Université Paris-Sud

École doctorale de mathématiques Hadamard (ED 574)
CMAP - École polytechnique (UMR 7641 CNRS)

Mémoire présenté pour l'obtention du

Diplôme d'habilitation à diriger les recherches

Discipline : Mathématiques

par

Amandine VÉBER

Modèles structurés de transmission

Rapporteurs :

Matthias BIRKNER
Amaury LAMBERT
Vlada LIMIC

Date de soutenance : 17 novembre 2017

Composition du jury :

Christophe GIRAUD	(Examinateur)
Yueyun HU	(Examinateur)
Amaury LAMBERT	(Rapporteur)
Catherine LAREDO	(Examinatrice)
Jean-François LE GALL	(Examinateur)
Sylvie MELEARD	(Examinatrice)

Modèles structurés de transmission

Cette thèse d'habilitation aborde deux thèmes. Le premier et principal est celui de la modélisation et de la compréhension de l'évolution de la diversité génétique d'une population, en particulier lorsque cette population a une structure spatiale continue. Dans le chapitre 2, nous présentons les résultats obtenus la plupart du temps dans le cadre très général du *processus* Λ*-Fleming-Viot spatial*. Ce processus à valeurs mesures est un outil très flexible pour étudier l'évolution d'une population vivant dans un espace continu de dimension d, $d = 2$ étant évidemment la dimension la plus pertinente pour les applications biologiques. Le chapitre 3 présente quelques résultats sur le *pédigré* d'une population avec et sans structure. Plus précisément, nous décrivons les relations généalogiques au sein d'un échantillon d'individus diploïdes du point de vue de leur parenté « physique », ces parents étant ensuite ou non des ancêtres génétiques de l'échantillon. Dans le chapitre 4, nous discutons une nouvelle approche pour la reconstruction de paramètres démographiques affectant la distribution des arbres généalogiques, et donc de la diversité génétique, de la même manière quel que soit l'endroit du génome considéré. Cette approche est fondée sur l'idée simple de simuler les arbres généalogiques à leur *résolution* optimale, qui dépend du type de données disponibles pour l'inférence.

Le second thème concerne la modélisation d'un certain type de partage de ressources dans des réseaux de communication. Ceci correspond au chapitre 5, dans lequel nous analysons le comportement en temps long et la stabilité d'un système de files d'attente régulées à travers un graphe d'incompatibilités de service. Par ailleurs, afin d'empêcher une file particulièrement grande de monopoliser la capacité de service, nous supposons que chaque file est servie à un taux proportionnel au logarithme de sa taille. Nous décrivons la *limite fluide* du système lorsque la taille de l'une des files tend vers l'infini, en détaillant le rôle de différentes échelles de temps intermédiaires dans le comportement asymptotique obtenu. Nous commençons par le cas d'un réseau d'incompatibilités correspondant au graphe complet (i.e., une seule file peut être servie à la fois), puis nous considérons le réseau en étoile (les files périphériques n'interférant qu'avec la file centrale).

Structured models of transmission

This habilitation thesis revolves around two topics. The first and major one is the modelling and understanding of the evolution of the genetic diversity in a population, in particular when this population has a continuous spatial structure. In Chapter 2, we review the different results obtained most of the time in the quite general framework of the *spatial Λ-Fleming-Viot process.* This measure-valued process is a very flexible tool to study the evolution of a population living in some continuous d-dimensional space, $d = 2$ being of course the most relevant dimension for biological applications. Chapter 3 presents some results on the *pedigree* of a population with and without structure. More precisely, we describe the genealogical relationships within a sample of diploid individuals at the level of their physical ancestry, these progenitors being then, or not being, genetic ancestors to the sample. In Chapter 4, we discuss a new approach for the reconstruction of demographic parameters affecting the distribution of the genealogical trees, and thus of the genetic diversities, across the whole genome. It is based on the simple idea of simulating the genealogical trees at their optimal *resolution*, which depends on the kind of data available for inference.

The second topic concerns the modelling of a certain type of resource sharing in communication networks. This corresponds to Chapter 5, in which we analyse the long-term behaviour and stability of a system of queues regulated through a network of service incompatibilities. Furthermore, in order to prevent a very busy node to monopolise the service capacity, we assume that each queue is served at a rate which is proportional to the logarithm of its current size. We describe the *fluid limit* of the system as the size of one of the queues tends to infinity, disentangling the roles of different intermediate timescales in the asymptotic behaviour obtained. This is done first in the case where the network of incompatibilities is the complete graph (i.e., only one queue can be served at a time), and then for the star network (the peripheral queues interfering only with the central queue).

Production scientifique

Modèles de génétique des populations et arbres aléatoires :

[TV09] J.E. Taylor et A. Véber (2009). Coalescent processes in subdivided populations subject to recurrent mass extinctions. *Electron. J. Probab.*, 14 : 242–288.

[BEV10] N.H. Barton, A.M. Etheridge et A. Véber (2010). A new model for evolution in a spatial continuum. *Electron. J. Probab.*, 15 : 162–216.

[EV12] A.M. Etheridge et A. Véber (2012). The spatial Lambda-Fleming-Viot process on a large torus : genealogies in the presence of recombination. *Ann. Applied Probab*, 22 : 2165–2209.

[BEV13a] N. Berestycki, A.M. Etheridge et A. Véber (2013). Large scale behaviour of the spatial Lambda-Fleming-Viot process. *Ann. Inst. H. Poincaré Probab. Statist.*, 49 : 374–401.

[BEV13b] N.H. Barton, A.M. Etheridge et A. Véber (2013). Modelling evolution in a spatial continuum. *JSTAT*, P01002.

[BEKV13a] N.H. Barton, A.M. Etheridge, J. Kelleher et A. Véber (2013). Inference in two dimensions : allele frequencies versus lengths of shared sequence blocks. *Theor. Pop. Biol.*, 87 : 105–119.

[BEKV13b] N.H. Barton, A.M. Etheridge, J. Kelleher et A. Véber (2013). Genetic hitchhiking in spatially extended populations. *Theor. Pop. Biol.*, 87 : 75–89.

[VW15] A. Véber et A. Wakolbinger (2015). The spatial Lambda-Fleming-Viot process : an event-based construction and a look-down representation. *Ann. Inst. H. Poincaré Probab. Statist.*, 51 : 570–598.

[SSV15] R. Sainudiin, T. Stadler et A. Véber (2015). Finding the best resolution for the Kingman-Tajima coalescent : theory and applications. *J. Math. Biol.*, 70 : 1207–1247.

[STV16] R. Sainudiin, B. Thatte et A. Véber (2016). Ancestries of a recombining diploid population. *J. Math. Biol.*, 72 : 363–408.

[KEVB16] J. Kelleher, A.M. Etheridge, A. Véber et N.H. Barton (2016). Spread of pedigree versus genetic ancestry in spatially distributed populations. *Theor. Pop. Biol.*, 108 : 1–12.

[SV16] R. Sainudiin et A. Véber (2016). A Beta-splitting model for evolutionary trees. *R. Soc. open sci.*, 3 :160016.

[BEV17] N.H. Barton, A.M. Etheridge et A. Véber (2017). The infinitesimal model : definition, derivation, and implications. *Theor. Pop. Biol.*, à paraître.

[SV17] R. Sainudiin et A. Véber (2017). Full likelihood inference from the site frequency spectrum of a non-recombining locus. *bioRxiv preprint 181412.*

[EVY17] A.M. Etheridge, A. Véber et F. Yu (2017). Rescaling limits of the spatial Lambda-Fleming-Viot process with selection. *En préparation.*

[PVWR17] J. Palacios, A. Véber, J. Wakeley et S. Ramachandran (2017). BESTT : Bayesian Estimation by Sampling Tajima's Trees. *En préparation.*

Processus de branchement :

[Véb09] A. Véber (2009). Quenched convergence of a sequence of superprocesses in $\mathbb{R}^d$ among Poissonian obstacles. *Stochastic Process. Appl.*, 119 : 2598–2624.
[LGV12] J.-F. Le Gall et A. Véber (2012). Escape probabilities for branching Brownian motion among mild obstacles. *J. Theor. Probab.*, 25 : 505–535.
[BFV13] C. Bouillaguet, P.-A. Fouque et A. Véber (2013). Graph-theoretic algorithms for the isomorphism of polynomials problem. *Eurocrypt 2013.*

Réseaux de communication :

[RV15] P. Robert et A. Véber (2015). A stochastic analysis of resource sharing with logarithmic weights. *Ann. Applied Probab.*, 25 : 2626–2670.
[RV16] P. Robert et A. Véber (2016). A scaling analysis of a star network with logarithmic weights. *arXiv preprint 1609.04180.*

Vulgarisation scientifique :

[Véb10] A. Véber (2010). Théorèmes limites pour des processus de branchement et de coalescence spatiaux. *MATAPLI*, 92 : 53–60.
[BMV13] V. Bansaye, S. Méléard et A. Véber (2013). Les différentes échelles de temps de l'évolution. *MATAPLI*, 100 : 101–116.

Thèse de doctorat :

[Véb09] A. Véber (2009). Théorèmes limites pour des processus de branchement et de coalescence spatiaux. *Thèse de doctorat*, Université Paris-Sud.

Remerciements

Pour commencer, je suis extrêmement reconnaissante envers Vlada Limic, Amaury Lambert et Matthias Birkner d'avoir accepté de rapporter ce mémoire d'habilitation. Leurs travaux sont depuis longtemps une source d'inspiration et leur avis sur mon travail m'importe beaucoup. J'aimerais également remercier chaleureusement Christophe Giraud, Yueyun Hu, Catherine Laredo, Jean-François Le Gall et Sylvie Méléard d'avoir accepté d'apporter leur expertise lors de la soutenance, c'est un honneur de vous compter parmi les membres du jury. Un grand merci aussi à Frédéric Paulin pour la diligence et l'enthousiasme avec lesquels il a veillé au bon déroulement de toutes les étapes, du projet d'HDR à la soutenance.

Bien entendu, c'est à mes collaborateurs et mes collègues que je dois ma principale source de questions et d'idées, ainsi qu'une grande partie du plaisir de faire de la recherche et d'enseigner. J'ai appris avec eux que même les pauses cafés peuvent être un moment de partage scientifique (ou non-scientifique, j'avoue...) et que la convivialité était un élément essentiel de la transmission des savoirs. Le CMAP est en cela un laboratoire particulièrement agréable et stimulant, un immense merci à tous (enseignants-/chercheurs de tous statuts et équipe administrative) ! Une pensée toute particulière va bien sûr aux membres de l'équipe PEIPS, qui porte très bien son nom, et tout particulièrement à Sylvie Méléard et Vincent Bansaye à qui je dois énormément de tous points de vue. J'espère que nous finirons par cosigner autre chose qu'un article de vulgarisation mais, en attendant, c'est toujours un plaisir d'échanger des idées, encadrer des étudiants ou monter un cours avec eux ! Merci également à Thierry Bodineau, Lucas Gerin, Carl Graham, Igor Kortchemski et Gaël Raoul, aux (post-)doctorants de l'équipe, ainsi qu'à tous les collègues avec qui j'ai l'occasion d'échanger régulièrement à travers la chaire Modélisation Mathématique et Biodiversité, chaire qui soutient généreusement mes activités de recherche et d'enseignement depuis mon arrivée au CMAP.

Je suis évidemment profondément reconnaissante envers Alison Etheridge et Nick Barton pour m'avoir embarquée dans cette grande aventure de la modélisation en génétique des populations ; j'apprends toujours énormément de nos interactions. Merci également à Nathanaël Berestycki, Raphaël Forien, Jerome Kelleher, Julia Palacios, Sohini Ramachandran, Philippe Robert, Raazesh Sainudiin, Tanja Stadler, Jay Taylor, Bhalchandra Thatte, John Wakeley, Anton Wakolbinger et Feng Yu, pour nos discussions riches et les papiers qui en ont résulté (merci également à tous ceux avec qui les discussions n'ont pas encore donné lieu à un article !). Merci pour finir à Jean-François Le Gall, pour tout ce que j'avais appris en thèse grâce à lui et qui me reste encore aujourd'hui (notamment un goût prononcé pour les exposés au tableau, même si l'élève n'égalera jamais le maître...).

Last but certainly not least, merci à mes amis, ma famille, mes amours (même les plus matinaux...) : les maths seraient bien moins belles sans vous.

Table des matières

Chapitre 1

Diversité génétique au sein d'une population structurée en espace

Modéliser l'évolution d'une population au cours des nombreuses générations ayant mené à son état actuel est une tâche très complexe, car de multiples facteurs peuvent avoir influencé, parfois de manière transitoire, la variabilité génétique observée dans un échantillon d'individus. En effet, la population peut avoir subi des fluctuations démographiques telles que des goulots d'étranglement ou des périodes d'expansion, des épisodes de *balayages sélectifs*, elle peut être (ou avoir été) structurée spatialement, ou soumise à des fluctuations de son environnement, ... L'un des principaux rôles des modèles mathématiques est d'étudier les effets de chacun de ces facteurs, ou d'une combinaison d'entre eux, pour mettre en évidence la signature qu'ils laisseraient dans la diversité génétique actuelle s'ils avaient contribué à l'évolution de la population à un niveau suffisant.

Bien entendu, de nombreux modèles existent déjà et le lecteur intéressé pourra consulter l'ouvrage [37] pour une présentation claire des plus classiques d'entre eux. La plupart de ces modèles n'ont pas pour but de décrire la biologie des organismes dans le détail, mais sont basés sur des approximations parfois très grossières de la manière dont les gènes sont transmis de parents à enfants. Cependant, leur force réside dans leur capacité à rendre de compte de l'évolution de la population au cours des dizaines, centaines, voire milliers de générations pendant lesquelles la diversité génétique actuelle de la population s'est construite.

1.1 Le modèle de Wright-Fisher et le coalescent de Kingman

Le modèle d'évolution génétique le plus utilisé est le modèle de Wright-Fisher [47, 118]. Il suppose que les individus sont *haploïdes* (i.e., ils n'ont qu'une seule copie de chaque chromosome), la population n'est structurée d'aucune manière (elle est *panmictique*) et l'évolution se produit par générations discrètes au cours desquelles la taille de la population reste constante égale à un certain (grand) nombre N. Si nous supposons en outre que le gène qui nous intéresse est neutre, au sens où aucune de ses versions possibles, ou *allèles*, ne confère d'avantage reproductif aux individus qui la portent, alors le mécanisme de reproduction s'exprime ainsi : pour former la génération $t+1$, chacun des N individus vivant à ce moment « choisit » un parent uniformément au hasard parmi la génération précédente, indépendamment les uns des autres, et chaque descendant hérite de l'allèle de son parent. La figure 1.1 montre un exemple dans lequel $N = 7$.

Si le gène a seulement deux allèles, notés A et a, il suffit alors de suivre la proportion $p^N(t)$

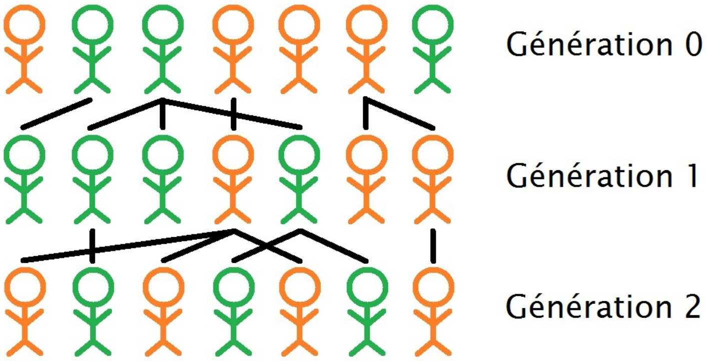


FIGURE 1.1 – Le modèle de Wright-Fisher avec $N = 7$. Les individus peuvent être de deux types génétiques (orange et vert ici) qui sont transmis de parents à descendants (symbolisé par les lignes noires).

de l'allèle A à la génération t pour décrire complètement l'évolution de la diversité génétique de la population. Dans le cas neutre présenté ici, par construction nous avons que pour tout $t \in \mathbb{N}$, conditionnellement à $p^N(t)$,

$$Np^N(t+1) \sim \text{Binomiale}\big(N, p^N(t)\big). \tag{1.1}$$

Le comportement en temps long de la chaîne de Markov $(p^N(t))_{t\in\mathbb{N}}$ s'obtient aisément : avec probabilité un, l'un des allèles *se fixe* dans (i.e., envahit) la population et la probabilité que l'allèle chanceux soit A est égale à sa proportion initiale $p^N(0)$. Cependant, obtenir des informations plus précises sur le chemin menant à la fixation, par exemple la loi du temps nécessaire pour que p^N atteigne 0 ou 1, devient un problème combinatoire de plus en plus difficile au fur et à mesure que la taille de la population augmente. Puisque nous nous intéressons à de grandes populations, l'astuce mathématique usuelle consiste à faire tendre N vers l'infini et à voir si nous pouvons obtenir un objet limite qui soit une bonne approximation de l'évolution d'une population de taille très grande mais finie.

En utilisant (1.1), nous voyons immédiatement que pour tout $t \in \mathbb{N}$, nous avons

$$\mathbb{E}\big[p^N(t+1)\,|\,p^N(t)\big] = p^N(t) \quad \text{et} \quad \text{Var}\big(p^N(t+1)\,|\,p^N(t)\big) = \frac{1}{N}\,p^N(t)\big(1-p^N(t)\big).$$

Par conséquent, si nous faisons tendre N vers l'infini et si nous faisons l'hypothèse que la proportion initiale $p^N(0)$ d'individus portant l'allèle A converge vers une constante $p \in [0,1]$, alors le processus p^N converge en loi vers le processus constant égal à p. En d'autres termes, les fréquences de chaque allèle ne varient plus dans la population limite de taille infinie. En réfléchissant un peu à ce qu'il se passe ici, nous voyons que la variance de la fluctuation de p^N sur une génération est d'ordre $1/N$. Par analogie avec la convergence d'une marche aléatoire changée d'échelle vers le mouvement brownien, nous nous attendons en fait à avoir à attendre un nombre de générations d'ordre N avant d'observer des variations macroscopiques de p^N. Ceci motive l'introduction du processus $\tilde{p}^N$ défini comme suit :

$$\tilde{p}^N_t := p^N(\lfloor Nt \rfloor), \quad t \in \mathbb{R}_+, \tag{1.2}$$

où $\lfloor x \rfloor$ est la partie entière de $x \in \mathbb{R}$, de sorte qu'une unité de temps pour $\tilde{p}^N$ correspond à N générations. Si nous faisons tendre N vers l'infini à nouveau, cette fois la suite de processus $(\tilde{p}^N)_{N\geq 1}$ converge en loi (dans l'espace de Skorokhod $D_{[0,1]}[0,\infty)$ des trajectoires càdlàg à valeurs dans $[0,1]$) vers la *diffusion de Wright-Fisher*, unique solution de l'équation différentielle stochastique

$$\mathrm{d}p_t = \sqrt{p_t(1-p_t)}\,\mathrm{d}B_t, \qquad p_0 = p, \tag{1.3}$$

où $(B_t)_{t\geq 0}$ est un mouvement brownien standard. Une preuve de cette convergence se trouve dans le chapitre 3.2 de [37]. Les fluctuations aléatoires décrites par (1.3) correspondent au phénomène appelé *dérive génétique* dans la littérature biologiste. Comme dans le cas d'une population finie, elles sont causées par le remplacement aléatoire d'individus d'allèles a par des descendants portant l'allèle A et inversement, ce qui explique que la variance infinitésimale soit proportionnelle au produit des fréquences des individus A et a. Grâce aux résultats généraux du calcul stochastique, nous pouvons alors montrer à nouveau que l'un des allèles se fixe dans la population en temps fini p.s., cet allèle étant A avec probabilité p_0, et de nombreuses autres propriétés sont maintenant accessibles (temps moyen de fixation, lois conditionnelles pour les trajectoires, ...).

En plus de caractériser le comportement à long terme d'une population infinie évoluant selon le modèle de Wright-Fisher, les résultats présentés dans le paragraphe précédent mettent en avant le fait que l'évolution peut n'avoir un réel impact que sur une échelle de temps très longue, de l'ordre de la taille de la population *dans ce modèle*. Nous verrons que trouver les échelles temporelles et spatiales sur lesquelles nous pourrions observer un comportement non trivial constituera une étape clé dans l'analyse des modèles étudiés dans les prochains chapitres (y compris le chapitre 5). Ceci pourrait ressembler à un jeu un peu artificiel. Cependant, la véritable question que cette approche pose est la suivante : *en supposant que nous observons une certaine forme de diversité génétique, quels sont les ordres de grandeur des différentes forces évolutionnaires qui pourraient l'expliquer ?*

Une approche particulièrement fructueuse pour l'étude du modèle de Wright-Fisher consiste à retourner la flèche du temps et à essayer de comprendre la forme de la généalogie de quelques individus échantillonnés au hasard dans la population actuelle. En effet, plus deux individus doivent remonter loin dans le passé pour trouver un premier ancêtre commun, plus il y a de temps pour que des mutations ou des recombinaisons se produisent, laissant potentiellement une trace détectable lorsque nous comparons leurs génomes. Sous les hypothèses du modèle de Wright-Fisher neutre avec une taille de population N, l'ancêtre commun le plus récent de deux individus échantillonnés uniformément au hasard dans la génération actuelle se trouve T_2^N générations dans le passé, où T_2^N suit une loi géométrique de paramètre $1/N$. Par conséquent, comme dans l'analyse des fréquences d'allèles les relations ancestrales entre les individus évoluent réellement sur l'échelle de temps $(Nt,\, t\geq 0)$ et en effectuant le même changement d'échelle que précédemment, nous obtenons que le *temps de coalescence* des lignées ancestrales de deux individus sur la nouvelle échelle, $\tilde{T}_2^N = T_2^N/N$, satisfait

$$\tilde{T}_2^N \xrightarrow{(\mathrm{d})} T_2 \quad \text{lorsque } N\to\infty, \tag{1.4}$$

où T_2 suit une loi exponentielle de paramètre 1.

Considérons à présent un échantillon de taille $n\geq 2$, pris à nouveau uniformément au hasard dans la génération actuelle. Pour simplifier les notations, nous appellerons cette génération 0. Puisque nous voulons décrire qui partage un ancêtre avec qui t générations dans le passé, une manière naturelle de représenter l'état des relations ancestrales au sein de l'échantillon à la génération $-t$ est d'utiliser une partition $\pi^N(t)$ de $[n] := \{1,\ldots,n\}$ telle que i et j

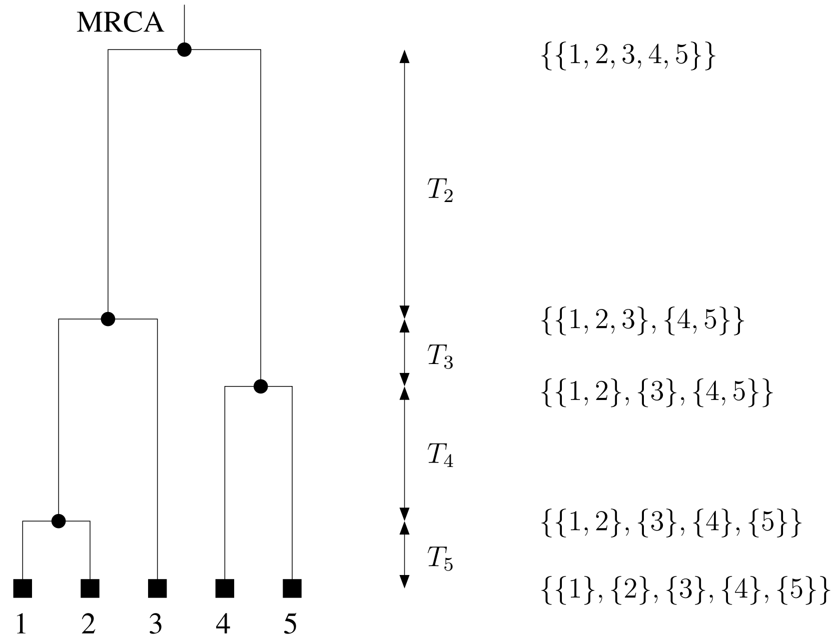


FIGURE 1.2 – Représentations en arbre et en partitions de la généalogie d'un échantillon de taille 5. Le temps généalogique va du bas vers le haut. Les carrés noirs représentent les individus échantillonnés dans la population actuelle. Pendant le laps de temps T_i, la partition ancestrale a i blocs, correspondant aux i arêtes (verticales) dans cette couche de l'arbre. Le Plus Récent Ancêtre Commun de l'échantillon est atteint au bout de $T_5 + \ldots + T_2$ unités de temps.

appartiennent au même bloc de $\pi^N(t)$ si et seulement si le $i$ème et $j$ème individu de l'échantillon ont le même ancêtre t générations dans le passé. La figure 1.2 montre un exemple. De cette manière, pour tout N nous obtenons une chaîne de Markov $(\pi^N(t))_{t\in\mathbb{N}}$ à valeurs dans l'ensemble $\mathcal{P}_n$ de toutes les partitions de $[n]$, partant de la partition en singletons $\{\{1\},\ldots,\{n\}\}$ et dont l'unique état absorbant est la partition triviale $\{\{1,\ldots,n\}\}$. Celle-ci correspond à l'état dans lequel tous les individus de l'échantillon partagent le même ancêtre et le nombre aléatoire T^N_{MRCA} de générations à remonter avant que cet événement ne se produise est appelé *temps du plus récent ancêtre commun* (*Most Recent Common Ancestor* en anglais). Si nous effectuons le même changement d'échelle de temps, il n'est pas difficile de montrer que

$$\tilde{\pi}^N := \big(\pi^N(\lfloor Nt \rfloor)\big)_{t\in\mathbb{R}_+} \xrightarrow{(\mathrm{d})} (\pi_t)_{t\in\mathbb{R}_+} \quad \text{lorsque } N \to \infty, \tag{1.5}$$

où $(\pi_t)_{t\geq 0}$ est le *coalescent de Kingman.* L'objet limite, formellement introduit dans l'article [67], est un processus de sauts markovien à valeurs dans $\mathcal{P}_n$ qui peut être décrit de la manière suivante : son état initial est la partition de $[n]$ en singletons, puis chaque paire de blocs tente de fusionner à taux 1 jusqu'à ce que l'état final $\{\{1,\ldots,n\}\}$ soit atteint. Par conséquent, lorsqu'il y a k blocs dans la partition ancestrale, le temps à attendre avant que la prochaine coalescence ne se produise suit une loi exponentielle de paramètre $\binom{k}{2}$ et cet événement résulte en la fusion d'une paire de blocs choisie uniformément au hasard parmi les $\binom{k}{2}$ paires possibles. De nombreuses propriétés du processus limite sont connues, cf. [34] et [110] par exemple. Une propriété particulièrement intéressante est la *relation de dualité* entre la diffusion de Wright-Fisher (1.3) et le coalescent de Kingman. En effet, si nous notons $|\pi|$ le nombre de blocs dans la partition π, alors pour tout $n \in \mathbb{N}$, $p_0 \in [0,1]$ et $t \in \mathbb{R}_+$,

$$\mathbb{E}_{p_0}\big[p_t^n\big] = \mathbb{E}_n\big[p_0^{|\pi_t|}\big], \tag{1.6}$$

où par abus de notation nous notons $\mathbb{P}_n$ la loi de $(\pi_t)_{t\geq 0}$ partant de $\{\{1\},\ldots,\{n\}\}$. Nous renvoyons au chapitre 6.2 de [37] pour une preuve de (1.6). Cette relation dit essentiellement que si nous échantillonnons n individus indépendamment et uniformément au hasard dans la population en vie au temps t, la probabilité qu'ils portent tous l'allèle A est égale à la probabilité que si nous retraçons la généalogie de l'échantillon pendant t unités de temps dans le passé (en remontant donc jusqu'au temps 0), les $|\pi_t|$ ancêtres à ce moment portent l'allèle A. La généralisation de la relation (1.6) au cas où la population a une structure spatiale sera un ingrédient primordial de la plupart des résultats présentés dans les chapitres 2.4 à 2.6.

Le coalescent de Kingman est à présent un modèle central en génétique des populations. L'une des raisons de son succès est qu'il est très robuste aux déviations des hypothèses d'haploïdie, panmixie, neutralité ou de taille de population constante, à condition de remplacer la « vraie » taille N de la population d'intérêt par une *taille de population effective (ou efficace)* N_e. En d'autres termes, par analogie avec (1.5), N_e est (en général) définie par la propriété que $(N_e t,\, t \geq 0)$ est l'échelle de temps sur laquelle deux lignées ancestrales fusionnent en un ancêtre commun à taux 1. En utilisant la relation entre le coalescent de Kingman et le modèle de Wright-Fisher, nous pouvons également l'interpréter comme le nombre effectif de parents potentiels à chaque génération. En pratique, la taille de population effective peut être bien plus faible que sa taille démographique actuelle. Par exemple, on estime que N_e est de l'ordre de 10^4 chez l'homme [106, 109] et de l'ordre de 10^6 chez *Drosophila melanogaster* [73], ces deux estimations étant bien plus petites que le nombre actuel d'individus dans chaque espèce. Ceci peut refléter l'occurrence de balayages sélectifs récurrents au cours de l'histoire de la population, réduisant le nombre de parents potentiels en favorisant ceux portant les allèles alors favorables. Cela pourrait également avoir été causé par une expansion démographique passée. Au contraire, une structure géographique (ou même une structure de population de manière générale) tend à faire augmenter la taille de population effective puisque deux lignées ancestrales doivent d'abord se rapprocher suffisamment avant d'avoir une chance de descendre du même individu et par là même de fusionner. Le résultat d'une combinaison de ces facteurs d'effets opposés n'est donc pas évident à prédire. La taille de population effective, ou la densité de population effective dans le cadre d'une population structurée en espace, sera une quantité essentielle dans les résultats exposés dans les chapitres 2.5 et 3. Dans le chapitre 4, nous supposerons que N_e fluctue au cours du temps (lorsque nous remontons dans le passé) et nous décrirons une approche permettant de reconstruire sa trajectoire à partir de la variabilité génétique observée dans un échantillon de séquences ADN.

1.2 Ajout d'une structure spatiale

La plupart des populations naturelles sont en fait disséminées sur une région géographique donnée et les individus sont *a priori* plus susceptibles de disperser leurs descendants dans un voisinage autour d'eux que dans l'intégralité du périmètre de l'espèce. Pour tenir compte de ce type de structure de population, la grande majorité des modèles existants fait l'hypothèse que la population est partagée en des sous-populations discrètes reliées entre elles par des migrations. Dans le *modèle d'îles* de Wright [119], chaque communauté est connectée à toutes les autres. Autrement dit, le graphe sous-jacent décrivant les migrations possibles est le graphe complet. Dans le modèle *stepping stone* de Kimura [64], le graphe sous-jacent est généralement

$\mathbb{Z}$ ou $\mathbb{Z}^2$ et le mécanisme de migration est encodé par un noyau $(m_{ij})_{i,j}$ décrivant le flot de gènes entre les sous-populations i et j. Dans ces deux cadres de travail, l'une des nombreuses variantes du modèle de Wright-Fisher décrit les reproductions internes à chaque communauté (supposée avoir une taille constante au cours du temps) et des descendants migrants sont échangés entre les sous-populations connectées.

Diverses déclinaisons de ces modèles et les généalogies qui leur correspondent ont été étudiées, par exemple dans les articles fondateurs [65, 76, 114] ou les études plus récentes [4, 25, 28, 54, 59, 75, 79, 108, 113, 122]. Dans le modèle d'îles, les individus sont à distance 0 ou 1 en fonction de s'ils appartiennent ou non à la même communauté. Par conséquent, l'impact de la structure spatiale de la population sur sa différentiation génétique peut être décrite par la statistique F_{ST} de Wright [120] :

$$F_{ST} = \frac{f_0 - \bar{f}}{1 - \bar{f}}, \tag{1.7}$$

où f_0 (resp., $\bar{f}$) est la probabilité que deux individus échantillonnés uniformément au hasard dans la même île (resp., dans la population totale) sont *identiques par descendance.* Ici, l'identité par descendance est la propriété que deux individus portent le même allèle hérité d'un ancêtre commun. Dans un modèle où les lignées ancestrales mutent indépendamment les unes des autres à un certain taux $\mu > 0$ et si nous notons T le temps de coalescence de deux lignées, nous pouvons réécrire F_{ST} ainsi :

$$F_{ST} = \frac{\mathbb{E}_0[e^{-2\mu T}] - \mathbb{E}_{\text{pop}}[e^{-2\mu T}]}{1 - \mathbb{E}_{\text{pop}}[e^{-2\mu T}]}, \tag{1.8}$$

où sous $\mathbb{P}_0$ (resp., $\mathbb{P}_{\text{pop}}$) les deux individus sont échantillonnés dans la même île (resp., dans la population totale). Lorsque μ est suffisamment petit pour que le produit μT soit également petit dans un sens approprié, nous pouvons utiliser l'approximation

$$F_{ST} \approx \frac{\mathbb{E}_{\text{pop}}[T] - \mathbb{E}_0[T]}{\mathbb{E}_{\text{pop}}[T]},$$

qui a l'avantage de ne pas nécessiter une estimation précise de μ. Les espérances apparaissant dans ce ratio peuvent ensuite être calculées en résolvant un système de deux équations linéaires reliant $\mathbb{E}_0[T]$ et $\mathbb{E}_{\text{pop}}[T]$. Nous ne le précisons pas ici car il dépend des détails du modèle considéré, mais nous renvoyons au chapitre 4.4 de [34] pour un exemple. Ceci ouvre la voie à l'inférence du petit nombre de paramètres caractérisant l'évolution de la population, et ce d'une manière relativement robuste au manque de précision dans l'estimation des taux de mutation. Cependant, bien qu'en effet les taux de mutation pour des gènes non-recombinants (car plutôt petits en nombre de paires de bases) soient généralement faibles, nous verrons que dans des modèles généraux de populations structurées l'espérance du temps de coalescence de deux lignées peut être très grande (voire infinie), cf. la remarque 2.6 de la section 2.5 en particulier.

A notre connaissance, F_{ST} est la statistique principale utilisée pour mesurer l'effet d'une structure spatiale sur la diversité génétique d'une population à un *locus* (i.e., une région d'intérêt sur le génome) sans recombinaison. Cependant, pour des populations disséminées sur un large territoire, nous nous attendons plutôt à ce que le temps de l'ancêtre commun le plus récent de deux individus échantillonnés à distance x soit une fonction croissante de x, de sorte que les corrélations entre les fréquences alléliques locales décroissent avec la distance (par un argument similaire à l'heuristique expliquant (1.6)). Ce phénomène, appelé *isolation par la distance* [119], n'est pas du tout pris en compte dans la définition (1.7) de F_{ST} et une étude approfondie de modèles plus complexes (tels que le modèle stepping stone) est donc nécessaire pour décrire l'empreinte laissée par une telle structure spatiale. Dans la section 2.5, nous introduirons une généralisation de F_{ST} qui en fait une fonction de la distance entre deux points d'échantillonnage et nous montrerons que cette nouvelle statistique peut être utilisée pour développer des méthodes d'inférence. Dans la perspective de reconstruire certains paramètres (éventuellement composés) caractérisant l'évolution de la population, il est également important de réfléchir aux plans d'échantillonnage et leurs liens avec les modèles mathématiques utilisés pour prédire la diversité observée dans l'échantillon. En particulier, lorsque la population vit dans un espace continu, il peut être difficile de la diviser artificiellement en des communautés discrètes. Dans ce cas, un modèle intégrant directement une structure spatiale continue est *a priori* plus souhaitable. C'est cette direction que nous poursuivrons dans le reste de ce chapitre.

Une généralisation naturelle à un espace continu du modèle de Wright-Fisher décrit dans le chapitre 1.1 serait d'ajouter un terme modélisant la diffusion spatiale des gènes. Si nous supposons à nouveau qu'il y a seulement deux allèles et que nous suivons la fréquence $p_t(x)$ de l'un d'entre eux à chaque site x et temps t, ceci nous donne l'équation suivante :

$$\mathrm{d}p_t = \frac{\sigma^2}{2}\,\Delta p_t\,\mathrm{d}t + \sqrt{\frac{1}{N_e}\,p_t(1-p_t)}\,W(\mathrm{d}t,\mathrm{d}x), \tag{1.9}$$

où σ^2 est le coefficient de diffusion, N_e une « densité locale de population » (que nous pouvons voir comme l'inverse du taux auquel deux lignées ancestrales situées au même endroit fusionnent) et W est un bruit blanc espace-temps. En une dimension, cette approche est valable : l'équation (1.9) admet une unique solution à condition initiale fixée, qui peut être obtenue comme la limite d'une suite de modèles stepping stone normalisés sur $\mathbb{Z}$. Cette convergence montre également que la généalogie duale est un système de mouvements browniens indépendants qui fusionnent deux à deux à un taux proportionnel au temps local qu'ils passent ensemble (i.e., au temps local en 0 de leur distance). Malheureusement, en dimension deux l'équation (1.9) n'admet pas de solution et la suite de modèles stepping stone normalisés sur $\mathbb{Z}^2$ converge vers la solution de l'équation de la chaleur, duale d'un système de mouvements browniens indépendants ne fusionnant jamais (notons que contrairement au cas de la dimension un, en deux dimensions deux mouvements browniens indépendants ne se rencontrent jamais). Nous renvoyons au chapitre 2.6.1 pour la preuve d'un résultat similaire. Par conséquent, cette approche n'est pas la bonne pour modéliser le phénomène de dérive génétique dans un espace continu de dimension deux.

Dans les années 1940, Wright et Malécot ont tenté de modéliser des populations vivant dans un continuum spatial [76, 119]. Leur modèle suppose que les individus sont disséminés dans $\mathbb{R}^2$ suivant un processus ponctuel de Poisson d'intensité constante λ. Le mécanisme de reproduction se veut proche de celui du modèle de Wright-Fisher : la population évolue en générations discrètes et le nombre de descendants de chaque individu suit une loi de Poisson de paramètre un. Les positions spatiales des descendants sont tirées indépendamment suivant une loi gaussienne centrée en la position de leur parent. En outre, le modèle incorpore un mécanisme de mutation : avec probabilité $\mu > 0$ un descendant, au lieu d'hériter de l'allèle de son parent, porte un nouvel allèle encore jamais vu dans la population. Wright et Malécot ont alors calculé la probabilité d'identité par descendance de deux individus échantillonnés à une séparation $x \in \mathbb{R}^2$. Notons $F(x)$ cette quantité. En utilisant une récurrence, Malécot a obtenu une approximation pour $F(x)$ faisant intervenir la fonction de Bessel modifiée du deuxième type d'ordre 0, K_0 :

$$F(x) \approx \frac{1}{\mathcal{N} + \ln(\ell/\kappa)} K_0\left(\frac{\|x\|}{\ell}\right), \quad \|x\| > \kappa, \tag{1.10}$$

où $\kappa > 0$ est une échelle locale sur laquelle nous supposons que $F(x)$ est constante (pour passer outre le problème de l'explosion de K_0 en 0), σ^2 est la variance de la distribution gaussienne qui détermine la position spatiale des descendants, $\ell = \sigma/\sqrt{2\mu}$ peut être vu comme une longueur caractéristique et $\mathcal{N}$ est la *taille de voisinage* de Wright qui, en essence, mesure le nombre de « parents potentiels » dans le voisinage de chaque descendant. L'approximation (1.10) est appelée *formule de Wright-Malécot.*

Malheureusement, la récurrence *backwards-in-time* conduisant à (1.10) est fondée sur l'hypothèse qu'à chaque génération, les positions des individus peuvent être décrites par un processus ponctuel de Poisson d'intensité λ sur $\mathbb{R}^2$. Ceci n'est pas cohérent avec l'évolution *forwards-in-time*, qui conduit la densité locale d'individus à exploser dans certaines régions tandis que d'autres régions se vident. Considérer un espace géographique compact au lieu de $\mathbb{R}^2$ ne résout pas ce problème, car alors la population de mécanisme de reproduction critique s'éteint en temps fini. C'est ce que Felsenstein a nommé *the pain in the torus* [46]. Cependant, comme discuté dans la section 2.3 de l'article de revue [BEV13b], la formule de Wright-Malécot avec des paramètres appropriés décrit étonnamment bien la décroissance avec la distance de la probabilité d'identité par descendance dans un modèle stepping-stone avec, par exemple, des migrations aux plus proches voisins. Rappelons que dans ce modèle les tailles des communautés sont supposées constantes, de sorte que l'approche récursive conduisant à (1.10) fonctionne. De manière évidente, dans notre modèle d'évolution en espace continu nous avons besoin d'un mécanisme garantissant la régulation locale de la densité de la population.

