci on the chromosome, that are potentially involved in reproductive incompatibilities. To visualise these evolutionary dynamics, **?** suggested the metaphor of adaptive landscapes. Adaptive landscapes represent individual fitness as a function defined on the genotype space, which is a multi-dimensional space representing all possible genotypes. Wright emphasised the idea of 'rugged' adaptive landscapes, with peaks of fitness representing species and valleys representing unfit hybrids. Speciation, seen as a population moving from one peak to another, implies a temporary reduction in fitness, which is not very likely to occur in large populations, where genetic drift is not important enough to counterbalance the effect of selection (see **?** for a more detailed discussion). However, **?** suggested the idea of 'holey' adaptive landscapes, where local fitness maxima can be partitioned into connected sets (called evolutionary ridges). Speciation is therefore seen as a population diffusing across a ridge, by neutral mutation steps, until it stands at the other side of a hole. Theoretical models, such as **?**, have shown, using percolation theory, that in high-dimensional genotype spaces, fit genotypes are typically connected by evolutionary ridges.

Our model (see Section 1.3) is built in this framework. In fact we will assume that, in large populations, deleterious mutations are washed away by selection at the micro-evolutionary timescale and describe the evolutionary dynamics for our set of incompatibility controlling loci as *neutral* (any genotype on the evolutionary ridge can be accessed by single mutation neutral steps). This is the idea behind the description of our model in Section 1.3.

Further, we consider that the evolutionary dynamics along the ridge are slow (as random mutations are very likely to be deleterious, mutations along the evolutionary ridge are assumed to be rarer than in the typical population genetics framework), which is why we study our model in a low mutation - low migration regime (see Section 1.4 for more details). This assumption is commonly made when studying speciation, for example in **?** or **?**.

1.3 An individual based model (IBM)

We model the metapopulation as a weighted directed graph with d vertices, corresponding to the different subpopulations. Each directed edge (i, j) is equipped with a migration rate in each direction. (In particular, if two subpopulations are not connected, we assume that the migration rates are equal to 0.) We assume the existence of two scaling parameters, γ and ϵ, that will converge to 0 successively (first $\gamma \to 0$ and then $\epsilon \to 0$, see Section 1.4 for more details).

Each subpopulation consists of n_i^ϵ individuals, $i \in E := \{1, \dots, d\}$. Each individual carries a single chromosome of length 1, which contains l^ϵ loci of interest (that are involved in reproductive incompatibilities). We assume that the vector of positions for those loci – denoted by $\mathcal{L}^\epsilon = \{x_1, \dots, x_{l^\epsilon}\}$ – is obtained by throwing l^ϵ uniform random variables on $[0, 1]$. The positions are chosen randomly at time 0, but are the same for all individuals and do not change through time. Recall that the upper indices (such as in l^ϵ and $\mathcal{L}^\epsilon$) are indices and not exponents).

Conditioned on $\mathcal{L}^\epsilon$, each subpopulation then evolves according to an haploid neutral Moran model with recombination.

- Each individual x reproduces at constant rate 1 and chooses a random partner y ($y \neq x$). Upon reproduction, their offspring replaces a randomly chosen individual in the population.
- The new individual inherits a chromosome which is a mixture of the parental chromosomes. Both parental chromosomes are cut into fragments in the following way: we assume a Poisson Point Process of intensity λ on $[0,1]$. Two loci belong to the same fragment iff there is no atom of the Poisson Point Process between them. For each fragment, the offspring inherits the fragment of one of the two parents chosen randomly.

To our Moran model we add two other types of events:

- **Mutation** occurs at rate $b^{\gamma,\epsilon}$ per individual, per locus according to an infinite allele model.
- **Migration** from subpopulation i to subpopulation j occurs at rate m_{ij}^γ. At each migration event, one individual migrates from subpopulation i to j, and replaces one individual chosen uniformly at random in the resident population. (We set $\forall i \in E,\ m_{ii}^\gamma = 0$.)

We define the genetic distance between two individuals x and y at time t as:

$$\delta_t^{\gamma,\epsilon}(x,y) = \frac{1}{l^\epsilon}\#\{\ k \in \{1,\dots,l^\epsilon\} :\ x \text{ and } y \text{ differ at locus } k\ \}.$$

Consider two subpopulations i and j and let $\{i_1,\dots,i_{n_i^\epsilon}\}$ be the individuals in population i and $\{j_1,\dots,j_{n_j^\epsilon}\}$ the individuals in population j. The genetic distance between subpopulations i and j at time t is defined as follows:

$$d_t^{\epsilon,\gamma}(i,j) = \left(\frac{1}{n_i^\epsilon}\sum_{x\in\{i_1,\dots,i_{n_i^\epsilon}\}}\min_{y\in\{j_1,\dots,j_{n_j^\epsilon}\}}\delta_t^{\gamma,\epsilon}(x,y)\right) \vee \left(\frac{1}{n_j^\epsilon}\sum_{y\in\{j_1,\dots,j_{n_j^\epsilon}\}}\min_{x\in\{i_1,\dots,i_{n_i^\epsilon}\}}\delta_t^{\gamma,\epsilon}(x,y)\right). \quad (1)$$

This corresponds to the so-called modified Haussdorff distance between subpopulations, as introduced by **?**. (This distance has the advantage of averaging over the individuals in each subpopulation, so introducing a single mutant or migrant would produce a smooth variation in the genetic distances.)

1.4 Slow mutation–migration and large population–dense site regime.

In this section, we start by describing in more details the slow mutation–migration regime alluded to in Sections 1.2 and 1.3.

It is well known that in the absence of mutation and migration, the neutral Moran model describing the dynamics at the local level reaches fixation in finite time i.e. after a finite amount of time the population becomes homogeneous. The average time to fixation for a single locus is of the order of the size of the subpopulation (**??**) (In our multi-locus model, it will also depend on the number of loci and on the recombination rate λ.) Heuristically, if we assume a low mutation - low migration regime, i.e. that

$$\forall i,j \in E, \quad \frac{1}{b^{\gamma,\epsilon}n_i^\epsilon},\ \frac{1}{m_{ij}^\gamma} \ \gg\ n_j^\epsilon,\ l^\epsilon \gg 1, \quad (2)$$

the average time between two migration events ($1/m_{ij}^\gamma$), and the average time between two successive mutations at a given locus ($1/(b^{\gamma,\epsilon}n_j^\epsilon)$) are much larger than the average time to fixation. This ensures that the fixation process is fast compared to the time-scale of mutation and migration, and, as a result, when looking at a randomly chosen locus, subpopulations are

homogeneous except for short periods of time right after a migration event or a mutation event. This suggests that if we accelerate time properly, we can neglect intra-subpopulation diversity and approximate our model by a population based model.

Inspired by these heuristics, we are going to take a low mutation - low migration regime, by making the mutation and migration rates depend on the scaling parameter γ in the following way:

$$\begin{aligned} m_{ij}^{\gamma} &= \gamma M_{ij} \quad \text{where } M_{ij} \geq 0 \text{ is a constant} \\ b^{\gamma,\epsilon} &= \gamma\epsilon\, b_\infty \quad \text{where } b_\infty > 0 \text{ is a constant.} \end{aligned}$$

Recall that we take a slow mutation-migration regime but the recombination rate is constant. This is consistent with the fact that in most species, mutation rates are very low (they vary from 10^{-6} to 10^{-8} per base per generation) compared to the recombination rates (for example, for the human genome there are approximatively 66 crossovers per generation).

In a second step, we will make an additional approximation: we will consider a large population - dense site limit. In fact, our second scaling parameter ϵ, corresponds to the inverse of a typical subpopulation size. The parameters of the model depend on ϵ in the following way (corresponding to the second inequality in (2):

$$\begin{aligned} &n_i^{\epsilon} = [N_i/\epsilon] \quad \text{where } N_i > 0 \text{ remains constant as } \epsilon \to 0 \\ &l^{\epsilon} \to \infty \ \text{ as } \epsilon \to 0 \end{aligned}$$

In this article, we are going to take the limits successively: first $\gamma \to 0$ and then $\epsilon \to 0$, in order to be consistent with the informal inequality (2). We are now ready to state the main result of this paper.

Theorem 1.1. *For each pair of subpopulations $i, j \in E$, let S^i and S^j be two independent random walks on E starting respectively from i and j and whose transition rate from k to p is equal to $\tilde{M}_{kp} := M_{pk}/N_k$. Finally, define $D_t(i,j)$ as*

$$\forall t \geq 0, \quad D_t(i,j) \;=\; 1 \;-\; \int_0^t e^{-2b_\infty s}\mathbb{P}(\tau_{ij} \in ds) \;-\; e^{-2b_\infty t}\mathbb{P}\left(\tau_{ij} > t\right),$$

where $\tau_{ij} = \inf\{t \geq 0 \ : \ S^i(t) = S^j(t)\}$.

If at time 0 the metapopulation is homogeneous (i.e. all the individuals in all subpopulations share the same genotype) then

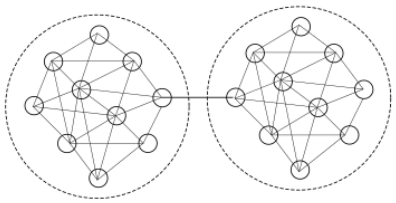
$\lim_{\epsilon\to 0}\lim_{\gamma\to 0}(d^{\gamma,\epsilon}_{t/(\gamma\epsilon)}(i,j),\ t \geq 0) \;=\; (D_t(i,j),\ t \geq 0)$ *in the sense of finite dimensional distributions (f.d.d.).*

In particular,

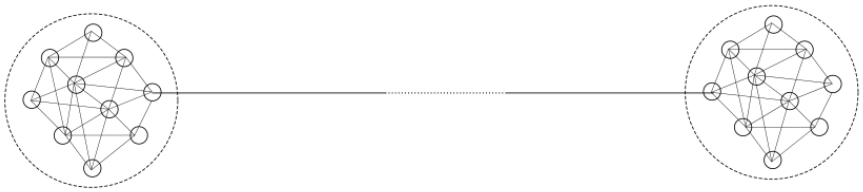
$$\lim_{t\to\infty} D_t(i,j) \;=\; 1 - \mathbb{E}(e^{-2b_\infty \tau_{ij}}).$$

This result can be seen as a law of large numbers over the chromosome. Although the loci are linked and they do not fix independently (and the recombination rate is constant), when considering a large number of them, they become decorrelated, regardless of the value of λ. (Note that the limiting process does not depend on λ.) The model behaves as if infinitely many loci evolved independently according to a Moran model with inhomogeneous reproduction rates (see Remark 2.4). The expression of the genetic distances has then a natural genealogical interpretation. S^i and S^j can be interpreted as the ancestral lineages starting from i and j, and our genetic distance is related to the probability that those lines meet before experiencing a mutation (or in other words, that i and j are Identical By Descent (IBD)).

Remark 1.2. *In Theorem 1.1, we considered a rather restrictive initial condition. In Section 5, we give a stronger version of this theorem, which works for a larger range of initial conditions, but that requires to introduce several cumbersome notations.*



(a) Geographic distances



(b) Genetic distances

Figure 1: Amplification of a geographic bottleneck in the genetic distance metrics (small value of c in Theorem 6.1). In this example, the metapopulation is formed of two complete graphs (not all edges are represented), connected by a single edge (a). If i and j are connected, $M_{ij} = 1/d$. In (a) all the edges are the same length. In (b), the genetic distances between pairs of vertices belonging to the same subgraph are smaller than the genetic distances between pairs of vertices belonging to different subgraphs.

1.5 Consequences of our result

One interesting consequence of our result is that the genetic distance does not coincide with the classical graph distance, but instead it depends on all possible paths between i and j in the graph, and all the migration rates (and not only the shortest path and the direct migration rates M_{ij} and M_{ji}), i.e., it does not only depend on the direct gene flow between i and j but on the whole metapopulation structure. In particular, this suggests that adding new subpopulations to the graph (which would correspond to colonisation of new demes), removing any edge (which could correspond to the emergence of a geographical or reproductive barrier between two subpopulations), or changing any migration rate (which could correspond to modifying the habitat structure, for example) can potentially modify the whole genetic structure of the population.

One striking illustration of the previous discussion is presented in Section 6, where we consider an example where a geographic bottleneck is dramatically amplified in our new metric. See Figure 1 and Theorem 6.1 for a more precise statement. If we consider, as in **?**, that two populations are different species if their genetic distance reaches a certain threshold, that will mean that this metapopulation structure promotes the emergence of two different species, each one corresponding to the population in one subgraph. Very often, parapatric speciation is believed to occur only in the presence of reduced gene flow. Our example shows that in the presence of a geographic bottleneck, genetic differentiation is manly driven by the geographical structure of the population, i.e., even if the gene flow between two neighbouring subpopulations is approximatively identical in the graph, the genetic distance is dramatically amplified at the bottleneck (see Figure 1).

We note that using the hitting time of random walks as a metric on graphs is not new, and has been a popular tool in graph analysis (see **?** and **?**). For example, the commute distance, which is the time it takes a random walk to travel from vertex i to j and back, is commonly used in many fields such as machine learning (**?**), clustering (**?**), social network analysis (**?**), image processing (**?**) or drug design (**??**). In our case the genetic distance is given by the

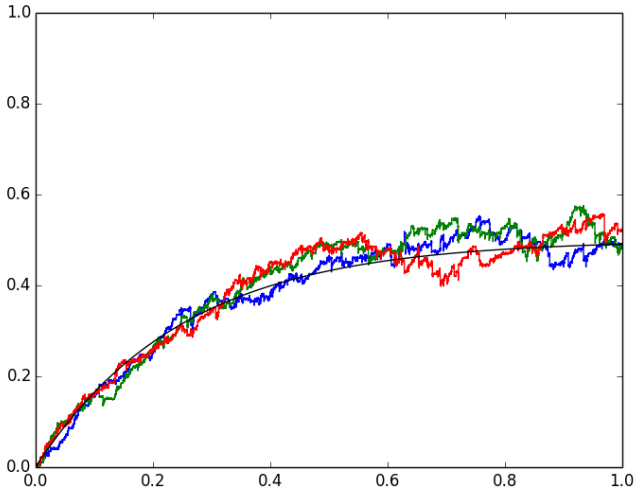


Figure 2: Simulation of the individual based model, for $d = 3$, $N_1 = N_2 = N_3 = 1$, $\epsilon = 0.005$, $\gamma = 2e^{-6}$, $l^\epsilon = 100$, $\lambda = 10$. The black curve corresponds to $D_t(i,j)$ (see Theorem 1.1). The blue, green and red curves correspond to the three genetic distances $d_t^{\epsilon,\gamma}(1,2), d_t^{\epsilon,\gamma}(2,3), d_t^{\epsilon,\gamma}(1,3)$.

Laplace transform of the hitting time between two random walks, which was already suggested as a metric on graphs by **?**. In that paper the authors claimed that this metrics preserves the cluster structure of the graph. In the example alluded to above (Section 6), we found that our metric reinforces the cluster structure of the metapopulation graph. In other words, a clustered geographic structure tends to increase genetic differentiation.

1.6 Discussion and open problems

As already mentioned above, the main result is obtained by: (i) proving that, in a low mutation - low migration regime (i.e., when $\gamma \to 0$), subpopulations are monomorphic most of the time and our individual based model converges to a population based model, (ii) showing that, under a large population - dense site limit (i.e. taking $\epsilon \to 0$), the genetic distances between subpopulations (for the population based model) converge to a deterministic process (defined in Theorem 1.1). Taking these two limits successively gives no clue on how the parameters should be compared to ensure the approximation to be correct. It would be interesting to take the limits simultaneously but it is technically challenging (for example we would need to characterise the time to fixation for l loci that do not fix independently, which is not easy).

As discussed in the previous paragraph, we can only show our results under some rather drastic constraints: subpopulations are asymptotically monomorphic. More generally, we believe that Theorem 1.1 should hold under relaxed assumptions, namely when the intra-subpopulation genetic diversity is low compared to the inter-subpopulation diversity (see Figure 2 for an example, where $\gamma = 2e^{-6}$ and $\epsilon = 5e^{-3}$). Technically, this would correspond to the condition that at a typical locus (i.e, a locus chosen uniformly at random) each subpopulations is monomorphic at that site with high probability (which is in essence (2)). Of course, proving such a result would be much more challenging, but would presumably correspond to a more realistic situation.

1.7 Outline

In Section 2, we show that in the rare mutation-rare migration regime (i.e. when $\gamma \to 0$ whereas ϵ remains constant), the individual based model (IBM) described above converges to a population

based model (PBM) (see Theorem 2.2). This PBM is a generalization of the model proposed by **??** in three ways. First, it is an extension of their model from two to an arbitrary number of subpopulations, which is not trivial from a mathematical point of view. Second, in **??**, the authors only assumed that the migrant alleles are fixed independently at every locus. To make the model more realistic, we took into account genetic linkage, which introduces a non-trivial spatial correlation between loci (along the chromosome). Finally, we suppose that the loci are distributed randomly along the chromosome (and not in a regular fashion). Section 2 is interesting on its own since it provides a theoretical justification of the model proposed by **??**.

In Section 3 and 4, we study the PBM in the large population - dense site limit (i.e. when $\epsilon \to 0$). We properly introduce the main tool used to study the population based model – the genetic partition probability measure – and show an ergodic theorem related to this process (see Theorem 3.1).

Finally, in Section 5 we prove our main result (Theorem 5.1 which is an extension of Theorem 1.1) by combining the results of the previous sections.

Section 6 proves the result related to the geographical bottleneck alluded to in Section 1.5 (see Proposition 6).

2 Approximation by a population based model

We now describe a population based model (PBM) that can be seen as the limit of the IBM presented above, when γ goes to 0 (whereas ϵ remains fixed) and time is rescaled by $1/(\gamma\epsilon)$. Consider a metapopulation where the individuals are characterised by a finite set of loci, whose positions are distributed as l^ϵ uniform random variables on $[0,1]$, and let $\mathcal{L}^\epsilon$ be the vector of the positions of the loci (as described in Section 1.3 for the IBM). We now describe the dynamics of the model, conditional on $\mathcal{L}^\epsilon = L^\epsilon$, with $L^\epsilon \in [0,1]^{l^\epsilon}$.

Before going into the description of our model, we start with a definition. It is well known that the Moran model reaches fixation in finite time, i.e., after a (random) finite time, every individual in the population carries the same genetic material, and from that time on, the system remains trapped in this configuration (see **?**, **?**).

Definition 2.1. *Consider a single population of size n_j^ϵ formed by a mutant individual (the migrant) and $n_j^\epsilon - 1$ residents, that evolves according to a Moran model with recombination at rate λ (as described in Section 1.3). We define $\mathcal{F}_j^{L^\epsilon,\lambda}$ as the (random) set of loci carrying the mutant type when the population becomes homogeneous. (Note that $\mathcal{F}_j^{L^\epsilon,\lambda}$ is potentially empty.)*

We are now ready to describe our PBM. We represent each subpopulation as a single chromosome, which is itself represented by the set of loci L^ϵ. The dynamics of the population can then be described as follows.

- For every $i \in E$: fix a new mutation in population i at rate $b_\infty l^\epsilon$, the locus being chosen uniformly at random along the chromosome.
- For every $i, j \in E$ and every $S \subseteq \{1, \dots, l^\epsilon\}$: at every locus in S, fix simultaneously the alleles from population i in population j at rate $\frac{1}{\epsilon} M_{ij} \mathbb{P}\left(\mathcal{F}_j^{L^\epsilon,\lambda} = S\right)$.

In the PBM (parametrised by ϵ), we define the genetic distance between subpopulations i and j at time t as follows:

$$d_t^\epsilon(i,j) \;=\; \frac{1}{l^\epsilon} \#\{\; k \in \{1, \dots, l^\epsilon\} \;:\; \text{subpopulations } i \text{ and } j \text{ differ at locus } k \text{ at time } t \;\}$$

as opposed to $d^{\gamma,\epsilon}$ which will refer to the genetic distances in the IBM as described in Section 1.3 (parametrised by γ and ϵ). We note that the definition of the genetic distance in the PBM

is consistent with the one in the IBM (see (1)) in the sense that if the subpopulations are homogeneous in the IBM, (1) is equal to the RHS of the previous equation. We are now ready to state the main result of this section.

Theorem 2.2. *Assume that, at time* 0*, the subpopulations in the IBM are homogeneous and that* $\forall i, j \in E$, $d_0^{\gamma,\epsilon}(i,j) = d_0^{\epsilon}(i,j)$. *Then, for every* $k \in \mathbb{N}$, $\forall\ 0 \leq t_1 < \cdots < t_k$,

$$\lim_{\gamma \to 0} \left(d^{\gamma,\epsilon}_{t_1/(\gamma\epsilon)}, \ldots, d^{\gamma,\epsilon}_{t_k/(\gamma\epsilon)}\right) \ = \ \left(d^{\epsilon}_{t_1}, \ldots, d^{\epsilon}_{t_k}\right) \ \textit{in distribution.} \tag{3}$$

Proof. Recall that the loci are distributed randomly along the chromosome. In the proof, we assume that the vector of the positions of the loci $\mathcal{L}^\epsilon$ is fixed and equals to $L^\epsilon \in [0,1]^{l^\epsilon}$ (and is the same in the IBM and in the PBM). We also consider that IBM and the PBM start from the same deterministic initial condition. The unconditional extension of the proof can be easily deduced from there.

We define a coupling between the IBM and a new PBM that is close (in distribution) to the PBM defined at the beginning of this section. The idea behind the coupling is that, when time is accelerated by $1/(\gamma\epsilon)$, and γ is small, in the IBM, the time to fixation after a mutation or migration event is short enough so that the population has become homogeneous before the next mutation or migration event takes place. Then, we can decompose the trajectories of the IBM into periods where the population is homogeneous (and waits for the next mutation or migration event to take place) and homogenization phases (where the dynamics of the population is described by a Moran model). See Figure 3 for an illustration of this concept.

More formally, let us consider $(Y_t^{\gamma,\epsilon}; t \geq 0)$ the process recording the genetic composition in the IBM (i.e. a matrix containing the sequences of the chromosomes of all the individuals in the metapopulation) *after rescaling time by* $\gamma\epsilon$ so that

1. For $i \in E$, mutation events on the subpopulation i occurs according to a Poisson Point Process (PPP) with intensity measure $b_\infty l^\epsilon n_i^\epsilon dt$.

2. For $i, j \in E$, migration events from i to j can be described in terms of a PPP with intensity measure $M_{ij}/\epsilon dt$.

Define $\mathcal{E}^{\gamma,\epsilon}$ the event that every time a subpopulation is affected by a mutation or a migration event on the interval $[0,T]$, the subpopulation is genetically homogeneous when the event occurs (as in Figure 3). In other words, there is no overlap between mutation and migration homogenization periods. The time to fixation in our (multi-locus) Moran model only depends on the number of individuals and the number of loci, so in our model it only depends on ϵ (but not on γ). As a consequence, $\mathbb{P}\left(\mathcal{E}^{\gamma,\epsilon}\right) \to 1$ as $\gamma \to 0$.

Next, let us consider $T_t^{\gamma,\epsilon}$ as the Lebesgue measure of

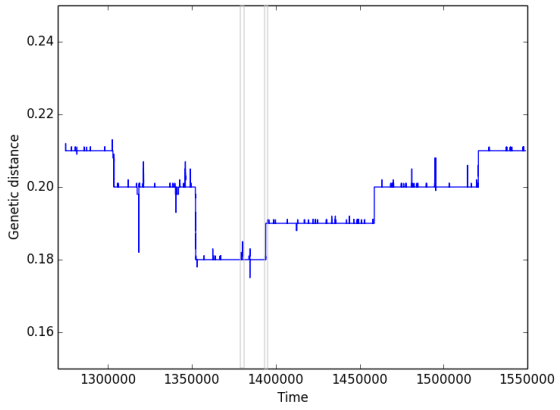
$$\{0 \leq s \leq t \ : \ \forall i \in \ E \text{ pop. } i \text{ is homogeneous at time } s\}.$$

In words, $(T_t^{\gamma,\epsilon}; t \geq 0)$ is the random clock which is obtained by skipping the homogenization period after a migration or mutation event (i.e. by skipping the red intervals in Figure 3). By arguing as in the previous paragraph, as $\gamma \to 0$, it is not hard to see that $(T_t^{\gamma,\epsilon}; t \geq 0)$ converges to the identity in the Skorokhod topology on $[0,T]$ for every $T \geq 0$.

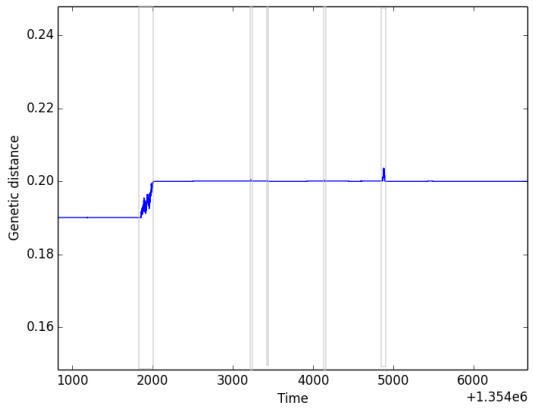
Let us now consider

$$Z_t^{\gamma,\epsilon} \ = \ Y^{\gamma,\epsilon}_{(T^{\gamma,\epsilon})_t^{-1}}, \text{ where } \ (T^{\gamma,\epsilon})_t^{-1} \ = \ \inf\{s \geq 0 \ : \ T_s^{\gamma,\epsilon} \geq t\}.$$

By construction, this process defines a PBM in the sense that at every time t, any subpopulation is composed by genetically homogeneous individuals. Further, since $(T_t^{\gamma,\epsilon}; t \geq 0)$ converges to



(a)



(b)

Figure 3: (a) Simulation of the individual based model, for $d = 3$, $N_1 = N_2 = N_3 = 1$, $\epsilon = 0.005$, $\gamma = 2e^{-6}$, $l^{\epsilon} = 100$, $\lambda = 10$. It is the same simulation as in Figure 2 (without rescaling time). Two examples of an homogenization phases are contained in the grey boxes. (b) Partial zoom on the curve.

the identity and mutation and migration events occur at Poisson times, the finite dimensional distributions of $Z^{\gamma,\epsilon}$ are good approximations of the ones for the IBM.