Dans cette perspective, plusieurs approches ont été tentées et nous renvoyons par exemple aux articles [6, 9, 116, 117] et aux références qui y sont données. Le modèle présenté dans la prochaine partie les unifie dans un cadre de travail flexible et qui se prête aisément à l'analyse.

1.3 Le processus Λ-Fleming-Viot spatial

Le modèle décrit dans ce sous-chapitre a été introduit dans les notes [38] puis formalisé dans l'article [BEV10]. La principale différence qui le sépare des modèles d'évolution précédents est que les reproductions ne sont pas basées sur des horloges individuelles, mais sur une suite aléatoire d'événements affectant chacun une zone donnée de l'espace. Au cours d'un tel événement, des parents sont choisis de manière aléatoire et leurs descendants remplacent une fraction de la population présente dans cette région. De cette manière, la densité de la population reste constante mais les fréquences locales d'allèles sont mises à jour en tenant compte du type génétique des parents et de la fraction des individus remplacés. Nous donnons ci-dessous un exemple particulier de mécanisme de reproduction, mais celui-ci peut être généralisé de nombreuses manières tant que nous conservons l'ingrédient essentiel d'un processus ponctuel de Poisson d'événements de reproduction spécifiant la zone géographique dans laquelle la diversité génétique locale va être modifiée. C'est ce que nous ferons par exemple lorsque nous ajouterons de la recombinaison (cf. chapitre 2.4.2), des mutations (cf. chapitre 2.4.4), de la sélection (cf. chapitre 2.6.1), ou même des inhomogénéités spatiales (cf. [49]).

Supposons que la population est uniformément répartie sur $\mathbb{R}^d$ ($d = 2$ étant évidemment la dimension la plus pertinente pour les populations biologiques) et que l'ensemble K de tous les allèles possibles est compact. Nous considérerons principalement $K = \{0, 1\}$ comme dans le modèle de Wright-Fisher, ou $K = [0, 1]$ pour permettre à un nombre arbitrairement grand d'allèles d'être présents dans la population entière. À un temps $t \geq 0$ donné, l'état de la population est représenté par une mesure $M_t(\mathrm{d}x, \mathrm{d}k)$ sur $\mathbb{R}^d \times K$ dont la première marginale est la mesure de Lebesgue sur $\mathbb{R}^d$. Notons $\mathcal{M}_\lambda$ l'ensemble de toutes ces mesures. Puisque toute $M_t \in \mathcal{M}_\lambda$ peut être décomposée en

$$M_t(\mathrm{d}x, \mathrm{d}k) = \mathrm{d}x\, \rho_t(x, \mathrm{d}k), \tag{1.11}$$

où $\rho_t : \mathbb{R}^d \to \mathcal{M}_1(K)$ est une fonction Lebesgue-mesurable à valeurs dans l'espace des mesures de probabilité sur K (voir le chapitre 2.4.4 pour un énoncé plus précis), cette représentation reflète bien la répartition uniforme des individus dans l'espace et $\rho_t(x, \mathrm{d}k)$ peut être vu comme la distribution de l'allèle d'un individu qui serait échantillonné au site x au temps t. En fait, l'espace d'états $\mathcal{M}_\lambda$ est relativement naturel lorsque nous pensons à la population représentée ici comme la limite d'une population d'individus discrets dont les positions spatiales forment un processus ponctuel de Poisson d'intensité $\lambda \mathrm{d}x$, lorsque λ tend vers l'infini. Cf. la remarque 1.2 ci-dessous.

Jusqu'à présent nous avons seulement spécifié la manière dont nous encodons la diversité génétique d'une population structurée. Nous avons maintenant besoin d'un mécanisme pour la faire évoluer dans le temps. Pour ce faire, fixons une mesure σ-finie μ sur $(0, \infty)$ et un ensemble $\{\nu_r,\ r > 0\}$ de mesures de probabilité sur $[0, 1]$. Soit Π un processus ponctuel de Poisson sur $\mathbb{R}_+ \times \mathbb{R}^d \times (0, \infty) \times [0, 1]$ d'intensité $\mathrm{d}t \otimes \mathrm{d}z \otimes \mu(\mathrm{d}r)\nu_r(\mathrm{d}u)$. Autrement dit,

$$\Pi = \big\{(t_i, z_i, r_i, u_i),\ i \in \mathbb{N}\big\}$$

est un ensemble dénombrable aléatoire d'événements de reproduction décrits par leur temps d'occurrence t_i, leur centre z_i, leur rayon r_i et leur *impact* u_i. Plus précisément, pour tout $i \in \mathbb{N}$, au temps t_i un événement de reproduction se produit dans la boule fermée $B(z_i, r_i)$. Un parent est choisi uniformément au hasard dans $B(z_i, r_i)$ et ses descendants, portant le même allèle, remplacent une fraction u_i de la population locale à chaque site de la boule (la fraction $1 - u_i$ restante n'étant pas affectée). En des termes plus mathématiques, ceci signifie qu'un allèle parental κ_i est choisi suivant la distribution

$$\frac{1}{\mathrm{Vol}(B(z_i, r_i))} \int_{B(z_i, r_i)} M_{t_i-}(y, \mathrm{d}k)\mathrm{d}y$$

des allèles dans $B(z_i, r_i)$ juste avant l'événement et à chaque site $y \in B(z_i, r_i)$, nous avons

$$\rho_{t_i}(y, \mathrm{d}k) = (1 - u_i)\rho_{t_i-}(y, \mathrm{d}k) + u_i \delta_{\kappa_i}(\mathrm{d}k).$$

La figure 1.3 montre un exemple avec $K = \{0, 1\}$ et $d = 1$; seules les fréquences locales d'individus de type 1 sont représentées.

La forme de la mesure d'intensité de Π impose que les événements de reproduction se produisent de manière uniforme en temps et en espace, tandis que les coordonnées r et u ont des distributions plus générales et *a priori* corrélées. Ceci nous permet par exemple de modéliser des événements de reproduction « réguliers » affectant de manière modérée des régions de petites tailles, en même temps que de rares catastrophes telles que des événements climatiques extrêmes, affectant des zones bien plus étendues et pendant lesquelles une fraction significative de la population s'éteint et est rapidement remplacée par les descendants d'un petit nombre d'individus survivants. Finalement, remarquons que dans le modèle formulé ainsi les individus ne migrent pas au cours de leur vie (seules leurs propagules se dispersent), mais cette généralisation n'est pas difficile grâce à la construction décrite dans le chapitre 2.4.4.

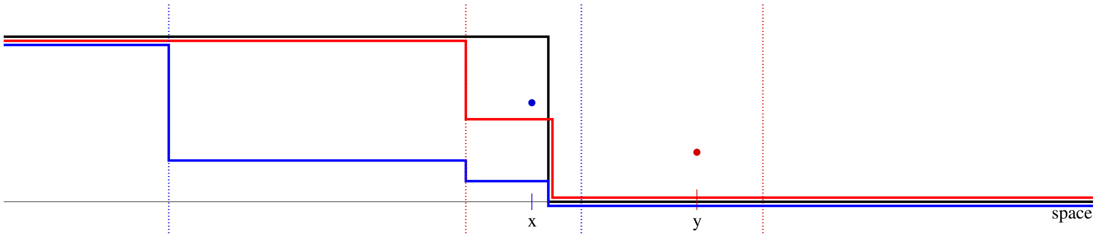


FIGURE 1.3 – Le processus Λ-Fleming-Viot spatial en une dimension. Les individus portent l'un des deux allèles 0 et 1 et la fréquence de l'allèle 1 à chaque site est représentée. La ligne noire correspond à l'état initial, dans lequel tous les individus à gauche portent l'allèle 1 et tous les individus à droite portent l'allèle 0. Un premier événement se produit dans la zone délimitée par les lignes verticales rouges, un parent d'allèle 0 est choisi au site y et une fraction $u_1 = 1/2$ de la population locale en tout point de la boule est remplacée par ses descendants, portant l'allèle 0 (ligne rouge). Quelques temps après, un événement se produit dans la zone délimitée par les lignes bleues, un parent est choisi dans la fraction des individus au site x qui portent à présent l'allèle 0 et une fraction $u_2 = 3/4$ de la population dans la boule est remplacée par ses descendants (ligne bleue).

L'existence et l'unicité du processus $(\rho_t)_{t\geq 0}$ correspondant à cette évolution ont été démontrées en premier lieu dans l'article [BEV10] en utilisant une technique de l'article [45] basée sur la caractérisation de son semi-groupe via une famille bien choisie de fonctions tests. Dans [VW15], ce processus est reformulé comme un processus aléatoire à valeurs mesures, dont l'existence et l'unicité sont prouvées au moyen d'arguments plus constructifs (voir le chapitre 2.4.4). Ces deux approches utilisent la relation de *dualité* entre l'évolution *forwards-in-time* des fréquences alléliques $(M_t)_{t\geq 0}$ et la généalogie d'un échantillon aléatoire d'individus, retracée *backwards-in-time*. Avant de donner un énoncé précis de cette relation, imaginons ce à quoi la généalogie de quelques individus devrait ressembler dans ce modèle.

Pour commencer, supposons qu'un individu est échantillonné au site x à un temps que nous appelons le présent. Pour simplifier les notations, nous notons ce temps 0 (et nous supposons que la coordonnée temporelle du processus ponctuel de Poisson Π d'événements de reproduction est à valeurs dans $\mathbb{R}$ au lieu de $\mathbb{R}_+$, de sorte que nous puissions revenir aussi loin dans le passé qu'il est nécessaire). Notre but est de retracer la position ξ_t de l'individu au temps $-t$ dont notre individu échantillonné descend. Pour qu'un individu donné soit né au cours d'un événement de reproduction, il doit se trouver dans la région affectée et appartenir à la fraction de la population remplacée. Dans ce cas, puisque le parent est choisi uniformément au hasard dans la boule où a lieu l'événement, sa position est distribuée uniformément dans la boule et cette position est précisément celle que prend à ce moment la lignée ancestrale que nous suivons. Puisque le processus ponctuel $\{(-t_i, z_i, r_i, u_i),\ i \in \mathbb{N}\}$ dont on a inversé la flèche du temps est également un processus ponctuel de Poisson d'intensité $\mathrm{d}t \otimes \mathrm{d}z \otimes \mu(\mathrm{d}r)\nu_r(\mathrm{d}u)$, nous pouvons écrire que le taux auquel la lignée ancestrale de notre individu saute en une nouvelle position lorsqu'elle se trouve en $x \in \mathbb{R}^d$ est donné par

$$\mathscr{J}_0 := \int_{\mathbb{R}^d}\int_0^\infty\int_0^1 \mathbf{1}_{\{x\in B(z,r)\}} u\,\nu_r(\mathrm{d}u)\mu(\mathrm{d}r)dz = \int_0^\infty\int_0^1 uV_r\,\nu_r(\mathrm{d}u)\mu(\mathrm{d}r), \tag{1.12}$$

où V_r est le volume d'une boule de rayon r en dimension d. En supposant que $\mathscr{J}_0$ est une quantité finie (ce que nous ferons dorénavant, bien que des processus de Lévy plus généraux puissent être obtenus sous des conditions plus faibles) et en notant $L_r(y)$ le volume de l'intersection $B(0,r) \cap B(y,r)$, le processus $(\xi_t)_{t\geq 0}$ est alors un processus de Poisson composé bien défini qui saute de x à $x+y$ avec une intensité donnée par

$$\begin{aligned} \mathcal{J}(y) &:= \int_{\mathbb{R}^d} \int_0^\infty \int_0^1 \mathbf{1}_{\{x\in B(z,r)\}} u \frac{\mathbf{1}_{\{x+y\in B(z,r)\}}}{V_r}\, \nu_r(\mathrm{d}u)\mu(\mathrm{d}r)\mathrm{d}z \\ &= \int_0^\infty \int_0^1 \frac{uL_r(y)}{V_r}\, \nu_r(\mathrm{d}u)\mu(\mathrm{d}r). \end{aligned} \tag{1.13}$$

En effet, pour qu'un tel événement se produise, la position actuelle x de la lignée et la position $x+y$ de son parent doivent toutes les deux se trouver dans la région $B(z,r)$ de l'événement de reproduction et la position du parent est tirée en $x+y$ avec une densité $1/V_r$.

Échantillonnons à présent un autre individu au site x' au temps 0. Appelons $(\xi'_t)_{t\geq 0}$ le processus retraçant la position de l'ancêtre de ce second individu t unités de temps dans le passé. *A priori* $(\xi_t)_{t\geq 0}$ et $(\xi'_t)_{t\geq 0}$ ne sont pas indépendants, puisqu'ils utilisent le même processus ponctuel de Poisson d'événements pour sauter. Si tous deux se trouvent dans la région d'un événement de reproduction donné, disons d'impact u, alors

- avec probabilité $(1-u)^2$ aucun des deux n'appartient à la population locale remplacée et aucune lignée ne bouge à cet instant,
- avec probabilité $u(1-u)$ l'ancêtre du premier individu appartient à la fraction remplacée mais pas l'ancêtre du second individu, auquel cas ξ saute en la position du parent et ξ' reste là où il est,
- avec probabilité $u(1-u)$ ξ' saute mais pas ξ,
- avec probabilité u^2 les deux ancêtres appartiennent à la descendance de l'unique parent choisi durant l'événement. Dans ce cas, les deux lignées ancestrales fusionnent en une seule, dont la position est distribuée uniformément dans la région de l'événement.

Nous pouvons déduire de cette observation que le taux auquel deux lignées ξ et ξ' fusionnent lorsqu'elles sont à une séparation $y \in \mathbb{R}^d$ s'écrit

$$\mathcal{C}(y) = \int_0^\infty \int_0^1 u^2 L_r(y)\, \nu_r(\mathrm{d}u)\mu(\mathrm{d}r). \tag{1.14}$$

Notons que cette quantité est bornée par le taux de saut $\mathcal{J}_0$ d'une seule lignée. Puisque $\mathcal{J}_0$ est supposé être fini, nous pouvons en déduire que le couple (ξ,ξ') forme un système de processus de Poisson composés corrélés qui fusionnent en une seule lignée à un taux instantané donné par (1.14).

La même analyse peut être effectuée pour un échantillon de n'importe quelle taille finie pris au temps 0 à un ensemble donné de positions. Pour décrire plus formellement le système de processus de sauts coalescents que nous obtenons, nous représentons à nouveau les relations ancestrales au sein de l'échantillon au temps $-t$ par une partition de $\{1,\ldots,n\}$. Cependant, contrairement au cas d'une population sans structure décrit dans le chapitre 1.1, ici il est nécessaire de garder en mémoire la position spatiale de chaque ancêtre. Par conséquent, nous utilisons des partitions *marquées* de $\{1,\ldots,n\}$ de la forme

$$\mathcal{A} = \big\{(b^1,x^1),\ldots,(b^k,x^k)\big\},$$

où les blocs $\{b^1,\ldots,b^k\}$ forment une partition de $[n]$ décrivant quels individus de l'échantillon ont un ancêtre en commun au temps d'intérêt et $x^j \in \mathbb{R}^d$ est la position de l'ancêtre à ce moment des individus dont les étiquettes appartiennent à b^j. Appelons $\mathcal{P}_n^s$ l'ensemble des partitions marquées de $[n]$. Le processus ancestral d'un échantillon de taille n est donc un processus de sauts markovien à valeurs dans $\mathcal{P}_n^s$, noté

$$(\mathcal{A}_t)_{t\geq 0} = \Big(\big\{(B_t^1, \xi_t^1), \dots, (B_t^{N_t}, \xi_t^{N_t})\big\}\Big)_{t\geq 0},$$

où N_t est le nombre de blocs (ou ancêtres distincts) t unités de temps dans le passé. Pour éviter des notations trop lourdes nous ne donnons pas une description complète de ses taux de saut, mais le lecteur intéressé trouvera tous les ingrédients nécessaires dans le paragraphe précédent.

Nous pouvons à présent dévoiler la relation de dualité entre le processus Λ-Fleming-Viot spatial $(M_t)_{t\geq 0}$ (ou $(\rho_t)_{t\geq 0}$, en se rappelant (1.11)) et le processus ancestral $(\mathcal{A}_t)_{t\geq 0}$. Notons $C(E)$ l'ensemble des fonctions continues $f : E \to \mathbb{R}$, $C_c(E)$ le sous-ensemble de ces fonctions qui sont à support compact, $\wp_n(\mathbf{x})$ la partition marquée $\{(\{1\}, x_1), \dots, (\{n\}, x_n)\}$ constituée de singletons donc les marques sont données par le vecteur $\mathbf{x} = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ et rappelons la notation $\mathcal{M}_\lambda$ pour l'espace des mesures sur $\mathbb{R}^d \times K$ dont la première marginale est la mesure de Lebesgue sur $\mathbb{R}^d$. Nous avons alors le résultat suivant, qui correspond au théorème 4.2 de [BEV10] ou au corollaire 2.4 de [VW15].

Theorem 1.1. *Supposons que la quantité $\mathcal{J}_0$ définie dans* (1.12) *soit finie. Alors il existe un unique processus de Hunt à valeurs dans $\mathcal{M}_\lambda$, noté $(M_t)_{t\geq 0}$, tel que pour tous $m \in \mathcal{M}_\lambda$, $t \geq 0$, $n \in \mathbb{N}$, $F \in C_c((\mathbb{R}^d)^n)$ et $g_1, \dots, g_n \in C(K)$,*

$$\begin{aligned}&\mathbb{E}_m\left[\int_{(\mathbb{R}^d\times K)^n} F(x_1,\dots,x_n)\left(\prod_{i=1}^n g_i(\kappa_i)\right) M_t^{\otimes n}(\mathrm{d}x_1, \mathrm{d}\kappa_1, \dots, \mathrm{d}x_n, \mathrm{d}\kappa_n)\right] \qquad (1.15)\\ &= \int_{(\mathbb{R}^d)^n} F(x_1,\dots,x_n)\mathbb{E}_{\wp_n(\mathbf{x})}\left[\int_{K^{N_t}} \prod_{j=1}^{N_t}\left(\prod_{i\in B_t^j} g_i(\kappa_j)\right)\rho\big(\xi_t^1, \mathrm{d}\kappa_1\big)\cdots\rho\big(\xi_t^{N_t}, \mathrm{d}\kappa_{N_t}\big)\right]\mathrm{d}x_1\cdots\mathrm{d}x_n,\end{aligned}$$

où nous avons utilisé la décomposition $m(\mathrm{d}x, \mathrm{d}\kappa) = \mathrm{d}x\rho(x, \mathrm{d}\kappa)$.

La relation (1.15) aura l'air moins intimidante lorsque nous travaillerons avec deux allèles seulement (cf. le chapitre 2.4.3). Ce qu'elle nous dit est que pour obtenir la distribution des allèles de n individus échantillonnés à des positions distinctes $x_1, \dots, x_n$ au temps t, nous pouvons retracer qui partage un ancêtre commun avec qui t unités de temps dans le passé (remontant ainsi jusqu'au temps 0) et échantillonner leur allèle commun suivant la distribution allélique au temps 0 à la position de l'ancêtre. Cette interprétation est à la base de la représentation particulaire décrite dans le chapitre 2.4.4. Ce chapitre présente également les propriétés générales de l'espace d'états $\mathcal{M}_\lambda$, certaines propriétés trajectorielles du processus $(M_t)_{t\geq 0}$ et un résultat sur la (non) *descente de l'infini* du processus généalogique $(\mathcal{A}_t)_{t\geq 0}$ partant d'un nombre infini d'individus.

Concluons cette partie par quelques remarques.

Remark 1.1. (Condition plus faible pour l'existence et l'unicité du processus). *Comme mentionné précédemment, l'hypothèse que $\mathcal{J}_0$ doive être fini est plus forte que nécessaire pour le résultat d'existence et d'unicité du théorème 1.1. La condition la plus générale que l'on puisse requérir pour utiliser la technique de l'article [45] est que le mouvement $(\xi_t)_{t\geq 0}$ d'une seule lignée ancestrale doive être un processus de Lévy, ce qui est vrai sous la condition plus faible*

$$\int_{\mathbb{R}^d} (1 \wedge |y|^2)\, \mathcal{J}(y)\mathrm{d}y < \infty,$$

où $\mathcal{J}(y)$ *est défini par* (1.13)*. Cependant, lorsque nous voulons écrire un générateur infinitésimal ou un problème de martingales pour identifier le processus* $(M_t)_{t\geq 0}$ *obtenu par dualité et le processus potentiel que nous avons défini en termes d'un processus ponctuel de Poisson d'événements de reproduction, des problèmes techniques (dans l'échange entre dérivation et espérance) apparaissent. Puisqu'il n'y a pas grand intérêt biologique à considérer le cas où les lignées accumulent une infinité de sauts microscopiques, nous supposerons toujours que* $\mathcal{J}_0 < \infty$*.*

Remark 1.2. (Modèle basé sur des individus). *Le processus* Λ*-Fleming-Viot spatial est en fait la limite en grande densité d'un modèle basé sur des individus et évoluant grâce au même processus ponctuel de Poisson* Π *d'événements de reproduction. Dans ce modèle, les individus sont disséminés dans* $\mathbb{R}^d$ *suivant un processus ponctuel de Poisson d'intensité* $\lambda > 0$*. Lorsque la région affectée par un événement de reproduction ne contient personne, cet événement est simplement annulé. Sinon, un parent est choisi uniformément au hasard parmi les individus présents dans cette zone, puis chacun de ces individus meurt avec probabilité* u *(l'impact de l'événement) indépendamment les uns des autres et finalement la région est repeuplée par un processus ponctuel de Poisson de descendants d'intensité* $u\lambda$*, tous portant l'allèle de leur parent. Dans l'article [12], les auteurs montrent que cette population aléatoire (sans allèle, comptant simplement le nombre d'individus dans chaque région de l'espace) survit avec probabilité* 1 *dès que* λ *est suffisamment grand et ils décrivent alors son comportement ergodique. En revanche, si* λ *est trop petit, la population s'éteint en temps fini p.s. Bien qu'il semble plus naturel de décrire une population d'individus discrets, ce modèle est très délicat à étudier tandis que sa limite en grande densité* $\lambda \to \infty$ *conduit à un modèle bien plus simple à analyser (en particulier, aucun événement n'est annulé puisqu'il y a toujours quelqu'un dans la région touchée). Une preuve de cette convergence peut être trouvée dans [42].*

Remark 1.3. (Modèle à base de noyaux gaussiens). *Une version du modèle à individus discrets et de sa limite en grande densité, utilisant des noyaux gaussiens (pouvant effectivement apparaître comme plus réalistes) au lieu de boules pour le choix du parent et les naissances/morts d'individus, a été étudié dans [8]. En particulier, les auteurs montrent que dans le modèle limite, lorsque les variances des noyaux gaussiens sont les mêmes pour tous les événements la probabilité d'identité par descendance de deux individus échantillonnés à une séparation* $x \in \mathbb{R}^2$ *est bien approchée par la formule de Wright-Malécot* (1.10) *avec des paramètres appropriés. Nous généralisons ce résultat dans l'article [BEKV13].*

Chapter 2

Genetic diversity in a spatially structured population

Modelling the evolution of a population over the many generations through which it went until now is a very difficult task, as many factors may have influenced, sometimes only for a short amount of time, the pattern of genetic variation observed within a sample of individuals. Indeed, the population may have undergone census fluctuations such as bottlenecks or demographic expansions, episodes of *selective sweeps*, it may be (or have been) spatially structured, or experience(d) environmental fluctuations, ... One of the major roles of mathematical models is thus to study the effects of each of these factors, or of a combination of them, to highlight the signatures they would leave in the current genetic diversity assuming that they contribute sufficiently strongly to the evolution of the population.

Of course a wealth of models already exist, and we refer to [37] for a clear exposition of the most classical ones. Most of them are not meant to describe the biology of the organisms in fine detail, but are based on rather crude approximations for the way genes are transmitted from parents to offspring. However, their strength is that they are able to render the evolution of the population over the tens, hundreds or even thousands of generations during which the current genetic diversity of the population was built.

2.1 The Wright-Fisher model and Kingman's coalescent

The most widely used model of genetic evolution is the Wright-Fisher model [47, 118]. It assumes that individuals are *haploid* (i.e., they have only one copy of each chromosome), the population exhibits no structure of any kind (it is *panmictic*), evolution happens in discrete generations over which the population size remains constant equal to some large number N. If we further assume that the gene in which we are interested is neutral, meaning that none of its possible versions, or *alleles*, provides a reproductive advantage to the individuals carrying it, then the reproduction mechanism is easily expressed: to form generation $t+1$, each of the N individuals living at that time chooses a parent uniformly at random within the previous generation, independently of each other, and the offspring inherit the alleles of their parents. See Figure 2.1 for an example with $N = 7$.

If there are only two alleles, say A and a, then it suffices to follow the proportion $p^N(t)$ of allele A in generation t to fully describe the evolution of the genetic diversity of the population. In the neutral case expounded here, it is straightforward to show that for every $t \in \mathbb{N}$,

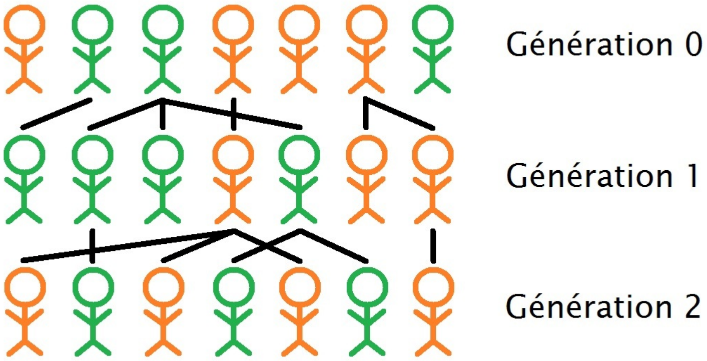


Figure 2.1 – The Wright-Fisher model with $N = 7$. Individuals can be of two genetic types (orange and green here) which are transmitted from parents to offspring (as symbolised by the black lines).

conditionally on $p^N(t)$ we have

$$Np^N(t+1) \sim \text{Binomial}\big(N, p^N(t)\big). \tag{2.1}$$

The long-term behaviour of the Markov chain $(p^N(t))_{t\in\mathbb{N}}$ is then easily obtained: with probability one, one of the alleles becomes *fixed* in (i.e., invades) the population and the probability of A being the lucky allele is equal to its initial proportion $p^N(0)$. However, obtaining more precise information on the path to fixation, for example the distribution of the time it takes for p^N to reach 0 or 1, becomes more and more combinatorially tricky as the population size increases. Since we are interested in large populations, the usual mathematical trick is to let N tend to infinity and see whether we can obtain a limiting object that may approximate the evolution of a population with a large but finite size.

Using (2.1), we readily see that for every $t \in \mathbb{N}$, we have

$$\mathbb{E}\big[p^N(t+1)\,|\,p^N(t)\big] = p^N(t) \quad \text{and} \quad \text{Var}\big(p^N(t+1)\,|\,p^N(t)\big) = \frac{1}{N}\,p^N(t)\big(1-p^N(t)\big).$$

As a consequence, if we let N tend to infinity and if we suppose that the initial proportion $p^N(0)$ of individuals with allele A tends to some constant $p \in [0,1]$, then the process p^N converges in distribution to the constant process equal to p. In other words, there are no fluctuations in allele frequencies in the limiting infinite population. Thinking a bit more about what happens here, we see that the variance of the change over one generation, $p^N(t+1) - p^N(t)$, is of the order of $1/N$. By analogy with the convergence of rescaled symmetric random walks to Brownian motion, we thus expect to have to wait of the order of N generations to observe macroscopic variations of p^N. This motivates us to define

$$\tilde{p}^N_t = p^N(\lfloor Nt \rfloor), \quad t \in \mathbb{R}_+, \tag{2.2}$$

where $\lfloor x \rfloor$ denotes the integer part of $x \in \mathbb{R}$, so that one unit of time for $\tilde{p}^N$ corresponds to N generations. Letting N tend to infinity, this time we obtain that the sequence of processes $(\tilde{p}^N)_{N\geq 1}$ converges in distribution (in the Skorokhod space $D_{[0,1]}[0,\infty)$ of all càdlàg paths with values in $[0,1]$) to the *Wright-Fisher diffusion*, unique solution to the stochastic differential equation

$$\mathrm{d}p_t = \sqrt{p_t(1-p_t)}\,\mathrm{d}B_t, \qquad p_0 = p, \tag{2.3}$$

where $(B_t)_{t\geq 0}$ denotes standard Brownian motion. A proof of this convergence can be found in Chapter 3.2 of [37]. The random fluctuations described in (2.3) are what is referred to as *genetic drift* in the biology literature. As in the case of a finite population, they are only caused by the random replacement of individuals with allele a by offspring carrying the allele A, and conversely, which explains that their infinitesimal variance is proportional to the product of the current frequencies of alleles A and a. Using the toolbox of stochastic calculus, we can show that again fixation of one allele occurs in finite time a.s., this allele being A with probability p_0; in fact many more properties are now within reach (expected time to fixation, conditional path distributions, ...).

Besides characterising the long-term evolution of an infinite population following the Wright-Fisher reproduction scheme, the results presented in the previous paragraph highlight the fact that evolution may actually occur on a very long timescale, of the order of the size of the population *in this model.* As we shall see, chasing the right time- and space-scales on which we may observe some non-trivial behaviour will be a key step in the analysis of the different models studied in the next chapters (including Chapter 5). This may look like a slightly artificial game. However, the real question after which we are is the following: *assuming that we observe a given pattern of diversity, how should the different evolutionary forces compare to each other to explain this pattern?*

A particularly fruitful approach in the study of the Wright-Fisher model is to reverse the arrow of time and to try to understand the genealogy of a few individuals sampled at random from the current population. Indeed, the longer it takes to two individuals to find a common ancestor in the past, the more time there is for mutation or recombination to occur, (potentially) leaving a detectable trace when we compare their genomes. Under the assumptions of the neutral Wright-Fisher model with population size N, the most recent common ancestor to two individuals sampled uniformly at random in the present generation can be found T_2^N generations in the past, where T_2^N has a geometric distribution with parameter $1/N$. Therefore, as in the analysis of the allele frequencies, here the action takes place on the timescale $(Nt,\, t\geq 0)$ and scaling time as before we obtain that the *coalescence time* of the ancestral lineages of the two individuals on the new timescale, $\tilde{T}_2^N = T_2^N/N$, satisfies

$$\tilde{T}_2^N \xrightarrow{(\mathrm{d})} T_2 \quad \text{as } N\to\infty, \tag{2.4}$$

where T_2 has an exponential distribution with parameter 1.

Let us now consider a sample of size $n\geq 2$, again taken uniformly at random from the present generation. To ease the notation, let us call this generation 0. Since we want to describe who shares an ancestor with whom t generations back in the past, a natural way of representing the state of the ancestral relations within the sample at generation $-t$ is through a partition $\pi^N(t)$ of $[n] := \{1,\ldots,n\}$ such that i and j belong to the same block of $\pi^N(t)$ if and only if the i-th and j-th individuals in the sample have the same ancestor t generations ago. See Figure 2.2 for an example. In this way, for any N we obtain a Markov chain $(\pi^N(t))_{t\in\mathbb{N}}$ with values in the set $\mathcal{P}_n$ of all partitions of $[n]$, starting at the partition into singletons $\{\{1\},\ldots,\{n\}\}$ and whose only absorbing state is the trivial partition $\{\{1,\ldots,n\}\}$. The latter corresponds to the state in which everyone in the sample share the same ancestor, and the random generation T_{MRCA}^N in the past at which this event occurs for the first time is called the *time to the most*

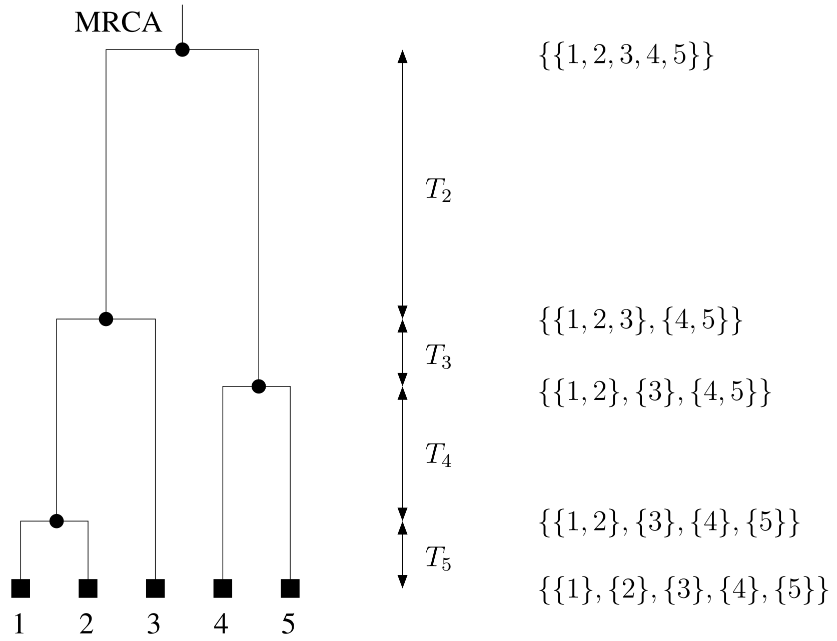


Figure 2.2 – Tree and partition representations of the genealogy of a sample of size 5. Genealogical time runs from bottom to top. The black squares represent the individuals sampled in the present population. During the time T_i, the ancestral partition has i blocks, corresponding to the i (vertical) edges in this layer of the tree. The Most Recent Common Ancestor to the sample is reached after $T_5 + \ldots + T_2$ units of time.

recent common ancestor. Performing the same change of timescale, it is not difficult to show that

$$\tilde{\pi}^N := \big(\pi^N\big(\lfloor Nt\rfloor\big)\big)_{t\in\mathbb{R}_+} \xrightarrow{(\mathrm{d})} (\pi_t)_{t\in\mathbb{R}_+} \quad \text{as } N\to\infty, \tag{2.5}$$

where $(\pi_t)_{t\geq 0}$ is *Kingman's coalescent*. The limiting object, formally introduced in [67], is a $\mathcal{P}_n$-valued Markov jump process that can be described as follows: it starts from the partition of $[n]$ into singletons, and then every pair of blocks tries to merge at rate 1 until the final state $\{\{1,\ldots,n\}\}$ is reached. Hence, when there are k blocks in the ancestral partition, the time it takes for the next coalescence event to occur is exponentially distributed with parameter $\binom{k}{2}$ and its outcome is the merger of a single pair of blocks chosen uniformly at random among all $\binom{k}{2}$ possible pairs. Many properties of the limiting process are known, as reviewed for example in [34] and [110]. A particularly interesting one is the *duality relation* between the Wright-Fisher diffusion (2.3) and Kingman's coalescent. Indeed, let us write $|\pi|$ for the number of blocks in the partition π. Then for every $n\in\mathbb{N}$, $p_0\in[0,1]$ and $t\in\mathbb{R}_+$,

$$\mathbb{E}_{p_0}\big[p_t^n\big] = \mathbb{E}_n\big[p_0^{|\pi_t|}\big], \tag{2.6}$$

where we have abused notation and denoted the law of $(\pi_t)_{t\geq 0}$ starting at $\{\{1\},\ldots,\{n\}\}$ by $\mathbb{P}_n$. We refer to Chapter 6.2 of [37] for a proof of (2.6). In essence, this relation says that if we sample n individuals independently and uniformly at random within the population alive at time t, the probability that they are all of allelic type A is equal to the probability that if we trace back the ancestry of the sample for t generations into the past (thus coming back to time 0), all $|\pi_t|$ ancestors carry the allele A. The extension of (2.6) to the case where the population is spatially structured will be a key ingredient in most of the results presented in Sections 2.4 to 2.6.

Kingman's coalescent is now a central model in population genetics. One of the reasons for its success is that it is very robust to deviations from the assumptions of haploidy, panmixia, neutrality or constant population size, up to replacing the 'true' size N of the population it approximates by some *effective population size* N_e. In other words, by analogy with (2.5), N_e is (usually) defined by the property that $(N_e t,\, t \geq 0)$ is the timescale on which two ancestral lines merge into a common ancestor at rate 1. By the relation between Kingman's coalescent and the Wright-Fisher model, it can also be interpreted as the effective number of potential parents at each generation. In practice, the effective population size can be much smaller than the current census size. For example, N_e is estimated to be of the order of 10^4 in humans [106, 109] and of the order of 10^6 in *Drosophila melanogaster* [73], both estimates being far lower than census numbers. This can reflect the occurrence of recurrent selective sweeps in the history of the population, reducing the number of potential parents by favouring those carrying the selected alleles. It may also be caused by a past population expansion. In contrast, geographical structure (or even population structure in general) tends to increase the effective population size since two ancestral lineages first need to come within a reasonable distance to each other before having a chance to descend from the same individual and thus merge. The outcome of a combination of these factors with opposite effects is therefore non trivial to predict. The effective population size, or the analogous effective population density in the context of spatially structured populations, will be a key quantity in the results exposed in Chapters 2.5 and 3. In Chapter 4, we shall instead assume that N_e is a function of time (as we come back in the past) that we wish to infer from the genetic variability observed in a sample of DNA sequences.

2.2 Adding a spatial structure

Most natural populations are distributed over some geographical area and individuals are *a priori* more likely to disperse their offspring in a neighbourhood around them than across the whole range. To account for this type of population structure, the vast majority of the existing models assume that the population is split into discrete subpopulations connected through migration. In Wright's *island model* [119], each deme is connected to every other deme. That is, the underlying graph on which the islands are organised is the complete graph. In Kimura's *stepping stone model* [64], the underlying graph is usually taken to be $\mathbb{Z}$ or $\mathbb{Z}^2$ and migration is encoded by a kernel $(m_{ij})_{i,j}$ describing the flow of genes between subpopulations i and j. In both frameworks, a kind of Wright-Fisher resampling takes place within each deme (assumed to have a constant size) and migrant offspring are exchanged between the subpopulations which are connected.

Many variants of these models and the corresponding genealogies have been investigated, see for example [65, 76, 114] for historical works and [4, 25, 28, 54, 59, 75, 79, 108, 113, 122] for more recent studies. In the island model, individuals are at distance 0 or 1 depending on whether they belong to the same deme or not. As a consequence, the impact of population subdivision on genetic differentiation can be described by Wright's F_{ST} statistics [120]:

$$F_{ST} = \frac{f_0 - \bar{f}}{1 - \bar{f}}, \tag{2.7}$$

where f_0 (resp., $\bar{f}$) is the probability that two individuals sampled uniformly at random within the same island (resp., in the whole population) are *identical by descent*. Here, identity by descent (or IBD) is the property that the two individuals carry the same allele that they inherited from a common ancestor. In a model where lineages mutate independently of each other at some rate $\mu > 0$, and if we denote the coalescence time of the two ancestral lines by T, we can rewrite F_{ST} as

$$F_{ST} = \frac{\mathbb{E}_0[e^{-2\mu T}] - \mathbb{E}_{\mathrm{pop}}[e^{-2\mu T}]}{1 - \mathbb{E}_{\mathrm{pop}}[e^{-2\mu T}]}, \tag{2.8}$$

where under $\mathbb{P}_0$ (resp., $\mathbb{P}_{\mathrm{pop}}$) the two individuals are sampled in the same deme (resp., in the whole population). When μ is small enough for the product μT to be small too in some appropriate sense, we have the approximation

$$F_{ST} \approx \frac{\mathbb{E}_{\mathrm{pop}}[T] - \mathbb{E}_0[T]}{\mathbb{E}_{\mathrm{pop}}[T]},$$

which has the advantage of not requiring a precise estimate of μ. The expectations involved in this ratio can then be computed by solving a system of two linear equations relating $\mathbb{E}_0[T]$ and $\mathbb{E}_{\mathrm{pop}}[T]$. We do not give it here as it depends on the detailed dynamics of the model considered, but see Chapter 4.4 of [34] for an example. This opens the door to the inference of the small number of parameters characterising the evolution of the population, in a way which is relatively robust to the lack of precision in the estimation of mutation rates. However, although mutation rates for non-recombining (hence rather short) genes are indeed usually small, we shall see that in more general models of population structure the coalescence time may have a very large expectation. See Remark 2.6 in Section 2.5 in particular.

To our knowledge, F_{ST} is the main statistics used to measure the effect of a spatial structure on the genetic diversity of a population at a single non-recombining *locus* (region of interest in the genome). However, for populations distributed over a large one- or two-dimensional range, we expect the time to the most recent common ancestor for two individuals sampled at distance x to increase as x increases, so that the correlations between local allele frequencies should decay with distance (by an argument similar to the heuristics explaining (2.6)). This phenomenon, called *isolation by distance* [119], is not at all taken in account in F_{ST} as defined in (2.7) and the proper study of more complex models (such as the stepping stone model) is definitely needed to describe the signature left by such a spatial structure. In Section 2.5, we shall introduce a generalisation of F_{ST} which makes it a function of the distance between the sampling points, and we shall show that this new statistics can be used for inference purposes. In this perspective of reconstructing some of the (compound) parameters characterising the evolution of the population, it is also important to think of sampling schemes and their relations to the mathematical model used to predict the diversity observed in the sample. In particular, when the population lives in a continuum it may be difficult to split it artificially into discrete demes. In this case, a model dealing directly with continuous space is *a priori* more desirable. This is the direction we shall pursue in the rest of this chapter.

A natural generalisation to continuous space of the Wright-Fisher model described in Section 2.1 would be to add a term modelling the spatial diffusion of genes. Assuming again that there are only two alleles and that we follow the frequency $p_t(x)$ of one of them at each site x and time t, this yields the following equation:

$$\mathrm{d}p_t = \frac{\sigma^2}{2}\,\Delta p_t\,\mathrm{d}t + \sqrt{\frac{1}{N_e}\,p_t(1-p_t)}\,W(\mathrm{d}t, \mathrm{d}x), \tag{2.9}$$

where σ^2 is a diffusion coefficient, N_e is a ‘local population density’ (of which we can think as the inverse of the rate at which two ancestral lineages at the same location coalesce) and W is a space-time white noise. In one dimension, this approach is valid: there is a unique solution to (2.9), which can be obtained as the limit of a sequence of rescaled stepping stone models on $\mathbb{Z}$. This convergence result also shows that its dual genealogy is a system of independent Brownian motions which coalesce pairwise at a rate proportional to the local time that they spend at the same location. Unfortunately, in two dimensions Equation (2.9) has no solution and the sequence of rescaled stepping stone models on $\mathbb{Z}^2$ converges to the solution to the heat equation, dual to a system of independent Brownian motions that never coalesce (note that contrary to what happens in one dimension, in two dimensions two independent Brownian motions never meet). See Section 2.6.1 for the proof of a similar result. As a consequence, this approach is not appropriate if we want to model genetic drift in a continuous two-dimensional space.

In the 1940s, Wright and Malécot attempted to model populations living in a spatial continuum [76, 119]. Their model assumes that individuals are dispersed according to a Poisson point process of constant intensity λ in $\mathbb{R}^2$. The reproduction mechanism mimics that of the Wright-Fisher model: the population evolves in discrete generations and the number of offspring of each individual is Poisson with mean one. The spatial locations of the offspring are sampled independently from a Gaussian distribution centred on the position of their parent. In addition, the model incorporates mutation: with probability $\mu > 0$ an offspring, rather than inheriting the allele of its parent, mutates to an allele never before seen in the population. Wright and Malécot computed the probability of identity by descent of two individuals sampled at separation $x \in \mathbb{R}^2$. Let us denote this quantity by $F(x)$. Using a recursion, Malécot obtained an approximation for $F(x)$ in terms of K_0, the modified Bessel function of the second kind of order 0:

$$F(x) \approx \frac{1}{\mathcal{N} + \ln(\ell/\kappa)} K_0\left(\frac{\|x\|}{\ell}\right), \quad \|x\| > \kappa, \tag{2.10}$$

where $\kappa > 0$ is a local scale over which we assume that $F(x)$ is constant (to overcome the problem of the explosion of K_0 at 0), σ^2 is the variance of the Gaussian distribution that determines the offspring location, $\ell = \sigma/\sqrt{2\mu}$ can be seen as a characteristic length, and $\mathcal{N}$ is Wright's *neighbourhood size* which, in essence, measures the number of 'potential parents' in the neighbourhood of each offspring. The approximation (2.10) is referred to as the *Wright-Malécot formula.*

Unfortunately, the backwards-in-time recursion leading to (2.10) is based on the assumption that in any generation, the individual locations can be described as a Poisson point process of intensity λ in $\mathbb{R}^2$. This is inconsistent with the forwards-in-time evolution, in which the population forms clumps of larger and larger sizes, while some areas in space become empty. Considering compact areas instead of $\mathbb{R}^2$ does not solve the problem, as in this case the population with a critical reproduction mechanism dies out in finite time. This is Felsenstein's *pain in the torus* [46]. However, as discussed in Section 2.3 of the review paper [BEV13b], the Wright-Malécot formula with appropriate parameters fits astonishingly well the decay with distance of the probability of identity by descent in a stepping-stone model with, for example, nearest neighbour migration. Recall that in this model deme sizes are assumed to be constant, and so the recursive approach leading to (2.10) works. Obviously, in our model of evolution in a spatial continuum we need a mechanism to ensure a local regulation of the population density.

In this respect, several approaches have been tried, see for example [6, 9, 116, 117] and the references given therein. The framework presented in the next paragraph unifies them in a tractable way.

2.3 The spatial Λ-Fleming-Viot process

The model described here was introduced in [38] and fully formalised in [BEV10]. The main difference with former models of evolution is that reproduction is not based on individual clocks, but instead on a random sequence of events each affecting a given area in space. During such an event, a few parents are chosen at random and their offspring replace a fraction of the current population in the area. In this way, the population density remains constant but the local allele frequencies are updated based on the genetic types of the parents and on the fraction of individuals replaced. Below we give a particular formulation for the reproduction mechanism, but it can be generalised in many ways as long as we keep the essential ingredient of a Poisson point process of reproduction events specifying the geographical area in which the local genetic diversity will be updated. This is what we shall do for example when we need to add recombination (cf. Section 2.4.2), mutation (cf. Section 2.4.4), selection (cf. Section 2.6.1), or even spatial inhomogeneities (cf. [49]).

Suppose that the population is uniformly distributed over $\mathbb{R}^d$ ($d = 2$ being of course the most relevant dimension for most biological populations) and that the set K of all possible alleles is compact. We shall mostly consider $K = \{0, 1\}$ as in the Wright-Fisher model, or $K = [0, 1]$ to allow for arbitrarily many alleles to be present in the whole population. At any given time $t \geq 0$, the state of the population is represented by a measure $M_t(\mathrm{d}x, \mathrm{d}k)$ on $\mathbb{R}^d \times K$ whose first marginal is Lebesgue measure on $\mathbb{R}^d$. We write $\mathcal{M}_\lambda$ for the set of all such measures. Since any $M_t \in \mathcal{M}_\lambda$ can be decomposed as

$$M_t(\mathrm{d}x, \mathrm{d}k) = \mathrm{d}x\, \rho_t(x, \mathrm{d}k), \tag{2.11}$$

where $\rho_t : \mathbb{R}^d \to \mathcal{M}_1(K)$ is a Lebesgue-measurable map with values in the space of all probability measures on K (see Section 2.4.4 for a more precise statement), this representation indeed reflects the uniform density of individuals in space and $\rho_t(x, \mathrm{d}k)$ can be seen as the distribution of the allele of an individual which would be sampled at site x at time t. In fact, the state space $\mathcal{M}_\lambda$ is rather natural when we think of the population represented here as the limit of a population with discrete individuals whose positions form a Poisson point process with intensity $\lambda \mathrm{d}x$, as λ tends to infinity. See Remark 2.2 below.

Up to now we have only specified the encoding of the genetic diversity of the population. We now need a mechanism to make it evolve in time. To this end, let us fix a σ-finite measure μ on $(0, \infty)$ and a collection $\{\nu_r,\ r > 0\}$ of probability measures on $[0, 1]$. Let Π be a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}^d \times (0, \infty) \times [0, 1]$ with intensity $\mathrm{d}t \otimes \mathrm{d}z \otimes \mu(\mathrm{d}r)\nu_r(\mathrm{d}u)$. Thus,

$$\Pi = \big\{(t_i, z_i, r_i, u_i),\ i \in \mathbb{N}\big\}$$

is a random countable set of reproduction events described by their time t_i, centre z_i, radius r_i and *impact* u_i. More precisely, for every $i \in \mathbb{N}$, at time t_i a reproduction event occurs in the closed ball $B(z_i, r_i)$. A parent is chosen uniformly at random within $B(z_i, r_i)$ and its offspring, carrying the same allele, replace a fraction u_i of the local population at every site of the ball (the remaining fraction $1 - u_i$ is unaffected). In more mathematical terms, this means that a parental allele κ_i is sampled according to the distribution

$$\frac{1}{\mathrm{Vol}(B(z_i, r_i))} \int_{B(z_i, r_i)} M_{t_i-}(y, \mathrm{d}k)\mathrm{d}y$$

of alleles within $B(z_i, r_i)$ just before the event, and at every location $y \in B(z_i, r_i)$ we have

$$\rho_{t_i}(y, \mathrm{d}k) = (1 - u_i)\rho_{t_i-}(y, \mathrm{d}k) + u_i \delta_{\kappa_i}(\mathrm{d}k).$$

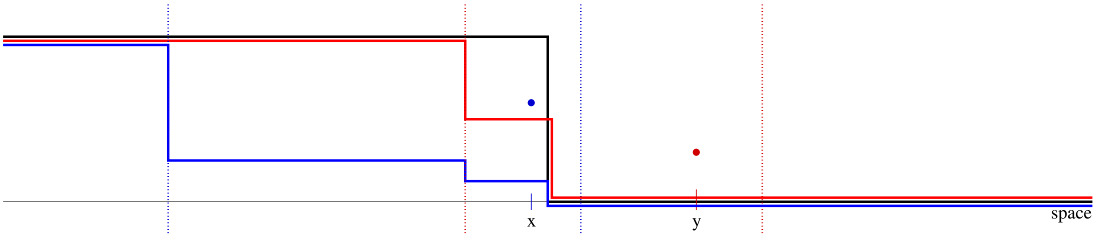


Figure 2.3 – The spatial Λ-Fleming-Viot process in one spatial dimension. Individuals carry one of the two alleles 0 and 1, and the local frequencies of allele 1 are represented. The black line shows the initial state, in which all individuals to the left carry the allele 1, and all individuals to the right carry the allele 0. A first event happens in the area delimited by the red vertical lines, a parent with allele 0 is chosen at location y and a fraction $u_1 = 1/2$ of the local population at every point in the ball is replaced by its offspring, carrying the allele 0 (red line). Some time later, an event occurs in the area delimited by the blue lines, a parent is chosen in the fraction of individuals at location x which are now of allele 0 and a fraction $u_2 = 3/4$ of the population in the ball is replaced by its offspring (blue line).

Figure 2.3 shows an example with $K = \{0, 1\}$ and $d = 1$; there the local frequencies of type 1 individuals are represented.

Observe from the intensity measure of Π that the reproduction events occur uniformly in time and space, while the coordinates r and u have more general distributions which are *a priori* correlated. This enables us for instance to model 'regular' reproduction events of small geographical extent and moderate impact at the same time as rare catastrophes, such as severe climatic events, affecting much larger areas and during which a significant fraction of the population goes extinct and is rapidly replaced by the descendants of a few lucky individuals. Finally, let us also remark that individuals do not migrate during their lifetimes in this formulation of the model (only their propagules disperse), but this generalisation is not difficult using the construction described in Section 2.4.4.

Originally, the existence and uniqueness of the process $(\rho_t)_{t\geq 0}$ corresponding to this evolution was shown in [BEV10] using a technique of [45] based on the characterisation of its semigroup through a well-chosen family of test functions. In [VW15], it is reformulated as a measure-valued process, whose existence and uniqueness is proved using more constructive arguments (see Section 2.4.4). Both techniques make use of the relation between the forwards-in-time evolution of allele frequencies $(M_t)_{t\geq 0}$ and the genealogy of a random sample of individuals, traced backwards-in-time. Before giving a precise statement of this duality relation, let us imagine what the ancestry of a few individuals should look like in this model.

First, suppose that we sample an individual at site x at a time that we call the present. For simplicity we denote this time by 0 (and assume that the time coordinate of the Poisson point process Π of reproduction events takes its values in $\mathbb{R}$ instead of $\mathbb{R}_+$, so that we may come back as far as we want in the past). Our aim is to trace back the position ξ_t of the individual at time $-t$ from which our sampled individual descends. For a given individual to be born during a reproduction event, it needs to lie within the area that is affected and to belong to the fraction of the population replaced. In this case, since the parent is chosen uniformly at random from the ball where the event takes place, its location is uniformly distributed over this ball and so is the new position of the ancestral lineage we are following. Since the law of the time-reversed point process $\{(-t_i, z_i, r_i, u_i),\ i \in \mathbb{N}\}$ is again that of a Poisson point process with intensity $\mathrm{d}t\otimes \mathrm{d}z\otimes \mu(\mathrm{d}r)\nu_r(\mathrm{d}u)$, we obtain that the rate at which the ancestral lineage of our individual jumps to a new location when it sits at $x\in\mathbb{R}^d$ is equal to

$$\mathcal{J}_0 := \int_{\mathbb{R}^d}\int_0^\infty\int_0^1 \mathbf{1}_{\{x\in B(z,r)\}} u\, \nu_r(\mathrm{d}u)\mu(\mathrm{d}r)dz = \int_0^\infty\int_0^1 uV_r\, \nu_r(\mathrm{d}u)\mu(\mathrm{d}r), \tag{2.12}$$

where V_r stands for the volume of a d-dimensional ball of radius r. If we further assume that $\mathcal{J}_0$ is finite (which we shall do in the rest of this chapter, though more general Lévy processes can be obtained under less stringent conditions) and if we write $L_r(y)$ for the volume of the intersection $B(0,r)\cap B(y,r)$, then the process $(\xi_t)_{t\geq 0}$ is a well-defined compound Poisson process which jumps from x to $x+y$ with an intensity given by

$$\begin{aligned}\mathcal{J}(y) &:= \int_{\mathbb{R}^d}\int_0^\infty\int_0^1 \mathbf{1}_{\{x\in B(z,r)\}} u\frac{\mathbf{1}_{\{x+y\in B(z,r)\}}}{V_r}\, \nu_r(\mathrm{d}u)\mu(\mathrm{d}r)\mathrm{d}z \\ &= \int_0^\infty\int_0^1 \frac{uL_r(y)}{V_r}\, \nu_r(\mathrm{d}u)\mu(\mathrm{d}r). \end{aligned}\tag{2.13}$$

Indeed, for such an event to happen, both the current location x of the lineage and the location $x+y$ of its parent should belong to the area $B(z,r)$ of the reproduction event, and then the parental position is chosen to be $x+y$ with density $1/V_r$.

Let us now sample another individual at site x' at time 0. Let us call $(\xi'_t)_{t\geq 0}$ the process tracing back the location of the ancestor of this second individual t units of time in the past. *A priori* $(\xi_t)_{t\geq 0}$ and $(\xi'_t)_{t\geq 0}$ are not independent, since they use the same Poisson point process of events to jump. If both of them happen to lie within the area of a given reproduction event, say with impact u, then

— with probability $(1-u)^2$ none of them belong to the local population replaced and the lineages do not move at that time,

— with probability $u(1-u)$ the ancestor of the first individual belongs to the fraction replaced but not the ancestor of the second individual, in which case ξ jumps to the location of the parent and ξ' stays where it is,

— with probability $u(1-u)$ ξ' jumps but not ξ,

— with probability u^2 both ancestors belong to the offspring of the (unique) parent chosen during the event. In this case, the two ancestral lineages merge into a single one, whose position is uniformly distributed over the area of the event.

Based on this observation, we see that the rate at which the two lineages ξ and ξ' merge when they are at separation $y\in\mathbb{R}^d$ is given by

$$\mathcal{C}(y) = \int_0^\infty\int_0^1 u^2L_r(y)\, \nu_r(\mathrm{d}u)\mu(\mathrm{d}r). \tag{2.14}$$

Note that this quantity is bounded by the jump rate $\mathcal{J}_0$ of a single lineage. Since $\mathcal{J}_0$ is supposed to be finite, we obtain that the pair (ξ,ξ') form a system of correlated compound Poisson processes which merge into a single lineage at an instantaneous rate given by (2.14).

The same analysis can be applied to any finite sample of individuals chosen at time 0 at a given set of locations. To describe the system of coalescing jump processes that we obtain more formally, we again resort to representing the ancestral relations within the sample at time $-t$ as a partition of $\{1,\ldots,n\}$. However, unlike the case of an unstructured population described in Section 2.1, here we need to keep track of the location of each ancestor. Hence, we use *marked* partitions of $\{1,\ldots,n\}$ of the form

$$\mathcal{A} = \big\{(b^1,x^1),\ldots,(b^k,x^k)\big\},$$

where the blocks $\{b^1, \dots, b^k\}$ form a partition of $[n]$ keeping track of which individuals in the sample have a common ancestor at the time of interest, and $x^j \in \mathbb{R}^d$ records the spatial location of the ancestor at that time of the individuals with labels in b^j. Let us denote the set of all marked partitions of $[n]$ by $\mathcal{P}_n^s$. The ancestral process of a sample of size n is thus a $\mathcal{P}_n^s$-valued Markovian jump process

$$(\mathcal{A}_t)_{t\geq 0} = \Big(\big\{(B_t^1, \xi_t^1), \dots, (B_t^{N_t}, \xi_t^{N_t})\big\}\Big)_{t\geq 0},$$

where N_t denotes the number of blocks (or distinct ancestors) t units of time in the past. To avoid cumbersome notation we do not give the full description of its jump intensity, but the interested reader will find all the necessary ingredients in the previous paragraph.

We can now unveil the duality relation between the spatial Λ-Fleming-Viot process $(M_t)_{t\geq 0}$ (or $(\rho_t)_{t\geq 0}$, recalling (2.11)), and the ancestral process $(\mathcal{A}_t)_{t\geq 0}$. Let us write $C(E)$ for the set of all continuous functions $f : E \to \mathbb{R}$, $C_c(E)$ for the subset of those which have compact support, $\wp_n(\mathbf{x})$ for the marked partition $\{(\{1\}, x_1), \dots, (\{n\}, x_n)\}$ made of singletons whose marks are given by the vector $\mathbf{x} = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$, and let us recall the notation $\mathcal{M}_\lambda$ for the space of all measures on $\mathbb{R}^d \times K$ whose first marginal is Lebesgue measure on $\mathbb{R}^d$. Then we have the following result, which corresponds to Theorem 4.2 in [BEV10] or Corollary 2.4 in [VW15].

Theorem 2.1. *Suppose that the quantity $\mathcal{J}_0$ defined in* (2.12) *is finite. Then there exists a unique $\mathcal{M}_\lambda$-valued Hunt process $(M_t)_{t\geq 0}$ such that for every $m \in \mathcal{M}_\lambda$, $t \geq 0$, $n \in \mathbb{N}$, $F \in C_c((\mathbb{R}^d)^n)$ and $g_1, \dots, g_n \in C(K)$,*

$$\mathbb{E}_m\bigg[\int_{(\mathbb{R}^d\times K)^n} F(x_1, \dots, x_n)\Big(\prod_{i=1}^n g_i(\kappa_i)\Big)\, M_t^{\otimes n}(\mathrm{d}x_1, \mathrm{d}\kappa_1, \dots, \mathrm{d}x_n, \mathrm{d}\kappa_n)\bigg] \tag{2.15}$$
$$= \int_{(\mathbb{R}^d)^n} F(x_1, \dots, x_n)\mathbb{E}_{\wp_n(\mathbf{x})}\bigg[\int_{K^{N_t}} \prod_{j=1}^{N_t}\Big(\prod_{i\in B_t^j} g_i(\kappa_j)\Big)\,\rho\big(\xi_t^1, \mathrm{d}\kappa_1\big)\cdots\rho\big(\xi_t^{N_t}, \mathrm{d}\kappa_{N_t}\big)\bigg]\,\mathrm{d}x_1\cdots\mathrm{d}x_n,$$

where we have used the decomposition $m(\mathrm{d}x, \mathrm{d}\kappa) = \mathrm{d}x\rho(x, \mathrm{d}\kappa)$.

The relation (2.15) will look less daunting when we have only two alleles (see Section 2.4.3). What it tells us is that to obtain the distribution of the alleles of n individuals sampled at distinct locations $x_1, \dots, x_n$ at time t, we can find who shared a common ancestor t units of time earlier (thus corresponding to *forward* time 0) and sample their common allele from the allelic distribution at time 0 at the location of the ancestor. This interpretation is at the basis of the particle representation described in Section 2.4.4. See also this section for general properties of the state space $\mathcal{M}_\lambda$, path properties of the process $(M_t)_{t\geq 0}$ and a result on the (not) *coming down from infinity* of the genealogical process $(\mathcal{A}_t)_{t\geq 0}$ starting with infinitely many individuals.

We end this section with a few remarks.

Remark 2.1. (Weaker conditions for existence and uniqueness of the process). *As mentioned earlier, the assumption that $\mathcal{J}_0$ should be finite is too strong for the existence and uniqueness result of Theorem 2.1. The most general condition required to use the technique of [45] is that the motion $(\xi_t)_{t\geq 0}$ of a single ancestral lineage should be a Lévy process, which is true under the weaker condition*

$$\int_{\mathbb{R}^d} (1 \wedge |y|^2)\,\mathcal{J}(y)\mathrm{d}y < \infty,$$

where $\mathcal{J}(y)$ was defined in (2.13)*. However, when we try to write down an infinitesimal generator or a martingale problem to identify the process $(M_t)_{t\geq 0}$ obtained by duality with the potential process that we have defined in terms of the Poisson point process of reproduction events, some technical problems (in the interchange of differentiation and taking expectations) appear. Since there is not much biological interest in the case where lineages accumulate infinitely many tiny jumps, we shall always suppose that $\mathcal{J}_0 < \infty$.*

Remark 2.2. (Individual-based model). *The spatial Λ-Fleming-Viot process is in fact the high-density limit of an individual-based model driven by the same Poisson point process Π of reproduction events. In this model, individuals are scattered on $\mathbb{R}^d$ according to a Poisson point process with some intensity $\lambda > 0$. When the region affected by a reproduction event contains no individuals, this event is simply cancelled. Otherwise, we choose a parent uniformly at random among the individuals present in the area, then each of these individuals dies with probability u (the impact of the event) independently of each other, and finally the region is repopulated by a Poisson point process of offspring with intensity $u\lambda$, all carrying the allele of the parent. In [12], the authors show that the population process (without alleles, just counting the number of individuals in any region of space) survives with probability* 1 *whenever λ is large enough, and they describe its ergodic behaviour. On the other hand, if λ is too small, the population becomes extinct a.s. Although it looks more natural to describe a population of discrete individuals, this model is very intricate to study. Taking the high-density limit $\lambda \to \infty$ leads to a much more tractable model (in particular, no events are cancelled since there is always someone in the affected area). A proof of this convergence can be found in [42].*

Remark 2.3. (Model with Gaussian kernels). *A version of the individual-based model and its high-density limit, using (arguably more realistic) Gaussian kernels instead of balls for the choice of the parent and the birth/death of individuals, was studied in [8]. In particular, the authors show that in the limiting model, when the variances of the Gaussian kernels are the same for all events, then the probability of identity by descent of two individuals sampled at separation $x \in \mathbb{R}^2$ is well approximated by the Wright-Malécot formula* (2.10) *with some appropriate parameters. We generalise this result in [BEKV13].*

2.4 Evolution under the hypothesis of neutrality

As a start, we assume that no alleles confer a reproductive advantage to the individuals carrying them. This is an important first step, in particular if we want to detect deviations from the null hypothesis of a neutral evolution.

2.4.1 Genealogies on a large torus

The first study carried out in the framework of the spatial Λ-Fleming-Viot process describes the long-term behaviour of the genealogy of a sample of individuals sitting far from each other on a continuous two-dimensional torus. It corresponds to Chapter 5 of my Ph.D. thesis [Véb09], and is also published in [BEV10]. It was motivated by some results on the collapse of structure in a stepping-stone model with finite-range migration on a discrete two-dimensional torus obtained in [26, 27] and then [25, 122]. In most of these works, the authors consider a torus of sidelength L. They show that under some appropriate conditions on the population size in each deme and on the migration kernel, the genealogy of a finite number of individuals sampled uniformly at random from the torus, when considered on the timescale $((L^2 \ln L)t,\, t \geq 0)$, converges to a time-changed Kingman coalescent as L tends to infinity. Indeed, this timescale corresponds to the mixing time of a random walk on the torus whose step distribution has finite range. Since any given number of individuals sampled independently and uniformly at random from the torus are at distance $\mathcal{O}(L)$ from each other with very high probability, well-known results on random walks ensure that no pairs of lineages can physically meet on a shorter timescale (see e.g. Section 2 in [24]). But after a time of the order of $L^2 \ln L$, the positions of the different lineages are uniformly distributed over the torus, and so each pair has the same chance to be the first one to meet (and merge very quickly after). Finally, since the probability that three independent random walks starting at distance $\mathcal{O}(L)$ gather at distance $\mathcal{O}(1)$ in $\mathcal{O}(L^2 \ln L)$ units of time tends to zero as L tends to infinity, in the limit we never see any merger of more than two lineages. The limiting genealogy is thus Kingman's coalescent.

In [BEV10], we complement these studies in several ways. First, we consider the more difficult case of continuous space. In this framework, two lineages cannot be at exactly the same position unless they have already merged, and coalescence occurs at a rate of the form (2.14) which is a function of the distance between the two (or more) lineages. Second, we consider reproduction events with bounded radii occurring at rate $\mathcal{O}(1)$, but also rare catastrophes covering a region of diameter L^α for some $\alpha \in (0,1]$. This gives rise to a wider class of possible genealogies in the limit as $L \to \infty$. The motivation for including such large scale extinction-recolonisation events comes from the observations that (a) on the timescale of evolution, i.e., hundreds to thousands of generations, many events like forest fires, epidemics, or severe climatic conditions may happen; (b) the exponential decay of the probability of identity by descent predicted by the Wright-Malécot formula (2.10) under the assumption that all reproduction events are local is a good approximation over relatively short distances, but fails to explain why local genetic diversities are correlated on much larger spatial scales than expected (see e.g. [36]). Therefore, it seems important to characterise the signature left by large but *a priori* rare catastrophes, to be able to compare it to that left by other evolutionary forces like natural selection.

Let us suppose that the population is distributed over a continuous two-dimensional torus $\mathbb{T}_L$ of sidelength L, which we can identify with $[-L/2, L/2]^2$ (with periodic boundary conditions) if needed. Since we shall concentrate on the genealogy of a sample of individuals, we do not specify the set of possible alleles. As explained above, we assume that there are two types of reproduction events, small and big. To describe the corresponding reproduction dynamics, we fix two measures μ^S and μ^B on $(0,\infty)$, with ranges bounded by some $R^S, R^B < \infty$, and two collections $\{\nu_r^S,\, r>0\}$, $\{\nu_r^B,\, r>0\}$ of probability measures on $[0,1]$. We also fix $\alpha \in (0,1]$ and a sequence $(\rho_L)_{L\geq 1}$ in $(0,\infty]$ tending to $+\infty$ as $L\to\infty$. Then:

— **Small events** are described by a Poisson point process Π^S on $\mathbb{R}\times\mathbb{T}_L\times(0,\infty)\times[0,1]$ with intensity $\mathrm{d}t\otimes\mathrm{d}z\otimes\mu^S(\mathrm{d}r)\nu_r^S(\mathrm{d}u)$. If $(t,z,r,u)\in\Pi^S$, then at time t a reproduction event occurs within $B(z,r)$. A parent is chosen uniformly at random in this region and its offspring replace a fraction u of the current population at every site of the ball.

— **Large events** are described by a Poisson point process Π^B on $\mathbb{R}\times\mathbb{T}_L\times(0,\infty)\times[0,1]$, independent of Π^S and with intensity $(\rho_L L^{2\alpha})^{-1}\mathrm{d}t\otimes\mathrm{d}z\otimes\mu^B(\mathrm{d}r)\nu_r^B(\mathrm{d}u)$. If $(t,z,r,u)\in\Pi^B$, then at time t a reproduction event occurs within $B(z,L^\alpha r)$. Again, a parent is chosen uniformly at random in the region and its offspring replace a fraction u of the current population there.

Note that since Π^S and Π^B are independent, we could have formulated the evolution in terms of a single Poisson point process $\Pi^S\cup\Pi^B$. In what follows, we are interested in the genealogy of a sample of n individuals picked uniformly at random from $\mathbb{T}_L$, which with high probability (when L is large) corresponds to considering sampling locations in

$$\Gamma(L,n) := \left\{\{x_1,\dots,x_n\} \in \mathbb{T}_L^n : |x_i - x_j| \geq \frac{L}{\ln L} \quad \forall i \neq j\right\}.$$

Using (2.13), we see that a given lineage makes jumps of size $\mathcal{O}(1)$ at rate $\mathcal{O}(1)$ as well as jumps of size $\mathcal{O}(L^\alpha)$ at rate $\mathcal{O}(1/\rho_L)$. Hence, denoting the position of the ancestral lineage t units of time in the past by ξ_t^L, it looks sensible to consider the rescaled process ℓ^L defined by

$$\ell_t^L := \frac{1}{L^\alpha}\,\xi_{2\rho_L t}^L, \quad t \geq 0. \tag{2.16}$$

The factor 2 in the rescaling of time comes from the fact that what we are really after is the difference between the positions of two ancestral lineages. Recall that R_B denotes our upper bound on the support of the intensity measure μ^B of radii for the large events. Since two ancestors cannot be hit by the same reproduction event unless they lie within distance $2R^BL^\alpha$ (in original units) of each other, they behave independently whenever they are far enough apart. As a consequence, ℓ^L is a finite-rate jump process with values in the torus $L^{-\alpha}\mathbb{T}_L$ of sidelength $L^{1-\alpha}$, starting at distance $\mathcal{O}(L^{1-\alpha})$ from the origin and evolving like a single rescaled lineage jumping twice as fast when it is not in the ball $B(0,2R^B)$ (how it evolves within this ball is slightly more delicate to describe). Using classical results on Poisson point processes (see Section 6.1 in [BEV10]), we can show that the contribution of small (resp., big) events to the variance of the displacement of ℓ^L over one unit of time (outside $B(0,2R^B)$) can be written

$$2\frac{\rho_L}{L^{2\alpha}}(\sigma_S^2 + o(1)), \qquad \text{resp., } 2\sigma_B^2 + o(1). \tag{2.17}$$

Comparing these contributions, we see that if $\rho_L/L^{2\alpha} \to 0$ as $L \to \infty$ only the big events drive the evolution, whereas if $\rho_L/L^{2\alpha} \to b^{-1} > 0$, both small and big events contribute to the motion. Finally, if $\rho_L/L^{2\alpha} \to +\infty$ it is the small events that drive the evolution and the rescaling (2.16) is not appropriate. In this case we work directly with the process $(\xi_{2t}^L)_{t\geq 0}$.

Assume first that $\alpha < 1$, so that big events are still of a negligible size compared to the whole population range. Then, an adaptation of the techniques developed in [24, 122] gives us that the timescale (in original time units) on which ancestral lineages meet at a distance which enables them to coalesce is $\rho_L L^{2(1-\alpha)}\ln(L^{1-\alpha})$ when big events are sufficiently frequent to have an impact on their motions, and $L^2\log L$ otherwise. Furthermore, once two lineages have managed to come sufficiently close to each other, they spend enough time at distance less than $2R^BL^\alpha$ (or $2R^S$ when only small events matter) to be affected by the same reproduction event and coalesce. Adapting again the results of [122] to show that no more than two lineages at a time can meet at distance less than $2R^BL^\alpha$, we obtain the following result. To set the notation, for every $n \in \mathbb{N}$ and every $L \in \mathbb{N}$, let $\mathcal{A}^{n,L}$ denote the process of marked partitions of $[n]$ describing the ancestry of n individuals sampled independently and uniformly at random over $\mathbb{T}_L$. Let also $\mathcal{A}^{n,L,u}$ denote the *unmarked* process recording only the blocks of $\mathcal{A}^{n,L}$. Finally, let $\mathcal{K}^n$ stand for Kingman's coalescent with sample size n (cf. Section 2.1), starting from the partition $\{\{1\},\dots,\{n\}\}$, and let $\Rightarrow$ denote weak convergence of càdlàg processes.

Theorem 2.2. (Th. 3.3 in [BEV10]). *Let $n \in \mathbb{N}$. As $L \to \infty$, we have*

$$(\mathcal{A}^{n,L,u}(\varpi_L t))_{t\geq 0} \Rightarrow \mathcal{K}^n,$$

where

$$\varpi_L = \begin{cases} \frac{(1-\alpha)\rho_L L^{2(1-\alpha)}\ln L}{2\pi\sigma_B^2} & \text{if } \frac{L^{2\alpha}}{\rho_L} \to +\infty, \\ \frac{(1-\alpha)L^2\ln L}{2\pi(\sigma_S^2+b\sigma_B^2)} & \text{if } \frac{L^{2\alpha}}{\rho_L} \to b \in [0,\infty) \text{ and } \frac{L^{2\alpha}\ln L}{\rho_L} \to +\infty, \\ \frac{L^2\ln L}{2\pi\sigma_S^2} & \text{if } \left(\frac{L^{4\alpha}}{\rho_L}\right)_{L\geq 1} \text{ is bounded or } \frac{L^2\ln L}{\rho_L} \to 0. \end{cases}$$

The quantities σ_S^2 and σ_B^2 are defined in (2.17).

As a consequence, as long as the largest events have a radius much smaller than the population range, the genealogy of a number of individuals sampled over the whole range is approximately described by the unstructured Kingman coalescent, when considered on the appropriate timescale.

If $\alpha = 1$, each big event affects a nonnegligible fraction of the population. In this case, the limiting ancestral process depends on how frequent these catastrophes are. Indeed, two lineages originally at distance $\mathcal{O}(L)$ need of the order of $\mathcal{O}(L^2 \ln L)$ units of time to come together and merge when they are affected only by small events. Thus, if $\rho_L \ll L^2 \ln L$, the lineages move around by little steps thanks to the small events and jump from time to time at distance $\mathcal{O}(L)$ due to a big event. Coalescence occurs after a finite number of big events. In this regime of parameters, the spatial structure still matters in the limit. If $\rho_L \propto L^2 \ln L$, the positions of the lineages have the time to homogenise over the torus before they are hit by a big event. Then any lineage present in the area of a big event can be affected by it with the same positive probability, and so we obtain a non-spatial coalescent with multiple mergers (or Λ-coalescent, see [91]) in the limit. If $\rho_L \gg L^2 \ln L$, the big events are so rare that Kingman-type coalescence events due to the small events reduce the ancestral process to a single lineage before any big event occurs. With the same notation as above for the marked and unmarked ancestral processes, we obtain:

Theorem 2.3. (Th. 3.7 in [BEV10]). *Let $n \in \mathbb{N}$ and for every $L \geq 1$, let*

$$\varpi_L = \begin{cases} \rho_L & \text{if } \frac{\rho_L}{L^2\ln L} \text{ has a finite limit,} \\ \frac{L^2\ln L}{2\pi\sigma_S^2} & \text{if } \frac{\rho_L}{L^2\ln L} \to +\infty. \end{cases}$$

Then as L tends to infinity,

(a) If $\rho_L L^{-2} \to b \in [0,\infty)$, the process $(\frac{1}{L}\mathcal{A}^{n,L}_{\varpi_L t})_{t\geq 0}$ with marks multiplied by $1/L$ converges weakly towards a process in which marks evolve according to independent Brownian motions with variance parameter $b\sigma_S^2$, and blocks coalesce whenever they are affected by (i.e., the corresponding ancestors belong to the fraction of the population replaced during) the same big event. Note that because of the rescaling of time, these big events happen at rate $\mathcal{O}(1)$ in the limit.

(b) If $\rho_L L^{-2} \to +\infty$, $\frac{2\pi\sigma_S^2\rho_L}{L^2\ln L} \to \beta \in [0,\infty)$ and if the total rate of occurrence of large events is finite (i.e., μ^B has finite mass), the unmarked process $(\mathcal{A}^{n,L,u}_{\varpi_L t})_{t\geq 0}$ converges weakly to a non-spatial Λ-coalescent in which lineages are involved in a multiple merger coalescence event at a rate equal to the rate at which an individual sampled uniformly at random from the torus belongs to the fraction of the population replaced during a big event (again, big events occur at rate $\mathcal{O}(1)$ on the timescale considered). Pairs of lineages also merge in Kingman-type events at rate β.

(c) If $\rho_L/(L^2 \ln L) \to +\infty$, the unmarked process $(\mathcal{A}^{n,L,u}_{\varpi_L t})_{t\geq 0}$ converges weakly towards Kingman's coalescent $\mathcal{K}^n$.

Here weak convergence always refers to weak convergence of càdlàg processes in the standard Skorokhod topology.

Thus, when the catastrophes can cover a significant part of the population range, depending on how fast these events occur we see again a collapse of structure (potentially with multiple mergers) when big events are sufficiently rare, or the influence of space remains in the limit.

Even when the genealogy of a random sample of individuals is well-approximated by the non-spatial Kingman coalescent, the influence of space, and large catastrophes in particular, is still felt through the timescale ϖ_L on which common ancestors are found. Recall from Section 2.1 that this timescale corresponds to an *effective population size* describing how much shorter or longer genealogies are compared to the panmictic case with a 'volume' L^2 of individuals. In the presence of small events only, the effective population size is of the order of $L^2 \ln L$ and the spatial structure gives rise to longer genealogical trees. When big events are sufficiently frequent, on the other hand, the effective population size ϖ_L can be much smaller than L^2 and genealogies are much shorter than expected in the panmictic case. These results open the door to the detection of the occurrence of rare but recurrent catastrophes of large geographic extent in the history of a population.

2.4.2 A more precise signature of the effect of space

In the previous section, we have seen that in most of the cases considered, the spatial structure of the population left a trace in the genealogy at a given locus of a random sample of individuals (recall that a locus is a region of interest in the genome). However, natural selection or fluctuations in census size could have similar effects on the timescale of common ancestry in a panmictic population. Hence, we need to look for a more detailed fingerprint of space to be able to detect its influence on the genetic diversity of the population.