Let us now show that $Z^{\gamma,\epsilon}$ (constructed from the IBM) is close in distribution to the PBM defined at the begining of this section. Conditioned on the event $\mathcal{E}^{\gamma,\epsilon}$ (whose probability goes to 1) and on the PPP's described in 1 and 2 above, the PBM $Z^{\gamma,\epsilon}$ can be described as follows. Define $p_{\Delta t,i}$ to be the probability for a mutant allele (at a given locus of a given individual) to fix in a population of size n_i^ϵ, *conditioned on the homogenization time to be smaller than* Δt. Then the distribution of the conditioned PBM $Z_t^{\gamma,\epsilon}$ can be generated as follows.

(a) At every mutation time t in subpopulation $i \in E$, choose a locus k uniformly at random and fix the mutation instantaneously with probability $p_{\frac{\Delta t}{\gamma\epsilon},i}$, where Δt is the time between t and the next mutation or migration event (in our new time scale). We note that if the mutation does not fix, then $Z^{\gamma,\epsilon}$ is not affected by the mutation event, and as a consequence "effective mutation" events in $Z^{\gamma,\epsilon}$ are obtained from the mutation events in the IBM after thinning each time with their respective probability $p_{\frac{\Delta t}{\gamma\epsilon},i}$.

(b) At every migration event t on subpopulation j, fix a random set S where S is chosen according to $\mathcal{F}_j^{L^\epsilon,\lambda,\Delta t/(\gamma\epsilon)}$, where Δt is defined as in the previous point, and where $\mathcal{F}_j^{L^\epsilon,\lambda,s}$ is the random variable $\mathcal{F}_j^{L^\epsilon,\lambda}$ conditioned on the homogenization to occur in a time smaller smaller than s.

Since fixation (of one of the alleles) occurs in finite time almost surely, and the distribution of the homogenization time only depends on ϵ, we have

$$\mathcal{F}_j^{L^\epsilon,\lambda,\Delta t/(\gamma\epsilon)} \underset{\gamma\to 0}{\Longrightarrow} \mathcal{F}_j^{L^\epsilon,\lambda}, \text{ and } \lim_{\gamma\to 0} p_{\Delta t/(\gamma\epsilon),i} = \frac{1}{n_j^\epsilon}$$

where the RHS of the second limit is the probability of fixation of a mutant allele in the absence of conditioning.

Putting all the previous observations together, one can easily show that the genetic distance in $Z^{\gamma,\epsilon}$ converges (in the finite dimensional distributions sense) to the ones of the PBM. In particular, we recover the mutation rate on subpopulation i in the PBM

$$\underbrace{b_\infty l^\epsilon n_i^\epsilon}_{\text{rate of mutation in the IBM}} \times \underbrace{\frac{1}{n_i^\epsilon}}_{\text{proba of fixation}} = b_\infty l^\epsilon,$$

which corresponds to the limiting "effective mutation rate" in the PBM $Z^{\gamma,\epsilon}$ (see (a) above) as $\gamma \to 0$. This completes the proof of Theorem 2.2. □

We note that Theorem 2.2 could be extended to the case where the subpopulations are not homogeneous in the IBM at time $t = 0$. Indeed, arguing as in the proof of Proposition 2.2, if we start with some non-homogeneous initial condition, then each island becomes homogeneous before experiencing any mutation or mutation event with very high probability. In order to get an efficient coupling between the IBM and the PBM, we simply choose the initial condition of the PBM as the (random) state of the IBM after this initial homogenization period.

Remark 2.3. *Choose a locus $k \in \{1, \dots, l^\epsilon\}$. We let the reader convince herself that in the PBM the genetic composition at a given locus k follows the following Moran-type dynamics:*

(mutation) "Individual" i takes on a new type (or allele) at rate b_∞.

(reproduction) *"Individual" j inherits its type from "individual" i at rate $(1/\epsilon)M_{ij}\mathbb{P}\left(k \in \mathcal{F}_j^{L^\epsilon,\lambda}\right)$. Further, in a neutral one-locus Moran model, the probability of fixation of a single allele in a resident population of size n_j^ϵ is equal to its initial frequency, which in our case is $1/n_j^\epsilon$. Thus*

$$\frac{1}{\epsilon}M_{ij}\mathbb{P}\left(k \in \mathcal{F}_j^{L^\epsilon,\lambda}\right) \;=\; \frac{1}{\epsilon}\; M_{ij}/n_j^\epsilon.$$

This dynamics is not dependent on the position of the locus under consideration.

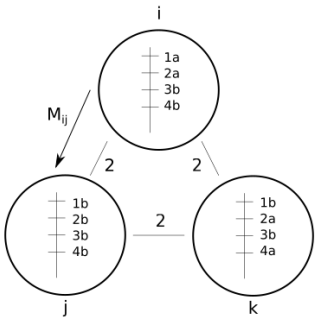
Remark 2.4. *Our model can be seen as a multi-locus Moran model with inhomogeneous reproduction rates. The main difficulty in analysing this model stems from the fact that there exists a non trivial correlation between loci. This correlation is induced by the fact that fixation of migrant alleles can occur simultaneously at several loci during a given migration event. In turn, the set of fixed alleles during a given migration event is determined by the local Moran dynamics described in the Introduction.*

3 Large population - dense site limit

In this Section, we study the PBM described in the Section 2, in the large population - dene site limit. In particular, we study the dynamics of the genetic distances and we state Theorem 3.1, that together with Theorem 2.2 implies the main result of this article, namely Theorem 5.1, that is a stronger version of Theorem 1.1 (see Section 5).

3.1 The genetic partition measure

The main difficulty in dealing with the genetic distance is that it lacks the Markov property, and as a consequence, it is not directly amenable to analysis. In fact, when $d > 2$, a migration event from i to j can potentially have an effect on the genetic distance between j and another subpopulation k (see Figure 4 for an example).



Locus fixed	$d^\epsilon(i,j)$	$d^\epsilon(i,k)$	$d^\epsilon(k,j)$
1	1	2	3
2	1	2	1
3	2	2	2
4	2	2	2

Figure 4: The three subpopulations (i,j,k) are characterised by a chromosome with four loci (1,2,3,4), with different alleles (1a, 1b, ...). The three genetic distances $(d^\epsilon(i,j), d^\epsilon(i,k), d^\epsilon(k,j))$ are equal to two (before migration). The table shows the new genetic distances after a migration event from i to j where one locus from i is fixed in population j. At locus 1, the allelic partition $\Pi_1^\epsilon(t)$ is equal to $\{i\}\{j,k\}$, whereas at locus 4, $\Pi_4^\epsilon(t) = \{i,j\}\{k\}$.

To circumvent this difficulty, we now introduce an auxiliary process – the genetic partition probability measure – from which one can easily recover the genetic distances (see (5) below), and whose asymptotical dynamics is explicitly characterised in Theorem 3.1 below.

Let $\mathcal{P}_d$ be the set of partitions of $\{1,\dots,d\}$. Fix $\pi \in \mathcal{P}_d$ and $i,j \in E$. Define $\mathcal{S}_i(\pi)$ as the element of $\mathcal{P}_d$ obtained from π by making i a singleton (e.g., $\mathcal{S}_2(\{1,2,3\}) = \{1,3\}\{2\}$). Define

$\mathcal{I}_{ij}(\pi)$ as the element of $\mathcal{P}_d$ obtained from π by displacing j into the block containing i (e.g., $\mathcal{I}_{2,3}(\{1,3\}\{2\}) = \{1\}\{2,3\}$).

At every locus $k \in \{1,\dots,l^\epsilon\}$ (ordered in increasing order along the chromosome) and every time t, the allele composition of the metapopulation induces a partition on E. More precisely, at locus k, two subpopulations are in the same block of the partition at time t iff they share the same allele at locus k.

In the following, fix $L^\epsilon \in [0,1]^{l^\epsilon}$, the vector containing the positions of the loci. In the PBM parametrised by ϵ, we condition on the loci being located at L^ϵ and for every $k \in \{1,\dots,l^\epsilon\}$ we let $\Pi_k^{\epsilon,L^\epsilon}(t)$ be the partition induced at locus k (see Figure 4). The vector $\Pi^{\epsilon,L^\epsilon}(t) = \left(\Pi_k^{\epsilon,L^\epsilon}(t);\ k \in \{1,\dots,l^\epsilon\}\right)$ describes the genetic composition of the population at time t. According to the description of our dynamics, $\Pi^{\epsilon,L^\epsilon}(t)$ is a Markov chain with the following transition rates:

- (mutation) For every $i \in E$, $k \in \{1,\dots,l^\epsilon\}$, define $\mathcal{S}_i^k$ to be the operator on $(\mathcal{P}_d)^{l^\epsilon}$ such that $\forall\ \Pi\ \in (\mathcal{P}_d)^{l^\epsilon}$

$$\mathcal{S}_i^k(\Pi) = (\tilde{\Pi}_1,\dots,\tilde{\Pi}_{l^\epsilon}) \ \text{ with } \begin{cases} \tilde{\Pi}_j = \Pi_j & \forall j \neq k \\ \tilde{\Pi}_k = \mathcal{S}_i(\Pi_j) & j = k \end{cases}.$$

  The transition rate of the process from state Π to $\mathcal{S}_i^k(\Pi)$ is given by b_∞.

- (migration from i to j) For every $i,j \in E$, $S \subset \{1,\dots,l^\epsilon\}$, define $\mathcal{I}_{ij}^S$ the operator on $(\mathcal{P}_d)^{l^\epsilon}$ such that $\forall\ \Pi\ \in (\mathcal{P}_d)^{l^\epsilon}$

$$\mathcal{I}_{ij}^S(\Pi) = (\tilde{\Pi}_1,\dots,\tilde{\Pi}_{l^\epsilon}) \ \text{ with } \begin{cases} \tilde{\Pi}_j = \Pi_j & \forall k \notin S \\ \tilde{\Pi}_k = \mathcal{I}_{ij}(\Pi_k) & \forall k \in S \end{cases}.$$

  The transition rate of the process from Π to $\mathcal{I}_{ij}^S(\Pi)$ is given by $\frac{M_{ij}}{\epsilon}\mathbb{P}\left(\mathcal{F}_j^{L^\epsilon,\lambda} = S\right)$.

To summarise, for any function $h : \mathcal{P}_d^{l^\epsilon} \to \mathbb{R}$, the generator of $\Pi^{\epsilon,L^\epsilon}$ can be written as

$$\begin{aligned} \mathbb{G}^{\epsilon,L^\epsilon}h(\Pi) \ &= \ \frac{1}{\epsilon}\sum_{i,j=1}^{d} M_{ij} \sum_{S\subset\{1,\dots,l^\epsilon\}} \mathbb{P}(\mathcal{F}_j^{L^\epsilon,\lambda} = S)\ [h(\mathcal{I}_{ij}^S(\Pi)) - h(\Pi)] \ + \\ & \qquad b_\infty \sum_{i=1}^{d}\sum_{k=1}^{l^\epsilon} \left(h(\mathcal{S}_i^k(\Pi)) - h(\Pi)\right). \end{aligned} \tag{4}$$

3.2 Some notation

Let $\mathcal{M}_d$ denote the space of signed finite measures on $\mathcal{P}_d$. Since $\mathcal{P}_d$ is finite, we can identify elements of $\mathcal{M}_d$ as vectors of $\mathbb{R}^{Bell_d}$, where $Bell_d$ is the Bell number, which counts the number of elements in $\mathcal{P}_d$ (the number of partitions of d elements). In particular, if π is a partition of E, and $\mu \in \mathcal{M}_d$, then $\mu(\pi)$ will correspond to the measure of the singleton $\{\pi\}$, or equivalently, to the "π^{th} coordinate" of the vector μ. We define the inner product $< \cdot,\cdot >$ as

$$\begin{aligned} < \cdot,\cdot >:\ \mathcal{M}_d \times \mathcal{M}_d \ &\to \ \mathbb{R} \\ m, v \ &\to \ \sum_{\pi\in\mathcal{P}_d} m(\pi)v(\pi). \end{aligned}$$

For every function $f : \mathcal{P}_d \to \mathcal{P}_d$, we define the operator $*$ s.t for every $m \in \mathcal{M}_d$, for every $\pi \in \mathcal{P}_d$, $f * m(\pi) = m(f^{-1}(\pi))$. In words, $f * m$ is the push-forward measure of m by f. Further, we will

also consider square matrices indexed by elements in $\mathcal{P}_d$. For such a matrix K and an element $m \in \mathcal{M}_d$, we define $Km(\pi) := \sum_{\pi' \in \mathcal{P}_d} K(\pi, \pi')m(\pi')$.