A natural approach is to consider two loci on the same chromosome and measure their *linkage disequilibrium*. More precisely, because the two loci are on the same chromosome, we may suppose that any pair of alleles (A, B) at these loci is transmitted as such from parent to offspring. However, in sexual populations the mechanism of *recombination* can break this link, and an individual with two parents of alleles $(A, B), (a, b)$ may inherit a combination (A, b) or (a, B) instead. (This description is in fact not quite true as recombination occurs between the parental chromosomes during meiosis, but when considering many generations this simplification is reasonable). The association between the alleles observed at the two loci in an individual picked at random from the population is thus looser when the recombination rate is high, the extreme being the case of no linkage in which alleles at each locus are inherited from one of the parents in an independent way. When there are only two alleles A, a at the first locus, and B, b at the second locus, several measures of linkage disequilibrium in the population are classically used. For example, the r^2 statistics is defined by

$$r^2 := \frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)},$$

where p_A (resp., p_B) is the frequency of individuals in the population carrying the allele A at the first locus (resp., B at the second locus) and p_{AB} is the frequency of individuals carrying the pair of alleles (A, B). In particular, $r^2 = 0$ when there is no linkage between the two loci. Deriving an analytic expression for the expectation of r^2 is already difficult in the neutral panmictic case, and is probably impossible for a spatially structured population. However, by analogy with the duality relating the moments of the Wright-Fisher diffusion and the number of blocks/ancestors in Kingman's coalescent, we may expect to be able to relate the expectation of the numerator and of the denominator in the expression of r^2 to the correlation between the coalescence times of two lineages at each locus. Indeed, in [82], McVean shows that considering the ratio

$$\sigma_d^2 = \frac{\mathbb{E}[(p_{AB} - p_A p_B)^2]}{\mathbb{E}[p_A(1-p_A)p_B(1-p_B)]}$$

of expectations instead of the expectation of the ratio r^2 is a reasonably good approximation as long as all allele frequencies are bounded away from 0 (larger than 0.1, for example), and furthermore that σ_d^2 can itself be approximated by

$$\frac{\mathrm{Cov}\big(T_A^{(1,2)}, T_B^{(1,2)}\big) - 2\mathrm{Cov}\big(T_A^{(1,2)}, T_B^{(1,3)}\big) + \mathrm{Cov}\big(T_A^{(1,2)}, T_B^{(3,4)}\big)}{\mathbb{E}\big[T_A^{(1,2)}\big]^2 + \mathrm{Cov}\big(T_A^{(1,2)}, T_B^{(3,4)}\big)},$$

where $\{1,2,3,4\}$ are the labels of four individuals picked at random from the current population and $T_A^{(i,j)}$ (resp., $T_B^{(i,j)}$) stands for the time to the most recent common ancestor of individuals i and j at the first (resp., second) locus. In [112], Wakeley and Lessard find analytical expressions for σ_d^2 in an island model with a potentially very large number of demes, and apply their results to data from human populations. The distribution of r^2 and of other measures of linkage disequilibrium in a stepping-stone model with finitely many demes is explored by simulations in [28].

To study linkage disequilibrium in a population living in a continuum, let us thus introduce recombination in the spatial Λ-Fleming-Viot process and let us describe the correlations between the ancestries of two individuals at two linked loci. The results we shall describe below can be readily extended to any finite number of loci and of individuals. They will also apply to a discrete stepping-stone model with equivalent reproduction and scattering mechanisms.

Since we need more than one parents for recombination to occur (excluding *selfing*, i.e. self-fertilisation), let us extend the model described in Section 2.4.1 by choosing several parents during a reproduction event and allowing recombinant offspring to arise. To this end, let us fix two distributions λ_S and λ_B on $\mathbb{N}$, such that $\lambda_S(\{1\}) < 1$. To slightly simplify the analysis, let us fix $R_S, R_B \in (0,\infty)$ and $u_S, u_B \in (0,1)$, and let us suppose that all small events have radius R_S and impact u_S, and all big events have radius $R_B L^\alpha$ and impact u_B. Let also $(r_L)_{L\geq 1}$ be a nonincreasing sequence with values in $(0,1]$. Because there are many possible asymptotic behaviours for the genealogy at a single locus, as described in Section 2.4.1, in the remaining of this section we shall focus on the case where

$$\alpha \in (0,1), \qquad \rho_L \geq \ln L \quad \text{for all } L \geq 1 \qquad \text{and} \qquad \frac{\rho_L}{L^{2\alpha}} \to C \in [0,\infty),$$

in which large events are driving the evolution of the genealogies, possibly with some contribution from the small events. As in Section 2.4.1, we consider two kinds of events:

- **Small events** given by a Poisson point process on $\mathbb{R} \times \mathbb{T}_L$ with intensity $\mathrm{d}t \otimes \mathrm{d}z$. During an event (t,z), we choose a number k according to the law λ_S and we sample k individuals (or alleles) independently and uniformly at random within $B(z,R_S)$. If $k=1$, the offspring of the single parent replace a fraction u_S of the population at every site $y \in B(z,R_S)$. If $k>1$, at each site of the ball a fraction $1-u_S$ of the population remains the same as just before the event, a fraction $u_S(1-r_L)$ is replaced by offspring of each parent in equal proportions and the remaining fraction $u_S r_L$ is replaced by recombinant offspring inheriting their loci from two distinct parents (again, all pairs of parents have descendants in equal proportions).

- **Big events** given by an independent Poisson point process on $\mathbb{R} \times \mathbb{T}_L$ with intensity $\frac{1}{\rho_L L^{2\alpha}} \mathrm{d}t \otimes \mathrm{d}z$. During an event (t, z), we choose a number k of parents according to λ_B and sample k parents independently and uniformly in $B(z, R_B L^\alpha)$. Their offspring replace a fraction u_B of the current population in the ball, without recombining.

The assumption of no recombination during the big events is not compulsory, but it makes the analysis easier. Analogous results would be obtained if we allowed recombination to occur during all events, with a different timescale of decorrelation.

Let us reformulate the evolution of allele frequencies during an event slightly more formally. Using the notation $M_t(\mathrm{d}x, \mathrm{d}\kappa) = \rho_t(x, \mathrm{d}\kappa)\mathrm{d}x$ of Section 2.3 with $\kappa = (\kappa(1), \kappa(2)) \in K_1 \times K_2$, during a small event we sample $k \sim \lambda_S$ pairs of alleles $\kappa_1, \ldots, \kappa_k \in K_1 \times K_2$, and for every y in the area of the event we have (assuming that $k \geq 2$)

$$\rho_t(y, \mathrm{d}\kappa) = (1 - u_S)\rho_{t-}(y, \mathrm{d}\kappa) + \frac{u_S(1 - r_L)}{k} \sum_{i=1}^{k} \delta_{\kappa_i} + \frac{u_S r_L}{k(k-1)} \sum_{i \neq j} \delta_{(\kappa_i(1), \kappa_j(2))}.$$

On the other hand, during a big event we sample $k \sim \lambda_B$ pairs of alleles $\kappa_1, \ldots, \kappa_k \in K_1 \times K_2$ and for every y in the area of the event,

$$\rho_t(y, \mathrm{d}\kappa) = (1 - u_B)\rho_{t-}(y, \mathrm{d}\kappa) + \frac{u_B}{k} \sum_{i=1}^{k} \delta_{\kappa_i}.$$

Let us now focus on the genealogies of a sample of individuals at both loci. Let $\beta \in (\alpha, 1]$ and suppose that we sample two individuals at a distance $\mathcal{O}(L^\beta)$ much larger than the radius of the biggest reproduction events. We write (A, B) and (a, b) for their respective genetic types, τ_{Aa}^L (resp., τ_{Bb}^L) for the coalescence time of the lineages ancestral to the two individuals at the first (resp., second) locus. As explained earlier, what we are interested in is the correlation between τ_{Aa}^L and τ_{Bb}^L. In particular, we ask the following question: Is there a distance D_L^* such that if the two individuals are sampled at a distance smaller than D_L^*, the genealogies at their two loci are correlated, whereas if they are sampled at a distance much larger than D_L^*, the two genealogies are essentially independent. The results we expose below give an answer in the asymptotic regime $L \to \infty$, which of course depends on the rate at which recombination occurs. Before stating them, let us give an extension of Theorem 2.2 which is very similar to what is proved in [26, 122] for the voter model and the stepping stone model with finite-range migration. To fix the notation, under $\mathbb{P}_{x_L}$ two individuals are sampled at distance x_L on the torus $\mathbb{T}_L$ of sidelength L.

Proposition 2.1. (Prop. 1.2 in [EV12]). *Let $(x_L)_{L \geq 1}$ be a sequence in $(0, \infty)$ such that $\frac{\ln x_L}{\ln L} \to \beta \in (\alpha, 1]$ as $L \to \infty$. Then*

(a) For all $t \in [\beta, 1]$,

$$\lim_{L \to \infty} \mathbb{P}_{x_L}\left[\tau_{Aa}^L > \rho_L L^{2(t-\alpha)}\right] = \frac{\beta - \alpha}{t - \alpha}.$$

(b) For all $t > 0$,

$$\lim_{L \to \infty} \mathbb{P}_{x_L}\left[\tau_{Aa}^L > \frac{1 - \alpha}{2\pi\sigma^2} \rho_L L^{2(1-\alpha)} (\ln L) t\right] = \frac{\beta - \alpha}{1 - \alpha} e^{-t},$$

where

$$\sigma^2 = \lim_{L \to \infty} \frac{\rho_L}{L^{2\alpha}} \sigma_S^2 + \sigma_B^2.$$

Indeed, if we consider again a single rescaled lineage $L^{-\alpha}\xi^L_{\rho_L\cdot}$, its mixing time on the rescaled torus of sidelength $L^{1-\alpha}$ is $L^{2(1-\alpha)}\ln(L^{1-\alpha})$. On a much shorter timescale, it does not feel that space is limited and thus behaves like in $\mathbb{R}^2$.

Proposition 2.1 gives the asymptotic coalescence time at a single locus. Let us now consider both loci, defining $\tau^L := \tau^L_{Aa}\wedge\tau^L_{Bb}$. There are two regimes, depending on how fast recombination occurs.

Theorem 2.4. (Th. 1.4 in [EV12]). *Let again $(x_L)_{L\geq 1}$ be a sequence in $(0,\infty)$ such that $\frac{\ln x_L}{\ln L}\to\beta\in(\alpha,1]$ as $L\to\infty$. Then if*

$$\limsup_{L\to\infty}\frac{\ln\big(1+(\ln\rho_L)/(r_L\rho_L)\big)}{2\ln L}\leq\beta-\alpha, \tag{2.18}$$

we have

(a) For all $t\in[\beta,1]$,

$$\lim_{L\to\infty}\mathbb{P}_{x_L}\big[\tau^L>\rho_L L^{2(t-\alpha)}\big]=\frac{(\beta-\alpha)^2}{(t-\alpha)^2}.$$

(b) For all $t>0$,

$$\lim_{L\to\infty}\mathbb{P}_{x_L}\left[\tau^L>\frac{1-\alpha}{2\pi\sigma^2}\rho_L L^{2(1-\alpha)}(\ln L)t\right]=\frac{(\beta-\alpha)^2}{(1-\alpha)^2}e^{-2t}.$$

Consequently, under Condition (2.18) (satisfied for example when $r_L\equiv r\in(0,1]$ does not depend on L), asymptotically the coalescence times at the two loci become independent. Note that, in fact, we would need to consider more general event probabilities such as $\mathbb{P}_{x_L}(\tau^L_{Aa}>\rho_L L^{2(t-\alpha)},\tau^L_{Bb}>\rho_L L^{2(t'-\alpha)})$ with $t,t'\in[\beta,1]$ to draw such a conclusion. However, Proposition 2.1, Theorem 2.4 and the Markov property are sufficient to give their asymptotic behaviours and show the decorrelation claimed here.

When recombination is slower, we have instead:

Theorem 2.5. (Th. 1.5 in [EV12]). *Suppose that $(x_L)_{L\geq 1}$ is as in Proposition 2.1 and that there exists $\gamma\in(\beta,1)$ such that*

$$\lim_{L\to\infty}\frac{\ln\big(1+(\ln\rho_L)/(r_L\rho_L)\big)}{2\ln L}=\gamma-\alpha, \tag{2.19}$$

Then:

(a) For all $t\in[\beta,\gamma]$,

$$\lim_{L\to\infty}\mathbb{P}_{x_L}\big[\tau^L>\rho_L L^{2(t-\alpha)}\big]=\frac{\beta-\alpha}{t-\alpha}.$$

(b) For all $t\in(\gamma,1]$,

$$\lim_{L\to\infty}\mathbb{P}_{x_L}\big[\tau^L>\rho_L L^{2(t-\alpha)}\big]=\frac{(\beta-\alpha)(\gamma-\alpha)^2}{(\gamma-\alpha)(t-\alpha)^2}.$$

(c) For all $t>0$,

$$\lim_{L\to\infty}\mathbb{P}_{x_L}\left[\tau^L>\frac{1-\alpha}{2\pi\sigma^2}\rho_L L^{2(1-\alpha)}(\ln L)t\right]=\frac{(\beta-\alpha)(\gamma-\alpha)^2}{(\gamma-\alpha)(1-\alpha)^2}e^{-2t}.$$

Hence, until 'time' $\rho_L L^{2(\gamma-\alpha)}$ the genealogies at the two loci remain fully correlated and conditional on not having coalesced by this timescale, they start behaving independently. Using the definition of γ given in (2.19), we obtain that the amount of time the two loci need to decorrelate is of the order of $\rho_L(1+\ln\rho_L/(r_L\rho_L))$. To phrase it differently and answer the question on the critical sampling distance formulated earlier, we see that

$$D_L^* := L^\alpha\sqrt{1+\frac{\ln\rho_L}{r_L\rho_L}}.$$

Indeed, $(1+\ln\rho_L/(r_L\rho_L))$ is the order of magnitude of the time that two rescaled lineages $L^{-\alpha}\xi^L_{\rho_L\cdot}$ need to start evolving independently of each other. Hence, if they start a (rescaled) distance much larger than $(1+\ln\rho_L/(r_L\rho_L))^{1/2}$, the probability that they meet and merge before the decorrelation timescale tends to 0 as L tends to infinity.

Below we give the elements of the proof of Theorems 2.4 and 2.5 which explain the expression for the timescale of decorrelation. These elements give some clear insights into the local mechanisms responsible for the correlation between nearby lineages and how to 'escape' them. In particular, this work led us to introduce the concept of *effective recombination*, which will be at the basis of the most promising approach presented in Section 2.5 to infer some key statistics of the evolution of a spatially structured population. Before that, let us draw some conclusions on the linkage between the two loci. Theorems 2.4 and 2.5 give the behaviour of the lengths of the genealogies at the two loci, when the lineages are sampled at a distance much larger than the radius of the biggest events but possibly much smaller than the population range. Let us now assume that mutations occur at some small rates $\theta_1,\theta_2>0$ along each ancestral lineage, independently between lineages and between loci. As in Section 2.2, the probability of identity by descent at the first locus (resp., at both loci) for two individuals sampled at some distance x_L is given by

$$\mathbb{E}_{x_L}\big[e^{-2\theta_1\tau^L_{Aa}}\big], \qquad \text{resp., } \mathbb{E}_{x_L}\big[e^{-2(\theta_1\tau^L_{Aa}+\theta_2\tau^L_{Bb})}\big].$$

Thanks to Theorems 2.4 and 2.5, approximate analytical expressions for these quantities can be computed, see Equations (4) and (5) in [EV12]. Figure 2.4 and 2.5 illustrate the impact of purely local reproduction events, as well as of rare but recurrent big catastrophes on the probability of identity by descent (at one and then two loci) as a function of sampling distance. We see from Figure 2.4 that already at a single locus, identity by descent decays much more slowly in the presence of large events, implying that the correlation between local allele frequencies at a given locus persists over longer spatial scales. Figure 2.5 compares the different curves obtained when decorrelation always happens ($\gamma\leq\alpha$), when the genealogies are always fully correlated ($\gamma\geq 1$) and when we have a transition between these two regimes ($\gamma\in(\alpha,1)$). As expected, we see that the probability of identity by descent at both loci is higher in the presence of large events (when $\rho_L\leq L^{2\alpha}$), and there may be correlation between the two loci even when individuals are sampled over large spatial distances.

Elements of the proofs of Theorems 2.4 and 2.5: The core of the proofs resides in Figure 2.6. Indeed, we need to understand why and in how much time the lineages ancestral to the alleles at the two loci of a single individual start behaving as if they were independent. To this end, let us consider the individual carrying the pair of alleles (A,B). At the beginning, the two lineages are merged since they start in the same individual. When the first recombination event occurs, this lineage splits into two distinct ancestors sitting at distance less than $2R_S$ of each other (recall that recombination occurs only during small reproduction events). Because they are geographically close, the two lineages are then very likely to merge again quickly

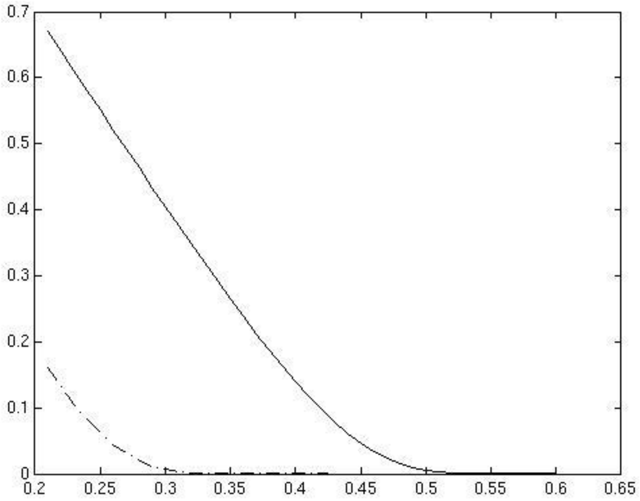


Figure 2.4 – Probability of identity by descent at a single locus, as a function of β. Here, $L = 10^5$, $\alpha = 0.1$, $\rho_L/L^{2\alpha} = 0.01$ and $\theta_1 = 10^{-3}$. The solid line corresponds to the case with small and large events, the dash-dot line to the case with only small events. Geographical correlations vanish around $\beta = 0.32$ without large events, and are positive up to $\beta = 0.52$ when large events occur.

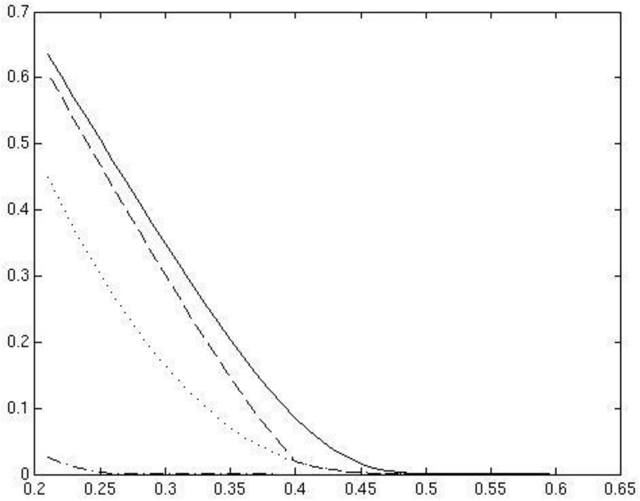


Figure 2.5 – Probability of identity by descent at both loci, as a function of β. As in Figure 2.4, $L = 10^5$, $\alpha = 0.1$, $\rho_L/L^{2\alpha} = 0.01$ and $\theta_1 = \theta_2 = 10^{-3}$. The solid line corresponds to the case $\gamma \geq 1$ (complete correlation for any β), the dotted line to the case $\gamma \leq \alpha$ (decorrelation for any β), and the dashed line to the intermediate case $\gamma = 0.4$. The dash-dot line corresponds to the case without large events.

before moving far apart. However, if at least one of them were affected by a big event at a time when the two lineages sit within two distinct individuals, their distance would become

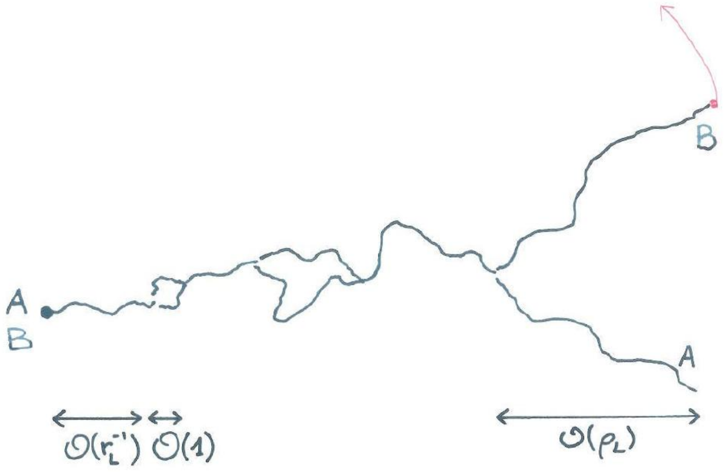


Figure 2.6 – Ancestral lineages of a single individual at two loci. Until a recombination event occurs, the ancestors at the two loci are identical and so the ancestral lines are merged into a single lineage. Then this lineage splits into two distinct lineages starting at nearby locations, which are therefore likely to coalesce again after a short time. The number of such excursions until they remain separated for a time long enough for at least one of them to be affected by a large event when the two lineages are distinct (the red arrow) is of the order of $\ln(\rho_L)$.

of the order of $\mathcal{O}(L^\alpha)$ and this would constitute an efficient first step towards independence. Let us thus define an *effective recombination* as a recombination event followed by the lineages becoming well-separated before merging again.

Let us find out how much time an effective recombination needs to take place. When the ancestral lineages are merged, the time to the next recombination event is of the order of $\mathcal{O}(r_L^{-1})$ (since small events hit the lineage at a rate $\mathcal{O}(1)$ and the recombination probability is r_L during each event). Then the two distinct lineages start a small excursion away from each other, during which they behave essentially like two independent finite-variance random walks on $\mathbb{R}^2$ until they come back at a small distance and merge again. For such an excursion to be long enough to see a big event, it needs to last at least $\mathcal{O}(\rho_L^{-1})$ units of time. Now, standard results on finite-variance random walks tell us that the probability that a given excursion lasts at least ρ_L units of time is of the order of $(\ln \rho_L)$, which means that we need about $\ln \rho_L$ excursions to see this happen. Considering that each 'small' excursion lasts $\mathcal{O}(1)$ units of time, we obtain that the time until the lineages jump far apart due to a big event is of the order of

$$(\ln \rho_L)\big(r_L^{-1}+1\big)+\rho_L \approx \rho_L\left(\frac{\ln \rho_L}{r_L\rho_L}+1\right).$$

Once they are separated, we can show that only a finite number of big events are sufficient to send the lineages far enough apart for them to start evolving independently. There only remains to add the second individual with alleles (a, b) in the picture, and to compare the decorrelation time of the lineages ancestral to A and B to the coalescence time of the lineages ancestral to A and a. When decorrelation does not have the time to occur, the lineages starting in the same individuals remain geographically close until they coalesce with the lineages ancestral to the other individual, implying that $\tau_{Aa}^L \approx \tau_{Bb}^L$. □

Note that the results presented in this section concern the genealogy of pairs of individuals sampled at very large distances. The separation of timescales between the very large amount of time that lineages starting in different individuals need to come sufficiently close and the remaining short time required to coalesce afterwards does not hold when the individuals are sampled too close together. Nevertheless, in Section 2.5 we shall see how to use the main ideas developped here to set up an inference method based on relatively local sampling to reconstruct two paramount statistics of the genetic evolution of the population, namely the *diffusion rate* σ^2 of lineages (or genes) through space and the *neighbourhood size* $\mathcal{N}$ summarising the competition between gene diffusion and local coalescence.

2.4.3 Evolution at an interface

Up to now we have been mainly concerned with the ancestral relations within a sample of individuals. Thanks to the duality relation (2.15) (which has not been explicitly given in the case with recombination because of its notational load), we can then translate the information we obtain on the system of coalescing marked partitions into properties of the more complicated measure-valued forwards evolution. This is what we do in this section, in a slightly different setting than in the two studies reviewed earlier. The results presented here correspond to the publication [BEV13a].

In Sections 2.4.1 and 2.4.2 we were interested in a population living in some large but finite range. In particular, we found the largest timescale on which several lineages sampled very far apart could find a common ancestor. However, when the range of the population is very large, it is likely that the past few hundreds of generations of evolution that we want to understand are not sufficient to reach this largest timescale, corresponding to the mixing time of ancestral lineages in the whole population range (that is, the amount of time after which a lineage has been able to cross the population range many times). We now want to focus on an intermediate time- and space-scale, during which ancestral lineages (and thus local genetic diversities) will not feel that space is limited. To start with a simple case, we assume that individuals can be of two types, or alleles, and we want to understand how the interface between two regions 'held' by different alleles evolves in the long run. Again, we consider two cases: when reproduction is only local, and when rare but large extinction-recolonisation events may happen.

In the notation of Section 2.3, we thus take $\mathbb{R}^d$ as the geographical space and $K = \{0, 1\}$ as the allele space. Since the density of individuals is constant across space and time, instead of the measure $M_t(\mathrm{d}x, \mathrm{d}\kappa) = \mathrm{d}x\rho_t(x, \mathrm{d}\kappa)$ we can simply consider the local frequencies of individuals carrying the allele 1, namely

$$w_t(x) = \rho_t(x, \{1\}). \tag{2.20}$$

Of course the mapping $x \mapsto \rho_t(x, \mathrm{d}\kappa)$ used in the decomposition of M_t is defined up to a Lebesgue nullset of $\mathbb{R}^d$, and so is the function $x \mapsto w_t(x)$. But what we are interested in is the convergence of a rescaling of the measure $w_t(x)\mathrm{d}x$, and so the definition (2.20) does make sense. We abuse notation and write again $\mathcal{M}_\lambda$ for the set of all mappings $w : x \to [0, 1]$ quotiented by the equivalence relation

$$w \sim w' \quad \text{if} \quad \mathrm{Vol}\{x \in \mathbb{R}^d : w(x) \neq w'(x)\} = 0.$$

In this case, the duality relation (2.15) takes a simpler form: For every $w_0 \in \mathcal{M}_\lambda$, $t \geq 0$, $k \in \mathbb{N}$

and $F \in C_c((\mathbb{R}^d)^k)$,

$$\mathbb{E}_{w_0}\Big[\int_{(\mathbb{R}^d)^k} F(x_1,\dots,x_k)\Big(\prod_{i=1}^k w_t(x_i)\Big)\mathrm{d}x_1\cdots\mathrm{d}x_k\Big]$$
$$= \int_{(\mathbb{R}^d)^k} F(x_1,\dots,x_k)\mathbb{E}_{\mathbf{x}}\Big[\prod_{j=1}^{N_t} w_0(\xi_t^j)\Big]\mathrm{d}x_1\cdots\mathrm{d}x_k, \tag{2.21}$$

where in somewhat sloppy notation $(\{\xi_t^1,\dots,\xi_t^{N_t}\})_{t\geq 0}$ is the system of coalescing jump processes recording the positions in $\mathbb{R}^d$ of the N_t ancestors, t units of time in the past, of a sample of k individuals taken at $\mathbf{x} = \{x_1,\dots,x_k\}$ under $\mathbb{P}_{\mathbf{x}}$.

Let us now introduce the two evolution rules in which we shall be interested. In both cases, for simplicity we assume that the impact of all events is fixed to some $u \in (0,1]$, but the results would hold in much greater generality.

— **Case A (Local evolution):** We also fix the radius of all events to some $R > 0$. That is, $\mu(\mathrm{d}r) = \delta_R(\mathrm{d}r)$ and the evolution happens through a Poisson point process of events of intensity $\mathrm{d}t \otimes \mathrm{d}z$ on $\mathbb{R}\times\mathbb{R}^d$. All events have radius R and impact u.

— **Case B (Rare large events):** We fix $\alpha \in (1,2)$ and set

$$\mu(\mathrm{d}r) = \frac{\mathbf{1}_{\{r>1\}}}{r^{d+\alpha+1}}\,\mathrm{d}r.$$

This time the evolution is described in terms of a Poisson point process on $\mathbb{R}\times\mathbb{R}^d\times(0,\infty)$ with intensity $\mathrm{d}t\otimes\mathrm{d}z\otimes\mu(\mathrm{d}r)$, all events having impact u.

Again, more general definitions of 'local' or 'rare large events' would lead to similar results, but we concentrate on the most notationally simple framework. Recalling the expression for the total rate of jump of a single lineage, $\mathscr{J}_0$ defined in (2.12), it is straightforward to check that the condition $\mathscr{J}_0 < \infty$ is satisfied in both cases. In Case B, this is guaranteed by the indicator function $\mathbf{1}_{\{r>1\}}$ which prevents the lineage from accumulating infinitely many tiny jumps (as it would do if μ was defined without the indicator function).

Recall from the beginning of this section that we want to understand the evolution on an 'intermediate' timescale of an interface between two regions originally occupied by individuals carrying different alleles. Thus, we define H as the half-space of positions in $\mathbb{R}^d$ whose first coordinates are negative and we set $w_0 = \mathbf{1}_H$. That is, all individuals in H carry the allele 1, while all individuals in $\mathbb{R}^d\setminus H$ carry the allele 0. In addition, we consider the timescale $(nt,\, t\geq 0)$, where $n\to\infty$ encodes what we mean by 'intermediate'. The game is then to find which spatial scales are relevant, in the sense that scaling space appropriately should lead to a nontrivial limit as we let the parameter n tend to infinity. To answer this question, the easiest approach is to think again of the genealogies of a sample of individuals. In Case A, when time is accelerated by a factor n, a given lineage jumps at rate nuV_R (where V_R is the volume of the d-dimensional ball of radius R) to a new location at distance $\mathcal{O}(1)$. Thus, if we scale down space by $\sqrt{n}$, we expect $n^{-1/2}\xi_{n\cdot}$ to converge to Brownian motion with a clock speed σ^2 given by the variance of the displacement of ξ over one (original) unit of time. Furthermore, using the same arguments as in the proof of Theorem 2.2, we can show that whenever two scaled lineages come within distance $2R/\sqrt{n}$ of each other, the additional time they need to coalesce on the new timescale is negligible. Before gathering at a distance that enables them to be in the range of the same reproduction events, the lineages evolve independently (again by the independence of Poisson point processes restricted to disjoint subsets). Putting all these ingredients together, and recalling that two independent Brownian motions meet only in one dimension, we obtain the following result.

Proposition 2.2. (Case A, Lem. 4.1 and 4.2 in [BEV13a]). *For every $k \in \mathbb{N}$ and every distinct $x_1, \ldots, x_k \in \mathbb{R}^d$, the scaled ancestral process*

$$\left(\left\{\frac{1}{\sqrt{n}}\xi_{nt}^1, \ldots, \frac{1}{\sqrt{n}}\xi_{nt}^{N_{nt}}\right\}\right)_{t\geq 0}$$

starting at $\{x_1, \ldots, x_k\}$ converges, in the sense of the convergence of the finite-dimensional distributions, towards a system of independent Brownian motions with clock speed σ^2. When $d = 1$, the Brownian motions coalesce instantaneously upon meeting; in higher dimension they never coalesce.

In fact we expect the sequence of scaled ancestral processes to be tight when starting from distinct positions, and hence the convergence stated above should hold also in the sense of weak convergence of càdlàg processes. But only the convergence of the finite dimensional distributions was necessary to obtain that of the corresponding forwards process, and so we did not prove tightness in [BEV13a].

Let us now consider Case B. The form of the jump intensity (2.13) of a single lineage that we obtain using the definition of μ makes us suspect that an α-stable process should appear with the right scaling of space. This time a direct analysis of the genealogies of a sample, as was done for the first case, is not easy. Instead, we write down the infinitesimal generator of the motion of a single lineage with time multiplied by n and space scaled down by $n^{1/\alpha}$, to find out that indeed our scaled lineage converges to a symmetric α-stable process under this scaling. Obtaining the limit of the generator of the scaled ancestral process with more than one lineage is more difficult, as the speed at which two nearby lineages coalesce through a very quick series of small events explodes as $n \to \infty$. However, if t_ε stands for the first time at which at least two scaled lineages lie at distance less than $\varepsilon > 0$ without having coalesced, we are able to show that for every set of distinct initial locations $\mathbf{x} = \{x_1, \ldots, x_k\}$, we have

$$\lim_{\varepsilon\to 0} \mathbb{P}_{\mathbf{x}}\big(t_\varepsilon < \infty\big) = 0.$$

This gives us that the martingale problem associated to the limiting generator has a unique solution whenever the initial state $\{x_1, \ldots, x_k\}$ is such that $\min_{i\neq j}|x_i - x_j| > 0$ (see Lemma 5.2 in [BEV13a]). This generator $\mathcal{G}^\alpha$ takes the form: for every f compactly supported and of class C^2, and every set $\mathbf{x}$ of distinct points in $\mathbb{R}^d$ with cardinality $|\mathbf{x}| < \infty$,

$$\begin{aligned}
\mathcal{G}^\alpha f(\mathbf{x}) = &\int_{\mathbb{R}^d} \mathrm{d}y \int_0^\infty \frac{\mathrm{d}r}{r^{\alpha+d+1}} \int_{B(y,r)} \frac{\mathrm{d}z}{V_r} \sum_{I\subset J(y,r,\mathbf{x}),|I|\geq 2} u^{|I|}(1-u)^{|J\setminus I|}\big[f\big(\Phi_I(\mathbf{x},z)\big) - f(\mathbf{x})\big] \\
&+ u\sum_{i=1}^{|\mathbf{x}|} \int_{\mathbb{R}^d} \mathrm{d}y \int_0^\infty \mathrm{d}r\, \frac{\mathbf{1}_{\{x_i\in B(y,r)\}}}{r^{\alpha+d+1}}\,(1-u)^{|J(y,r,\mathbf{x})|-1} \\
&\qquad\times \int_{B(y,r)} \frac{\mathrm{d}z}{V_r}\Big[f\big(\Phi_{\{i\}}(\mathbf{x},z)\big) - f(\mathbf{x}) - \big\langle z - x_i, \nabla_i f(\mathbf{x})\big\rangle \mathbf{1}_{\{|z-x_i|\leq 1\}}\Big] \\
+u\sum_{i=1}^{|\mathbf{x}|} &\int_{\mathbb{R}^d} \mathrm{d}y \int_0^\infty \mathrm{d}r\, \frac{\mathbf{1}_{\{x_i\in B(y,r)\}}}{r^{\alpha+d+1}}(1-u)^{|J(y,r,\mathbf{x})|-1} \int_{B(y,r)} \frac{\mathrm{d}z}{V_r}\,\big\langle z - x_i, \nabla_i f(\mathbf{x})\big\rangle \mathbf{1}_{\{|z-x_i|\leq 1\}},
\end{aligned}$$

where $J(y, r, \mathbf{x}) = B(y, r) \cap \mathbf{x}$ is the number of points in $\mathbf{x}$ sitting in the area $B(y, r)$ of an event, $\Phi_I(\mathbf{x}, z)$ is obtained from $\mathbf{x}$ by withdrawing all the points of $\mathbf{x}\cap I$ and adding the single point z, and $\langle\cdot,\cdot\rangle$ is the scalar product in $\mathbb{R}^d$. The last two terms describe the motion of a single lineage due to an event affecting it but not affecting the other lineages (even the $|J(y,r,\mathbf{x})|-1$ others lying in the region of the event). The first term describes the merger of several lineages and the uniform choice of the parental location over the area $B(y,r)$. We can now phrase the analogue of Proposition 2.2 in the case of rare but large events.

Proposition 2.3. (Case B, Prop. 5.1 in [BEV13a]). *For every $k \in \mathbb{N}$ and every distinct $x_1,\ldots,x_k \in \mathbb{R}^d$, the scaled ancestral process*

$$\left(\left\{\frac{1}{n^{1/\alpha}}\xi^1_{nt},\ldots,\frac{1}{n^{1/\alpha}}\xi^{N_{nt}}_{nt}\right\}\right)_{t\geq 0}$$

starting at $\{x_1,\ldots,x_k\}$ converges, in the sense of the convergence of the finite-dimensional distributions, towards the system of coalescing symmetric α-stable Lévy processes solution to the martingale problem associated to $(\mathcal{G}^\alpha,\mathbf{x})$. Furthermore, in the limiting ancestral process the number of lineages reaches 1 in finite time a.s.

The last property says that any finite sample of individuals will find a common ancestor in finite time a.s. This is not easy to see directly from the generator $\mathcal{G}^\alpha$, but finding a lower bound on the probability that two lineages coalesce before their distance doubles or is divided by two, and observing that the time before any such event occurs when the two lineages are at distance x is of the order of x^α, we can show the following result.

Lemma 2.1. (Lem. 5.3 in [BEV13a].) *Suppose two individuals are sampled at distance $x > 0$ and their genealogy is described by the limiting process of Proposition 2.3. Let τ be the coalescence time of their ancestral lineages. Then $\tau < \infty$ a.s., and there exists a random variable Z, a.s. finite and independent of x, such that*

$$\tau \preceq x^\alpha Z,$$

where $\preceq$ stands for stochastic domination.

Let us now see how these results translate into properties of the landscape of allele frequencies. First, inspired by our analysis of the ancestries, let us set $\alpha = 2$ in Case A and define w^n for all $n \in \mathbb{N}$ by

$$w^n_t(x) := w_{nt}\big(n^{1/\alpha}x\big), \qquad x \in \mathbb{R}^d, t \geq 0.$$

Since $w_0 = \mathbf{1}_H$, we also have $w^n_0 = \mathbf{1}_H$ for all n. A simple change of variable shows that the duality relation (2.21) holds also between w^n and the scaled ancestral processes. Now the set of functions of w considered in (2.21) is sufficient to characterise the convergence of a process with values in $\mathcal{M}_\lambda$, and Theorem 4.1 in [45] (or a straightforward generalisation of this result in Case B) asserts that there exists a unique $\mathcal{M}_\lambda$-valued process dual to each limiting ancestry. We can therefore conclude from Propositions 2.2 and 2.3 that in both cases A and B, the sequence of scaled spatial Λ-Fleming Viot processes converges to the dual to the corresponding limiting ancestral processes, in the sense of convergence of the finite dimensional distributions (in fact the one-dimensional distributions, but the finite-dimensional distributions can be obtained by considering a heterochronous ancestral process in which some subsamples are taken at different times). But the relation (2.21) does not readily give an explicit description of the local allele frequencies. For example, we cannot obtain the moments of a given $w_t(x)$ since the k sampling locations are a.s. distinct. However, the following general result enables us to fully describe the forwards processes in these cases.

Lemma 2.2. (Lem. 3.2 in [BEV13a]). *Suppose that $(w_t)_{t\geq 0}$ is an $\mathcal{M}_\lambda$-valued process dual to an exchangeable and consistent system of coalescing Markov processes $(\Xi_t)_{t\geq 0}$ through the relations (2.21). Let $(\xi_t)_{t\geq 0}$ denote the Markov process followed by a single lineage, and suppose that the initial condition of w is such that for every $t>0$, the map $z\mapsto \mathbb{E}_z[w_0(\xi_t)]$ is continuous on $\mathbb{R}^d$.*

(a) If for every $\varepsilon>0$ we have

$$\lim_{|y-x|\to 0}\mathbb{P}\big[\textit{lineages 1 and 2 have not coalesced by time } \varepsilon\,|\,\xi_0^1=x,\xi_0^2=y\big]=0,$$

where the convergence is uniform with respect to $x\in\mathbb{R}^d$, then for every $t>0$ and a.e. $x\in\mathbb{R}^d$, $w_t(x)$ is a Bernoulli random variable with parameter $\mathbb{E}_x[w_0(\xi_t)]$.

(b) If $(\Xi_t)_{t\geq 0}$ is a system of independent Markov processes which never coalesce whenever they start from distinct locations, then for every $t>0$ and a.e. $x\in\mathbb{R}^d$, $w_t(x)$ is deterministic and equal to $\mathbb{E}_x[w_0(\xi_t)]$.

Here, 'exchangeable' means that the law of Ξ is independent of the way we label the lineages; 'consistent' means that for any $j\in\mathbb{N}$, if Ξ starts with $j+1$ lineages but we only follow the evolution of the first j of them, we obtain a system of coalescing Markov processes which has the same law as Ξ started with only j lineages. These two properties hold in the framework of the spatial Λ-Fleming-Viot process. The proof of Lemma 2.2 is based on the fact that thanks to the continuity of $z\mapsto\mathbb{E}_z[w_0(\xi_t)]$, we can recover the moments of $w_t(x)$ by looking at the moments of the average frequency of allele 1 in small balls around x.

Using Lemma 2.2 and the different properties we have shown on the limiting ancestries, we can finally conclude that:

Theorem 2.6. (Case A, Th. 1.1 in [BEV13a]). *There exists an $\mathcal{M}_\lambda$-valued process $(w_t^{(2)})_{t\geq 0}$ such that*

$$w^n \longrightarrow w^{(2)} \qquad \textit{as } n\to\infty,$$

in the sense of weak convergence of the (temporal) finite-dimensional distributions. Furthermore, at every time $t\geq 0$, the local density of individuals of allele 1 can be described as follows. If X denotes standard d-dimensional Brownian motion and

$$p_t^2(x) := \mathbb{P}_x\big[X_{\sigma^2 t}\in H\big], \qquad t\geq 0, x\in\mathbb{R}^d,$$

then

(a) If $d=1$, for every $t\geq 0$ and a.e. $x\in\mathbb{R}$, $w_t^{(2)}(x)$ is a Bernoulli random variable with parameter $p_t^2(x)$. The correlations between their values at distinct sites of $\mathbb{R}$ are non-trivial and can be derived using the duality relation (2.21) with the ancestral process obtained in Proposition 2.2.

(b) If $d\geq 2$, for every $t\geq 0$ and a.e. $x\in\mathbb{R}^d$, $w_t^{(2)}(x)$ is deterministic and equal to $p_t^2(x)$.

Hence, in one dimension two alleles almost surely do not coexist at any given point, whereas in higher dimensions the two alleles 0 and 1 do coexist at every site instantaneously. When large reproduction events can happen, we obtain instead:

Theorem 2.7. (Case B, Th. 1.5 in [BEV13a]). *There exists an $\mathcal{M}_\lambda$-valued process $(w_t^{(\alpha)})_{t\geq 0}$ such that*

$$w^n \longrightarrow w^{(\alpha)} \qquad \textit{as } n\to\infty,$$

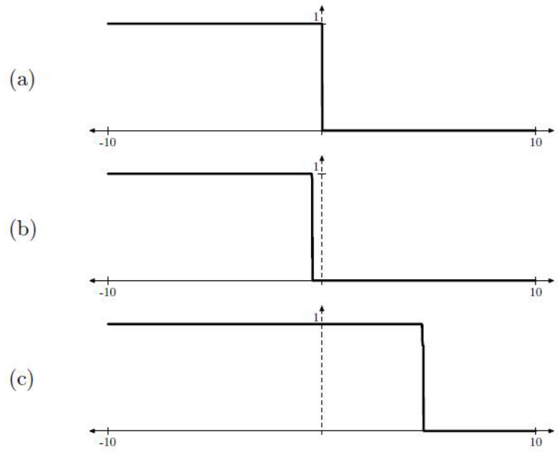


Figure 2.7 – Fixed radius in $d = 1$ on a line of length 20. (a) initial conditions; (b) after 10^5 events; (c) after 10^7 events. The model parameters are $u = 0.8$, $r = 0.033$, $n = 10^3$.

in the sense of weak convergence of the (temporal) finite-dimensional distributions. Furthermore, there exists a symmetric α-stable process X^α such that if

$$p_t^\alpha(x) := \mathbb{P}_x\big[X_t^\alpha \in H\big], \qquad t \geq 0, x \in \mathbb{R}^d,$$

then for every $t \geq 0$ and a.e. $x \in \mathbb{R}$, $w_t^{(\alpha)}(x)$ is a Bernoulli random variable with parameter $p_t^\alpha(x)$. The correlations between their values at distinct sites of $\mathbb{R}^d$ are non-trivial and can be derived using the duality relation (2.21) *with the ancestral process obtained in Proposition 2.3.*

This time segregation between alleles occurs in any dimension. Comparing the results of Theorems 2.6 and 2.7, we see that very large extinction-recolonisation events create correlations between local genetic diversities over much larger spatial scales ($n^{1/\alpha} \gg \sqrt{n}$) than purely local reproduction events. We also see that even in the dimensions where the α-stable process describing the spatial motion of a single lineage does not hit points, the large events are frequent enough to bring together two lineages which are far apart and make them coalesce in finite time. Let us end this section with a few simulations, performed by Jerome Kelleher (University of Oxford).

In Figure 2.7, which shows an example of evolution with purely local reproduction in one dimension, we can observe that starting from a half-line of individuals with allele 1, at any later time we recover the same pattern. The only difference is that the boundary has moved. This is due to the fact that in the limiting genealogies, lineages cannot jump over each other without coalescing, and so we cannot find an individual with allele 0 inside a patch of individuals of allele 1. In this case, the boundary moves according to Brownian motion with clock speed σ^2. On the other hand, when large events occur the lineages may jump over each other without coalescing, and at any time $t > 0$, $w_t^{(\alpha)}$ is the indicator function of a complex set. Figures 2.8 and 2.9 show an example in dimensions 1 and 2. In both figures, the initial state mimicks the occurrence of a large event, and the simulations show how quickly small events make the local allele frequencies come back to 0 or 1 afterwards. It would definitely be of interest, at least mathematically, to understand the fractal properties of the set of sites where individuals carry the allele 1.

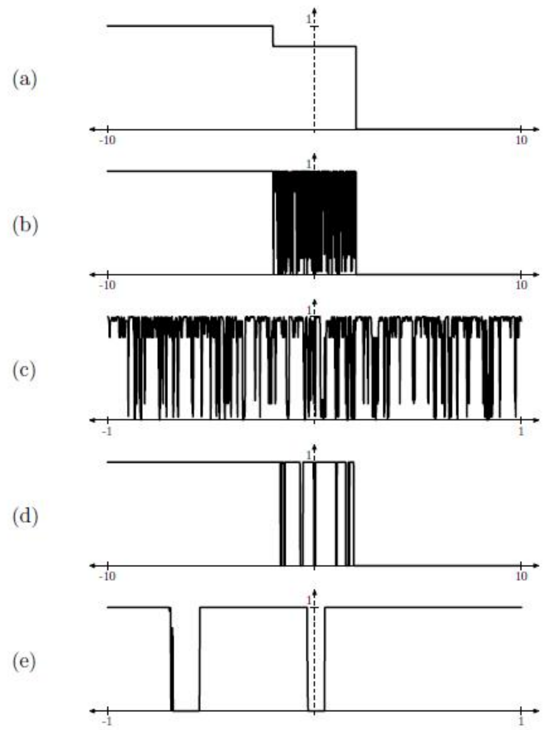


Figure 2.8 – Variable radius in $d = 1$ on a line of length 20. (a) initial conditions; (b) after 100 events, full range; (c) after 100 events, zooming in; (d) after 10^6 events, full range; (e) after 10^6 events, zooming in. The model parameters are $u = 0.8$, $n = 10^4$ and $\alpha = 1.3$.

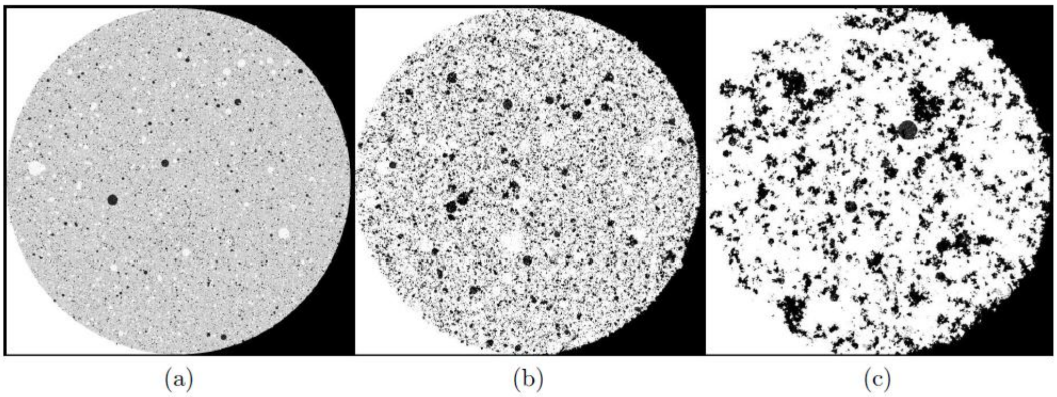


Figure 2.9 – Model in $d = 2$ after (a) 10^5; (b) 10^6; and (c) 10^7 events. We have a square range of edge 8, and the initial patch is a circle of radius 4 with frequency 0.8 (white is frequency 1, black is 0). The model parameters are $u = 0.8$, $\alpha = 1.3$ and $n = 10^3$.

2.4.4 The spatial Λ-Fleming-Viot process and its different sources of randomness

Until now we have remained a bit vague about the state space of the spatial Λ-Fleming-Viot process (SLFV), the associated topology and possible path properties of the process or its genealogies. Clarifying these points is the first goal of this section. Second, in this model different sources of randomness combine to make the local genetic diversities evolve in time. Indeed, once the Poisson point process of events is realised, the genealogies of a set of individuals are still random, since the lineages present in the area of an event decide independently of each other, and at random, whether they are affected by this event or not. When we want to have an interpretation of the model in terms of a population of infinitely many individuals evolving through time, this additional randomness should thus appear. We shall carry out a *quenched* particle construction of the spatial Λ-Fleming-Viot process which will disentangle the different sources of randomness and will enable us to add mutations in a straightforward manner. However, this construction is not dynamic, in the sense that we recover the value of the SLFV at a fixed time $t \geq 0$ as the empirical distribution of a countably infinite collection of particles tagged by their alleles, but there are no links between the discrete populations at two distinct times s and t. Our third goal is thus to turn the main ideas of the individual-based construction into a system of interacting paths with *levels*, giving rise to a *look-down* construction of the SLFV. These results correspond to the publication [VW15].

State space of the SLFV

Let us place ourselves in the general framework of Section 2.3. Recall that the geographical space on which the population is distributed is $\mathbb{R}^d$ (for example) and the compact space of possible alleles is denoted by K. The state space of the SLFV is the set $\mathcal{M}_\lambda$ of all nonnegative Radon measures on $\mathbb{R}^d \times K$ whose first marginals are Lebesgue measure on $\mathbb{R}^d$. A standard disintegration theorem (see e.g. [61], p.561) gives us that $\mathcal{M}_\lambda$ is in one-to-one correspondence with the quotient of the space of all Lebesgue-measurable maps $\rho : \mathbb{R}^d \to \mathcal{M}_1(K)$ by the equivalence relation

$$\rho \sim \rho' \qquad \Leftrightarrow \qquad \mathrm{Vol}\big(\{x \in \mathbb{R}^d :\ \rho(x, d\kappa) \neq \rho'(x, d\kappa)\}\big) = 0$$

(recall that $\mathcal{M}_1(K)$ denotes the set of all probability measures on K). As mentioned several times already, this correspondence is given by

$$m(dx, d\kappa) = dx\rho(x, d\kappa).$$

We endow $\mathcal{M}_\lambda$ with the topology $\mathcal{T}_v$ of vague convergence and the associated Borel σ-field. For every $k \in \mathbb{N}$, $F \in C_c((\mathbb{R}^d)^k)$, and $g_1, \ldots, g_k \in C(K)$, we write $G_{\mathbf{g}}(\kappa_1, \ldots, \kappa_k) := \prod_{j=1}^k g_j(\kappa_j)$ and define the function $I_k(\cdot\,; F; g_1, \ldots, g_k)$ on $\mathcal{M}_\lambda$ by

$$I_k(m\,; F; g_1, \ldots, g_k) := \big\langle m^{\otimes k}, F \otimes G_{\mathbf{g}}\big\rangle. \tag{2.22}$$

The following lemma gives some properties of the state space of the SLFV.

Lemma 2.3. (Lem. 2.1 in [VW15]).

(a) The space $(\mathcal{M}_\lambda, \mathcal{T}_v)$ is compact.

(b) For every $k \in \mathbb{N}$, $F \in C_c((\mathbb{R}^d)^k)$, and $g_1, \ldots, g_k \in C(K)$, the function $I_k(\cdot\,; F; g_1, \ldots, g_k)$ is $\mathcal{T}_v$-continuous on $\mathcal{M}_\lambda$.

(c) The linear span of the set of constant functions and of functions of the form $I_k(\cdot\,; F; g_1, \ldots, g_k)$, $k \in \mathbb{N}$, $F \in C_c((\mathbb{R}^d)^k)$ and $g_1, \ldots, g_k \in C(K)$ is dense in $C(\mathcal{M}_\lambda)$.

In particular, the set of functions of the form I_k constitute a wide enough family of tests functions to show the vague convergence of a sequence of measures. In addition, we can put a metric on the space $D_{\mathcal{M}_\lambda}[0, \infty)$ of all càdlàg paths with values in $\mathcal{M}_\lambda$, which will be useful for instance in the *look-down* construction carried out below. The following lemma can be found for example in Section 1 of [31].

Lemma 2.4. (Lem. 2.2 in [VW15]). *There exists a sequence $(f_n)_{n\geq 1}$ of uniformly bounded functions in $C_c(\mathbb{R}^d \times K)$ which separates points in $\mathcal{M}_\lambda$. Furthermore, if $(f_n)_{n\geq 1}$ is such a sequence, then*

$$d(m,m') := \sum_{n=1}^{\infty} \frac{1}{2^n} \left|\langle m, f_n\rangle - \langle m', f_n\rangle\right|, \qquad m, m' \in \mathcal{M}_\lambda,$$

defines a metric for the topology of vague convergence on $\mathcal{M}_\lambda$, while

$$\Delta\big((m_t),(m'_t)\big) := \int_0^\infty e^{-t} d\big(m_t, m'_t\big)\,\mathrm{d}t$$

is a metric for the topology of locally uniform convergence on $D_{\mathcal{M}_\lambda}[0,\infty)$.

The* quenched *SLFV and its genealogies

Having specified the state space on which we want to work, let us now define again the spatial Λ-Fleming-Viot process as in Section 2.3, but in a more constructive way. Indeed, to make the link with biological populations (in particular for statistical purposes), we would like to see the individuals of the population at some time $t \geq 0$, or a large subset of them, as a Poissonian sample from the measure M_t describing the landscape of local allele frequencies. The functional duality relation (2.15) suggests that the alleles of these individuals may be constructed from the initial state $M_0 \in \mathcal{M}_\lambda$ in three steps. First, we fix a realisation of the sequence of reproduction events; second, we trace back the genealogy of the countably many individuals by choosing at random which lineages are affected or not by a given reproduction event; lastly, every individual in the sample from time t receives the allele of its ancestor at time 0, drawn from the allele distribution $\rho_0(\xi_t)$ at its location. Seen in this way, it is then easy to add mutations in the evolution: Instead of inheriting the allele of its ancestor at time 0, each individual carries an allele obtained by starting a given mutation process from the ancestral allele, and letting it run along the path in the genealogy leading to our individual. A related motivation for the *quenched* approach (with respect to the law of the Poissonian sequence of events) adopted below is that, if we now consider several loci or even a fraction of the genome, the genetic diversities observed at these sites are correlated first and foremost by the fact that they all flow through the same sequence of events. Hence, any information on this sequence brought by the analysis of the diversity at one locus yields some constraints on the genetic diversity at other loci. A good understanding of these correlations is thus needed to devise statistical tests of the occurrence of massive extinction-recolonisation events or of the effect of natural selection.

Let us introduce some slightly different notation, only for this section. Let μ be a σ-finite measure on $(0,\infty)$ and $\{\nu_r,\, r>0\}$ be a collection of probability measures on $[0,1]$, satisfying

$$\int_0^\infty \int_0^1 u V_r \nu_r(\mathrm{d}u)\mu(\mathrm{d}r) = \mathcal{J}_0 < \infty.$$

(Again, V_r is the volume of the d-dimensional ball of radius r.) Let Π be a Poisson point process on $\mathbb{R} \times \mathbb{R}^d \times (0,\infty) \times [0,1]$ with intensity measure $\mathrm{d}t \otimes \mathrm{d}z \otimes \mu(\mathrm{d}r)\nu_r(\mathrm{d}u)$. We write $\mathbb{P}$ for the law of Π. The condition $\mathcal{J}_0 < \infty$ ensure that $\mathbb{P}$ assigns full measure to the set Ω of point configurations $\omega = (t_i, z_i, r_i, u_i)_{i\in\mathbb{N}}$ with the properties that $t_i \neq t'_i$ for $i \neq i'$ and that for all $s < t \in \mathbb{R}$ and every bounded subset B of $\mathbb{R}^d$,

$$\sum_{i: s\leq t_i \leq t, z_i \in B} r_i^d u_i < \infty. \tag{2.23}$$

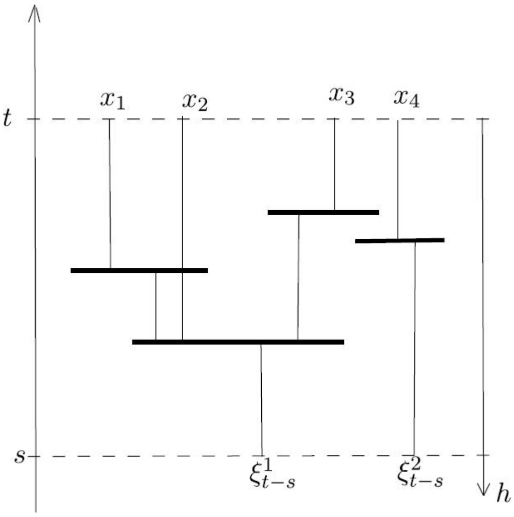


Figure 2.10 – An example of quenched genealogies. The area covered by each event is represented by a black horizontal line, whose height gives the time of occurrence of this event. Four individuals are sampled at present time (corresponding to forward time t) at locations $x_1, \dots, x_4$. Coming back into the past, the lineage ancestral to the first individual is affected by the first event it encounters but not the second lineage, which is also in the area of the event. However, they are both affected by the next event they encounter, and at that time they merge into a single lineage starting from the uniformly chosen location of the parent.

In what follows, we call any configuration ω an *environment*. We want to construct the random SLFV measure M_t given an environment ω as the empirical distribution of the locations and alleles of a set of countably many individuals. To this end, we start by introducing the *quenched* genealogy of a sample.

Recall the notation $\wp_k(\mathbf{x}) = \{(\{1\}, x_1), \dots, (\{k\}, x_k)\}$ for the marked partition representing a set of k distinct lineages whose positions are described by the vector $\mathbf{x} = (x_1, \dots, x_k)$. For $\mathbb{P}$-a.a. $\omega \in \Omega$ and $t \in \mathbb{R}$, let $P^{\omega,t}_{\wp_k(\mathbf{x})}$ denote the law of a system $(\mathcal{A}_h)_{h\geq 0}$ of coalescing marked partitions evolving as follows:

- The evolution of $\mathcal{A}$ starts at time $h = 0$ and uses (backwards in time) only the events $(t_i, z_i, r_i, u_i) \in \omega$ such that $t_i \leq t$.
- Whenever one or more lineages belong to the range of an event, each of the lineages within $B(z_i, r_i)$ takes part in this event with probability u_i, or remains unaffected with probability $1 - u_i$, independently of each other. All those lineages which are affected merge into a single lineage whose location is uniformly distributed over $B(z_i, r_i)$. Then $\mathcal{A}$ remains constant equal to its new value $\mathcal{A}_{t-t_i}$ until the next event of ω in the past which encompasses at least one of the lineages *and* for which at least one of these lineages takes part in the merger.

See Figure 2.10 for an example.

The condition $\mathcal{J}_0 < \infty$ guarantees that for any given $t \in \mathbb{R}$, for $\mathbb{P}$-a.e. environment ω, with probability 1 no lineage in $\mathcal{A}$ started at 'forwards' time t has an accumulation point of jumps in the time interval $[0, \infty)$. Hence, for $\mathbb{P}$-a.a. ω we can define $P^{\omega,t}_{\wp_k(\mathbf{x})}$ (for all $k \in \mathbb{N}$ and $dx^{\otimes k}$-a.a. $\mathbf{x} \in (\mathbb{R}^d)^k$) on the space $\mathcal{D}$ of coalescing marked partitions whose mark processes have càdlàg paths in $\mathbb{R}^d$. We write $P^t_{\wp_k(\mathbf{x})}$ for the joint distribution on $\Omega \times \mathcal{D}$ defined by

$$P^t_{\wp_k(\mathbf{x})}(\mathrm{d}\omega, \mathrm{d}a) := \mathbb{P}(d\omega)P^{\omega,t}_{\wp_k(\mathbf{x})}(\mathrm{d}a).$$

For $a \in \mathcal{D}$ and $h > 0$, we write $a_{[h]}$ for the restriction of a to the time interval $[0, h]$.

Let us now introduce our final ingredient, namely a mutation mechanism. Let $(\mathcal{K}_t)_{t\geq 0}$ be a Feller process with values in K, defined on some probability space $(\tilde{\mathcal{D}}, \tilde{\mathcal{F}}, \mathbb{Q})$. For every $\kappa \in K$ and every genealogical tree a, rooted in a single individual and having n leaves at some time $h > 0$, let us write

$$\mathbb{Q}^a_\kappa\Bigg[\prod_{j=1}^n g_j\big(\mathcal{K}^j_h\big)\Bigg], \qquad g_1, \ldots, g_n \in C(K)$$

to characterise the distribution of the alleles at the leaves when the root is of allele κ and alleles evolve along the branches of a according to the mutation process $\mathcal{K}$ (we assume that this evolution occurs independently along distinct subtrees emanating from the same vertex). We can now define the *quenched* spatial Λ-Fleming-Viot process with mutation.

Theorem 2.8. (Th. 1 in [VW15]). *For $\mathbb{P}$-almost all ω, there exists a unique $\mathcal{M}_\lambda$-valued time-inhomogeneous Hunt process $(M_t)_{t\in\mathbb{R}}$ whose two-parameter semigroup is characterised as follows: For every $s \leq t = s + h \in \mathbb{R}$, $m \in \mathcal{M}_\lambda$, $k \in \mathbb{N}$, $F \in C_c((\mathbb{R}^d)^k)$ and $g_1, \ldots, g_k \in C(K)$,*

$$\begin{aligned}&E^\omega_{s,m}\big[\big\langle M^{\otimes k}_t, F \otimes G_{\mathbf{g}}\big\rangle\big]\\ &= \int_{(\mathbb{R}^d)^k} F(x_1, \ldots, x_k)E^{\omega,t}_{\wp_k(\mathbf{x})}\Bigg[\int_{K^{N_h}} \prod_{i=1}^{N_h} \mathbb{Q}^{\mathcal{A}^i_{[h]}}_{\kappa_i}\Bigg[\prod_{j\in B^i_h} g_j\big(\mathcal{K}^j_h\big)\Bigg]\rho\big(\xi^1_h, \mathrm{d}\kappa_1\big)\cdots\rho\big(\xi^{N_h}_h, \mathrm{d}\kappa_{N_h}\big)\Bigg]\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathrm{d}x_1\cdots\mathrm{d}x_k,\end{aligned}$$

where $m = \mathrm{d}x\rho(x, \mathrm{d}\kappa)$ and $\mathcal{A}^1_{[h]}, \ldots, \mathcal{A}^{N_h}_{[h]}$ is the forest of N_h trees describing the genealogy of the sample between times 0 and h (in the past, starting from the 'forwards' time t).

This is just a formalisation of the construction given in words a few paragraphs earlier. Taking $s = 0$ and using the time homogeneity of the Poisson point process of events, we obtain the *annealed* SLFV as a direct corollary to Theorem 2.8.

Corollary 2.1. (Cor. 2.4 in [VW15]). *There exists a unique $\mathcal{M}_\lambda$-valued Hunt process $(M_t)_{t\geq 0}$ such that for every $m \in \mathcal{M}_\lambda$, $t \geq 0$, $k \in \mathbb{N}$, $F \in C_c((\mathbb{R}^d)^k)$ and $g_1, \ldots, g_k \in C(K)$,*

$$\begin{aligned}&E_m\big[\big\langle M^{\otimes k}_t, F \otimes G_{\mathbf{g}}\big\rangle\big]\\ &= \int_{(\mathbb{R}^d)^k} F(x_1, \ldots, x_k)E_{\wp_k(\mathbf{x})}\Bigg[\int_{K^{N_t}} \prod_{i=1}^{N_t} \mathbb{Q}^{\mathcal{A}^i_{[t]}}_{\kappa_i}\Bigg[\prod_{j\in B^i_t} g_j\big(\mathcal{K}^j_t\big)\Bigg]\rho\big(\xi^1_t, \mathrm{d}\kappa_1\big)\cdots\rho\big(\xi^{N_t}_t, \mathrm{d}\kappa_{N_t}\big)\Bigg]\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathrm{d}x_1\cdots\mathrm{d}x_k.\end{aligned}$$

When there are no mutations, i.e., when $\mathcal{K}$ is the constant process equal to its initial value, we recover the definition of the SLFV given in (2.15).

Before expounding the main ideas of the proof of Theorem 2.8, let us give a few properties of the *quenched* SLFV which can be proved using this construction. First, Theorem 2.8 tells us that it is a strong Markov process with càdlàg paths. In the absence of mutation, it has even stronger paths properties, due to the $\mathbb{P}$-a.s. property (2.23).

Lemma 2.5. (Lem. 2.6 in [VW15]). *For $\mathbb{P}$-a.e. environment ω, the* quenched *SLFV without mutation has paths of finite variation P^ω-a.s.*

Finally, again using the analysis carried out in the proof of Theorem 2.8, we can also show an interesting property of the genealogical process. A coalescent is said to *come down from infinity* if, starting with countably many lineages, there exists a time in the past at which the number of ancestors is finite. For non-spatial (exchangeable) coalescents, it is known that whenever the quantity corresponding to the impact u here is always less than 1, then either the coalescent comes down from infinity instantaneously with probability 1, or the number of ancestors remains infinite for all times a.s. (see, e.g., [100] and references therein). The question is more difficult when the population has a spatial structure, since migration could separate the lineages before they have an occasion to merge in the same local population. Several results of coming down or not coming down from infinity have been obtained in different models, for instance in [4, 50, 75]. In the framework of the SLFV, we can show that whenever the condition

$$\int_0^\infty \nu_r(\{1\})\mu(\mathrm{d}r) = 0 \tag{2.24}$$

holds, a countable sample of individuals taken from some region of space has infinitely many ancestors at any time in the past, with probability 1. In other words, when the impact parameter of an event is always less than 1 (which is the meaning of Condition (2.24)), the structured coalescent describing the genealogy of an infinite sample never comes down from infinity. More formally:

Proposition 2.4. (Prop. 5.2 in [VW15]). *Suppose that* (2.24) *holds. Let $\mathcal{N}$ be a Poisson point process on $B(0,1)\times[0,\infty)$ with intensity measure $\mathbf{1}_{B(0,1)}(x)\mathrm{d}x\otimes d\ell$ (i.e., Lebesgue measure for both coordinates), and let us use $\{(\{i\},x_i) : (x_i,\ell_i) \in \mathcal{N}\}$ as the initial value for the ancestral process corresponding to the* quenched *SLFV. Then for $\mathbb{P}$-a.e. environment ω, at any time $t > 0$ the set of ancestral locations at time t contains a Poisson point process on $B(0,1)$ with infinite intensity. In particular, the spatial Λ-coalescent never comes down from infinity.*

The proof of this result reveals that, more precisely, at any time in the past a positive fraction of the lineages have not yet been affected by an event (and therefore are still singletons).

Elements of the proof of Theorem 2.8: an individual-based construction

The key concept in the proof of Theorem 2.8 is that of *parental skeleton*. To fix ideas, take $s = 0$ and fix an environment ω. We want to construct a countably infinite population at time $t > 0$, uniformly spread over $\mathbb{R}^d$, in which the allele of each individual is determined by the allele of its ancestor at time 0 (possibly modified by a mutation process). Now, to each event of ω occurring in the time interval $[0,t]$ corresponds a parent, whose location z_i is chosen uniformly at random in the area of the event. Identifying the parent with its location and reproduction time, we can thus start an ancestral lineage from each of these space-time points (z_i,t_i) and make them be affected and potentially merge, or not, during each event in the past in which they sit. See Figure 2.11. We stop this process once we have come back to forwards time 0. To allocate an allele to every parent, it now suffices to sample the alleles of their ancestors at time 0 according to the initial distribution ρ_0 at their locations, and to run the mutation process $\mathcal{K}$ along the branches of the genealogies. This gives us a parental skeleton labelled by the alleles of all parents.

Let now $\tilde{\mathcal{N}}_1$ be a Poisson point process on $\mathbb{R}^d$ with intensity dx, representing countably many individuals living at time t. To decide of their alleles, we simply have to see whether

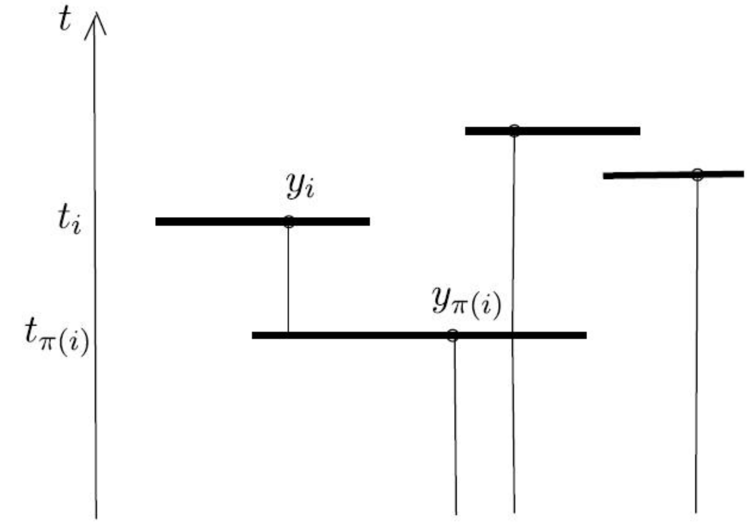


Figure 2.11 – Parental skeleton. Each horizontal line represents the area and time of a reproduction event. The location of the corresponding parent is represented by a circle. The ancestral lineage of this parent remains at its position until the first event in the past during which it belongs to the fraction of offspring created. At this time, it jumps to the location of the parent in this event. See the paragraph around Figure 2 in [VW15] for a full description of the notation.

and when their ancestral lines joins the parental skeleton. Indeed, if a given lineage is affected for the first time by an event occurring at time $t_i \in [0,t]$, then we start the mutation process from the allele of the parent during this event, and run it for $t - t_i$ units of time to obtain the allele of the corresponding individual. If it is not affected by an event in $[0,t]$, then we sample an allele from the allelic distribution ρ_0 at its location and run the mutation process, starting from this allele, for t units of time. In this way, we can define a random point measure $\mathcal{N}_1$ on $\mathbb{R}^d \times K$ by considering each pair of $x \in \tilde{\mathcal{N}_1}$ and the associated allele. Letting now $\tilde{\mathcal{N}_1}, \tilde{\mathcal{N}_2}, ...$ be an i.i.d. sequence of Poisson point processes and performing the same construction to obtain a sequence $(\mathcal{N}_n)_{n\in\mathbb{N}}$ of point measures on $\mathbb{R}^d \times K$, we see that conditionally on the labelled parental skeleton, the $\mathcal{N}_n$'s are i.i.d. and so it is straightforward to conclude from the strong law of large numbers that $\frac{1}{n}(\mathcal{N}_1 + \cdots + \mathcal{N}_n)$ converges almost surely, for the topology of vague convergence, towards a random limit $M_{0,t}$. There remains to check that this random measure satisfies the duality relation stated in Theorem 2.8, which can be done by using simple moment properties of Poisson point processes.

Remark 2.4. *We see from this sketch of proof that we may have chosen many other reproduction mechanisms based on a Poisson point process* Π *of events. For example, an event could be specified by its time, centre, the variance of the Gaussian kernel from which we draw a parental location (around the centre of the event), and the variance of the Gaussian kernel used to kill and replace individuals depending on their distance to the centre of the event. As detailed in Section 4.3 of [VW15], many variants of the SLFV process presented in this thesis can be constructed by a similar procedure.*

Remark 2.5. *We have to note that in our model, mutations occur between the reproduction events, and not during them as we could expect. But the construction carried out here is only* *an example of the many ways in which mutations may be included in the evolution of the population. It corresponds to a situation where most reproductions have a microscopic effect on the local allele frequencies but allow mutations to occur and the events of ω model the rarer events having a macroscopic effect. We could instead decide that the process of mutations along the branches of the tree jumps only at the internal nodes, which would lead to another SLFV with mutations.*

A look-down *construction*

As we mentioned earlier, the individual-based construction carried out in the proof of Theorem 2.8 does not give a dynamic picture of a population of countably many individuals reproducing and dying as time goes on, since the marked Poisson point processes $\mathcal{N}_n$ constructed for any given pair (s,t) of times have nothing to do with those corresponding to another pair (s',t') of times. A now standard way of obtaining such a global particle representation, constructing the measure-valued forwards process $(M_t)_{t\geq 0}$ and the backwards ancestral process $(\mathcal{A}_t)_{t\geq 0}$ on the same probability space, is to set up a *look-down* construction. This type of construction was introduced in [31, 32] and since then they have been used in various contexts, see for example [13, 14, 71]. Introducing a spatial structure complicates the matter as a lot of the exchangeability required between the ancestral lineages is lost due to the dependence of their behaviours on their spatial locations. It was done in [30] for continuum-sites stepping-stone models in which the genealogy consists in Feller processes which coalesce upon meeting, and in [42] to obtain (as a particular case) the convergence of the individual-based Poissonian model mentioned in Remark 2.2 to the spatial Λ-Fleming-Viot process as the density of individuals tends to infinity. These two approaches are rather different from that described below.

The main ideas of our construction are the following. Let us fix an environment ω and an initial measure $m \in \mathcal{M}_\lambda$. We start at time $t=0$ with a Poisson configuration of particles on $\mathbb{R}^d \times [0,\infty)$ with intensity measure $\mathrm{d}x \otimes \mathrm{d}\ell$. The first component is the particle's position, and the second will be called the particle's *level*. While the levels stay fixed in time, the locations ζ_t^j, given the environment $\omega = (t_i, z_i, r_i, u_i)_{i\in\mathbb{N}}$, perform independent jump processes: At each time t_i such that $\zeta_{t_i-}^j \in B(z_i,r_i)$, the particle at level ℓ_j tosses a coin with success probability u_i, independently of everything else. If the coin comes up with 'success', the particle jumps to a location $\zeta_{t_i}^j$ which is chosen uniformly at random within $B(z_i,r_i)$ (again independently of everything else). This mimicks the death of the individual sitting at $\zeta_{t_i-}^j$ and the birth of a new individual at some uniformly distributed location.

More formally, let us define the *forwards in time* motion $(\zeta_t)_{t\geq 0}$ of the (sequence of) individual(s) occupying a given level as above. Let us write P_x^ω for the probability measure on $\mathcal{D}$ under which ζ starts at $x \in \mathbb{R}^d$ (where $\mathcal{D}$ defined earlier is seen here as the space of coalescing càdlàg paths with values in $\mathbb{R}^d$). Let us now define a Poisson point process $\mathcal{N}$ on $\mathbb{R}^d \times \mathcal{D} \times [0,\infty)$ with intensity measure $\mathrm{d}x P_x^\omega(\mathrm{d}\zeta) \otimes \mathrm{d}\ell$. In words, we define a Poisson point process of trajectories starting at a Poissonian set of points of $\mathbb{R}^d$ and jumping thanks to the events of ω. In addition, each of these trajectories receives a label in $\mathbb{R}_+$, which will serve later to decide who reproduces. Because Lebesgue measure is invariant under the dynamics of ζ, at any time $t \geq 0$ the set of pairs (ζ_t^j, ℓ_j) still forms a Poisson point process on $\mathbb{R}^d \times [0,\infty)$ with intensity $\mathrm{d}x \otimes \mathrm{d}\ell$. This result holds true in particular at the fixed times of the events of ω, which guarantees that the spatial location of the individual with lowest level that jumps during a given event is uniformly distributed over the ball of the event. Thus, let us decree that this individual is the one which reproduces during the event, and that all 'newborns' (i.e., levels j such that ζ^j jumps during the event) *look down* on this lowest level and adopt its allele. Of course we have to assign an allele to each of the individuals living at time 0, which we do by sampling this allele from the distribution at time 0 at its location. In this way, we obtain a well-defined system of countably many paths in $\mathbb{R}^d \times K$, labelled by levels in $\mathbb{R}_+$. Then it is not difficult to show that the genealogical process of a sample of individuals taken at time t, i.e. the process of ancestral partitions marked by the locations of the individuals in the past from which our sample inherited their alleles, has the same distribution as the genealogical process of the quenched SLFV between times 0 and t. Consequently, as in the proof of Theorem 2.8, the empirical measure of the set of pairs of locations and alleles of all individuals alive at time t and with levels in $[0, n]$,

$$M_t^n := \frac{1}{n} \sum_{j:\ell_j \leq n} \delta_{(\zeta_t^j, \mathcal{K}_t^j)},$$

converges in distribution as $n \to \infty$ to the quenched SLFV at time t, M_t (with initial value $M_0 = m$). Note that to simplify the exposition there are no mutations in this construction, but it suffices to let the mutation process run independently on each level between two events to obtain the quenched SLFV with mutation.

In fact, the convergence holds in a pathwise manner. Recall from Lemma 2.4 the topology of uniform convergence over compact time intervals with which $D_{\mathcal{M}_\lambda}[0, \infty)$ is equipped.

Theorem 2.9. (Th. 2 in [VW15]). *For $\mathbb{P}$-a.e. environment ω, the sequence $(M^n)_{n\geq 1}$ converges P^ω-a.s. towards a process $(M_t^\infty)_{t\geq 0}$ which has the same law as the quenched spatial Λ-Fleming-Viot process of Theorem 2.8 with initial condition $M_0 = m$. This convergence is uniform over compact time intervals.*

The proof of Theorem 2.9 is inspired by the technique of [14]. The facts that we work with a fixed configuration of events and that reproduction occurs only locally in space introduce some technical issues that are overcome by controlling the number of events occurring and of particles present in a given area over a fixed interval of time (see the proof of Lemma 4.5 in [VW15]).

2.5 Inference in two dimensions

The spatial Λ-Fleming-Viot process, like other detailed models of evolution for a spatially structured population, is formulated in terms of a set of parameters. Even when there is only a small number of them, trying to reconstruct all of them may not be the most relevant goal to pursue. Indeed, these parameters describe the idealised way in which the individuals reproduce and transmit their genes *in the model*, but as we saw with the Wright-Fisher model for instance, they are not necessarily meant to reflect the biology of the organisms. Hence, to gain some understanding of the evolution of the population, instead of focusing on the reconstruction of the whole collection of parameters (among which there may be some auxiliary quantities), we should rather try to find a few summary statistics describing the fluctuations in local genetic diversities in a quantitative way, and devise statistical methods to infer them from data. This is our aim in this section, which corresponds to the publication [BEKV13a]. We focus on the most biologically relevant case of a two-dimensional spatial distribution.