Define

$$\begin{aligned} X \ : \ (\mathcal{P}_d)^{l^\epsilon} &\to \mathcal{M}(\mathcal{P}_d) \\ \Pi &\to \frac{1}{l^\epsilon} \sum_{k \le l^\epsilon} \delta_{\Pi_k}, \end{aligned}$$

i.e., $X(\Pi)$ is the empirical measure associated to the "sample" $\Pi_1, \ldots, \Pi_{l^\epsilon}$. In the following, we define

$$\xi_t^{\epsilon, L^\epsilon} \ := \ X(\Pi^{\epsilon, L^\epsilon}(t))$$

will be referred to as the (empirical) genetic partition probability measure of the population, conditional on the l^ϵ loci to be located at L^ϵ. We also define

$$\xi_t^\epsilon \ \equiv \ \xi_t^{\epsilon, \mathcal{L}^\epsilon} \ = \ X(\Pi^{\epsilon, \mathcal{L}^\epsilon}(t)) \text{ where } \mathcal{L}^\epsilon \sim \mathcal{U}([0,1]^{l^\epsilon})$$

will be referred to as the (empirical) genetic partition probability measure of the population.

The genetic distance between i and j at time t can then be expressed in terms of ξ_t^ϵ as follows:

$$d_t^\epsilon(i,j) = 1 - \xi_t^\epsilon(\{\pi \in \mathcal{P}_d : \ i \sim_\pi j)\}). \tag{5}$$

In the following, we identify the process $(\xi_t^\epsilon, t \ge 0)$ to a process in the set of the càdlàg functions from $\mathbb{R}^+$ to $\mathbb{R}^{Bell_d}$, equipped with the standard Skorokhod topology.

3.3 Convergence of the genetic partition probability measure

Following Remark 2.3, for every $k \in \{1, \ldots, l^\epsilon\}$, the process $(\Pi_k^{\epsilon, L^\epsilon}(t); t \ge 0)$ – the partition at locus k – obeys the following dynamics:

1. (reproduction event) j is merged in the block containing i at rate $M_{ij} \frac{1}{\epsilon n_j^\epsilon}$.
2. (mutation) Individual i takes on a new type at rate b_∞.

The generator associated to the allelic partition at locus k is then given by

$$G^\epsilon g(\pi) = \sum_{i,j=1}^d M_{ij} \frac{1}{\epsilon n_j^\epsilon} \left(\ g(\mathcal{I}_{ij}(\pi)) - g(\pi) \ \right) \ + \ b_\infty \sum_{i=1}^d \left(\ g(\mathcal{S}_i(\pi)) - g(\pi) \ \right). \tag{6}$$

Recall that the expression of the generator associated to the allelic partition at a given locus k is independent on the position of locus k and on λ. Also recall that $\epsilon n_i^\epsilon \to N_i$ and that, by definition, $\forall k, p \in E, \ \tilde{M}_{kp} = M_{pk}/N_k$. Thus

$$G^\epsilon g(\pi) \to Gg(\pi) := \sum_{i,j=1}^d \tilde{M}_{ji} \left(\ g(\mathcal{I}_{ij}(\pi)) - g(\pi) \ \right) \ + \ b_\infty \sum_{i=1}^d \left(\ g(\mathcal{S}_i(\pi)) - g(\pi) \ \right) \quad \text{as } \epsilon \to 0. \tag{7}$$

Direct computations yield that ${}^tG$, the transpose of the matrix G satisfies

$$\forall m \in \mathcal{M}_d, \quad {}^tGm \ = \ \sum_{i,j=1}^d \tilde{M}_{ji} (\mathcal{I}_{ij} * m - m) + b_\infty \sum_{i=1}^d (\mathcal{S}_i * m - m). \tag{8}$$

In the light of (7), the following theorem can be interpreted as an ergodic theorem. We show that the (dynamical) empirical measure constructed from the allelic partitions along the

chromosome converges to the probability measure of a single locus. Although in the IBM the different loci are linked and do not fix independently (as already mentioned in Remark 2.4), as the number of loci tends to infinity, they become decorrelated. In the large population - dense site limit, the following result indicates that the model behaves as if infinitely many loci evolved independently according to the (one-locus) Moran model with generator G provided in (7).

In the following " $\Longrightarrow$" indicates the convergence in distribution. Also, we identify $(\xi_t^\epsilon;\ t \geq 0)$ to a function from $\mathbb{R}^+$ to $\mathbb{R}^{Bell_d}$; and convergence *in the weak topology* means that for every $T > 0$, the process $(\xi_t^\epsilon;\ t \in [0,T])$ converges in the Skorohod topology $D([0,T], \mathbb{R}^{Bell_d})$.

Theorem 3.1 (Ergodic theorem along the chromosome)**.** *Assume that ξ_0^ϵ is deterministic and there exists a probability measure $P^0 \in \mathcal{M}_d$ such that the following convergence holds:*

$$\xi_0^\epsilon \underset{\epsilon \to 0}{\longrightarrow} P^0. \tag{9}$$

Then

$$(\xi_t^\epsilon;\ t \geq 0) \underset{\epsilon \to 0}{\Longrightarrow} (P_t;\ t \geq 0) \text{ in the weak topology,}$$

where P solves the forward Kolmogorov equation associated to the aforementioned Moran model, i.e.,

$$\frac{d}{ds} P_s \ = \ {}^tG P_s$$

with initial condition $P_0 = P^0$ and where ${}^tG$ denotes the transpose of G (see (8)).

4 Proof of Theorem 3.1

The idea behind the proof is to condition on $\mathcal{L}^\epsilon = L^\epsilon$, and then decompose the Markov process $(< \xi_t^{\epsilon, L_\epsilon}, v >; t \geq 0)$ into a drift part and a Martingale part. We show that the drift part converges to the solution of the Kolmogorov equation alluded to in Theorem 3.1 and that the Martingale part vanishes when $\epsilon \to 0$. The main steps of the computation are outlined in the next subsection. We leave technical details (tightness and second moment computations) until the end of the section.

4.1 Main steps of the proof

Fix $L^\epsilon \in [0,1]^{l^\epsilon}$. Recall the definition of $\mathbb{G}^{\epsilon, L^\epsilon}$, the generator of the process $(\Pi^{\epsilon, L^\epsilon}(t);\ t \geq 0)$, given in (4). Let $f : \mathbb{R} \to \mathbb{R}$ be a Borel bounded function and fix $v \in \mathcal{M}_d$. Define $h(\Pi) \ = \ f(\langle X(\Pi), v \rangle)$. Then, it is straightforward to see from (4) that

$$\begin{aligned} \mathbb{G}^{\epsilon, L^\epsilon} h(\Pi) \ &= \ \frac{1}{\epsilon} \sum_{i,j=1}^{d} M_{ij}\ \mathbb{E}_{\lambda, L^\epsilon, j} \left(f(\langle X(\mathcal{I}_{ij}^S(\Pi)), v \rangle) - f(\langle X(\Pi), v \rangle) \right) \\ &+ \ b_\infty l^\epsilon \sum_{i=1}^{d} \ \mathbb{E}_{l^\epsilon} \left(f(\langle X(\mathcal{S}_i^K(\Pi)), v \rangle) - f(\langle X(\Pi), v \rangle) \right), \end{aligned} \tag{10}$$

where in the first line $\mathbb{E}_{\lambda, L^\epsilon, j}$ is the expected value taken with respect to the random variable S, distributed as $\mathcal{F}_j^{L^\epsilon, \lambda}$ as defined in Definition 2.1. In the second line, $\mathbb{E}_{l^\epsilon}$ is the expected value taken with respect to K, distributed as a uniform random variable on $\{1, \ldots, l^\epsilon\}$.

Lemma 4.1. *Define $v \in \mathcal{M}_d$, $L^\epsilon \in [0,1]^{l^\epsilon}$, $g(\Pi) := \langle X(\Pi), v \rangle$. Then $\mathbb{G}^{\epsilon, L^\epsilon} g(\Pi) \ = \ \left\langle {}^tG^\epsilon X(\Pi), v \right\rangle$ where ${}^tG^\epsilon$ is the transpose of G^ϵ – the generator of the allelic partition at a single locus as defined in (6) – i.e.,*

$$\forall m \in \mathcal{M}_d, \quad {}^tG^\epsilon m \ = \ \sum_{i,j=1}^{d} M_{ij} \frac{1}{\epsilon n_i^\epsilon} (\mathcal{I}_{ij} * m - m) + b_\infty \sum_{i=1}^{d} (\mathcal{S}_i * m - m). \tag{11}$$

Proof. We define the two following signed measures:

$$\partial_{ij}^S X(\Pi) \;=\; X(\mathcal{I}_{ij}^S(\Pi)) \;-\; X(\Pi), \;\; \partial_i^K X(\Pi) \;=\; X(\mathcal{S}_i^K(\Pi)) \;-\; X(\Pi). \tag{12}$$

In words, $\partial_{ij}^S X(\Pi)$ is the change in the genetic partition measure $X(\Pi)$ if we merge j in the block of i at every locus in S, and $\partial_i^K X(\Pi)$ is the change in $X(\Pi)$ if we single out element i at locus K. Using those notations, for our particular choice of g, (10) writes

$$\mathbb{G}^{\epsilon,L^\epsilon} g(\Pi) \;=\; \frac{1}{\epsilon}\sum_{i,j=1}^{d} M_{ij}\; \mathbb{E}_{\lambda,L^\epsilon,j}\left(\left\langle \partial_{ij}^S X(\Pi), v\right\rangle\right) \;+\; b_\infty l^\epsilon \sum_{i=1}^{d} \mathbb{E}_{l^\epsilon}\left(\left\langle \partial_i^K X(\Pi), v\right\rangle\right).$$

We now show that for every $v \in \mathcal{M}_d$,

$$\begin{aligned}
\mathbb{E}_{\lambda,L^\epsilon,j}\left(\left\langle \partial_{ij}^S X(\Pi), v\right\rangle\right) &= \frac{1}{n_j^\epsilon}\left\langle \mathcal{I}_{ij} * X(\Pi) - X(\Pi), v\right\rangle, \\
\mathbb{E}_{l^\epsilon}\left(\left\langle \partial_i^K X(\Pi), v\right\rangle\right) &= \frac{1}{l^\epsilon}\left\langle \mathcal{S}_i * X(\Pi) - X(\Pi), v\right\rangle.
\end{aligned} \tag{13}$$

We only prove the first identity. The second one can be shown along the same lines. Again, we let Π_k be the k^{th} coordinate of Π. By definition, the vector $\mathcal{I}_{ij}^S(\Pi)$ is only modified at the coordinates belonging to S, and thus

$$\begin{aligned}
\left\langle \partial_{ij}^S X(\Pi), v\right\rangle &= \sum_{\pi\in\mathcal{P}_d} \frac{v(\pi)}{l^\epsilon}\left(\left|\{k \le l^\epsilon \;:\; (\mathcal{I}_{ij}^S(\Pi))_k = \pi\}\right| \;-\; \left|\{k \le l^\epsilon : \Pi_k = \pi\}\right|\right) \\
&= \sum_{\pi\in\mathcal{P}_d} \frac{v(\pi)}{l^\epsilon}\left(\left|\{k \in S \;:\; \mathcal{I}_{ij}^S(\Pi_k) = \pi\}\right| - \left|\{k \in S : \Pi_k = \pi\}\right|\right).
\end{aligned} \tag{14}$$

Secondly, for every $j \in E$,

$$\mathbb{E}_{\lambda,L^\epsilon,j}(\left|\{k \in S \;:\; \Pi_k = \pi\}\right|) \;=\; \mathbb{E}_{\lambda,L^\epsilon,j}(\sum_{k\in S} 1_{\{\Pi_k=\pi\}}) \;=\; \sum_{k\le l^\epsilon} 1_{\{\Pi_k=\pi\}}\mathbb{E}_{\lambda,L^\epsilon,j}(1_{\{k\in S\}}).$$

As S is distributed as $\mathcal{F}_j^{L^\epsilon,\lambda}$, we can use the fact that $\mathbb{P}\left(k \in \mathcal{F}_j^{L^\epsilon,\lambda}\right) \;=\; M_{ij}\frac{1}{n_j^\epsilon}$ (see Remark 2.3), and then

$$\mathbb{E}_{\lambda,L^\epsilon,j}(\left|\{k \in S \;:\; \Pi_k = \pi\}\right|) \;=\; \frac{1}{n_j^\epsilon}\left|\{k \le l^\epsilon,\; \Pi_k = \pi\}\right| = \frac{l^\epsilon}{n_j^\epsilon} X(\Pi)(\pi). \tag{15}$$

Furthermore, by applying (15) for every $\pi' \in \mathcal{I}_{ij}^{-1}(\pi)$ and then taking the sum over every such partitions, we get

$$\mathbb{E}_{\lambda,L^\epsilon,j}(\left|\{k \in S \;:\; \mathcal{I}_{ij}(\Pi_k) = \pi\}\right| \;=\; \frac{l^\epsilon}{n_j^\epsilon} X(\Pi)(\mathcal{I}_{ij}^{-1}(\pi)) = \frac{l^\epsilon}{n_j^\epsilon}\, \mathcal{I}_{ij} * X(\Pi)(\pi).$$

This completes the proof of (13). From this result, we deduce that

$$\mathbb{G}^{\epsilon,L^\epsilon} g(\Pi) \;= \frac{1}{\epsilon}\sum_{i,j=1}^{d} M_{ij}\frac{1}{n_j^\epsilon}(\mathcal{I}_{ij} * X(\Pi) - X(\Pi)) \;+\; b_\infty \sum_{i=1}^{d} (\mathcal{S}_i * X(\Pi) - X(\Pi)).$$