To start with, recall from (2.10) the Wright-Malécot approximation for the probability $F(x)$ of identity by descent of two individuals sampled at some separation $x \in \mathbb{R}^2$ when the individual mutation rate is $\mu > 0$ (which can also be seen as the Laplace transform of the coalescence time T of their ancestral lineages):

$$F(x) = \mathbb{E}_x\big[e^{-2\mu T}\big] \approx \tilde{F}(x) := \frac{1}{\mathcal{N} + \ln(\ell/\kappa)} K_0\bigg(\frac{\|x\|}{\ell}\bigg), \quad \|x\| > \kappa,$$

where K_0 is the modified Bessel function of the second kind of order 0, σ^2 is the speed of the diffusion that describes the motion of a single lineage in the long run, $\ell = \sigma/\sqrt{2\mu}$ is a characteristic length, $\mathcal{N}$ is Wright's neighbourhood size which compares gene diffusion to local coalescence rates (see below), and finally κ is the local scale over which we consider the probability of identity by descent as being approximatively constant. (In practice, κ is not a very interesting parameter, as it is only a mathematically convenient way to cope with the explosion of the Bessel function at 0, and does not have a real biological meaning.) This approximation holds in a quite general class of models, in which the motions of two ancestral lineages have finite variance and are independent whenever they are far enough from each other, and coalescence can occur only when the lineages are 'reasonably' close together. For instance, in a stepping stone model with Wright-Fisher resampling within each deme of fixed size $2N$ and migration according to a discretised Gaussian kernel with variance s^2, it holds with $\sigma^2 = s^2$ and $\mathcal{N} = 2\pi\sigma^2 \times (2N)$ (where the $2N$ should be read as the inverse of the local coalescence rate $1/(2N)$). In the SLFV with all events having a fixed radius R, a fixed impact $u \in (0,1)$ and such that a fixed number $\nu \geq 1$ of individuals reproduce during each event, we have $\sigma^2 = u\pi R^4/2$ and $\mathcal{N} = \nu/u$ (see Figure 2.12). Since we suppose that the probability of identity by descent is approximately constant over distances less than κ, in both examples κ can be obtained (as a function of σ^2, μ and $\mathcal{N}$) by solving $\hat{F}(\kappa) = \hat{F}(0)$, where $\hat{F}(0)$ is the supposedly observable probability of identity by descent for two individuals sampled close-by (here we have abused notation by seeing $\hat{F}$ as a function of distance instead of separation in $\mathbb{R}^2$). The quantity $\hat{F}(0)$ can be obtained by simulation if we want to fit the probability of identity by descent in a given model with its Wright-Malécot approximation, or it can be estimated by sampling pairs of individuals locally when we want to apply this approximation to some real population. More generally, if σ_e^2 is the effective variance of the long-term motion of a single lineage and if we define ρ_e, the *effective population density*, as half of the inverse of the integral over all $x \in \mathbb{R}^2$ of the rate at which two lineages at separation x coalesce, we have $\mathcal{N} = 4\pi\sigma_e^2\rho_e$. Note that ρ_e could be considerably lower than the actual population census size, as we already discussed in the context of the Wright-Fisher model without space. We refer to [6] and Appendix A of [BEKV13a] for the derivation of this result and more examples.

We thus see that the two (or three) compound parameters σ^2 and $\mathcal{N}$ (and κ) summarise the long-term evolution of the population somewhat independently of the fine details of the two-dimensional structure of the population. Hence, let us suppose that our population evolves in such a way that the Wright-Malécot formula is a good approximation for the probability of identity by descent of two individuals sampled at some 'intermediate' distance. We want to devise some statistical methods to reconstruct σ^2 and $\mathcal{N}$ from different types of data. The first approach presented below uses measures of the local allele frequencies. We shall see that, unfortunately, it enables us to infer $\mathcal{N}$ and an analogue of κ, but not σ^2. Our second approach is based on the typical length of a *block of conserved sequences*. More precisely, instead of sampling many individuals to estimate the local allele frequencies at a given locus at distinct locations, we sample a moderate number of individuals at several pairwise distances and consider a long stretch of their DNA. For each pair sampled at distance r, we look at the typical length of a connected region of DNA at which the two individuals' sequences are identical because they were inherited from a recent common ancestor.

In all that follows, we suppose that we sample our individuals in a large region of space $\mathcal{D}$ such that we can find a reference time t^* (of the order of a few hundred generations ago) satisfying

$$\frac{\mathrm{diam}(\mathcal{D})^2}{\sigma^2} \ll t^* \ll \frac{1}{\mu}, \tag{2.25}$$

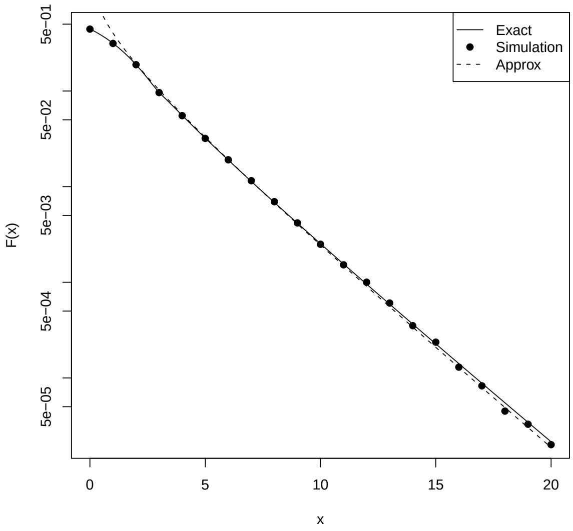


Figure 2.12 – Probability of identity by descent plotted against distance for the spatial Λ-Fleming-Viot model with parameters $\nu = 1$, $R = 1.5$, $\lambda = 1$, $u = 0.5$ with a mutation rate $\mu = 10^{-4}$ on a torus of diameter 64 (Here we assume that each lineage is hit by a mutation at rate μ). The numerical solution to the integro-differential equation satisfied by the probability of IBD, simulations and the Wright-Malécot solution (with $\kappa \approx 1.34$), are shown. Simulation results report the mean identity over 10^5 replicates.

where σ and μ are as in the Wright-Malécot formula and diam($\mathcal{D}$) stands for the diameter of $\mathcal{D}$. The left part of Condition (2.25) ensures that over the last t^* units of time in the past, the lineages of a sample of individuals have the time to homogenise their positions in $\mathcal{D}$ or to leave the area for a very long while. As in the study presented in Section 2.4.1, this enables us to assume that the allelic distribution in $\mathcal{D}$ (i.e., the average allele frequencies over the whole region $\mathcal{D}$) has remained approximately constant to some quasi-equilibrium over this lapse of time. Moreover, if an escapee finally comes back and finds some common ancestor with the rest of the sample, the time it took it to come back allows for the apparition of mutations differentiating its allele from the others, rendering its long excursion observable. On the other hand, the right part of Condition (2.25) ensures that those lineages which did not escape coalesce before mutating. In particular, the alleles present in a sample are all distinguished from each other by mutations much deeper than t^* in the genealogy. This makes our two methods robust to the deep and possibly complex history of the population, since they are based only on the dichotomy quick coalescence/long excursion away separated by our intermediate reference time t^*.

Inference based on allele frequencies

This first approach is based on a variant of Wright's F_{ST} statistic (2.7). Suppose we observe m alleles in the region $\mathcal{D}$, and write $p_i(x)$ for the frequency of the i-th allele at location x at sampling time. Let also $\bar{p}_i$ be the frequency of the i-th allele in $\mathcal{D}$ (assumed to have remained at a quasi-equilibrium over the last few hundred generations). We want to compare the correlations in allele frequencies between two sampling sites at a specific distance r to those between two sampling sites chosen uniformly at random within $\mathcal{D}$. To fix the notation, let us write $\mathbb{P}_r$ for the distribution of the paths in $\mathbb{R}^2$ followed by the ancestral lineages of two individuals sampled at distance r and T for their coalescence time. Under $\mathbb{P}_{\mathcal{D}}$, the two individuals are sampled independently and uniformly over $\mathcal{D}$. In accordance with our assumption of local equilibrium in $\mathcal{D}$, we assume that if the two lineages have not coalesced by time t^*, the chance that they are of different alleles is independent of their initial separation and is given by $H(t^*)$, the heterozygosity at time t^* in the past. Hence, if

$$H_r = 1 - \mathbb{E}\Big[\sum_{i=1}^{m} p_i(x)p_i(y)\Big]$$

is the heterozygosity between two sites x, y sitting at distance r, we have

$$H_r = \mathbb{P}_r(T > t^*)H(t^*)$$

and if we now define our distance-dependent analogue of Wright's F_{ST} by

$$F'(r) := \frac{H_{\mathcal{D}} - H_r}{H_{\mathcal{D}}}, \tag{2.26}$$

(where $H_{\mathcal{D}}$ is the average heterozygosity over pairs of locations in $\mathcal{D}$), we obtain that

$$F'(r) = \frac{\mathbb{P}_{\mathcal{D}}(T > t^*) - \mathbb{P}_r(T > t^*)}{\mathbb{P}_{\mathcal{D}}(T > t^*)}$$

is independent of $H(t^*)$.

Remark 2.6. *Note that the classical approach that we have seen in particular in the island model and* (2.7) *is based on a comparison of the coalescence times with the timescale of mutation instead of the reference time t^*, which corresponds to replacing t^* by an exponential random variable with parameter 2μ in the definition of $F'(r)$. However, the return time to a neighbourhood of the origin for a finite variance symmetric random walk in two dimensions has infinite expectation, and so has the coalescence time of two lineages. As a consequence, the approximation*

$$\frac{\mathbb{E}_r[e^{-2\mu T}] - \mathbb{E}_{\mathcal{D}}[e^{-2\mu T}]}{1 - \mathbb{E}_{\mathcal{D}}[e^{-2\mu T}]} \approx \frac{\mathbb{E}_{\mathcal{D}}[T] - \mathbb{E}_r[T]}{\mathbb{E}_{\mathcal{D}}[T]}$$

does not hold and the statistic we would obtain in this way would depend on our estimate of the mutation rate μ (usually not available).

Using some estimates on continuous-space random walks to compare $\mathbb{P}_r(T > t^*)$ to the reference $\mathbb{P}_0(T > t^*)$, we can write that

$$F'(r) \approx \tilde{F}'(r) := \frac{\ln(\bar{r}/r)}{\mathcal{N} + \ln(\bar{r}/\kappa')},$$

where $\bar{r}$ is the geometric mean of the distance of all possible pairs of individuals sampled from $\mathcal{D}$, $\mathcal{N}$ is as in the Wright-Malécot formula and κ' is a local scale chosen so that

$$\widehat{F'}(0) \approx \tilde{F}'(\kappa') = \frac{\ln(\bar{r}/\kappa')}{\mathcal{N} + \ln(\bar{r}/\kappa')},$$

where as before $\widehat{F'}(0)$ is the supposedly observable value of F' for very small sampling distances. Here again, the parameter κ' is a model- (or data-)dependent distance over which F' is assumed to be nearly constant (its relation to the parameter κ appearing in the Wright-Malécot formula is not clear).

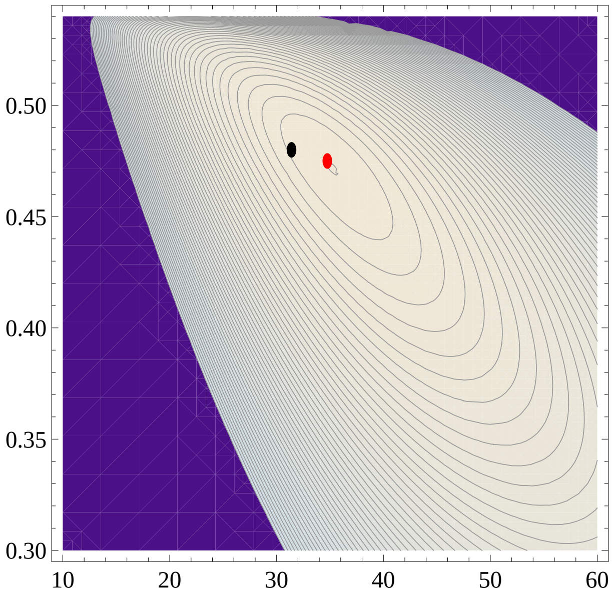


Figure 2.13 – Likelihood surface based on ten 'loci' sampled from a 10×10 patch within the 40×40 population; log likelihood is plotted against $\mathcal{N}$ (x-axis) and κ' (y-axis). The MLE (red dot) is $\mathcal{N} = 34.75$, $\kappa' = 0.475$; the true $\mathcal{N} = 31.4$, $\kappa' = 0.48$ (black dot), with log-likelihood lower by 1.5. Contours are spaced at 2 units of log-likelihood, so that the inner circle indicates the support limits for each parameter.

Remark 2.7. *Observe that, unfortunately, the approximation for $F'(r)$ is independent of σ which will thus not be reconstructible by such a method. By analogy with the Wright-Malécot formula for which we usually do not know the mutation timescale $1/\mu$, and thus cannot estimate σ^2 from the knowledge of $\mathcal{N}$, κ and $\hat{F}(0)$, here we do not know exactly the timescale to which the coalescence times are compared (the dichotomy between quick coalescence and long excursions away is very crude) and $\bar{r}$ only implicitly contains some information on σ^2 (via the condition* $\mathrm{diam}(\mathcal{D})^2 \ll \sigma^2 t^*$*).*

On the other hand, if for every pair of sampling sites x and y we define

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

then using the quasi-equilibrium assumption for the average allele frequencies and the fact that $H_r = \mathbb{P}_r(T > t^*)H(t^*)$, we can show that

$$\mathbb{E}\big[\mathcal{F}(x,y)\big] = F'(\|x-y\|),$$

so that $\mathcal{F}(x,y)$ provides a statistic for $F'(r)$. A maximum likelihood approach based on these results and on the assumption that the current local allele frequencies $p_i(x)$ are small Gaussian deviations from their means $\bar{p}_i$ is presented in Section 4.2 of [BEKV13a]. Figure 2.13 shows an example of likelihood surface for the parameters $\mathcal{N}$ and κ' obtained by implementing this method. Since it does not allow us to reconstruct σ^2, we do not give more details but rather turn to the more promising second approach.

Inference based on lengths of conserved blocks

The previous method fails to identify σ^2 because it is only based on the dichotomy between short or long coalescence times, and therefore does not enable us to date the recent coalescence events sufficiently precisely to obtain an estimate of how quickly two lineages initially at some distance r took to meet and coalesce. However, if we consider long stretches of DNA instead of a set of independent nonrecombining loci, recombination will have occurred relatively often in the recent past and will provide an additional evolutionary clock. Indeed, two genomes that share an ancestor t generations in the past will share a portion 2^{-t} of their genomes, in blocks of map length $\sim 1/t$. Here and thereafter, lengths of blocks of DNA are measured in Morgans, a Morgan being defined as the sequence length over which we see on average one recombination per generation. Thus, sharing exceptionally long blocks indicates recent common ancestry and the block size gives an approximate date for that ancestry. This idea is exploited by [93] to identify recent shared ancestry in a sample of 2257 Europeans (the POPRES dataset).

There is a subtlelty that we must take into account if we want to exploit recombination in this way. If a recombination event occurs, then at that moment in time, the two resulting ancestral lineages are adjacent to one another. As we have seen in Section 2.4.2, often they will coalesce again before time t^* and, since we are assuming that there will be no mutations over that period, the recombination event will not be visible in our data. However, with some probability they do not coalesce by time t^*. If this happens, this is typically because they have escaped far from one another. As a result, they only coalesce in the distant past and we do expect to see some mutations occur before that time. It is these recombination events that we expect to be able to detect and in keeping with Section 2.4.2, we call them *effective* recombination events. Because an 'ineffective' recombination can change the genealogy of the sample in a way which is not detectable in the data, the resulting distribution of detectable blocks is very complex. We thus consider only two particular regimes, (a) when neighbourhood size is so large that two recombinants never coalesce again before t^* (i.e., all recombination events are effective) and (b) when neighbourhood size is small enough that if they do not manage to escape from each other, two recombinants coalesce back very quickly and this 'ineffective recombination' does not change the genealogy (this case corresponds to the regime of parameters studied in Section 2.4.2 over much larger spatial scales, and with the additional occurrence of large events). In what follows we suppose that we sample two individuals at some distance r (still in the region $\mathcal{D}$) and look at the length of a block of DNA which is identical in the two individual sequences. Again, we write T for the coalescence time of the ancestral lineages of the sample.

Let us start with the case of large $\mathcal{N}$. Recall that recombination occurs at rate 1 on every unit length of sequence measured in Morgans, independently between the two lineages we consider. Thus, if we fix a focal locus and move along the genome in a given direction from there, conditionally on the coalescence time T at the focal locus the length B of the portion of genomes shared by the two individuals is an exponential random variable with parameter $2T$. Now, the Wright-Malécot formula gives us an approximation for the Laplace transform of the distribution of T, and so from this it is easy to obtain that

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

Since recombinants never coalesce again, distinct 'half-blocks' of shared sequence are independent and so the above probability can be inferred from the tail distribution of the empirical CDF of the size of a half-block.

In the more complicated case of small neighbourhood size, we need to fix a model to compute an approximation for the escape probability of two nearby lineages. The expression we obtain below is actually not exactly in terms of σ^2, $\mathcal{N}$ and κ, but we shall argue that the tail of the distribution of the conserved blocks depends only on these three parameters. Hence, let us work with the SLFV in which all events have radius $R > 0$, impact $u \in (0, 1)$ and $\nu \geq 2$ parents are chosen each time. We also suppose that loci are discrete (instead of continuous as in the previous case) and linearly organised on the genome, and that during a reproduction event seen backwards in time the probability that two neighbouring loci are inherited from distinct parents is $\rho \in (0, 1)$. Assume again that we sample two genomes at distance r, supposed to be large compared to the radius R of a reproduction event. As we already mentioned, for a given *individual*, whenever two recombinant lineages are created (both of which being ancestral to that individual but at different loci), either they coalesce very quickly or they manage to escape from one another and only coalesce in the distant past, by which time they have accumulated many mutations. For a given *locus*, in order to see identity by descent at this locus in our sample of size two, the lineages ancestral to the two individuals must have coalesced in the recent past. Thus, a block of consecutive loci that are identical by descent will have length b if an effective recombination happens for one of the individuals between the b-th and $b+1$st loci, the (very few) lineages ancestral to the two individuals at the b loci in IBD coalesce quickly, while the lineages ancestral to the two individuals at the $b+1$st locus escape from one another instead of coalescing quickly. To be a bit more precise, let us define a coalescence to be *quick* or *early* for two lineages initially at distance r if it takes place before an exponentially distributed time with mean $r^2/2\sigma^2$ (the minimal amount of time required for two lineages at distance $r \gg R$ to meet and possibly coalesce). The reason for this definition is that we then implicitly suppose that if a coalescence is quick, it happens in this minimal exponentially distributed amount of time indeed; we then expect these ‘early’ coalescence events to generate the largest blocks of identity by descent. Let us also define a recombination to be effective if the recombinant lineages become separated by distance r before they coalesce again. Because the probability that a recombination is effective is very small, we can make the approximation that effective recombination events between two neighbouring loci occur at rate $\rho_{\text{eff}}(r) := (\pi R^2 u\rho)\alpha(r)$, where $\pi R^2 u\rho$ is (truly) the rate at which recombination occurs between these loci in the SLFV and $\alpha(r)$ is the probability that the resulting lineages move away at distance r instead of coalescing. As shown in Appendix C2 of [BEKV13a], the function α is solution to an equation that can be solved numerically. Putting all these ingredients together, we obtain that the probability that an effective recombination occurs between the b-th and $b+1$st loci in one of the two individuals' ancestry is approximately equal to

$$\frac{2\rho_{\text{eff}}(r)}{2\rho_{\text{eff}}(r) + 2\sigma^2/r^2}.$$

(Note that the quantity $\zeta(r)$ appearing in the statement of Theorem 1 in [BEKV13a] is not $r^2/(2\sigma^2)$ as it is unfortunately defined in the paragraph just before this statement, but instead is $\zeta(r) = (\sigma/r)^2$ as defined in Appendix C1.) Finally, the probability that the lineages at the $b+1$st locus, which have reached distance r, do not coalesce quickly is obtained from the Wright-Malécot formula with $\mu = \sigma^2/r^2$ (and $\mathcal{N} = \nu/u$):

$$\mathbb{P}_r(\text{no early coalescence}) = 1 - \frac{K_0(\sqrt{2})}{\mathcal{N} + \ln(r/(\kappa\sqrt{2}))}.$$

Combining the above, we obtain the following result.

Theorem 2.10. *Suppose we sample two individuals at distance r and let X be the length of a block of consecutive loci at which the two individuals are identical by descent because of an early coalescence. Then X follows approximately a geometric distribution with parameter $\gamma(r)$ given by*

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

where all the quantities have been defined above.

In Figure 2.14, we see that the geometric block length distribution predicted here is reasonably accurate if we sample from far enough apart. There the red curve represents the average length of a block due to early coalescence, defined as above as a block of consecutive loci at which coalescence happens before (a given realisation of) an exponentially distributed time with parameter $2\sigma^2/r^2$. Such a block may have undergone some non-effective recombination events, and so two loci belonging to the same 'early' block may not have exactly the same (but highly correlated) coalescence times. If we now focus on the average length of a block due to early coalescence and such that all loci should have exactly the same coalescence time (the blue curve), then somewhat surprisingly, we see that the geometric distribution of Theorem 2.10 fits the empirical distribution of these 'equal' block lengths better. We have no explanation for this fact, but it should be noted that the discrepancy between early coalescence and equal coalescence vanishes as the sampling distance grows.

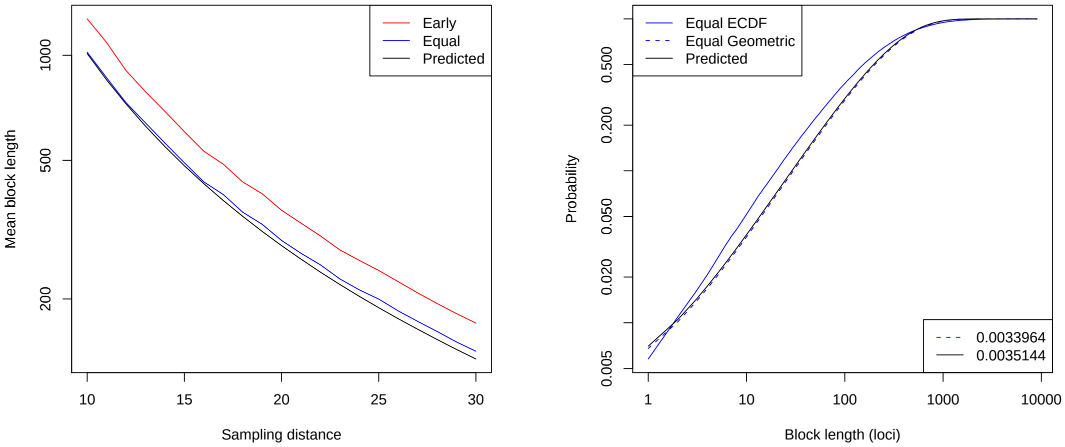
Finally, let us comment on the different terms appearing in the expression of $\gamma(r)$. First, $\pi R^2 u\rho$ is the total rate at which two neighbouring loci recombine. When dealing with data, this compound term should be replaced by an estimate of the recombination rate. Second, $\alpha(r)$ is the probability that two nearby lineages (at initial distance $\mathcal{O}(R)$) separate at distance r before coalescing. If r is large compared to R, this probability is essentially the same as the probability that two lineages starting at distance κ do not coalesce before the time $\mathcal{O}(r^2/\sigma^2)$ that they need to travel a distance r. Hence, using the Wright-Malécot formula with $2\mu = \sigma^2/r^2$, we arrive at

$$\alpha(r) \approx 1 - \frac{\ln(r/(\kappa\sqrt{2}))}{\mathcal{N} + \ln(r/(\kappa\sqrt{2}))},$$

which depends only on $\mathcal{N}$ and κ. Thus, for sufficiently large r the quantity $\gamma(r)$ is really a function of σ^2, $\mathcal{N}$ and κ alone.

Perspectives: a real inference method based on lengths of shared sequences

The ideas developped in this section, and especially the second part of it, are essentially proofs of concept. In [94], the authors set up a practical inference method based on the results of the second part of [BEKV13a] and assuming a large neighbourhood size. Because two recombinant lineages essentially never coalesce again quickly in this regime, the lineage 'freed' by recombination will move far away before coming back and coalescing with the ancestry of the second individual. This implies that with high probability it will experience a mutation which will render this event observable, and furthermore that we can make the approximation that distinct blocks of shared sequence are independent. Their empirical distribution is thus a good proxy for the distribution of a given block length of conserved sequence. In contrast, when neighbourhood size is small, only a very few of the recombination events are effective and two blocks of shared sequences cannot be considered as independent. Indeed, the occurrence of a very long block is due to very recent coalescence, and so many portions of the genome are still carried by the same ancestor at that time. Hence, the formation of such a long block is rare, but when it occurs several long blocks are produced at the same time. Consequently,



(a) (b)

Figure 2.14 – Block lengths due to early coalescence; (a) plots mean block length against sampling distance r and (b) shows the distribution of block lengths for $r = 20$. Simulations trace the ancestry of two individuals sampled at distance r until time T_ζ in the past, where T_ζ is an exponentially distributed value with rate $2\sigma^2/r^2$ chosen independently for each replicate. The length of blocks of loci are then calculated in two different ways: we have *early* and *equal* blocks. An early block is defined as a set of contiguous loci that have coalesced by time T_ζ. An equal block is a set of contiguous loci that have coalesced by T_ζ and have equal coalescence times. In (a) we have $\sim 10^6$ early blocks and $\sim 1.2 \times 10^6$ equal blocks from 7911 independent simulations (many simulations have no early coalescences); there is an excellent correspondence between the predicted block length $1/\gamma(r)$ and the length of equal blocks. Panel (b) shows the CDF of the length of equal blocks for $r = 20$, and compares the empirical CDF of block lengths from simulations with a geometric distribution with parameter $\gamma(20)$. Also shown is the geometric distribution with parameter estimated from simulation data; this agrees very closely with the predicted value.

the empirical distribution of lengths of blocks which are identical by descent between pairs of individuals in a sample cannot *a priori* be used to estimate $\gamma(r)$. The first step in overcoming this issue is to investigate about the correlations between adjacent blocks of shared sequence thoroughly and build up on this knowledge to understand the full sequence of conserved blocks along the genome.

Of course the approaches expounded here can also be complemented by the MCMC methods set up recently in the context of the SLFV, see Chapter 2.5 in [70] and [55] (the first one being based on the look-down construction reviewed in Section 2.4.4). These techniques are quite interesting and have been applied for example to the estimation of dispersal distance and population density of the influenza virus in [55]. However, they remain computationally intensive (despite the smart simulation algorithms for the SLFV devised by Kelleher and his co-authors [62, 63]), and it is still desirable to think of which summary statistics are important to reconstruct and to try to turn fine analytical results into reliable inference methods.

2.6 The effects of natural selection

Up to now we have only considered neutral evolution, in which no alleles gave a reproductive advantage to the individuals carrying them. This was a mandatory first step, at least for comparison purposes. Indeed the reduction in effective population sizes observed in many species, compared to what would be expected under the absolute null model of neutral panmictic evolution, could be caused by a spatial structure with or without recurrent catastrophes, natural selection, fluctuations in population sizes, etc., or some combination of these factors. It is thus important to understand the signature left by each evolutionary scenario to be able to detect the main forces at work.

In this section, we study the effect of natural selection in a spatially structured population. The first work reviewed here is again based on the spatial Λ-Fleming-Viot process with and without large-scale extinction-recolonisation events, generalised to incorporate selection. The second tries to understand the effect of a selective sweep at an associated neutral locus, in a population with a discrete or continuous geographical structure and when reproduction events are local. Natural selection and population structure are also at the core of the third work presented below, but in a more hidden way. There we find conditions on the pedigree of a population (which will typically be influenced by selection on a given set of traits in the main application of these results) for the *infinitesimal model* to be an accurate description of the phenotypic evolution of the population over a few tens or hundreds of generations. See Section 2.6.3 for a definition of all these terms.

2.6.1 Large-scale effects of a weak selection pressure

The results expounded in this section correspond to the work currently in progress [EVY17]. They are in the same spirit as those presented in Section 2.4.3, with the notable difference that the impact u of an event will now tend to zero as the scaling parameter n tends to infinity. This will of course affect the shape of the genealogies. We shall confer a selective advantage to the individuals carrying the allele 1, this effect becoming weaker and weaker as $n \to \infty$. The main questions we want to address are: For which range of parameters and over which time- and space-scales can we observe an evolution of the local allele frequencies which is well-approximated by the solution to the Fisher-KPP equation (possibly with noise)? How are these different ranges modified by the occurrence of rare but large extinction-recolonisation events?

Fisher introduced what is now referred to as the Fisher-KPP equation [48, 69] to model the wave of advance of a favourable allele in a continuously structured population. If we assume that individuals can be of only two allelic types and if $w(t,x)$ stands for the fraction of individuals at site $x \in \mathbb{R}^d$ at time t that carry the *unfavoured* allele, then this equation can be written as

$$\frac{\partial w}{\partial t} = \frac{\sigma^2}{2}\,\Delta w - sw(1-w), \tag{2.27}$$

where σ^2 is the speed at which alleles diffuse locally in space, in a symmetric way, and s quantifies the strength of the selective (dis)advantage. (In the usual form of this equation, $w(t,x)$ stands for the frequency of the *favourable* allele, i.e. $1 - w(t,x)$ here, and so the last term is $+sw(1-w)$.) Note in particular the product $w(t,x)(1-w(t,x))$ appearing in the term modelling the effect of selection, which encodes the fact that the favourable allele wins over the other one by being more often that carried by an individual reproducing and replacing someone at random at its current position. We thus expect this effect to be proportional to the local rate of encounter between individuals of the two types, although many other sorts of selective advantages could be considered instead.

Equation (2.27) has solutions in any dimension, which have been extensively studied. In particular, it admits travelling wave solutions with constant speed. In one dimension we can add a noise term modelling the local effect of random drift. The Fisher-KPP equation with noise reads

$$\frac{\partial w}{\partial t} = \frac{\sigma^2}{2}\,\Delta w - sw(1-w) + \sqrt{\frac{1}{N_e}\,w(1-w)}\dot{\mathcal{W}}, \tag{2.28}$$

where by analogy with the Wright-Fisher diffusion (2.3), N_e can be seen as an effective local population size and $\mathcal{W}$ denotes space-time white noise. Equation (2.28) has no solutions in more than one dimensions. In one dimension, it does have a solution and the effects of a small noise (or large N_e) have received a lot of attention (see [84] and references therein). Because of the success of this equation in the modelling of the spatial expansion of a favourable allele, we want to see whether we can recover it from our model, and if so under which conditions on the parameters.

As a last remark before describing the framework adopted in this work, let us emphasise that the regime of parameters considered below will indeed look very particular, especially when we allow large-scale extinction-recolonisation events to happen. However, as we already discussed, the identification of the orders of magnitude of the parameters which lead to the desired asymptotic behaviour is an essential step to understand how the different factors (spatial diffusion, selection, genetic drift) should compare to give rise to such a limit. Other studies have considered different regimes and obtained a variety of limiting evolutions depending on the assumptions made on the strength of selection and on the local population density (modelled through the fraction of individuals replaced during an event). See [39, 40, 41].

Suppose that our population is distributed over $\mathbb{R}^d$ and that individuals can be of two allelic types, 0 and 1. Allele 1 is the one favoured by natural selection. For every $t \geq 0$ and $x \in \mathbb{R}^d$, let $w_t(x)$ denote the fraction of individuals carrying the *unfavoured* allele 0 at site x at time t (later this will depend on our scaling parameter n, but we shall not report it to ease the notation). Let us fix $u \in (0,1]$ and $s > 0$, and for every $n \geq 1$, let us define

$$u_n = \frac{u}{n^\gamma}, \qquad s_n = \frac{s}{n^\delta}, \qquad \text{and} \quad w_t^n(x) = w_{nt}(n^\beta x) \tag{2.29}$$

for some $\beta, \gamma, \delta > 0$ whose values will depend on the cases considered. In fact, because $w_t(x)$ is only defined up to a Lebesgue nullset of $\mathbb{R}^d$, we shall instead work with its local average around each point x:

$$\bar{w}_t^n(x) = \frac{1}{V_1 n^{-\beta d}} \int_{B(x,n^{-\beta})} w_t^n(y)\mathrm{d}y. \tag{2.30}$$

In words, we want to look at very high local population densities and very small selective effects, averaged over large spatial and temporal scales. As in Section 2.4.3, we consider two cases for the sequence of reproduction events.

- **Local evolution:** We fix some $R > 0$ and suppose that all events have radius R. That is, $\mu(\mathrm{d}r) = \delta_R(\mathrm{d}r)$ in our previous notation.
- **Rare large events:** We fix $\alpha \in (1,2)$ and choose as an intensity measure for the radii

$$\mu(\mathrm{d}r) = \frac{\mathbf{1}_{\{r>1\}}}{r^{d+\alpha+1}}\,\mathrm{d}r.$$

Now for every fixed n and each of these scenarii, we define two independent Poisson point processes corresponding to *neutral* and *selective* reproduction events.

- **Neutral events:** Let Π_n^N be a Poisson point process on $\mathbb{R}\times\mathbb{R}^d\times(0,\infty)$ with intensity $\mathrm{d}t\otimes\mathrm{d}z\otimes\mu(\mathrm{d}r)$. During such an event (t,z,r), we choose one allele $\kappa\in\{0,1\}$ according to the allele distribution in $B(z,r)$ just before the event. Then for every $y\in B(z,r)$, we set

$$w_t(y) = (1-u_n)w_{t-}(y) + u_n\delta_{\{\kappa=0\}}.$$

- **Selective events:** Let Π_n^S be a Poisson point process on $\mathbb{R}\times\mathbb{R}^d\times(0,\infty)$ with intensity $s_n\mathrm{d}t\otimes\mathrm{d}z\otimes\mu(\mathrm{d}r)$, independent of Π_n^N. During such an event (t,z,r), we choose two alleles κ,κ' independently and according to the allele distribution in $B(z,r)$, and for every y in this ball we set

$$w_t(y) = (1-u_n)w_{t-}(y) + u_n\delta_{\{\kappa=\kappa'=0\}}.$$

Again this can be generalised in many ways, but we concentrate on the cases which minimise the notation. Observe that the selective events favour allele 1 indeed, since the offspring are of type 0 if and only if both 'potential parents' are of type 0. Since selective events occur s_n times as fast as neutral events, the parameter s_n tunes the relative frequency at which type 1 individuals have a reproductive advantage over the others.

To get a handle on the evolution of the process of local allele frequencies, it is useful to think of the ancestry of a sample of individuals. Let us start with the simpler case of local reproduction, and with a single individual taken from position 0, say, at some time $t > 0$. First, the neutral events make the lineage ancestral to this individual (again traced backwards in time) jump at rate $\mathcal{O}(u_n)$ to a new position at distance $\mathcal{O}(1)$. If time is run n times as fast and space is scaled down by n^β, we thus expect the rescaled lineage to converge to Brownian motion whenever

$$n\times n^{-\gamma}\propto n^{2\beta}, \quad \text{i.e.} \quad 1-\gamma = 2\beta. \tag{2.31}$$

Now let us consider what happens during a selective event. By construction there are two potential parents, and we need to know the alleles carried by both of them to determine the allelic type of the offspring. We thus need to track the ancestry of both potential parents, creating a branching event in the ancestral process of our sample. Now the two new *rescaled* lineages emanating from this branching event are born at a separation of order $\mathcal{O}(n^{-\beta})$. If we are to 'see' the event, they must move apart to a separation of order one before (perhaps) coalescing. The number of excursions they must make away from the region in which they can both be affected by an event (and thus coalesce) before we can expect to see such a 'long' excursion is of order $\mathcal{O}(1)$ in $d \geq 3$, of order $\mathcal{O}(\log n)$ in $d = 2$ and of order $\mathcal{O}(n^{\beta})$ in $d = 1$. On the other hand, when they are sufficiently close together that they can be hit by the same event, given that one of them jumps, there is a probability of order $u_n = u/n^{\gamma}$ that the other one is affected by the same event and so they coalesce. So the number of times they come close to one another before they coalesce is of order $\mathcal{O}(n^{\gamma})$. Thus, in the limit as $n \to \infty$, for each branching event in the dual, in dimensions at least 2, the probability that there is a long excursion before coalescence (and so we 'see' the event) tends to one. Moreover, the same argument tells us that we shall never see the coalescence of any other lineages in our system. In one dimension, we can expect to see both branching and coalescence provided that the number of excursions we expect to wait before seeing a coalescence and the number we expect to wait before the lineages escape to a distance of order one are comparable, that is

$$\frac{1}{u_n} \propto n^{\beta}, \quad \text{or} \quad \gamma = \beta. \tag{2.32}$$

Combining with Condition (2.31), we find $\beta = \gamma = 1/3$. Finally, selection events occur at a rate proportional to nu_ns_n in the rescaled process, and so we choose

$$1 - \gamma - \delta = 0, \tag{2.33}$$

i.e., $\delta = 2/3$ to make this order one. Note that only Equations (2.31) and (2.33) need to be satisfied in dimension more than 1 (since there are no parameter values for which we would see coalescence occur anyway) to obtain the results stated below.

Let us turn to the case of rare large events. As before, we first consider the motion of a single rescaled lineage and we see that if we choose $nu_n = n^{\alpha\beta}$, then in the limit as $n \to \infty$ its motion will converge to a symmetric stable process. Now let us consider selection. Since $u_n \to 0$ as $n \to \infty$, although it is now the case that two lineages can always be affected by the same event, 'most of the time' they will not and the motions are almost independent. Moreover, since 'small' events are so much more frequent than 'big' events, selection events are almost always 'small' and, moreover, lineages only have a realistic chance of coalescing when they are close together. We now use the same argument as before. The number of excursions away from each other before they are 'visible' under our rescaling is of order $\mathcal{O}(1)$ in $d \geq 2$ and of order $\mathcal{O}(n^{(\alpha-1)\beta})$ in $d = 1$. Equating this to the number of visits together before we expect to see a coalescence event yields $\gamma = (\alpha - 1)\beta$. In order to see any selection events at all, we need nu_ns_n to be of order one, so $1 - \gamma - \delta = 0$. We now have three equations in three unknowns (in one dimension) and solving yields

$$\beta = \frac{1}{2\alpha - 1}, \quad \gamma = \frac{\alpha - 1}{2\alpha - 1}, \quad \text{and } \delta = \frac{\alpha}{2\alpha - 1}. \tag{2.34}$$

As in the neutral case, we can show that a duality relation holds (for any n and in much greater generality than the framework considered here) between the spatial Λ-Fleming-Viot process with selection as defined above, and the system $(\Xi_t)_{t\geq 0}$ of branching and coalescing jump processes tracking backwards in time the positions of the potential ancestors of a sample of individuals. This is where we use the fact that w_t describes the frequencies of the less favoured allele. Indeed, by construction, for an individual to be of allelic type 0, all its potential ancestors must carry the allele 0. This yields the same form of duality relation as in (2.21):

Proposition 2.5. (Prop. 2.2 in [EVY17]). *For every $w_0 \in \mathcal{M}_\lambda$, $t \geq 0$, $k \in \mathbb{N}$ and $F \in C_c((\mathbb{R}^d)^k)$, we have*

$$\mathbb{E}_{w_0}\bigg[\int_{(\mathbb{R}^d)^k} F(x_1,\dots,x_k)\bigg(\prod_{i=1}^k w_t(x_i)\bigg)\mathrm{d}x_1\cdots\mathrm{d}x_k\bigg] \\ = \int_{(\mathbb{R}^d)^k} F(x_1,\dots,x_k)\mathbb{E}_{\mathbf{x}}\bigg[\prod_{j=1}^{N_t} w_0(\xi_t^j)\bigg]\mathrm{d}x_1\cdots\mathrm{d}x_k.$$

The analysis carried out in the previous paragraph suggests that when reproduction is purely local, the rescaled potential ancestry of a sample of individuals defined by

$$\Xi_t^n := \bigg\{\frac{\xi_{nt}^1}{n^\beta},\dots,\frac{\xi_{nt}^{N_{nt}}}{n^\beta}\bigg\}, \quad t \geq 0, \tag{2.35}$$

should converge to a system of independent Brownian motions that branch into two at rate usV_R (the two resulting lineages being born at the location of their 'parent') and coalesce pairwise in one dimension at a rate which is proportional to the local time at zero of their distance. When large extinction-recolonisation events happen, because large selective events can be neglected both for branching and coalescence (since the probability that they are of size at least n^ε adds another $n^{-\alpha\varepsilon}$ in their rate of occurrence), we expect the rescaled potential ancestry to converge to a system of independent symmetric α-stable Lévy processes that branch into two at a rate proportional to us (the two lineages starting at the location of their 'parent') and again coalesce pairwise in one dimension at a rate which is proportional to the local time at zero of their distance. In $d \geq 2$, this result can indeed be proved by working directly with the rescaled system of coalescing and branching jump processes corresponding to each n. However, this approach does not allow us to characterise the coalescence mechanism in one dimension. Thus, in contrast with the usual method of proof for these complex measure-valued evolutions, this time we first show the convergence of the process of rescaled local allele frequencies, and then use this convergence to deduce that of the ancestral processes. This gives us the following results in the case of local reproduction. Recall the definition of $\bar{w}^n$ given in (2.30).