This completes the proof of Lemma 4.1. □

For every $L^\epsilon \in [0,1]^{l^\epsilon}$, for every $v \in \mathcal{M}_d$, define

$$M_t^{\epsilon,L^\epsilon,v} := \left\langle \xi_t^{\epsilon,L^\epsilon}, v \right\rangle \; - \; \int_0^t \left\langle {}^tG^\epsilon \xi_s^{\epsilon,L^\epsilon}, v \right\rangle ds, \quad B_t^{\epsilon,L^\epsilon,L^\epsilon,v} \; := \; \int_0^t \left\langle {}^tG^\epsilon \xi_s^{\epsilon,L^\epsilon}, v \right\rangle ds.$$

Since $\left\langle \xi_t^{\epsilon,L^\epsilon}, v \right\rangle$ is bounded, the previous result implies that $M^{\epsilon,L^\epsilon,v}$ is a martingale with respect to $(\mathcal{H}_t^{L^\epsilon})_{t\geq 0}$, the filtration generated by $(\Pi^{\epsilon,L^\epsilon}(t); t \geq 0)$. Further, the semi-martingale $\left\langle \xi_t^{\epsilon,L^\epsilon}, v \right\rangle$ admits the following decomposition:

$$\left\langle \xi_t^{\epsilon,L^\epsilon}, v \right\rangle \; = \; M_t^{\epsilon,L^\epsilon,v} \; + \; B_t^{\epsilon,L^\epsilon,v}.$$

Lemma 4.2. *For every $v \in \mathcal{M}_d$, for every $L^\epsilon \in [0,1]^{l^\epsilon}$,*

$$\left\langle M^{\epsilon,L^\epsilon,v} \right\rangle_t \; = \; \int_0^t \; m^{\epsilon,L^\epsilon,v}(\Pi^{\epsilon,L^\epsilon}(s))ds$$

with

$$m^{\epsilon,L^\epsilon,v}(\Pi) = \frac{1}{\epsilon} \sum_{i,j=1}^d \; M_{ij} \; \mathbb{E}_{\lambda,L^\epsilon,j} \left(\left\langle \partial_{ij}^S X(\Pi), v \right\rangle^2 \right) \; + \; b_\infty l^\epsilon \sum_{i=1}^d \mathbb{E}_{l^\epsilon} \left(\left\langle \partial_i^K X(\Pi), v \right\rangle^2 \right)$$

and where $\left\langle M^{\epsilon,L^\epsilon,v} \right\rangle_t$ denotes the quadratic variation of $M^{\epsilon,L^\epsilon,v}$.

Proof. For $v \in \mathcal{M}_d$, define $h(\Pi) = \langle X(\Pi), v \rangle^2$. Then, by (10)

$$\begin{aligned} \mathbb{G}^{\epsilon,L^\epsilon} h(\Pi) \; &= \; \frac{1}{\epsilon} \sum_{i,j=1}^d M_{ij} \mathbb{E}_{\lambda,L^\epsilon,j} \left(\left\langle X(\mathcal{I}_{ij}^S(\Pi), v) \right\rangle^2 \; - \; \langle X(\Pi), v \rangle^2 \right) \\ &\quad + \; b_\infty l^\epsilon \sum_{i=1}^d \mathbb{E}_{l^\epsilon} \left(\left\langle X(\mathcal{S}_i^K(\Pi)), v \right\rangle^2 - \langle X(\Pi), v \rangle^2 \right). \end{aligned}$$

Since

$$\begin{aligned} \left\langle X(\mathcal{I}_{ij}^S(\Pi)), v \right\rangle^2 \; - \; \langle X(\Pi), v \rangle^2 \; &= \; \left\langle X(\mathcal{I}_{ij}^S(\Pi)) - X(\Pi), v \right\rangle^2 \; + \; 2 \left\langle X(\mathcal{I}_{ij}^S(\Pi)) - X(\Pi), v \right\rangle \; \langle X(\Pi), v \rangle \\ \left\langle X(\mathcal{S}_i^k(\Pi)), v \right\rangle^2 \; - \; \langle X(\Pi), v \rangle^2 \; &= \; \left\langle X(\mathcal{S}_i^k(\Pi)) - X(\Pi), v \right\rangle^2 \; + \; 2 \left\langle X(\mathcal{S}_i^k(\Pi)) - X(\Pi), v \right\rangle \; \langle X(\Pi), v \rangle, \end{aligned}$$

the previous identities yield

$$\begin{aligned} \mathbb{G}^{\epsilon,L^\epsilon} h(\Pi) \; &= \; 2\mathbb{G}^{\epsilon,L^\epsilon} g(\Pi) \, \langle X(\Pi), v \rangle \; + \\ &\quad \frac{1}{\epsilon} \sum_{i,j=1}^d M_{ij} \mathbb{E}_{\lambda,L^\epsilon,j} \left(\left\langle \partial_{ij}^S X(\Pi), v \right\rangle^2 \right) + \; b_\infty l^\epsilon \sum_{i=1}^d \mathbb{E}_{l^\epsilon} \left(\left\langle \partial_i^K X(\Pi), v \right\rangle^2 \right), \end{aligned}$$

where $g(\Pi) = \langle X(\Pi), v \rangle$. As a consequence

$$\left\langle X(\Pi^{\epsilon,L^\epsilon}(t)), v \right\rangle^2 \; - \; 2 \int_0^t \; \mathbb{G}^{\epsilon,L^\epsilon} g(\Pi^{\epsilon,L^\epsilon}(s)) \left\langle X(\Pi^{\epsilon,L^\epsilon}(s)), v \right\rangle ds$$

$$- \int_0^t \; \frac{1}{\epsilon} \sum_{i,j=1}^d M_{ij} \mathbb{E}_{\lambda,L^\epsilon,j} \left(\left\langle \partial_{ij}^S X(\Pi^{\epsilon,L^\epsilon}(s)), v \right\rangle^2 \right) ds - \; \int_0^t \; b_\infty l^\epsilon \sum_{i=1}^d \mathbb{E}_{l^\epsilon} \left(\left\langle \partial_i^K X(\Pi^{\epsilon,L^\epsilon}(s)), v \right\rangle^2 \right) ds$$

is a martingale. Further using Itô's formula, the process

$$\left\langle X(\Pi^{\epsilon,L^\epsilon}(t)), v \right\rangle^2 - 2 \int_0^t \; \mathbb{G}^{\epsilon,L^\epsilon} g(\Pi^{\epsilon,L^\epsilon}(s)) \left\langle X(\Pi^{\epsilon,L^\epsilon}(s)), v \right\rangle ds \; - \; \left\langle M^{\epsilon,L^\epsilon,v} \right\rangle_t$$

is also a martingale. Combining the two previous results completes the proof of Lemma 4.2. □

Proposition 4.3.

$$\lim_{\epsilon \to 0} \mathbb{E}(\sup_{\Pi \in (\mathcal{P}_d)^{l^\epsilon}} m^{\epsilon, \mathcal{L}^\epsilon, v}(\Pi)) = 0,$$

where the expected value is taken with respect to the random variable $\mathcal{L}^\epsilon$.

Proposition 4.4. *For* $T > 0$*, the family of random variables* $(\xi^\epsilon;\ \epsilon > 0)$ *is tight in the weak topology* $D([0,T], \mathbb{R}^{Bell_d})$.

We postpone the proof of Propositions 4.3 and 4.4 until Sections 4.2 and 4.3 respectively.

Proof of Theorem 3.1 based on Proposition 4.3 and 4.4. Since $(\xi^\epsilon;\ \epsilon > 0)$ is tight, we can always extract a subsequence converging in distribution (for the weak topology) to a limiting random measure process ξ. We will now show that ξ can only be the solution of the Kolmogorov equation alluded to in Theorem 3.1. From (11), for every probability measure m on $\mathcal{P}_d$, for every $v \in \mathcal{M}_d$,

$$\left| \left\langle {}^t G^\epsilon m, v \right\rangle \right| \ \leq \ (2 \sum_{i,j=1}^{d} M_{ij} \ \frac{1}{\epsilon n_i^\epsilon} + 2 b_\infty d) ||v||_\infty, \tag{16}$$

where $||v||_\infty := \max_{\pi \in \mathcal{P}_N} v(\pi)$. Since as $\epsilon \to 0$, $n_i^\epsilon \epsilon \to N_i$, the term between parentheses also converges, and thus the RHS is uniformly bounded in ϵ. Finally, the bounded convergence theorem implies that for every $v \in \mathcal{M}_d$,

$$\mathbb{E}\left(\langle \xi_t, v \rangle - \int_0^t \left\langle {}^t G \xi_s, v \right\rangle ds \right)^2 \ = \ \lim_{\epsilon \to 0} \mathbb{E}\left(\left(\langle \xi_t^\epsilon, v \rangle - \int_0^t \left\langle {}^t G^\epsilon \xi_s^\epsilon, v \right\rangle ds \right)^2 \right),$$

where we used the fact that ${}^t G^\epsilon m \ \to \ {}^t G m$ for every $m \in \mathcal{M}_d$ (where G is defined as in Theorem 3.1). On the other hand, since

$$\begin{aligned}
\mathbb{E}\left(\langle \xi_t^\epsilon, v \rangle - \int_0^t \left\langle {}^t G^\epsilon \xi_s^\epsilon, v \right\rangle ds \right)^2 &= \mathbb{E}\left(\mathbb{E}\left(\left(\left\langle \xi_t^{\epsilon, \mathcal{L}^\epsilon}, v \right\rangle - \int_0^t \left\langle {}^t G^\epsilon \xi_s^{\epsilon, \mathcal{L}^\epsilon}, v \right\rangle ds \right)^2 \mid \mathcal{L}^\epsilon \right) \right) \\
&= \mathbb{E}(\mathbb{E}(\left\langle M^{\epsilon, \mathcal{L}^\epsilon, v} \right\rangle_t \mid \mathcal{L}^\epsilon)) \\
&= \mathbb{E}(\mathbb{E}(\int_0^t m^{\epsilon, \mathcal{L}^\epsilon, v}(\Pi^{\epsilon, \mathcal{L}^\epsilon}(s)) ds \mid \mathcal{L}^\epsilon)) \\
&\leq t \ \mathbb{E}(\sup_{\pi \in (\mathcal{P}_d)^{l^\epsilon}} m^{\epsilon, \mathcal{L}^\epsilon, v}(\Pi)).
\end{aligned}$$

Lemma 4.2 and Proposition 4.3 imply that

$$\mathbb{E}\left(\langle \xi_t, v \rangle - \int_0^t \left\langle {}^t G \xi_s, v \right\rangle ds \right)^2 \ = \ 0,$$

which ends the proof of Theorem 3.1. □

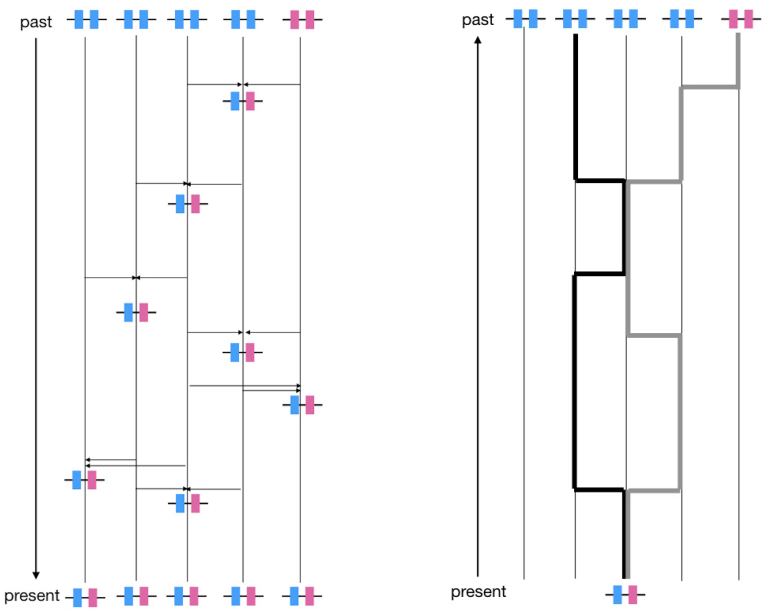
4.2 Proof of Proposition 4.3

Our first step in proving Proposition 4.3 is to prove the following result.

Lemma 4.5. $\forall\ j\ \in E, \forall \lambda > 0,$

$$\lim_{\epsilon \to 0} \ \frac{1}{(l^\epsilon)^2 \epsilon} \ \mathbb{E}\left(\mathbb{E}_{\lambda, \mathcal{L}^\epsilon, j}(\sum_{k=1}^{l^\epsilon} \sum_{k'=1}^{l^\epsilon} 1_{k \in S} 1_{k' \in S} \mid \mathcal{L}^\epsilon) \right) \ = \ 0.$$

Before turning to the proof of this result, we recall the definition of the ancestral recombination graph (ARG) (see also **?**, **?**, **?**) for the case of two loci. Fix $L^\epsilon = \{\ell_1, \dots \ell_{l^\epsilon}\}$ the positions of the loci in the chromosome, λ the recombination rate, and choose two loci k and k' among the l^ϵ loci. In order to compute the probability that for both loci the allele from the migrant is fixed in the host population – $\mathbb{P}(k, k' \in \mathcal{F}_j^{L^\epsilon,\lambda})$ – we follow backwards in time the genealogy of the corresponding alleles carried by a reference individual in the present population, assuming that a migration event occurred in the past (sufficiently many generations ago, so that we can assume that the present population is homogeneous). More precisely, at locus k, we consider the ances-



(a) Moran model with recombination

(b) ARG with 2 loci

Figure 5: Realisation of a Moran model with recombination and the corresponding ARG. The population size is equal to 5 and $l^\epsilon = 2$. The colors of the loci represent their origin: loci from the resident subpopulation are represented in blue whereas loci from the migrant are represented in pink. In (a) the origins of the arrows indicate the parents, and the tips of the arrows point to their offspring. Time goes from top to bottom as indicated by the arrow on the left. In (b) the black (resp. grey) line corresponds to the ancestral lineages of the first (resp. second) locus in the chromosome of an individual sampled in the extant population. Time goes backwards, from bottom to top. The first locus has been inherited from the resident subpopulation whereas the second locus has been inherited from the migrant (i.e. $1 \notin S,\ 2 \in S$).

tral lineage of a reference individual (chosen uniformly at random) in the extant population. We envision this lineage as a particle moving in $\{1, \dots, n_j^\epsilon\}$: time $t = 0$ corresponds to the present, and the position of the particle at time t – denoted by $A_k^{L^\epsilon,\lambda,j}(t)$ – identifies the ancestor of locus k, t units of time in the past (i.e., at locus k, the reference individual in the extant population inherits its genetic material from individual $A_k^{L^\epsilon,\lambda,j}(t)$ at time $-t$) (see Figure 5).

The recombination rate between the two loci, k and k', $r_{k,k'}^{L^\epsilon,\lambda}$ corresponds to the probability that there is at least one Poisson point between ℓ_k and $\ell_{k'}$ and the two fragments are inherited from different parents and is given by

$$r_{k,k'}^{L^\epsilon,\lambda} := \frac{1}{2}\left(1 - \exp(-\lambda|\ell_k - \ell_{k'}|)\right). \tag{17}$$

$\mathbf{A}^{L^\epsilon,\lambda,j} = (A_k^{L^\epsilon,\lambda,j},\ A_{k'}^{L^\epsilon,\lambda,j})$ defines a 2-dimensional stochastic process on $\{1, \dots, n_j^\epsilon\}$. At time

0, the two particles have the same position (they coincide at a randomly chosen individual as in Figure 5) and then evolve according to the following dynamics:

- When both particles are occupying the same location z, the group splits into two at rate $r_{k,k'}^{L^\epsilon,\lambda}$ (see (17)). Forward in time this corresponds to a reproduction event where z is replaced by the offspring of x and y. Each individual x reproduces at rate 1 (chooses a random partner y), and with probability $1/n_j^\epsilon$ his offspring replaces individual z. There are n_j^ϵ possible choices for x. Following (17), the probability that both loci are inherited from different parents is $r_{k,k'}^{L^\epsilon,\lambda}$, so the rate of fragmentation for loci k, k' is given by $n_j^\epsilon \,.\, \frac{1}{n_j^\epsilon} \,.\, r_{k,k'}^{L^\epsilon,\lambda}$.

- When the two particles are occupying different positions, they jump to the same position at rate $2/n_j^\epsilon$. Forwards in time, this corresponds to a reproduction event where the individual located at $A_k^{L^\epsilon,\lambda,j}$ (resp. $A_{k'}^{L^\epsilon,\lambda,j}$) replaces the one at $A_{k'}^{L^\epsilon,\lambda,j}$ (resp. $A_k^{L^\epsilon,\lambda,j}$), and the offspring inherits the allele at locus k' (resp. k) from this parent. A reproduction event where the individual located at $A_k^{L^\epsilon,\lambda,j}$ (resp. $A_{k'}^{L^\epsilon,\lambda,j}$) replaces the one at $A_{k'}^{L^\epsilon,\lambda,j}$ (resp. $A_k^{L^\epsilon,\lambda,j}$) occurs at rate $2/n_j^\epsilon$ (as the individual at $A_k^{L^\epsilon,\lambda,j}$ –resp. $A_{k'}^{L^\epsilon,\lambda,j}$– can be the mother or the father); and the probability that the offspring inherits the locus k' (resp. k) from this parent is $1/2$. The total rate of coalescence is $2 \,.\, \frac{2}{n_j^\epsilon} \,.\, \frac{1}{2}$.

Since we assume that the migration event occurred far back in the past, the following duality relation holds:

$$\mathbb{P}(k, k' \in \mathcal{F}_j^{L^\epsilon,\lambda}) \;=\; \lim_{t\to\infty} \mathbb{P}\Big(A_k^{L^\epsilon,\lambda,j}(t) = A_{k'}^{L^\epsilon,\lambda,j}(t) = 1\Big). \tag{18}$$

In other words, assuming that the migrant is labelled 1, the set on the RHS corresponds to the set of loci inheriting their genetic material from the migrant.

Proof of Lemma 4.5. Define $(Y^{L^\epsilon,\lambda,j}(t) := 1_{A_k^{L^\epsilon,\lambda,j}(t)=A_{k'}^{L^\epsilon,\lambda,j}(t)}; t \geq 0)$. It is easy to see from the previous description of the dynamics that Y is a Markov chain on $\{0,1\}$ with the following transition rates:

$$q_{1,0} = r_{k,k'}^{L^\epsilon,\lambda}, \quad q_{0,1} = \frac{2}{n_j^\epsilon}$$

and further

- conditional on $Y^{L^\epsilon,\lambda,j}(t) = 1$, the two lineages $(A_k^{L^\epsilon,\lambda,j}(t), A_{k'}^{L^\epsilon,\lambda,j}(t))$ occupy a common position that is distributed as a uniform random variable on $\{1, \ldots, n_j^\epsilon\}$.

- conditional on $Y^{L^\epsilon,\lambda,j}(t) = 0$, $(A_k^{L^\epsilon,\lambda,j}(t), A_{k'}^{L^\epsilon,\lambda,j}(t))$ are distinct and are distributed as a two uniformly sampled random variables (without replacement) on $\{1, \ldots, n_j^\epsilon\}$.

We have:

$$\mathbb{P}(A_k^{L^\epsilon,\lambda,j}(t) = A_{k'}^{L^\epsilon,\lambda,j}(t) = 1) = \mathbb{P}\left(Y^{L^\epsilon,\lambda,j}(t) = 1\right) \frac{1}{n_j^\epsilon} \,.$$

Furthermore, it is straightforward to show that

$$\lim_{t\to\infty} \mathbb{P}\left(Y^{L^\epsilon,\lambda,j}(t) = 1\right) \;=\; \frac{2}{n_j^\epsilon} \frac{1}{r_{k,k'}^{L^\epsilon,\lambda} + \frac{2}{n_j^\epsilon}}.$$

From (18), we get that,

$$\mathbb{E}_{\lambda,L^\epsilon,j}\Big(\sum_{k,k'\in\{1,\ldots,l^\epsilon\}} 1_{k\in S} 1_{k'\in S}\Big) \;=\; \lim_{t\to\infty} \sum_{k,k'\in\{1,\ldots,l^\epsilon\}} \mathbb{P}(A_k^{L^\epsilon,\lambda,j}(t) = A_{k'}^{L^\epsilon,\lambda,j}(t) = 1)$$

$$= \;\frac{2}{(n_j^\epsilon)^2} \sum_{k,k'\in\{1,\ldots,l^\epsilon\}} \frac{1}{\frac{1}{2}(1 - e^{-\lambda|\ell_k - \ell_{k'}|}) + \frac{2}{n_j^\epsilon}}.$$

One can then easily check that, $\exists\ \alpha > 0$ such that, for every $L^\epsilon = \{\ell_1, \dots \ell_{l^\epsilon}\}$,

$$\mathbb{E}_{\lambda, L^\epsilon, j}\Big(\sum_{k,k' \in \{1,\dots,l^\epsilon\}} 1_{k \in S} 1_{k' \in S}\Big) \ \leq\ \frac{2}{(n_j^\epsilon)^2} \sum_{k,k' \in \{1,\dots,l^\epsilon\}} \frac{1}{\alpha|\ell_k - \ell_{k'}| + \frac{2}{n_j^\epsilon}}. \tag{19}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left(\mathbb{E}_{\lambda, \mathcal{L}^\epsilon, j}\Big(\sum_{k=1}^{l^\epsilon}\sum_{k'=1}^{l^\epsilon} 1_{k \in S} 1_{k' \in S} \mid \mathcal{L}^\epsilon\Big)\right) &\leq \frac{2}{(n_j^\epsilon)^2} \int_{[0,1]^{l^\epsilon}} dx_1, \dots dx_{l^\epsilon} \sum_{k,k' \in \{1,\dots,l^\epsilon\}} \frac{1}{\alpha|x_k - x_{k'}| + \frac{2}{n_j^\epsilon}} \\
&\leq \frac{2}{(n_j^\epsilon)^2} \sum_{k,k' \in \{1,\dots,l^\epsilon\}} \int_{[0,1]^2} \frac{dx_k dx_{k'}}{\alpha|x_k - x_{k'}| + \frac{2}{n_j^\epsilon}}.
\end{aligned}$$

In addition, using the fact that $n_j^\epsilon = [N_j/\epsilon]$,

$$\begin{aligned}
& \frac{1}{(l^\epsilon)^2 \epsilon} \mathbb{E}\left(\mathbb{E}_{\lambda, \mathcal{L}^\epsilon, j}\Big(\sum_{k=1}^{l^\epsilon}\sum_{k'=1}^{l^\epsilon} 1_{k \in S} 1_{k' \in S} \mid \mathcal{L}^\epsilon\Big)\right) \\
\leq\ & \frac{2\epsilon}{(N_j - \epsilon)^2} \int_{[0,1]^2} \frac{dt ds}{\alpha|t - s| + 2\epsilon/N_j} \\
=\ & \frac{4\epsilon}{(N_j - \epsilon)^2} \int_0^1 ds \int_o^s \frac{dt}{\alpha|t - s| + 2\epsilon/N_j} \\
=\ & \frac{4\epsilon}{\alpha(N_j - \epsilon)^2} \int_0^1 \log\left(\frac{\alpha N_j}{2\epsilon} s + 1\right) ds \\
=\ & \frac{4\epsilon}{\alpha(N_j - \epsilon)^2} \left((1 + \frac{2\epsilon}{\alpha N_j}) \log\left(\frac{\alpha N_j}{2\epsilon} + 1\right) - 1\right) \\
\underset{\epsilon \to 0}{\longrightarrow}\ & 0.
\end{aligned}$$

□

We are now ready to prove Proposition 4.3.

Proof of Proposition 4.3. Using the definition given in Lemma 4.2,

$$m^{\epsilon, L^\epsilon, v}(\Pi) = \frac{1}{\epsilon} \sum_{i,j=1}^{d} M_{ij}\ \mathbb{E}_{\lambda, L^\epsilon, j}\left(\left\langle \partial_{ij}^S X(\Pi), v\right\rangle^2\right) \ +\ b_\infty l^\epsilon \sum_{i=1}^{d} \mathbb{E}_{l^\epsilon}\left(\left\langle \partial_i^K X(\Pi), v\right\rangle^2\right).$$

To bound the second term on the RHS, we note that, by definition, $\mathcal{S}_i^k(\Pi)$ and Π only differ in one component, so from the definition of $\partial_i^K X(\Pi)$ (see (12)), it is not hard to see that

$$\left\langle \partial_i^K X(\Pi), v\right\rangle^2 \ \leq\ \frac{4}{(l^\epsilon)^2} ||v||_\infty^2.$$

It follows that,

$$b_\infty l^\epsilon\ \mathbb{E}_{l^\epsilon}(\left\langle \partial_i^K X(\Pi), v\right\rangle^2) \leq \frac{4 b_\infty}{l^\epsilon} ||v||_\infty^2. \tag{20}$$

Since $l^\epsilon \to \infty$ as $\epsilon \to 0$, this term converges and can be bounded from above, uniformly in Π and $\epsilon \in (0, 1)$. Note that this bound does not depend on the choice of L^ϵ.