Theorem 2.11. (Th. 1.3 in [EVY17] - Local evolution). *Let $\beta = \gamma = 1/3$ and $\delta = 2/3$. Suppose that $\bar{w}_0^n$ converges in $\mathcal{M}_\lambda$ to some w^0. Then as $n \to \infty$, the process $(\bar{w}_t^n)_{t\geq 0}$ converges weakly in $D_{\mathcal{M}_\lambda}[0,\infty)$ towards a process $(w_t^\infty)_{t\geq 0}$ with initial value $w_0^\infty = w^0$. Furthermore,*

(a) When $d = 1$, $(w_t^\infty)_{t\geq 0}$ is the unique process for which, for every $f \in C_c^\infty(\mathbb{R})$,

$$\langle w_t^\infty, f\rangle - \langle w_0^\infty, f\rangle - \int_0^t \bigg\{\frac{u\Gamma_R}{2}\langle w_s^\infty, \Delta f\rangle - 2Rus\,\langle w_s^\infty(1-w_s^\infty), f\rangle\bigg\}\mathrm{d}s$$

is a zero-mean martingale with quadratic variation

$$4R^2u^2\int_0^t \langle w_s^\infty(1-w_s^\infty), f^2\rangle\,\mathrm{d}s,$$

where

$$\Gamma_R = \frac{1}{dV_R}\int_{B(0,R)}\int_{B(x,R)} \|z\|^2\mathrm{d}z\mathrm{d}x.$$

(b) When $d \geq 2$, $(w_t^\infty)_{t\geq 0}$ is the unique deterministic process for which, for every $f \in C_c^\infty(\mathbb{R}^d)$,

$$\langle w_t^\infty, f\rangle = \langle w_0^\infty, f\rangle + \int_0^t \bigg\{\frac{u\Gamma_R}{2}\langle w_s^\infty, \Delta f\rangle - usV_R\,\langle w_s^\infty(1-w_s^\infty), f\rangle\bigg\}\mathrm{d}s,$$

where $\Gamma_R > 0$ is defined as above.

In other words, in one space dimension, the limiting process $(w_t^\infty)_{t\geq 0}$ is a weak solution to the stochastic partial differential equation

$$\frac{\partial w}{\partial t} = \frac{u\Gamma_R}{2}\,\Delta w - 2Rusw(1-w) + 2Ru\sqrt{w(1-w)}\,\dot{\mathcal{W}},$$

with $w_0 = w^0$, and $\mathcal{W}$ a space-time white noise. In dimension $d \geq 2$, on the other hand, the noise term disappears in the limit and $(w_t^\infty)_{t\geq 0}$ is a weak solution to the deterministic Fisher-KPP equation

$$\frac{\partial w}{\partial t} = \frac{u\Gamma_R}{2}\,\Delta w - usV_R\,w(1-w), \qquad w_0 = w^0.$$

The proof of this result is not yet complete. It goes through the standard steps of (i) showing the tightness of $(\bar{w}^n)$, which is easily done by controlling the frequency and effects of the reproduction events affecting a given compact region of space and using the Aldous-Rebolledo criterion, and (ii) showing that the generator of $\bar{w}^n$ converges to that of the limit w^∞ when applied to the appropriate test functions. The convergence of the terms which are linear in $\bar{w}_t^n$ follows from the martingale problem formulation. What remains to show is that the part of the martingale problem which is 'quadratic' in $\bar{w}_t^n$ converges to a process which can be expressed in terms of $(w_t^\infty)^2$. We expect this to hold true thanks to the continuity estimates which can be obtained from the duality with Ξ^n.

Tightness of the sequence of rescaled ancestral processes can also be proved by controlling the frequency and effects of the reproduction events. Assuming that Theorem 2.11 is proved, the duality relation described in Proposition 2.5 enables us to show the following convergence result.

Theorem 2.12. (Th. 2.5 in [EVY17] - Local evolution). *For every $n \in \mathbb{N}$, let $(\xi_t^1, \dots, \xi_t^{N_t})_{t\geq 0}$ be the system of branching and coalescing jump processes which is dual to the unscaled process $(w_t)_{t\geq 0}$ with parameters $\mu = \delta_R$, fixed impact u_n and selection strength s_n. Define the rescaled process $(\Xi_t^n)_{t\geq 0}$ as in* (2.35), *and suppose that the initial condition Ξ_0^n converges weakly towards some Ξ_0 as $n \to \infty$. Then, if $d \geq 2$, as $n \to \infty$, $(\Xi_t^n)_{t\geq 0}$ converges in distribution (as a càdlàg process) to a branching Brownian motion $(\Xi_t^\infty)_{t\geq 0}$, in which individuals follow independent Brownian motions with variance parameter $u\Gamma_R$, which branch at rate usV_R into two new particles, started at the location of the parent. When $d = 1$, the corresponding object is a branching and coalescing system with the same diffusion constant and branching rate, but in addition each pair of particles, independently, also coalesces at rate $4R^2u^2$ times the local time at zero of their distance.*

In fact the identification of the limit is not completely obvious from the duality relation with the limiting forwards process. It is based on the fact that if we guess the limit of Ξ^n (from our heuristics) and if we use a generalisation of the construction of Chapter 7 of [74] to obtain the corresponding forwards in time evolution, we find that the unique $\mathcal{M}_\lambda$-valued process dual to the limiting ancestry in Theorem 2.12 is indeed the limit of $\bar{w}^n$ obtained in Theorem 2.11.

Using the same chain of arguments, together with some technical continuity estimates on $\bar{w}^n$, we (should) obtain the following results in the case of rare large events.

Theorem 2.13. (Th. 1.5 in [EVY17] - Rare large events). *Let $\beta = 1/(2\alpha-1)$, $\gamma = (\alpha-1)/(2\alpha-1)$ and $\delta = \alpha/(2\alpha-1)$. Suppose that $\bar{w}_0^n$ converges weakly to some $w^0 \in$*

$\mathcal{M}_\lambda$. Then, as $n \to \infty$, the process $(\bar{w}^n_t)_{t\geq 0}$ converges weakly in $D_{\mathcal{M}_\lambda}[0,\infty)$ towards a process $(w^\infty_t)_{t\geq 0}$ with initial value w^0. Furthermore, there exists a symmetric α-stable Lévy process X^α such that if $\mathcal{D}^\alpha$ denotes the generator of X^α, then

(i) When $d = 1$, $(w^\infty_t)_{t\geq 0}$ is the unique process for which, for every $f \in C^\infty_c(\mathbb{R})$,

$$\langle w^\infty_t, f\rangle - \langle w^\infty_0, f\rangle - \int_0^t \left\{ \langle w^\infty_s, \mathcal{D}^\alpha f\rangle - \frac{2us}{\alpha}\, \langle w^\infty_s(1-w^\infty_s), f\rangle \right\} \mathrm{d}s$$

is a zero-mean martingale with quadratic variation

$$\frac{4u^2}{\alpha - 1} \int_0^t \langle w^\infty_s(1-w^\infty_s), f^2\rangle\, \mathrm{d}s.$$

(ii) When $d \geq 2$, $(w^\infty_t)_{t\geq 0}$ is the deterministic process for which, for every $f \in C^\infty_c(\mathbb{R}^d)$,

$$\langle w^\infty_t, f\rangle = \langle w^\infty_0, f\rangle + \int_0^t \left\{ \langle w^\infty_s, \mathcal{D}^\alpha f\rangle - \frac{usV_1}{\alpha}\, \langle w^\infty_s(1-w^\infty_s), f\rangle \right\} \mathrm{d}s.$$

Theorem 2.14. (Th. 2.6 in [EVY17] - Rare large events). *For every $n \in \mathbb{N}$, let $(\xi^1_t, \ldots, \xi^{N_t}_t)_{t\geq 0}$ be the system of branching and coalescing jump processes which is dual to the unscaled process $(w_t)_{t\geq 0}$ corresponding to the case with rare large events. Define the rescaled process $(\Xi^n_t)_{t\geq 0}$ as in (2.35). Then, if the initial condition Ξ^n_0 converges weakly towards some Ξ_0 as $n \to \infty$, $(\Xi^n_t)_{t\geq 0}$ converges in distribution (as a càdlàg process) to a system $(\Xi^\infty_t)_{t\geq 0}$ of independent symmetric α-stable processes, which branch at rate usV_1/α into two particles starting at the location of their parent. The motion of a single particle has the same law as the process X^α defined in Theorem 2.13. In addition, when $d = 1$ each pair of particles, independently, coalesces at rate $4u^2/(\alpha - 1)$ times the local time at zero of their distance.*

We see that in this regime of parameters, the effect of the occurrence of rare large events is to create correlations between local allele frequencies over much larger scales ($n^{1/(2\alpha-1)} \gg n^{1/3}$ when $\alpha \in (1,2)$), as in the neutral case. These large events do not contribute to selection and coalescence, which remain local in the limit, but the spatial diffusion of alleles is now described by a fractional Laplacian. In one dimension it is known that travelling wave solutions to this equation exist, but the front position now moves exponentially in time [20].

2.6.2 Selective sweeps in spatially extended populations

As we have seen in the previous section, the Fisher-KPP equation can describe the wave of advance of a favourable allele over large spatial and temporal scales, when the selective advantage of this allele is weak and local population density (proportional to the inverse of the impact parameter in the SLFV) is very high. However, in a real population distributed over some large region of space, it is unlikely that local population densities reach such extremely large values. In fact, the noise term in the stochastic version of the Fisher-KPP equation was added in part to model the effect of actually finite (still reasonably large) local populations. It is thus natural to enquire about the genetic consequences of the random fluctuations at the wave front, assuming the Fisher-KPP equation with noise is a good approximation for the way a favourable mutation *sweeps to fixation* in a spatially extended population. This section describes the results of the publication [BEKV13b].

When the favourable allele goes to fixation in the population, it is likely to boost the frequency of the allele to which it was originally associated at a given linked neutral locus. This is called *genetic hitchhiking*, as the 'lucky' neutral allele gets a lift from the successful mutation at the locus under selection. During the course of the sweep, recombination may break the link between the alleles at the two loci, and so in general some diversity remains at the neutral locus. Already in a panmictic population, the effects of a selective sweep on the allele frequencies or on the genealogies at a linked locus are delicate to describe [5, 35, 43, 80, 101, 104]. The longer the sweep takes, the more time there is for recombination to occur and dissociate the neutral alleles from the selected one. The length of the sweep depends on the strength of selection and on the size of the population to invade.

When the population has a spatial structure, it is not at all obvious whether the net effect of a sweep on the genetic diversity at a linked neutral locus will be stronger or weaker than in a panmictic population. On the one hand, fixation will take much longer, extending the timescale over which recombination can happen; on the other hand, local founder effects at the wave front may greatly increase genetic drift and lead to faster local fixation. To quantify the strength of hitchhiking in such a population, we focus on the net rate of coalescence of neutral genes due to sweeps at linked loci passing through the population. We mainly consider the case of a one-dimensional spatial structure, assuming that the wave of advance of the favourable allele is well described by the Fisher-KPP equation with noise. These results can be extended to two dimensions by assuming that the invasion front is linear, but simulations shows that this approximation is poor due to the complex transverse fluctuations in the wave front.

Recall the model in one dimension from (2.28):

$$\frac{\partial w}{\partial t} = \frac{\sigma^2}{2}\Delta w + sw(1-w) + \sqrt{\frac{1}{\rho}\,w(1-w)}\dot{\mathcal{W}}, \tag{2.36}$$

where to match the notation of [BEKV13b] we write ρ for the population density to which the variance of the noise is inversely proportional, and to confuse the reader we have changed our point of view and now consider the equation satisfied by the frequencies of the *favourable* allele. In the absence of noise, it is well-known that this equation has a whole family of travelling wave solutions, and that if we start from any nonnegative initial condition that looks like $\mathbf{1}_{\{x<0\}}$, then the solution to the deterministic Fisher-KPP equation converges to the nonnegative travelling wave of the smallest possible velocity $c_\infty = \sigma\sqrt{2s}$. If we write the corresponding travelling wave solution as $w(t,x) = w_{c_\infty}(x - c_\infty t)$, then as Fisher showed, at the front (i.e., where $w_{c_\infty}(z)$ is small) it can be approximated by $\exp(-c_\infty z/\sigma^2)$.

Three consequences of the presence of small amounts of genetic drift have been explored, mostly by analysing related models meant to mimick noisy Fisher waves. First, the rate of advance of the wave is slowed down by a factor proportional to $1/(\ln(\rho\sigma\sqrt{s/2}))^2$, as shown in [18, 85]. That is, if we set

$$\eta := \rho\sigma\sqrt{\frac{s}{2}} \tag{2.37}$$

and write c_η for the speed of the asymptotic wave solution to the stochastic Fisher-KPP equation with noise (2.36), we have

$$c_\eta \approx c_\infty\left(1 - \frac{A}{(\ln\eta)^2}\right) \tag{2.38}$$

for some $A > 0$. Observe in passing that if $\tilde{w}$ is solution to the 'canonical' form of the equation

$$\frac{\partial \tilde{w}}{\partial t} = \Delta\tilde{w} + \tilde{w}(1-\tilde{w}) + \sqrt{\frac{1}{\eta}\,\tilde{w}(1-\tilde{w})}\dot{\mathcal{W}},$$

then for every $t \geq 0$ and $x \in \mathbb{R}$, we have

$$w(t,x) = \tilde{w}\bigg(st, \frac{\sqrt{2s}}{\sigma}\, x\bigg).$$

This allows us to translate all the results obtained for the canonical equation into properties of the solutions to (2.36). Second, the shape of the wave front of the stationary travelling wave solution is well approximated by that of the deterministic wave w_{c_η} corresponding to c_η, truncated at the smallest position where it hits 0 (recall that since $c_\eta < c_\infty$, the corresponding travelling wave does not remain nonnegative), c.f. [85]. In fact, the analysis carried out in [11] on a related model suggests that at least for large η, this approximation is good most of the time, but every $\mathcal{O}((\ln \eta)^3)$ units of time there are appreciable fluctuations. These fluctuations are essentially due to the fact that an individual manages to get significantly ahead of the bulk of the wave and reproduce without competition until the wave catches up again. As a result of the occurrence of these transiently prolific individuals, the third effect of the small noise component is that the genealogy at the locus under selection of a sample from the wave front will be dominated by founder effects resulting from rare but large fluctuations in the wave front. In the appropriate timescale $((\ln \eta)^3 t,\, t \geq 0)$ it is approximated by the Bolthausen-Sznitman coalescent [17] (with multiple mergers) instead of Kingman's coalescent [11, 18]. In particular, one of the main differences between the travelling wave solutions to the equation with and without noise is that if we start from an initial condition of the form $\mathbf{1}_{\{x<0\}}$ (or more generally from an initial condition in which the set of locations such that $w(0,x) \notin \{0,1\}$ is bounded), then in the stochastic case the region $\{x : w(t,x) \notin \{0,1\}\}$ remains bounded for all times [86].

Based on these three points, it is not difficult to see how a single lineage at a linked neutral locus, sampled after the sweep has passed through a (stationary) travelling wave, should behave. Initially, everyone around it carries the beneficial allele at the selected locus, and so the lineage moves around according to Brownian motion. But backwards in time, the wave moves back towards its origin at a linear speed, and thus at some point the lineage becomes caught by the wave and starts experiencing a drift. Since most of the time the (deterministic) truncated w_{c_η} is a good approximation for the actual stochastic wave, the drift term is given by $\sigma^2 w'_{c_\eta}/w_{c_\eta}$. As a consequence, if we write X_t for the position of the lineage relative to the wave front, we obtain that $(X_t)_{t\geq 0}$ is ('most of the time') solution to

$$\mathrm{d}X_t = \bigg(\frac{\sigma^2 w'_{c_\eta}(X_t)}{w_{c_\eta}(X_t)} - c_\eta\bigg)\,\mathrm{d}t + \sigma \mathrm{d}B_t,$$

where $(B_t)_{t\geq 0}$ denotes standard Brownian motion. Using the theory of one-dimensional diffusion we find that if X has a stationary distribution, then its density $f(x)$ is proportional to

$$w_{c_\eta}^2(x) \exp\bigg(-\frac{2c_\eta x}{\sigma^2}\bigg) \tag{2.39}$$

(assuming this function is integrable). In this case, the lineage becomes trapped within a narrow front whose width is of the order of $\sigma^2/c_\infty \propto \sqrt{\sigma^2/s}$, until it manages to recombine with an individual of the unfavoured allele and thereby escape the wave. The effects of the rare fluctuations in the front are not at all clear.

Let us now consider a sample taken from the population, again after the wave passed. By the same arguments as above, we can anticipate that their ancestral lineages will move around in space according to independent Brownian motions until they are (one by one) caught by the backward wave. From then on, they may recombine away from the wave. In the absence of fluctuations in the wave front, each pair of lineages may also coalesce at a rate proportional to $1/(\rho w_{c_\eta}(x))$ upon meeting at a relative position to the front equal to x. However, the work of [11, 18] suggests instead that for large η, the rare large fluctuations in the front create multiple merger events that dominate the genealogy of the sample. The resulting genealogy on the timescale $((\ln \eta)^3 t,\ t \geq 0)$ mentioned earlier should then be described by the Bolthausen-Sznitman coalescent, except for those lineages who manage to escape by recombining.

Our aim is to see whether these heuristics hold true in a population with the moderate local densities that we would expect from natural populations. These questions are investigated through simulations of a discrete stepping stone model with Wright-Fisher resampling within each deme of finite size $\rho \in \mathbb{N}$ and nearest neighbour migration with migration probability $\sigma^2 \in (0,1)$, and we restrict our attention to large but biologically realistic parameter values: $\rho = 10, \ldots, 10^6$, $\sigma^2 = 0.25$ and $s = 0.01, 0.05, 0.1$, corresponding to values of η ranging from 10^{-1} to 10^5. We expect the same conclusions to hold in more general models of discrete or continuous population structure sharing the same characteristics.

The first conclusions we can draw from these simulations is that the speed and shape of the resulting stationary wave indeed behave as expected under the 'large η' assumption, except at the very tip where the frequency of the favourable allele tends to be substantially larger than predicted. Because these results are obtained by calculating average allele frequencies (at each relative separation from the front) over 10^5 generations, this could be due to occasional large fluctuations taking the front well ahead of the centre of mass of the wave. See Figures 2-4 in [BEKV13b]. There we measure the position of the wave front by its centre of mass and we use twice the total number of heterozygotes, that is

$$W(t) = 4 \int w(t,x)(1 - w(t,x))\mathrm{d}x,$$

as a proxy for its width. This way, the expected rate of advance $c(t) = \mathrm{d}/\mathrm{d}t \int w(t,x)\mathrm{d}x$ (of which we can make sense because the area over which $w(t,\cdot)$ is not constant has finite size) is equal to $sW(t)/4$.

In Figure 2.15, the positions of ancestral lineages relative to the centre of mass of the wave are plotted. We superimpose on this the approximation for the stationary distribution predicted by (2.39), with w_{c_η} determined by the truncated deterministic Fisher wave and c_η obtained by fitting the constant A in (2.38) empirically. The fit is quite good, and it seems that fluctuations in the wave front are too rare to distort this distribution. Next, in Figure 2.16 we show the locations of coalescence events within the front as predicted by the pairwise merger rate $f^2/(\rho w_{c_\eta})$ (where f was defined just above (2.39) as the density of the stationary distribution of a single lineage in the front). We see that the distribution of coalescence events is close to the prediction based on the actual locations of ancestral chromosomes, although slightly more spread towards the front. In particular, for η not too large the expected rare but large fluctuations in the front do not seem to dominate the genealogy. In fact for moderately large values of η, the lineages do not even seem to be trapped within the front and we expect a fraction of the coalescence to occur within the bulk of the wave.

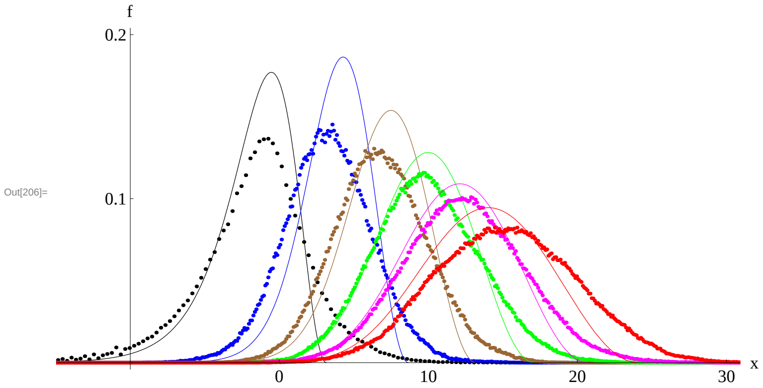


Figure 2.15 – The dots show the distribution of locations of ancestral lineages, relative to the centre of mass, for $\rho = 10, 100, \ldots, 10^6$ (left to right); $s = 0.05$, $m = 0.25$. The curves show the predicted locations, $f \propto w_{c_\eta}^2 e^{2c_\eta x/\sigma^2}$, where the allele frequency is calculated using the deterministic Fisher-KPP equation. For each ρ, four replicate lineages were propagated back through 10^5 generations, using a single realisation of the forwards process. Ancestors tend to be ahead of the deterministic prediction, which may be because of the upturn in allele frequency which we attribute to random fluctuations.

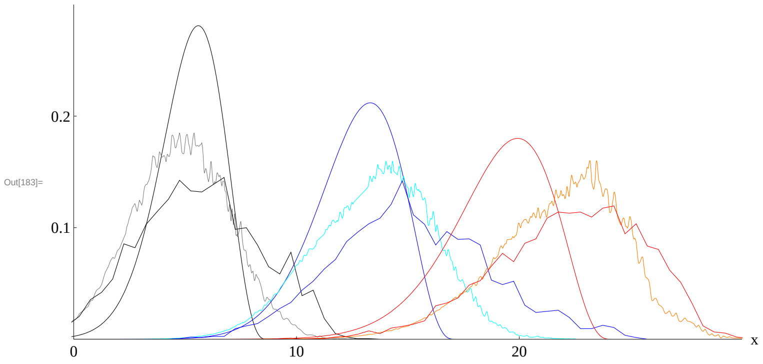


Figure 2.16 – The distribution of coalescence events, plotted against the position x relative to the centre of mass of the cline. Three sets of curves are shown for $\rho = 100$, 10^4, 10^6 (black, blue, red). Each set is based on a single forwards sweep of 10^5 generations,with $s = 0.5$, $m = 0.25$. The smooth curve is the prediction $f^2/(\rho w_{c_\eta}) \propto w_{c_\eta}^3 e^{2c_\eta x/\sigma^2}/\rho$. The middle jagged curve, drawn in a lighter shade, is the prediction $f^2/\rho w_{c_\eta}$ obtained by using the actual distribution of the location of ancestral lineages in the simulation for f. The broadest curve in each set is the observed distribution of coalescence events. Each is based on sampling one pair of lineages, replicated 400 times from each of 10 time points at $t = 0$, $10^4, \ldots, 9 \times 10^4$ (counting backwards). The distribution of coalescence events is close to the prediction based on the actual locations of ancestral locations, though somewhat further out to the front. The prediction based on the deterministic Fisher-KPP equation is sharper and lies further back - especially for the largest deme size, $\rho = 10^6$ (red curves at the right).

Genealogies with recombination at the wave front

Up to now we have left aside recombination. Let us now investigate its effects on genealogies as a function of the separation r on the genetic map between the locus under selection and the neutral locus of interest. Recall that the distance between two loci is measured in Morgans, so that recombination happens at rate r between two sites at genetic distance r. In all that follows, we suppose that the range of the population is finite (still one-dimensional) and that its size L satisfies

$$L \gg \ell := \sigma\sqrt{2/s},$$

where ℓ is the characteristic length of the wave front in the stationary deterministic wave. This assumption allows us to suppose that from its inception, the travelling wave front of favoured alleles is in its stationary form. Furthermore, we shall consider that lineages diffuse in space according Brownian motion with variance σ^2 before being caught by the wave. To set the notation, let us write λ for the approximate coalescence rate within the front

$$\lambda := \int_0^\infty \frac{f(x)^2}{\rho w_{c_\eta}(x)}\,\mathrm{d}x, \tag{2.40}$$

disregarding the coalescence due to large fluctuations. In Appendix B of [BEKV13b], we show that λ takes the form $2g(\eta)/(\rho\ell)$ (although we have no explicit form for the function g). Note that $\rho\ell$ is the order of magnitude of the 'number' of individuals in the wave front which are of the favoured allele. The term $g(\eta)$ gives a correction to this effective population density which accounts for the impact of the fluctuations in the front.

Figure 2.17 gives an overview of how recurrent selective sweeps, originating from random points in space and at random locations over a genetic map of length R, drive the genealogy of a sample. Consider one selective sweep. Suppose the wave is travelling at speed c and that a lineage is currently at distance y behind the wave front. Then the time before the lineage becomes caught by the wave can be approximated by the hitting time of 0 of Brownian motion with drift $-c$ and diffusion coefficient σ^2. The mean of this time is thus y/c. Assume next that the selected mutation arose at distance x away from the location where the lineage is taken by the wave. The expected remaining time until the origin of the favourable mutation is x/c. Then assuming that the lineage can recombine into the unfavourable background at a rate which is essentially r, the map distance between the neutral locus considered and the selected one, we obtain that the lineage will move with the wave front, and towards the origin of the beneficial mutation, by a distance which is the minimum between an exponentially distributed distance with parameter c/r and the distance x to the origin of the sweep. In particular, the probability that the lineage returns to the ancestral background without recombining is approximately $\exp(-rx/c)$. If we now assume that sweeps are uniformly distributed over space and over the genetic map, and if the overall rate of sweeps is Λ, and if finally we assume that $RL/c \gg 1$, then we can compute that the mean square displacement per unit time of a lineage due to hitchhiking, relative to that without hitchhiking, is

$$\frac{\sigma^2_{\text{eff}}}{\sigma^2} \approx \frac{8}{3}\,\frac{L}{\ell}\,\frac{\Lambda}{R}.$$

Now let us consider two lineages. Because population density is high, in the timescale we consider we can neglect coalescence due to events outside the front of the wave. Thus, the lineages will diffuse independently until they are both caught by the wave, at which point they start having the possibility of coalescing, or of recombining away from the favourable background (again, at rate r given by the map length between the selected site and the neutral

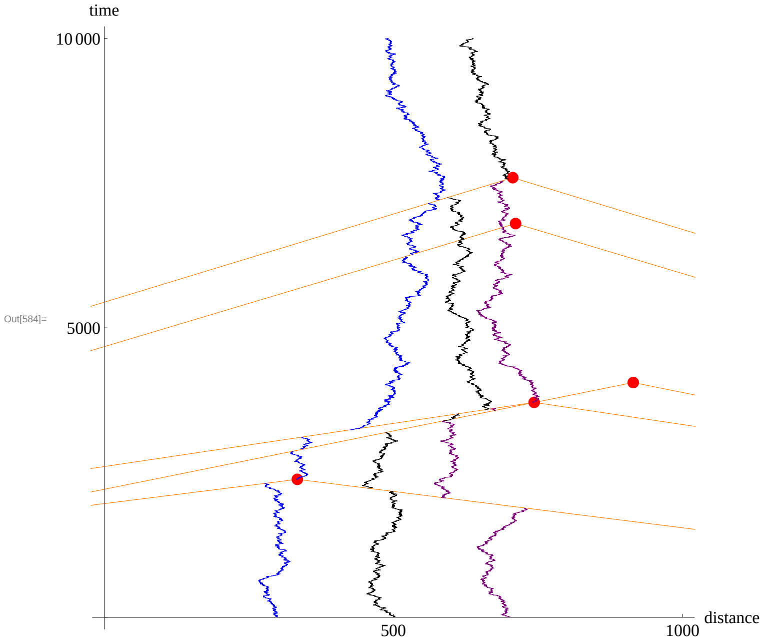


Figure 2.17 – The effect of 5 successive selective sweeps on the ancestry of three genes, sampled at $x = 300, 500, 700$; genes diffuse at a rate $\sigma^2 = 1$. The origin of five selective sweeps is shown by the red dots, and the advancing wavefronts by the orange lines. The origin, speed and position on the genetic map were drawn from a uniform distribution; in this example, the map location, relative to the focal locus, is -0.90cM, $+0.80$cM, -0.77cM, $+0.83$cM, -0.19cM. When a lineage hits a wavefront, it is carried back towards the favourable mutation (red dot) but may escape by recombination. The population is assumed very dense, so that coalescence only occurs within the wavefront. In this example, the purple and black lineages lie within the same wavefront ($x \sim 600 - 750$, $t \sim 3500$), but escape from it without coalescing. The only coalescence event occurs at the origin of the oldest selective sweep ($x \sim 700$, $t \sim 6800$), when both the purple and black lineages are carried back to coalesce on the genome that carried the favourable mutation.

site considered). We neglect the probability that they recombine again into the favourable background, since at the front the frequency of the favourable allele is very low. If they were stuck for an infinite amount of time in the wave, the two lineages would coalesce before one of them recombines with a probability approximately equal to $\lambda/(\lambda + 2r)$, where λ is the approximate coalescence rate defined in (2.40). In this case we can compute the probability of coalescence of two lineages sampled at distances x_1, x_2 behind the wave front. These expressions need to be corrected to account for the fact the lineages may trace back to the origin of the sweep without recombining or coalescing, but this has a small chance to happen if $Lr \gg c$. As we would expect, the probability that the two lineages actually meet in the front (instead of the first lineage caught by the wave front recombining away before the second is caught) depends on the scaled distance from the front, x_i/ℓ, and on the rate of recombination relative to selection r/s. If linkage is tight, i.e. $r \ll s$, the lineages are more likely to indeed meet at the wave front if their initial separation is not too big. In contrast, when linkage is loose ($r/s \approx 1$) this probability is low and very sensitive to the original scaled distance $|x_1 - x_2|/\ell$. The mean rate of coalescence, averaged over random locations of loci and sweeps, and over a genetic map of length $R \gg 1$ is

$$2\frac{\Lambda}{R}\frac{c}{L}h\left(\frac{L\lambda}{c}\right),$$

for a function $h(\theta)$ which tends to 0 as $\theta \to 0$ and is equivalent to

$$2\left(\ln\left(\frac{\theta}{2}\right) + \gamma - 1\right)$$

for large θ, where γ is Euler's gamma. Since h increases only logarithmically with θ, the net rate of coalescence is rather insensitive to the rate of coalescence within the front. The limiting factor is the rate at which tightly linked sweeps occur.

The average coalescence rate per sweep per map length is inversely proportional to L/c, which is roughly the time a sweep takes to cross the population's range. Since this is typically large, and Λ/R is expected to be small, the rate of coalescence due to hitchhiking will be very small. However, if the local density of individuals is high, then genetic drift will be negligible, so that hitchhiking may still be the main cause of coalescence [51, 80]. The effective size of a neutral haploid one-dimensional population is $\rho L + \frac{1}{12}\left(\frac{L}{\sigma}\right)^2$, where ρL is the total number of haploid individuals [22]. The contribution of hitchhiking will be larger than this essentially if

$$2\frac{\Lambda}{R}\frac{c}{L}h\left(\frac{L\lambda}{c}\right) > \frac{1}{\rho L},$$

The factor $c\rho$ that appears when we reorganise this condition is proportional to $\rho s W$, where W is twice the total number of heterozygotes which we took as a proxy for the width of the wave. We expect W to be proportional to $\ell = \sigma\sqrt{2/s}$, and so $\rho s W$ will proportional to what we defined to be η in (2.37). Thus, in one dimension, if the product of the rate of sweeps per map length and the parameter η is large, hitchhiking will be the main cause of coalescence.

Finally, comparing the probabilities that a sweep causes two lineages to coalesce with or without space (see Section 4.2.2 and Appendix C.1 in [BEKV13b] for their computations), we can conclude that the effect of a sweep on coalescence falls away much faster on the genetic map in a panmictic population than in one spatial dimension ($\sim e^{-r/s}$ vs. $\sim 1/r$ for two nearby lineages). However, if two individuals are sampled at different locations, there will be an additional time of order $|x_1 - x_2|/c$ during which only one of their two lineages will be able to recombine, the other having not yet been trapped in the wave front. Thus, the effectiveness of hitchhiking in a spatially structured population will be greater than in a panmictic population for samples taken at small separations, but smaller for far-fetched individuals.

Two spatial dimensions

We can ask the same questions when the geographical structure is two-dimensional. Remember that Equation (2.36) has no solution in two dimensions, and so instead we suppose that the wave spreads with a linear wave front. Hence, on top of meeting in the coordinate orthogonal to the front, the lineages diffusing in two dimensions also have to meet in the coordinate transverse to the front. We can then build upon the results obtained in one dimension. Unfortunately, if this putative analysis were correct, the diffusion of lineages transverse to the wave would be so slow that genetic hitchhiking would be remarkably ineffective in two dimensions. Simulations as those displayed in Figure 2.18 show that coalescence events tend to be strongly clustered, contrasting with what you would expect from a system of two-dimensional Brownian motions trapped in a narrow front and coalescing at a given rate upon meeting. This suggests that fluctuations in the wave front produce multiple mergers that dominate the ancestry of a sample, even when (and in fact all the more so as) η is of moderate size.

2.6.3 The infinitesimal model

In this section, we consider the effect of selection and population structure on the evolution of the distribution of a phenotypic character (which we shall also call a *trait*), whose value depends on a very large number of genetic loci each having an *infinitesimal* effect. With in mind possible applications to animal breeding, the kind of traits we may want to consider are milk production, resistance to some pests, etc. In fact, we shall more generally condition on the pedigree of the population, that is the ‘physical’ ancestral relations in a sexual population where each individual has two parents. Knowing this pedigree and the parental traits, which are themselves influenced by selection, structure, or other factors, we want to know the distribution of the traits of the offspring within a given family. Our aim is to justify the model known in quantitative genetics as the *infinitesimal model*, which states that under some not so restrictive (but always elusive) conditions, the distribution of the trait of an offspring is a normal distribution centred on the mean of the parental traits and with a variance independent of these parental traits. Furthermore, the allelic distribution at every locus is barely distorted by conditioning on the resulting trait, which justifies that selection acting at the level of the phenotype does not affect much (at least for a few generations) the genetic diversity of the population at each locus. This corresponds to the preprint [BEV17].

More precisely, we find some conditions on the allelic distributions and trait values seen in the population so that if the trait of interest is the sum of M genetic components all of order $M^{-1/2}$, then as the number M of loci tends to infinity, the values of the traits of a given number of offspring of the same pair of parents converges to a Gaussian multivariate distribution with mean the average trait of the parents, and a variance that depends only on the trait variance in some reference (ancestral) population and on the matrix of probability of identity by descent of the parental alleles at a given locus. Here we present only the simplest case of an additive trait, which is by definition the sum of the allelic effects corresponding to each locus, but in [BEV17] we extend our analysis to the presence of *epistasis*. The latter refers to the fact that some groups of alleles may interfere to add some nonlinear component to the trait value. Of course these interactions need to be of the right order of magnitude, which is detailed in Section 3.2 of [BEV17]. See also this preprint for more motivations, applications and an extended historical review of this model, which has been extensively used in quantitative genetics over the last 50

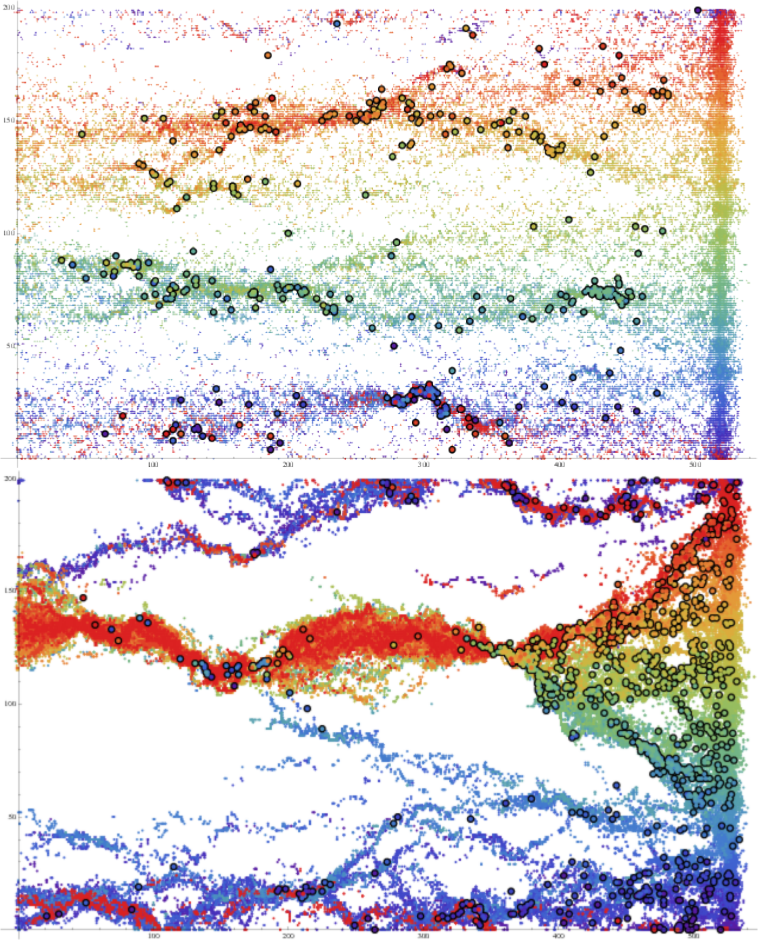


Figure 2.18 – Coalescence events in two dimensions cluster onto three paths, where most ancestors are located. The habitat is a cylinder with circumference 200 demes (vertical axis); the sweep runs from left to right, moving $\sim$ 520 demes in 4000 generations; $s = 0.05$, $m = 0.25$, a) $\rho = 10^6$, b) $\rho = 100$. Coalescence was simulated in two ways. First, 100 pairs of genes were started at each of 20 locations $y = 0, 10 \ldots, 190$ at right; coloured blue. . .red); large dots show where these coalesce. By 4000 generations, $461/2000 = 23\%$ had coalesced. Second, 50 single ancestral lineages were propagated back from $y = 0, 10 \ldots, 190$ (coloured in the same way); small dots show where any pair that started from the same location coincide in the same deme. The lower panel shows the same for $\rho = 100$. In that case 85% had coalesced by 4000 generations.

years without any thorough investigation of the assumptions under which it holds and of the accuracy of this approximation.

Thus, suppose that we consider a population of, say, haploid individuals, each with two parents (selfing is allowed, so that the two parents could in fact be identical). The population in each generation is supposed to be finite, with N_t individuals in generation t. Assume also that each individual has a trait Z which is given by some average $\bar{z}_0$ plus the sum over M loci of allelic effects of the form $\eta_l/\sqrt{M}$. A last term contributing to the value of the trait is an environmental noise, independent of the allelic components, which we take to be Gaussian so that the law of the observed trait in the limit $M \to \infty$ is Gaussian too. This form of noise is not compulsory, it only makes the calculations simpler. Let us describe the ingredients of the model in more detail.