For the second term on the RHS, we simply note that expanding $\frac{1}{\epsilon}\mathbb{E}_{\lambda, L^\epsilon, j}\left(\left\langle \partial_{ij}^S X(\Pi), v\right\rangle^2\right)$ (see (14)), yields a sum of four terms that can be upper bounded by

$$\frac{||v||_\infty^2}{(l^\epsilon)^2 \epsilon} \mathbb{E}_{\lambda, L^\epsilon, j}(|k \in S,\ \Pi_k \in p_1|\ |k \in S,\ \Pi_k \in p_2|),$$

where p_1 and p_2 are alternatively replaced by $\{\pi\}, \mathcal{I}_{ij}^{-1}(\pi)$ with $\pi \in \mathcal{P}_d$. Finally, $\forall L^\epsilon \in [0,1]^{l^\epsilon}$,

$$\begin{aligned}
&\frac{\mathbb{E}_{\lambda,L^\epsilon,j}(|k \in S,\ \Pi_k \in p_1|\ |k \in S,\ \Pi_k \in p_2|)}{(l^\epsilon)^2\epsilon} \\
= &\ \frac{1}{(l^\epsilon)^2\epsilon}\mathbb{E}_{\lambda,L^\epsilon,j}(\sum_{k=1}^{l^\epsilon} 1_{\Pi_k\in p_1}1_{k\in S}\sum_{k'=1}^{l^\epsilon} 1_{\Pi_{k'}\in p_2}1_{k'\in S}) \\
= &\ \frac{1}{(l^\epsilon)^2\epsilon}\sum_{k=1}^{l^\epsilon}\sum_{k'=1}^{l^\epsilon} 1_{\Pi_k\in p_1}1_{\Pi_{k'}\in p_2}\mathbb{E}_{\lambda,L^\epsilon,j}(1_{k\in S}1_{k'\in S}) \\
\leq &\ \frac{1}{(l^\epsilon)^2\epsilon}\mathbb{E}_{\lambda,L^\epsilon,j}(\sum_{k=1}^{l^\epsilon}\sum_{k'=1}^{l^\epsilon} 1_{k\in S}1_{k'\in S})
\end{aligned} \tag{21}$$

randomising the positions of the loci and using Lemma 4.5 the term on the RHS also converges and can also be bounded from above, which completes the proof. □

Remark 4.6 (Magnitude of the stochastic fluctuations)**.** *Lemma 4.1 and the proof Proposition 4.3 entail that:*

$$\forall v \in \mathcal{M}_d,\quad \mathbb{E}(\left\langle M^{\epsilon,\mathcal{L}_\epsilon,v}\right\rangle_t)\ \leq \epsilon\log(1/\epsilon)C + \frac{1}{l^\epsilon}C'$$

where C, C' are constants. This suggests that the order of magnitude of the fluctuations should be of the order of $\max(\sqrt{\epsilon\log(1/\epsilon)}, \sqrt{(1/l^\epsilon)})$.

In **?***, the authors proposed a diffusion approximation (only for the case of two subpopulations). Their approximation is based on the simplifying hypothesis that loci are fixed independently on each other – the number of fixed loci (after each migration event) follows a binomial distribution–, and the hypothesis that the number of loci l is s.t. $l >> \frac{1}{\epsilon}$. They found that the magnitude of the stochastic fluctuations was $\sqrt{\epsilon}$.*

In summary, the previous heuristics suggest that taking into account correlations between loci increases the magnitude of the stochastic fluctuations.

4.3 Tightness: Proof of Proposition 4.4

We follow closely **?**. It is sufficient to prove that for every $v \in \mathcal{M}_d$, the projected process $(\langle \xi^\epsilon, v\rangle\,;\ \epsilon > 0)$ is tight. To this end, we use Aldous criterium (see **?**). In the following, we define

$$M_t^{\epsilon,v} := \langle \xi_t^\epsilon, v\rangle\ -\ \int_0^t \left\langle {}^tG^\epsilon\xi_s^\epsilon, v\right\rangle ds,\quad B_t^{\epsilon,v}\ :=\ \int_0^t \left\langle {}^tG^\epsilon\xi_s^\epsilon, v\right\rangle ds.$$

We first note that

$$\sup_{t\in[0,T]} |\,\langle \xi_t^\epsilon, v\rangle\,|\ \leq\ ||v||_\infty,$$

which implies that that for every deterministic $t \in [0,T]$, the sequence of random variables $(\langle \xi_t^\epsilon, v\rangle\,; \epsilon > 0)$ is tight. Thus, the first part of Aldous criterium is sastified. Next, fix $\delta > 0$, and take two stopping times τ^ϵ and σ^ϵ with respect to $(\mathcal{H}_t^\epsilon)_{t\geq 0}$ the filtration generated by $(\Pi_t^{\epsilon,\mathcal{L}^\epsilon}, t \geq 0)$, such that $0 \leq \tau^\epsilon \leq \sigma^\epsilon \leq \tau^\epsilon + \delta \leq T$. Since $\langle \xi_t^\epsilon, v\rangle\ =\ M_t^{\epsilon,v}\ +\ B_t^{\epsilon,v}$, it is enough to show that the quantities

$$\mathbb{E}\,(|M_{\sigma^\epsilon}^{\epsilon,v} - M_{\tau^\epsilon}^{\epsilon,v}|)\quad\text{and}\quad \mathbb{E}\,(|B_{\sigma^\epsilon}^{\epsilon,v}\ -\ B_{\tau^\epsilon}^{\epsilon,v}|)$$

are bounded from above by two functions in δ (uniformly in the choice of $\tau^\epsilon, \sigma^\epsilon$ and ϵ) going to 0 as δ goes to 0.

The rest of the proof is dedicated to proving those two inequalities. We start with the martingale part. First,

$$\mathbb{E}\left(|M^{\epsilon,v}_{\sigma^\epsilon} - M^{\epsilon,v}_{\tau^\epsilon}|\right)^2 \;\leq\; \mathbb{E}\left((M^{\epsilon,v}_{\sigma^\epsilon} - M^{\epsilon,v}_{\tau^\epsilon})^2\right)$$

Recall that $\forall\ L^\epsilon \in [0,1]^{l^\epsilon}$, $M^{\epsilon,L^\epsilon,v}$ is a martingale. Thus, $M^{\epsilon,v}$ is a martingale with respect to $(\mathcal{G}^\epsilon_t)_{t\geq 0} = (\mathcal{H}^\epsilon_t)_{t\geq 0} \vee \sigma(\mathcal{L}^\epsilon)$, where $(\mathcal{H}^\epsilon_t)_{t\geq 0}$ is the filtration generated by $(\Pi^{\epsilon,\mathcal{L}^\epsilon}_t)$. As $(\mathcal{H}^\epsilon_t)_{t\geq 0} \subset (\mathcal{G}^\epsilon_t)_{t\geq 0}$, τ^ϵ and σ^ϵ are also stopping times for the filtration $(\mathcal{G}^\epsilon_t)_{t\geq 0}$, so that

$$\begin{aligned}
\mathbb{E}\left(|M^{\epsilon,v}_{\sigma^\epsilon} - M^{\epsilon,v}_{\tau^\epsilon}|\right)^2 &\leq \mathbb{E}\left(\mathbb{E}(\left(M^{\epsilon,\mathcal{L}^\epsilon,v}_{\sigma^\epsilon} - M^{\epsilon,\mathcal{L}^\epsilon,v}_{\tau^\epsilon}\right)^2 \mid \mathcal{L}^\epsilon)\right) \\
&\leq \mathbb{E}\left(\mathbb{E}\left(\left\langle M^{\epsilon,\mathcal{L}^\epsilon,v}\right\rangle_{\sigma^\epsilon} - \left\langle M^{\epsilon,\mathcal{L}^\epsilon,v}\right\rangle_{\tau^\epsilon} \mid \mathcal{L}^\epsilon\right)\right) \\
&= \mathbb{E}\left(\int_{\sigma^\epsilon}^{\tau^\epsilon} m^{\epsilon,\mathcal{L}^\epsilon,v}(\Pi^\epsilon(s))ds\right).
\end{aligned}$$

where $m^{\epsilon,L^\epsilon,v}(\Pi)$ was defined in Lemma 4.2 and where the second line follows from the fact that τ^ϵ and σ^ϵ are stopping times for the filtration $(\mathcal{G}^\epsilon_t)_{t\geq 0}$.

If there exists C_1 such that

$$\sup_{L^\epsilon\in[0,1]^{l^\epsilon}} \sup_{\Pi\in(\mathcal{P}_d)^{l^\epsilon}} m^{\epsilon,L^\epsilon,v}(\Pi) \leq C_1, \tag{22}$$

then,

$$\mathbb{E}\left(|M^{\epsilon,v}_{\sigma^\epsilon} - M^{\epsilon,v}_{\tau^\epsilon}|\right) \;\leq\; \sqrt{C_1}\sqrt{\delta},$$

thus showing the desired inequality for the martingale part $M^{\epsilon,v}$. To prove (22), we recall the definition of $m^{\epsilon,L^\epsilon,v}(\Pi)$,

$$m^{\epsilon,L^\epsilon,v}(\Pi) = \frac{1}{\epsilon}\sum_{i,j=1}^{d} M_{ij}\ \mathbb{E}_{\lambda,L^\epsilon,j}\left(\left\langle \partial^S_{ij}X(\Pi), v\right\rangle^2\right) \;+\; b_\infty l^\epsilon \sum_{i=1}^{d}\mathbb{E}_{l^\epsilon}\left(\left\langle \partial^K_i X(\Pi), v\right\rangle^2\right).$$

The second term on the RHS can be bounded as in the proof of Proposition 4.3 (see (20)). For the first term on the RHS, we use the bound given by (21). We only need to prove that $\frac{1}{(l^\epsilon)^2\epsilon}\ \mathbb{E}_{\lambda,L^\epsilon,j}(\sum_{k=1}^{l^\epsilon}\sum_{k'=1}^{l^\epsilon} 1_{k\in S}1_{k'\in S})$ is bounded. Using (19),

$$\frac{1}{(l^\epsilon)^2\epsilon}\ \mathbb{E}_{\lambda,L^\epsilon,j}(\sum_{k=1}^{l^\epsilon}\sum_{k'=1}^{l^\epsilon} 1_{k\in S}1_{k'\in S}) \leq \frac{1}{(l^\epsilon)^2\epsilon}\ \frac{(l^\epsilon)^2}{n^\epsilon_j} \underset{\epsilon\to 0}{\longrightarrow} N_j,$$

so (22) is proved.

We now turn to the drift part. First, for every $L^\epsilon \in [0,1]^{l^\epsilon}$,

$$\left|B^{\epsilon,L^\epsilon,v}_{\sigma^\epsilon} \;-\; B^{\epsilon,L^\epsilon,v}_{\tau^\epsilon}\right| \;\leq\; \int_{\tau^\epsilon}^{\sigma^\epsilon} |\left\langle {}^tG^\epsilon X(\Pi^{\epsilon,L^\epsilon}(s)), v\right\rangle| ds.$$

We already showed in (16), that the integrand on the RHS is uniformly bounded in ϵ. Thus, there exists C_2 such that, for every $L^\epsilon \in [0,1]^{l^\epsilon}$:

$$\left|B^{\epsilon,L^\epsilon,v}_{\sigma^\epsilon} \;-\; B^{\epsilon,L^\epsilon,v}_{\tau^\epsilon}\right| \;\leq\; \delta C_2.$$

So,

$$\mathbb{E}\left(|B^{\epsilon,v}_{\sigma^\epsilon} \;-\; B^{\epsilon,v}_{\tau^\epsilon}|\right) \;=\; \mathbb{E}\left(\mathbb{E}\left(\left(|B^{\epsilon,\mathcal{L}^\epsilon,v}_{\sigma^\epsilon} \;-\; B^{\epsilon,\mathcal{L}^\epsilon,v}_{\tau^\epsilon}|\right) \mid \mathcal{L}^\epsilon\right)\right) \leq \delta C_2.$$

which is the desired inequality. This completes the proof of Proposition 4.4.

Remark 4.7. *Notice that the tightness (and the convergence) does not depend on the recombination rate. However, for small values of λ, or if L^ϵ is such that the positions of the loci are all very close to each other, correlations between loci are very high. This means that, when a migration event takes place, either no locus will be fixed (with high probability), or almost all loci from the migrant will be fixed. Therefore, if we let $\lambda \to 0$ be the process of the genetics distances converges to a process that increases continuously (due to mutation) and has negative jumps (due to migration events). See Figure 6 for a numerical simulation.*

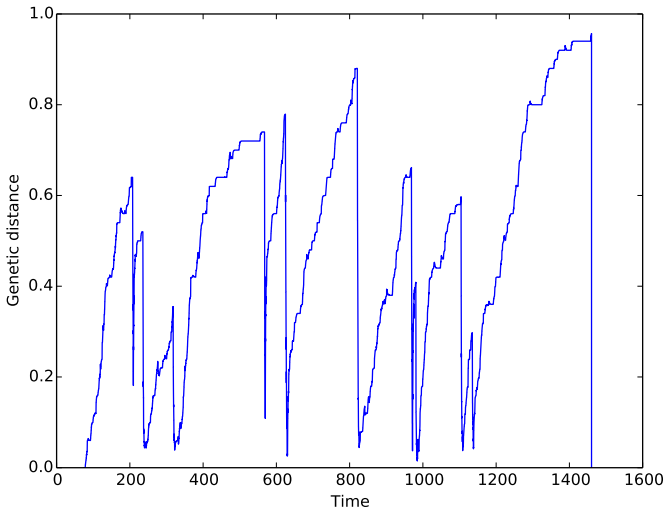


Figure 6: Simulation of the individual based model, for $d = 2$, $N_1 = N_2 = 1$, $\epsilon = 0.01$, $\gamma = 0.005$, $l^\epsilon = 100$, $\lambda = 0.5$. With this set of parameters, Theorem 1.1, predicts that the genetic distance at equilibrium should be 0.5. In this simulation, the mean genetic distance is 0.5.

5 Proof of Theorem 1.1 and more

In this section we state and prove a stronger version of Theorem 1.1 (see Theorem 5.1 below). As in Theorem 1.1, we consider, for each pair of subpopulations $i, j \in E$, S^i and S^j, two independent random walks on E starting respectively from i and j and whose transition rate from k to p is equal to $\tilde{M}_{kp}$. We have the following generalization of of Theorem 1.1.

Theorem 5.1. *Assume that*

- *At time 0, in the IBM, subpopulations are homogeneous and that, the genetic partition measure of the population (in the associated PBM) is given by ξ_0^ϵ, a deterministic probability measure in $\mathcal{P}_d$.*
- *There exists a probability measure $P^0 \in \mathcal{M}_d$ such that the following convergence holds:*

$$\xi_0^\epsilon \underset{\epsilon \to 0}{\longrightarrow} P^0. \tag{23}$$

For every $t \geq 0$, define

$$D_t(i,j) \;:=\; 1 \;-\; \int_0^t e^{-2b_\infty s}\mathbb{P}(\tau_{ij} \in ds) \;-\; \int_\pi e^{-2b_\infty t}\mathbb{P}\left(\tau_{ij} > t, S^i(t) \sim_\pi S^j(t)\right) P^0(d\pi)$$

where $\tau_{ij} = \inf\{t \geq 0 \ : \ S^i(t) = S^j(t)\}$. *Then,*

$\lim_{\epsilon\to 0} \lim_{\gamma\to 0}(d^{\gamma,\epsilon}_{t/(\gamma\epsilon)}(i,j),\ t \geq 0) \ = \ (D_t(i,j),\ t \geq 0)$ *in the sense of finite dimensional distributions.*

In particular,

$$\lim_{t\to\infty} D_t(i,j) \ = \ 1 - \mathbb{E}(e^{-2b_\infty \tau_{ij}}).$$

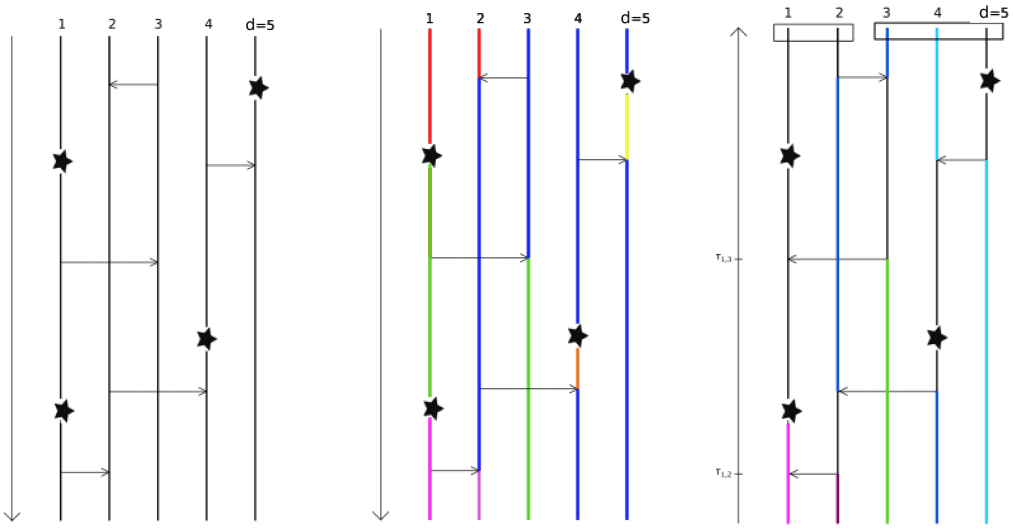
Proof. We start by proving that,

$$(d^\epsilon_t(i,j); t \geq 0) \underset{\epsilon\to 0}{\Longrightarrow} (D_t(i,j); t \geq 0) \ \text{ in the weak topology,} \tag{24}$$

where $(D_t(i,j); t \geq 0)$ is the deterministic process defined in Theorem 1.1.

From equation (5) and Theorem 3.1 we get that $\forall i,j \in E$, $(d^\epsilon_t(i,j);\ t \geq 0)$ converges in distribution in the weak topology to $(1 - P_t(\pi \in \mathcal{P}_d, i \sim_\pi j);\ t \geq 0)$. It remains to show that this expression is identical to the one provided in Theorem 1.1. This is done in a standard way by using the graphical representation associated to the one-locus Moran model whose generator is specified by G (defined in (7)). It is well known that such a Moran model is encoded by a graphical representation that is generated by a sequence of independent Poisson Point Processes as follows:

- B^i, with intensity measure $b_\infty dt$, that corresponds to mutation events at site i. If $b^i \in B^i$, at (i, b^i) we draw a $\star$ in the graphical representation (Figure 7(a)).
- $T^{i,j}$, with intensity measure $\tilde{M}_{ji}dt$, that corresponds to reproduction events, where j is replaced by i. If $t^{i,j} \in T^{i,j}$, we draw an arrow from $(i, t^{i,j})$ to $(j, t^{i,j})$ in the graphical representation to indicate that lineage j inherits the type of lineage i.



(a) Graphical representation of the Moran model

(b) Genetic partitioning process

(c) Ancestral lineages

Figure 7: Realisation of the genetic partitioning model and its dual. In Figure (b), colours indicate genetic types (that induce the partitions). In Figure (c), colours represent the ancestral lineages.

We now give a characterisation of the dual process starting at t. We define $(S^1_t, S^2_t, \ldots, S^d_t)$ a sequence of piecewise continuous functions $[0,t] \to E$, where $\forall i \in E$, S^i_t represents the ancestral

lineage of individual i (sampled at time t). $S_t^i(t) = i$ and as time proceeds backwards, each time S_t^i encounters the tip of an arrow it jumps to the origin of the arrow. It is not hard to see that S_t^i is distributed as a random walk started at i and with transition rates from k to p equal to $\tilde{M}_{kp}$ and that $(S_t^1, S_t^2, \ldots, S_t^d)$ are distributed as coalescing random walks running backwards in time, i.e. they are independent when appart and become perfectly correlated when meeting each other. In Figure 7(c), $S_t^1, S_t^2, \ldots, S_t^d$ are represented in different colours.

Define $\tau_{ij}^t = \inf\{s \geq 0,\ S_t^i(t-s) = S_t^j(t-s)\}$. By looking carefully at Figures 7(b) and 7(c), we let the reader convince herself that two individuals i and j have the same type at time t iff:

(i) $\tau_{ij}^t \leq t$ and there are no $\star$ in the paths of S_t^i and S_t^j before τ_{ij}^t, or

(ii) $\tau_{ij}^t \geq t$ and $S^i(0) \sim_{\pi_0} S^j(0)$ and S_t^i and S_t^j their is no $\star$ in their paths, where π_0 is the initial genetic partition of the metapopulation, that is random partition of law P^0.

From here, it is easy to check that:

$$\begin{aligned} D_t(i,j) &= 1 - P_t(\{\pi, i \sim_\pi j) \\ &= 1 - \int_0^t e^{-2b_\infty s}\mathbb{P}(\tau_{ij} \in ds)ds - \int_\pi e^{-2b_\infty t}\mathbb{P}\left(\tau_{ij} > t, S^i(t) \sim_\pi S^j(t)\right) P^0(d\pi). \end{aligned}$$

As $\forall i, j \in E$, $(D_t(i,j))$ is continuous, the fact that $(d_t^\epsilon(i,j))$ converges in distribution (in the weak topology) to $(D_t(i,j))$ (24) implies (by the continuous mapping theorem) that $(d_t^\epsilon(i,j))$ converges to $(D_t(i,j))$ in the sense of finite dimensional distributions, as $\epsilon \to 0$.

This result, combined with Theorem 2.2, also implies that:

$$\lim_{\epsilon\to 0}\lim_{\gamma\to 0}(d^{\gamma,\epsilon}_{t/(\gamma\epsilon)}(i,j),\ t \geq 0) = D_t(i,j),\ t \geq 0) \text{ in the sense of finite dimensional distributions.}$$

The fact that $\lim_{t\to\infty} D_t(i,j) = 1 - \mathbb{E}(e^{-2b_\infty \tau_{ij}})$ is a direct consequence of the definition of $(D_t(i,j); t \geq 0)$ and the dominated convergence theorem.

This completes the proof of Theorem 1.1. □

6 An example: a population with a geographic bottleneck

Fix $d \in \mathbb{N} \setminus \{0\}$. We let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two complete graphs of d vertices. We link the two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ by adding an extra edge (v_1, v_2), where $v_k, k = 1, 2$ is a given vertex in $\mathcal{G}_k$. We call $\mathcal{G}$ the resulting graph. We equip $\mathcal{G}$ with the following migration rates: if i is connected to j, then $M_{ij} = 1/d$ (so that the emigration rate from any vertex i is 1 if $i \neq v_1, v_2$ and $1 + \frac{1}{d}$ otherwise). We also assume that $N_i = 1$, so that $\tilde{M}_{ij} = 1/d$.

We think of $\mathcal{G}$ as two well-mixed populations connected by a single geographic bottleneck.

Theorem 6.1. *Fix $c > 0$, $b_\infty = \frac{c}{d}$. Then for any two neighbours $i, j \in \mathcal{G}$*

$$1 - \mathbb{E}\left(\exp(-2b_\infty \tau_{ij})\right) = \begin{cases} \frac{c}{1+c} + o(1) & \text{if } i,j \in \mathcal{G}_1, \text{ or if } i,j \in \mathcal{G}_2 \\ 1 - \frac{1}{d} + o(\frac{1}{d}) & \text{if } i = v_1 \text{ and } j = v_2. \end{cases}$$

Before going into the details of the proof, we present some heuristics for the formulae. First, if i and j belong to the same subgraph, we can assume that most of the time the two random walks hit each other before hitting the other subgraph. So we can consider $\tilde{S}^i$ and $\tilde{S}^j$, two random walks on a complete graph with d vertices. If J is the number of jumps made by $\tilde{S}^i$ and $\tilde{S}^j$ before hitting each other, J follows a geometric distribution of parameter $1/d$, which properly renormalized converges to an exponential distribution of parameter 1 (when $d \gg 1$). In addition, the mean time between two consecutive jumps (of S^i or S^j) is $2 \times 1/d$ so the distribution of τ_{ij} can be approximated by an exponential distribution of parameter $2/d$. If e

is an exponentially distributed random variable, with parameter $2/d$, then $\mathbb{E}\left(\exp(-2b_\infty e)\right) = 1/(1+2b_\infty/d) = 1/(1+c)$ which gives the desired result. Second, $i = v_1$ and $j = v_2$, with probability $1/(d+1) \simeq 1/d$, the first jump is from v_1 to v_2 (or v_2 to v_1), so the two random walks hit very fast, and the genetic distance is close to 0. Otherwise, each random walk "gets lost" in its subgraph and the hitting time becomes very large. In that case, the genetic distance is approximatively 1.

Proof. We give a brief sketch of the computations since the method is rather standard. We start with some general considerations. Consider a general meta-population with $\bar{d}$ subpopulations. Define $a(i,j) = \mathbb{E}\left(\exp(-2b_\infty \tau_{ij})\right)$. By conditioning on every possible move of the two walks on the small time interval $[0, dt]$, it is not hard to show that the $a(i,j)$'s satisfy the following system of linear equations: $\forall i \in \{1,\dots,\bar{d}\}$, $a(i,i) = 1$ and $\forall i,j \in \{1,\dots,\bar{d}\}$ with $i \neq j$:

$$0 \;=\; \sum_{k=1}^{\bar{d}} \left(a(k,j)\tilde{M}_{ik} + a(i,k)\tilde{M}_{jk} \right) \;-\; a(i,j) \left(\sum_{k=1}^{\bar{d}} (\tilde{M}_{ik} + \tilde{M}_{kj}) + 2b_\infty \right). \tag{25}$$

Let us now go back to our specific case (in particular $\bar{d} = 2d$). We distinguish between two types of points: the boundary points (either v_1 or v_2), and the interior points of the subgraphs $\mathcal{G}_1$ and $\mathcal{G}_2$ (points that are distinct from v_1 and v_2). For (i,j), with $i \neq j$, we say that (i,j) is of type

- (II) if the vertices belong to the interior of the same subgraph (either $\mathcal{G}_1$ or $\mathcal{G}_2$).
- $(I\bar{I})$ if the vertices belong to the interior of distinct subgraphs.
- (IB) if one of the vertex is in the interior of a subgraph, and the other vertex belongs to the boundary point of the same subgraph.
- $(I\bar{B})$, $(B\bar{B})$ are defined analogously.

By symmetry, $a(i,j)$ is invariant in each of those classes of pairs of points. We denote by $a(II)$ the value of $a(i,j)$ for (i,j) in (II). $a(I\bar{I}), a(IB), a(I\bar{B}), a(B\bar{B})$ are defined analogously. From this observation, we can inject those quantities in (25): this reduces the dimension of the linear problem from $\bar{d}(\bar{d}-1)$ to only 5. The system can then be solved explicitly and straightforward asymptotics yield Theorem 6.1.

□

Acknowledgements

We would like to thank the Editor and the reviewers for their useful comments on the previous version of the manuscript. We also thank Florence Débarre and Amaury Lambert for helpful discussions.