1. *Allelic effect at locus l.* We denote the allelic effect at locus l in the jth individual by $\eta_l^j/\sqrt{M}$. We centre η_l^j relative to the mean allelic effect at locus l in the ancestral population. The scaling of $1/\sqrt{M}$ ensures that the additive genetic variance is of order one. The random variable η_l^j is assumed to be uniformly bounded over all loci, with $|\eta_l^j| \leq B$. We sometimes refer to it as the *scaled* allelic effect.

2. *Genetic component of the trait value and inheritance.* The genetic component of the trait value in the jth individual in the present generation will be denoted by Z^j. It will always be written as $\bar{z}_0$, its average value in the ancestral population, plus a sum over loci of allelic effects.

   That is, in the notation just defined, the genetic component of the trait of the jth individual is
$$Z^j = \bar{z}_0 + \sum_{l=1}^{M} \frac{1}{\sqrt{M}} \eta_l^j. \tag{2.41}$$

   We suppose that the loci are unlinked, that is the jth allelic effect of an individual is inherited from one of its two parents chosen uniformly at random, independently between loci.

3. *Environmental noise and observed trait value.* We suppose that the observed trait value is
$$\widetilde{Z}^j = Z^j + E^j,$$
   where the E^j are independent normally distributed random variables with mean zero and variance σ_E^2.

4. *Ancestral population.* Although it is not strictly necessary, we assume that in generation zero, the individuals that found the pedigree are unrelated. They are sampled from an ancestral population in which all loci are assumed to be in linkage equilibrium.

   The genetic component of the trait value in the jth individual in generation zero is written as
$$Z^j = \bar{z}_0 + \sum_{l=1}^{M} \frac{1}{\sqrt{M}} \widehat{\eta}_l^j, \tag{2.42}$$
   where the $\widehat{\eta}_l^j$ are independent for different values of j, with the same distribution as $\widehat{\eta}_l$ where $\mathbb{E}[\widehat{\eta}_l] = 0$ for all l. The random variables $\widehat{\eta}_l$ are assumed to be independent but not necessarily identically distributed.

   We shall write
$$\widehat{\sigma}_M^2 = \frac{1}{M} \sum_{l=1}^{M} \mathrm{Var}(\widehat{\eta}_l)$$
   and assume that $\widehat{\sigma}_M^2$ converges to a finite limit $\widehat{\sigma}^2$ as $M \to \infty$.

5. *Parents.* To distinguish the parents of an individual we order them. The symbols [1] and [2] will refer to the first and second parents of an individual, respectively. Thus $\eta_l^{j[1]}$ is the scaled allelic effect at locus l in the first parent of the jth individual. Similarly, $Z^{j[1]}$ will denote the genetic component of the trait value of the first parent of individual j. Note that we allow selfing, in which case parents 1 and 2 are identical.

Let $\mathcal{P}(t)$ denote the pedigree relationships between all individuals up to and including generation t, and let $\tilde{Z}(t)$ denote the observed traits of all individuals in the pedigree from the ancestral population to (and including) generation t. We shall condition on knowing $\mathcal{P}(t)$ and $\tilde{Z}(t-1)$, and derive the conditional distribution of the traits of the individuals in generation t in the limit $M \to \infty$.

Theorem 2.15. (Additive case in [BEV17]). *Fix a generation $t \geq 0$ and define the vector*

$$\left(R^{j,M}\right)_{j=1,\dots,N_t} := \left(Z^j - \frac{Z^{j[1]} + Z^{j[2]}}{2}\right)_{j=1,\dots,N_t}.$$

(We do not report the dependence on M in the r.h.s. to alleviate the notation.) Then conditional on $\mathcal{P}(t)$ and $\tilde{Z}(t-1)$, as $M \to \infty$ the vector $(R^{j,M})$ converges in distribution to a multivariate normal random variable with mean zero and diagonal covariance matrix Σ_t, with the jth diagonal entry $(\Sigma_t)_{jj}$ given by the segregation variance

$$\hat{\sigma}^2\big\{1 - \mathbb{P}\big(j[1] \textit{ and } j[2] \textit{ identical by descent at locus } 1\big)\big\} \tag{2.43}$$

among offspring of the parents of individual j. More precisely, let Φ stand for the cumulative distribution function of a standard normal random variable and define the function $C(\sigma^2, x)$ by

$$C(\sigma^2, x) = \frac{C'x}{\sqrt{\sigma_E^2 + \sigma^2}} + \frac{C''}{\sigma p(\sigma_E^2 + \sigma^2, x)}\left(1 + \frac{1}{\sigma^2}\right),$$

where C', C'' are two constants whose values depend only on B (the bound on the η's) and $p(\sigma_E^2 + \sigma^2, x)$ is the density at x of a $\mathcal{N}(0, \sigma_E^2 + \sigma^2)$ distributed random variable. Then for every $j \in \{1, \dots, N_t\}$,

$$\left|\mathbb{P}\left[\frac{Z^j - \frac{Z^{j[1]}+Z^{j[2]}}{2}}{\sqrt{(\Sigma_t^M)_{jj}}} \leq y \,\middle|\, \mathcal{P}(t), \tilde{Z}(t-1) = \underline{z}\right] - \Phi(y)\right| \leq \frac{t}{\sqrt{M}}\, C\big(\underline{\Sigma}_t^M, \bar{\Delta}_t^M(\underline{z})\big),$$

where

$$(\Sigma_t^M)_{jj} = \frac{1}{4M}\sum_{l=1}^{M} \mathbb{E}\Big[\big(\eta_l^{j[1]} - \eta_l^{j[2]}\big)^2 \Big| \mathcal{P}(t)\Big] \tag{2.44}$$

is the segregation variance among the offspring of the parents $j[1]$ and $j[2]$ of individual j in generation t conditional only on the pedigree (not the traits), $\underline{\Sigma}_t^M$ is the minimum segregation variance of any family in the pedigree up to generation t, and $\bar{\Delta}_t^M(\underline{z})$ is the maximum over the pedigree up to time $t-1$ of

$$\left|z^j - \frac{z^{j[1]} + z^{j[2]}}{2}\right|.$$

Theorem 2.15 can be extended to include mutation, a small amount of linkage (allelic effects being inherited by packets of $\mathcal{O}(1)$ loci) and epistasis, see [BEV17]. Even in its simplest form this result is very informative. First, it says that indeed the infinitesimal model is a good approximation for the trait distribution in a population under some explicit and rather general conditions on the genetics underlying the phenotype of interest. Even in the presence of selection acting on the trait, or of population structure, if many loci contribute each to a tiny fraction of the trait value then the distribution of the genetic component of the trait is Gaussian, with mean the average of the genetic components of the parental traits. Furthermore, the variance of this distribution is independent of the parental traits and is equal to the genetic variance in the ancestral population, reduced by a factor that depends on the pedigree only through the probability that the two parents are identical by descent at a given locus (this probability is constant across loci, since we assumed no linkage between loci in the ancestral population or during the process of inheritance). From (2.44) it is easy to see why this probability of identity by descent comes in the segregation variance. Indeed, identity by descent implies that the two parental allelic effects are identical at each concerned locus and thus reduces the variability among the siblings at some fraction of the M loci.

Second, the refined statement shows that the convergence of the vector $(R^{j,M})_{j=1,\dots,N_t}$ happens at a rate proportional to $t/\sqrt{M}$. Hence, the accuracy of the infinitesimal model is good over $\mathcal{O}(\sqrt{M})$ generations, at least theoretically. In fact, simulations suggest that convergence could be as fast as $1/M$ is some cases.

Third, the accuracy of the approximation is also dictated by the two quantities $\underline{\Sigma}_t^M$ and $\bar{\Delta}_t^M(\underline{z})$. This shows that the infinitesimal model will fail if either the genetic variability within a given family becomes too low, so that all their descendants inherit traits that are more and more alike as identity by descent erodes what remains of the additive variance; or if at least one trait seen in the population becomes too extreme, so that knowing the value of the trait gives some nonnegligible information on the allelic effects that built it.

Notice that we have not given the asymptotic behaviour of the *observed* traits, only their genetic components. The environmental component is necessary to the proof of Theorem 2.15 because it 'smoothes' the state space of the trait. Indeed, if for instance each allelic effect can only take the values 0 and 1, then for any finite M the genetic component of the trait value can take only a finite number of values too and this impedes the use of the local central limit theorem that we sketch below. Because the environmental noise has a continuous distribution, the observed trait value is also continuously distributed. Now, in the presence of this component, the observed trait is the sum of the genetic and of the environmental contributions and we cannot observe each separately. Instead, let us consider the vector

$$\big(\Delta\tilde{Z}^j\big)_{i=1,\dots,N_t} := \left(\tilde{Z}^j - \frac{\tilde{Z}^{j[1]}+\tilde{Z}^{j[2]}}{2}\right)_{i=1,\dots,N_t} = \left(R^j + E^j - \frac{E^{j[1]}+E^{j[2]}}{2}\right)_{i=1,\dots,N_t}.$$

From Theorem 2.15 we know that $(R^j)_{j=1,\dots,N_t}$ is a multivariate Gaussian vector which is approximately independent of $\tilde{Z}(t-1)$, and by construction the same holds true for $(E^j)_{j=1,\dots,N_t}$. The parental environmental noises are not independent of $\tilde{Z}(t-1)$, but a classical result on conditional multivariate distributions gives us the distribution of the vector $(E^j)_{j=1,\dots,N_{t-1}}$ conditional on $\mathcal{P}(t)$ and $\tilde{Z}(t-1)$ (see Appendix F in [BEV17]). We can thus obtain the conditional law of $\tilde{Z}(t)$, which is also multivariate Gaussian and a recursion propagates the result to later generations.

Finally, let us sketch the main argument in the proof of Theorem 2.15. We proceed by recursion. In generation 0 the result is straightforward using an extension of the Central Limit Theorem which was established in [95]. Note that the Central Limit Theorem cannot be used directly as the allelic effects at different loci are assumed to be independent but not necessarily identically distributed. Suppose that we have our result for generation $(t-1)$. The key step is then to show that for individual j in generation t, conditioning on knowing $\mathcal{P}(t)$ and $\widetilde{Z}(t-1)$ provides negligible information on the values $\eta_l^{j[1]}$, $\eta_l^{j[2]}$ of the scaled allelic effects of locus l in its parents. Through an application of Bayes' rule, this will essentially boil down to showing that

$$\left|\frac{\mathbb{P}\big[\widetilde{Z}^{j[1]}=z_1\big|\eta_l^{j[1]}=x,\mathcal{P}(t)\big]}{\mathbb{P}\big[\widetilde{Z}^{j[1]}=z_1\big|\mathcal{P}(t)\big]}-1\right|\leq\frac{t}{\sqrt{M}}\,C\big(\underline{\Sigma}_t^M,\bar{\Delta}_t^M(\underline{z})\big),\tag{2.45}$$

where, since $\widetilde{Z}$ is a continuous random variable, the ratio on the left should be interpreted as a ratio of probability density functions, and the function $C(\sigma^2,x)$ was defined earlier. This result is due to the fact that the contribution of $\eta_l^{j[1]}/\sqrt{M}$ to $\widetilde{Z}^{j[1]}$ is so small that both $\widetilde{Z}^{j[1]}-\eta_l^{j[1]}/\sqrt{M}$ and $\widetilde{Z}^{j[1]}$ converge to the same normal distribution. The proof depends crucially on knowing the rate of convergence of the distribution of the parental trait values to a multivariate normal.

What (2.45) allows us to deduce (via Bayes' rule) is that knowing the trait of an individual gives very little information about the allelic state at a single locus. As we discussed earlier, although intuitively clear this result will break down if the segregation variance somewhere in our pedigree is small or if a trait in the pedigree is too extreme. Armed with (2.45), we can approximate the distribution of the allelic effects conditioned on $\mathcal{P}(t)$ and $\widetilde{Z}(t-1)$ by those conditioned just on $\mathcal{P}(t)$ and then it is an easy matter to identify the limiting variance-covariance matrix of the random variables $(R^j)_{j=1,\dots,N_t}$ in generation t.

Convergence of the vector of the genetic components of the trait values in generation t to a multivariate normal is then an application of the extended Central Limit Theorem. Knowing the rate of this convergence allows us to prove the analogue of (2.45) for generation $(t+1)$, and so on.

2.6.4 Perspectives

As we have seen in Section 2.6.1, many possible ranges of parameters could be investigated and would give rise to different long-term evolutions for the genetic diversity of the population. Without drawing up a catalogue of possible behaviours, it is important to look for detectable signatures of selection in a spatially structured population. This can be done by analysing the pattern of allele frequencies at the locus under selection. But this pattern can be very flat if selection is sufficiently strong for the favourable allele to have already swept to fixation in a large region of space. Instead, or to complement this first approach, we can also try to understand the effects of the selection pressure on linked neutral loci, as was initiated in the work presented in Section 2.6.2. In particular, the study carried out there is much more intricate in two dimensions because of the fluctuations in the front of the 'invasion' wave. This is a very natural and biologically relevant question that needs to be addressed in the future.

Talking about invasions, let us also mention that taking the selection strength to infinity leads to a model in which only one type of individuals reproduces and invades space in a wave-like manner. This kind of processes can model the expansion of a species into a new habitat. Some heuristics on such a population expansion and its links to classical growth models (like the Eden model) already exist, but they appear to be very difficult to justify rigourously. Of particular interest is the phenomenon of *gene surfing*, whereby deleterious mutations can remain present in the population for a large amount of time by literally surfing on the front of the wave of advance of the population, where there is only limited competition for space (so that carrying a deleterious allele is not such a burden). See for example [57] and references therein.

Chapter 3

Some models of pedigrees

In this chapter, we review two works investigating the shape of the 'physical' ancestry of a number of individuals taken in a panmictic sexual population with two-parent Wright-Fisher resampling first, and in a spatially extended sexual population next. The last work presented here is slightly different. It introduces a model of random trees which generalises a construction of Blum and François [15], and could be used to model species trees, transmission trees of infections, or cell lineage diagrams, for example.

3.1 Ancestries of a panmictic sexual population

In the celebrated paper [21], Chang considers a modification to the Wright-Fisher model presented in Section 2.1, in which each of the N individuals forming the next generation chooses *two* parents uniformly and independently at random within the current generation. The main motivation is to model the pedigree of a diploid population, in which each individual inherits its genes from two parents. Considering large population sizes, Chang proves the two following results. For a given population size N (constant through time), let $\mathcal{T}_N$ denote the number of generations we need to come back in the past to find the first individual which is an ancestor to the *whole* present population. That is, we trace back the bi-parental ancestry of every individual living in the current generation, and stop when we reach the first generation in the past in which we see an individual belonging to the ancestry of everyone. Observe that in contrast with the standard Wright-Fisher model, ancestral lines can not only merge but also branch into two since each individual has two parents. As a consequence, there may be more than one such common ancestors in a given generation.

Theorem 3.1. (Th. 1 in [21]). *As $N \to \infty$,*

$$\frac{\mathcal{T}_N}{\log_2 N} \xrightarrow{\text{(prob)}} 1.$$

Furthermore, if $\mathcal{U}_N$ denotes the number of generations to come back to find the first generation in which every individual is either an ancestor to everyone in the present generation, or an ancestor to no present-day individuals, we have the following asymptotic result.

Theorem 3.2. (Th. 2 in [21]). *Let γ denote the smaller of the two numbers satisfying the equation $\gamma e^{-\gamma} = 2e^{-2}$, and let $\zeta = -1/(\log_2 \gamma) \approx 0.7698$. Then as $N \to \infty$,*

$$\frac{\mathcal{U}_N}{(1+\zeta)\log_2 N} \xrightarrow{\text{(prob)}} 1.$$

Finally, the fraction of individuals in generation $-\mathcal{U}_N$ which are common ancestors to the whole present population is $1 - \gamma/2$, where γ is defined in the statement of Theorem 3.2 and corresponds to twice the extinction probability of a Galton-Watson process with offspring numbers drawn from a Poisson(2) random variable.

Chang's results quickly received a lot of attention and were discussed in [33] by some of the most famous population geneticists. Indeed, they showed that contrary to the ancestry at a non recombining locus in which the most recent common ancestor to the population is found after coming back at least $\mathcal{O}(N)$ generations, in the biparental pedigree describing the (potential) ancestry at the level of the whole genome, common ancestors can be found in much less time, of the order of $\log_2 N$. In about $1.77 \log_2 N$ generations, everyone shares the same ancestors. Note in passing that in any generation earlier than $\mathcal{U}_N$, again every individual is also either a common ancestor to every present-day individual, or to none of them. Of course being a pedigree ancestor does not necessarily imply being a genetic ancestor, especially if we consider only a restricted part of the genome. In the extreme case where this stretch of genome never recombines, every individual inherits its DNA from one of the two parents only and we are back in the standard haploid Wright-Fisher framework to describe the genetic ancestry at this locus. Even in the presence of recombination, the genetic contribution of a pedigree ancestor may become lost by a series of recombination events and this ancestor then becomes a *ghost* (genetically speaking). The relation between pedigree ancestry and genetic ancestry has been the object of several studies, which focus on the fate of independent non-recombining loci [78, 111] or that of the genetic content of an individual within a pedigree [7, 52].

We instead focus on a region of DNA in which there can be recombination with probability $r \in [0, 1]$ or not during each reproduction event, and we study the shape of the genealogy made of all ancestors potentially contributing to the genetic state in each individual of the present population. That is, in the version of the Wright-Fisher model we consider, each individual has two parents with probability r, which are then chosen independently and uniformly at random in the previous generation, or only one parent with probability $1-r$, again chosen at random in the previous generation. We are thus at an intermediate level where we take (the absence of) recombination into account to keep only the ancestors that are likely to transmit some of their genes, but do not model the fact that the genetic contribution of such ancestors could vanish because of a 'bad' series of recombinations. This model can also be applied to other phenomena such as *paternal leakage* in mitochondrial DNA, which corresponds to paternal mitochondria entering the egg cytoplasm at fertilisation (mitochondria and consequently mitochondrial DNA are predominantly maternally inherited). Such biparental mitochondrial inheritance has been documented in mammals, birds, reptiles, fish, molluscs, nematodes and arthropods, and is the norm in some bivalves, see [115]. The parameter r in this case is the probability of paternal leakage per generation.

Remark 3.1. *The ancestry in the model we consider here is the discrete analogue of the ancestral recombination graph (see for instance Chapter 7 of [110]). In fact, Theorem 4 in [STV16] states that if the recombination probability r_N tends to 0 as $N \to \infty$ in such a way that $Nr_N \to \rho \geq 0$, then as $N \to \infty$ the ancestral process of a finite sample of individuals on the timescale $(\lfloor Nt \rfloor, t \geq 0)$ converges in distribution to the ancestral recombination graph with recombination rate ρ and pairwise coalescence rate 1. The proof is rather combinatorial and seems to have never been done before.*

Recall the notation $\mathcal{T}_N$ for the (backward) generation of the first common ancestor to the whole present population, and $\mathcal{U}_N$ for the number of generations, counting back in time from the present, to the first generation at which each individual is either a common ancestor to all present-day individuals or to none of them. Notice that the case $r = 0$ corresponds to the haploid Wright-Fisher model with Kingman-like genealogies, while $r = 1$ corresponds to Chang's biparental Wright-Fisher model. Hence, there remains to consider the case $r \in (0, 1)$. Generalising the method of Chang, in [STV16] we show the following results.

Theorem 3.3. (Th. 2 in [STV16]). *Let $r \in (0, 1)$. As $N \to \infty$, we have*

$$\frac{\mathcal{T}_N}{C(r)\ln N} \overset{\text{(prob)}}{\longrightarrow} 1,$$

where

$$C(r) := \frac{1}{\ln(1+r)} - \frac{1}{\ln(1-r)}.$$

Theorem 3.4. (Th. 3 in [STV16]). *Let $r \in (0, 1)$. Let also $\rho = \rho(r)$ be the unique solution in $(0, 1)$ to the equation $x = e^{-(1+r)(1-x)}$. Then for every $\varepsilon > 0$,*

$$\lim_{N\to\infty} \mathbb{P}\bigg[(1-\varepsilon)\bigg(C(r) - \frac{1}{\ln((1+r)\rho)}\bigg)\ln N \le \mathcal{U}_N \\ \le (1+\varepsilon)\bigg(C(r) - \frac{1}{\ln((1+r)\rho)} - \frac{1}{\ln(1-r)}\bigg)\ln N\bigg] = 1.$$

The result of Theorem 3.4 is less sharp than the corresponding result of Chang. As explained below, this is due to the fact that in the case $r \in (0, 1)$ the extinction of the set of 'non-descendants' of a given individual in the past ends much more slowly and with a greater variability than in the case $r = 1$. This phenomenon was overlooked in the conjecture that $C(r)$ should be equal to $1/\ln(1 + r)$ proposed in [33]. The same mistake was made in a few biology papers like [72], and so it seemed important to provide a correct statement for this result.

Finally, concerning the fraction of individuals in generation $-\mathcal{U}_N$ who are common ancestors to the whole present population, we have:

Proposition 3.1. (Cor. 1 in [STV16]). *Let $r \in [0, 1]$. The fraction of individuals living $\mathcal{U}_N$ generations ago that are common ancestors to the current population converges in probability to $1 - \rho$ as $N \to \infty$, where ρ was defined in the statement of Theorem 3.4 and corresponds to the extinction probability of a Galton-Watson process with offspring distribution Poisson$(1+r)$.*

The main idea of the proof of these results is to start from some ancestral generation '0' and to consider the evolution of the family size of a given individual forwards in time. Let us fix an individual in generation 0 and for any $t \in \mathbb{N}$, let G_t denote the number of descendants of this individual living t generations later. Because an individual in generation $t + 1$ belongs to the family of size G_{t+1} if and only if at least one of its parents (or its single parent if there was no recombination) belong to the descendants in generation t, the Wright-Fisher resampling mechanism enables us to write that conditionally on G_t,

$$G_{t+1} \sim \mathrm{Bin}\bigg(N, (1-r)\frac{G_t}{N} + r\bigg(1 - \bigg(1 - \frac{G_t}{N}\bigg)^2\bigg)\bigg) = \mathrm{Bin}\bigg(N, (1+r)\frac{G_t}{N} - r\frac{G_t^2}{N^2}\bigg).$$

We also have $G_0 = 1$. As a consequence, when G_t is small, it essentially behaves like the supercritical Galton-Watson process with offspring law the Poisson$(1+r)$-distribution. On the other hand, when G_t is big enough it is nearly equal to its mean. Making these two statements precise are key ingredients in the proof, see Lemmas 2 and 13 in [STV16]. Assuming that

G_t does become large (instead of being absorbed at 0), we then consider the number B_t of individuals in generation t which are *not* descendants. This time, conditionally on B_t we have

$$B_{t+1} \sim \text{Bin}\left(N, (1-r)\frac{B_t}{N} + r\frac{B_t^2}{N^2}\right),$$

and again when B_t is large it is very close to its mean whereas once it has become small, it evolves approximately like the subcritical Poisson$(1-r)$ Galton-Watson process.

Let us now briefly sketch the proof of Theorem 3.3. It consists in splitting the creation and evolution of a successful family into 4 stages. We fix an $\varepsilon > 0$ small.

— **Stage 1:** there exists an individual I in generation 0 for which $(G_t^I)_{t\in\mathbb{N}}$ reaches $(\ln N)^2$ in less than $3(\ln\ln N)^2/\ln(1+r)$ generations with probability tending to 1. To see this, let us look at the family of individual 1. At the beginning it behaves like a supercritical Galton-Watson process for which we know that the probability that the size of the population reaches $(\ln N)^2$ in less than $3(\ln\ln N)/\ln(1+r)$ is asymptotically bounded from below by the survival probability $1-\rho$. If the family of individual 1 becomes larger than $(\ln N)^2$ in less than $3(\ln\ln N)/\ln(1+r)$ generations we have our successful individuals, otherwise we consider individual 1 in generation $3(\ln\ln N)/\ln(1+r)+1$ and start again the same reasoning. Proceeding incrementally, the number of attempts before we find an 'individual 1' whose family manages to grow sufficiently fast is stochastically bounded by a geometric random variable with success probability $1-\rho-\delta$ (for some δ sufficiently small). The probability that this number is larger than $\ln\ln N$ thus tends to 0 as $N\to\infty$. From now on we consider the family of this successful individual.

— **Stage 2:** G^I grows from $(\ln N)^2$ to more than $N/2$ in less than $\tau_N^{(2)}$ generations, where

$$\mathbb{P}\left(\tau_N^{(2)} > \left(1+\frac{\varepsilon}{2}\right)\frac{\ln N}{\ln(1+r)}\right) = o\left(\frac{1}{N}\right),$$

and

$$\mathbb{P}\left(\tau_N^{(2)} < \left(1-\frac{\varepsilon}{2}\right)\frac{\ln N}{\ln(1+r)}\right) = o\left(\frac{1}{N}\right).$$

These results are obtained by using a Bernstein inequality recalled in Lemma 2 of [STV16]. In effect, we have to split this stage (and the next one) into two steps in the full proof of Theorem 3.3, but this is only for an uninteresting technical reason.

— **Stage 3:** The family of non-descendants of I, of size B_t^I, goes down from less than $N/2$ to less than $(\ln N)^2$ in $\tau_N^{(3)}$ generations, where

$$\mathbb{P}\left(\tau_N^{(3)} > \left(1+\frac{\varepsilon}{2}\right)\frac{-\ln N}{\ln(1-r)}\right) = o\left(\frac{1}{N}\right),$$

and

$$\mathbb{P}\left(\tau_N^{(2)} < \left(1-\frac{\varepsilon}{2}\right)\frac{-\ln N}{\ln(1-r)}\right) = o\left(\frac{1}{N}\right).$$

— **Stage 4:** B^I starting from less than $(\ln N)^2$ is absorbed at 0 in less than $-3(\ln\ln N)/\ln(1-r)$ generations with probability tending to 1. This result is obtained by comparing B^I to a Poisson$(1-r)$ Galton-Watson process.

Based on these four results we easily obtain that $\mathcal{T}_N \leq (1+\varepsilon)C(r)\ln N$ with probability tending to 1. Since the probabilities described in Stages 2 and 3 are of order $o(1/N)$, summing over all individuals in generation 0 tells us that the probability that at least one of the corresponding families reach size N in less than $(1-\varepsilon)C(r)\ln N$ tends to 0 as $N \to \infty$.

As concerns the result of Theorem 3.4, the (positive) term $-(\ln N)/\ln((1+r)\rho)$ describes the additional time needed for *all* the families of the N individuals in generation 0 to either go extinct or be successful and reach size $(\ln N)^2$. The term $-(\ln N)/\ln(1-r)$ is an upper bound on the time required for all the successful families to eventually reach size N. In Chang's case $r = 1$, the rate of decay of a family of non-descendants is quadratic in N, and so Stage 4 and this last amount of time for the complete success of all successful families are negligible compared to $\ln N$. In the case $r \in (0,1)$, it is difficult to quantify the amount of 'slow' successful families, whose sizes reach $(\ln N)^2$ in about $(\ln N)(\frac{1}{\ln(1+r)} - \frac{1}{\ln((1+r)\rho)})$ generations, and their reaching N could in fact take much less than $-(\ln N)/\ln(1-r)$ generations to happen.

Finally, Proposition 3.1 is proved by observing from the proof of Theorem 3.3 that the fraction of individuals in generation 0 that will reach the state of common ancestor $\mathcal{U}_N$ generations later is made of those individuals whose families are initially successful.

3.2 Pedigree vs. genetic ancestry in a spatial population

The last section was concerned with the pedigree of a panmictic population of sexually reproducing individuals. A spatial structure will inevitably modify the form of the pedigree, and this is what we enquire here. This section corresponds to the publication [KEVB16].

Let us suppose that the population of interest is distributed over some large one-dimensional space. In [KEVB16], we work with three models of spatial structure: (*i*) a spatial Λ-Fleming-Viot process with local reproductions only, two parents being drawn at random during each event, and the population evolving on a continuous torus of length L (sufficiently large to mimick $\mathbb{R}$ over the timescale of interest in the simulations); (*ii*) a stepping stone model with biparental Wright-Fisher resampling within each of the L demes in the discrete torus $\mathbb{Z}/L\mathbb{Z}$, and with nearest neighbour migration; (*iii*) another stepping stone model on $\mathbb{Z}/L\mathbb{Z}$ with nearest neighbour migration but biparental Moran resampling within each deme (that is, only a pair of individuals reproduces at a time). We only present the results based on the SLFV, but the same approach leads to analogous results in the three models. A particularly interesting point, though, is that the statistics describing the wave of pedigree ancestors are heavily dependent on the details of the model, whereas the wave of genetic ancestors depends only on a few compound parameters which have the same interpretation in all three models (among which σ^2, the variance parameter of the long-term diffusion of a single lineage).

Let us also suppose that individuals are haploid, each with two parents. We consider a finite number l of linearly arranged loci, with recombination occurring with some probability ρ between any pair of neighbouring loci during a reproduction event. In what follows we shall consider the case $\rho = 1/2$ of free recombination (equivalent to each locus 'choosing' independently the parent from which it inherits its allele), but this concerns only the genetic ancestry and can be generalised. In general, we assume that the population size is large but regulated in such a way that it remains approximately constant locally (as in the three models mentioned above).

Pedigree ancestry

We sample a single individual, say at location 0, and we first trace back the spatial distribution of its ancestors as a function of time (running backwards into the past). Recall that an individual always has two parents, and so the pedigree ancestry is made of a set of lineages which branch into two nearby lineages at some given rate, but may also coalesce when they reach a common ancestor. During both types of events, the lineages also jump to new locations, which are the spatial positions of the corresponding ancestors. Initially, the set of ancestral lineages behaves essentially like a branching random walk and coalescence has only a small effect. However, the local density of ancestors cannot grow indefinitely because of our assumption of regulated local population densities. At this time, either the lineages manage to escape to less populated areas further away from the origin, or they coalesce locally. We thus expect the population of pedigree ancestors to develop into a travelling wave advancing in space at some speed that depends on the characteristics of the model. In Sections 3.1 and 3.2 of [KEVB16], we find some analytic approximations for the wave of advance of the pedigree ancestors. Although they capture the essence of the evolution of this population of ancestors, the approximation is still too crude to derive an estimate of the most important summary statistics of the wave, such as its speed, width, or front shape. Instead we resort to simulations. Defining $p_t(x)$ to be the size of the population at x relative to the mean size of the population at the origin, all this considered at time t in the past, we can use again the definition of the wave centre z_t and width W_t made in Section 2.6.2:

$$z_t = \int p_t(x)\mathrm{d}x, \qquad W_t = 4\int p_t(x)(1-p_t(x))\,\mathrm{d}x.$$

We want to compare our three models, as well as to compare them to the classical Fisher-KPP travelling wave. Thus, as in Section 2.5 we define the dispersal rate σ^2 as the variance of the Brownian motion which approximates the long-term behaviour of a single lineage, and the effective population density ρ_e by

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

in the continuum, and

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

for the discrete space Wright-Fisher and Moran stepping stone models. Here $h(x)$ is the instantaneous coalescence rate of two lineages at separation x and the integral or sum is over all possible separations (in $\mathbb{R}$) between two lineages. We then choose our model parameters so that these two statistics are identical between the models.

Figure 3.1 shows the mean wave shapes in the three models and compares them to that predicted by the Fisher-KPP equation. We see that they are all quite different, which suggests that the shape of the wave is very sensitive to the details of the model. Figure 3.2 shows simulations of the wave centre and width for the three models and three different effective population densities. We see that after an initial period of time during which the wave establishes, the position z_t of the wave moves linearly in time at a speed which is independent of ρ_e (like the travelling wave solution to the Fisher-KPP equation, we expect the wave of ancestors to be a *pulled* wave and so its speed should be independent of the population density in the bulk). Likewise, the width of the wave depends on the model but not on ρ_e.

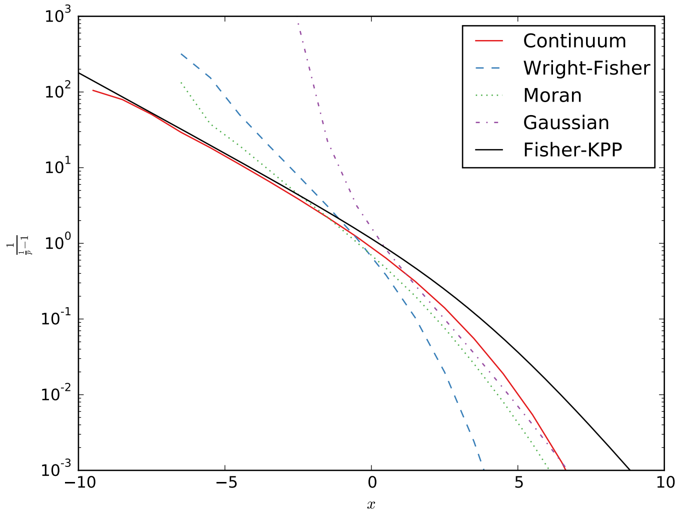


Figure 3.1 – Mean wave shapes in the simulated continuum, Wright-Fisher and Moran models with $\rho_e = 100$ and $\sigma = 1/\sqrt{2}$, along with theoretical predictions. On the x-axis is distance relative to the front centre, and on the y-axis (on a logarithmic scale) is $p(x)/(1-p(x))$, where $p(x)$ is the ancestral population size relative to the mean size at the origin. (A logistic curve would be linear.) Also shown is the predicted wave shape for the Gaussian approximation along with a solution to the Fisher-KPP equation.

Genetic ancestry

Recall that we consider l linearly arranged loci in the genome and free recombination between them (i.e., $\rho = 1/2$). Let us now keep track of the pedigree ancestors carrying genetic material ancestral to our sampled individual. The number of such ancestors is limited, since there cannot be more than l of them at any given time in the past. Since there cannot either be more genetic ancestors than pedigree ancestors, we expect the early development of the population of genetic ancestors to be essentially the same as that of the population of pedigree ancestors. However, after some time recombination and diffusion (and the slow coalescence rates due to our assumption of large ρ_e) will have split the genetic ancestry into approximately l lineages moving around in space like finite-variance random walks. For reasonably large times t, the area in which these ancestors can be found has a width of the order of $2\sigma\sqrt{t}$ and the genetic ancestors are well spread therein, evolving nearly independently of each other. Thus, we expect the wave of genetic ancestors to be much slower than that of pedigree ancestors, which is confirmed by simulations. Figure 3.3 shows the same summary statistics as in Figure 3.2, but for the wave of genetic ancestors. This time we see that these statistics are insensitive to the details of the model (recall that we only imposed that σ^2 and ρ_e should be the same in all models).

Two spatial dimensions

The analysis and comparisons through simulations expounded above can also be carried out in the more biologically relevant case of a two-dimensional population range. This is summarised in Section 5 of [KEVB16], where we show that essentially the same results hold true in this case.

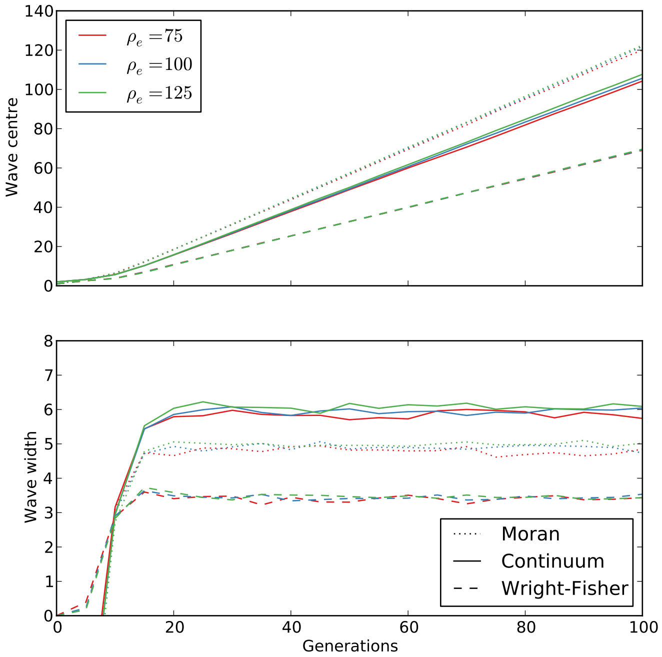


Figure 3.2 – Estimated pedigree wave centre and width for a range of effective densities in the continuum, Wright-Fisher and Moran models. The mean wave centre and width are estimated from 1000 replicate simulations. For each replicate, we estimate the centre and width independently and then take the mean of these values over all replicates.

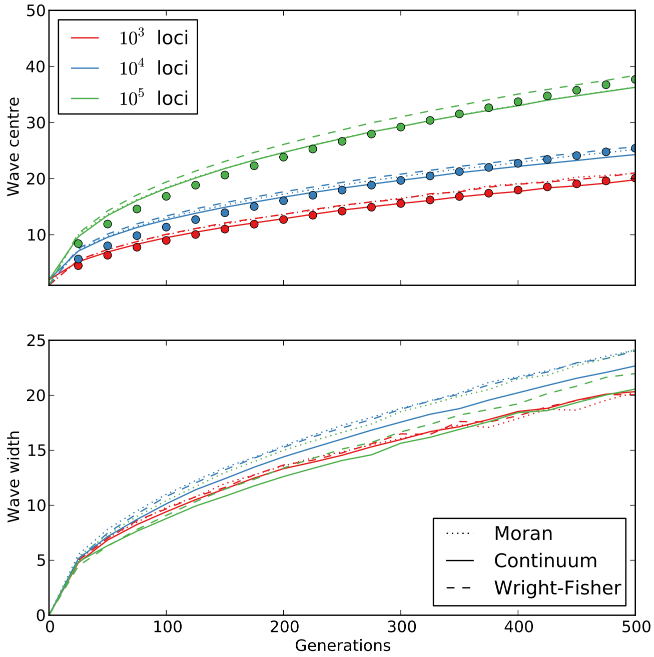


Figure 3.3 – Estimated wave centre and width for the genetic ancestors for different numbers of loci under free recombination ($\rho_e = 100$). Also shown here in dots is $a\sqrt{t}$ fitted for each value of l.

3.3 A model of asymmetric ancestral relationships

In this section, we present a model of random tree topologies which can encode the potential asymmetry between the two child nodes of a given internal node. For example, a speciation event could be seen as the divergence of one new species from an ancestral species which carries on existing. In this case, the two 'new' species play different roles and it may be relevant to include this asymmetry in the model. As an other example, in the transmission tree of an epidemics (which records who infected whom in which order), *a priori* the infector carries on being infectious after transmitting the disease to a susceptible individual, and it could be of interest to keep track of who plays which role. This example is developped in [98], where the disease transmission tree is constrained by the network over which the population of infectors and infectees is distributed. There the authors show that the parameters α, β introduced below can serve to classify the underlying individual networks into classes of (oriented, weighted) graphs which are equivalent with respect to the transmission of a disease.

Disregarding the modelling of the times between the events, which can usually be added later as a topology-dependent process, we wish to provide a family of random topologies characterised by a small number of parameters and offering a large panel of tree shapes. Indeed, the topology of Kingman's coalescent, which happens to be the same as the topology of the Yule process [121] (extensively used in phylogenetics, for example), is by far the most commonly used model for the topology of the tree describing the ancestral relationships between a sample of individuals within a population, or a sample of species within a clade. But many processes such as natural selection or population structure at the population level, or adaptive radiation at the species level, may give rise to more balanced or unbalanced trees. Therefore, a model in which the balance of the tree can be tuned by a parameter would enable us to quantify the effects of such processes on the shape of the tree of evolutionary relationships, or to test the adequacy of the standard null model of the Kingman-Yule topology.

In [2], Aldous introduces a one-parameter family of random cladograms, called the *Beta-splitting model*. Here a cladogram is defined as a binary tree shape with a specified number of leaves in which there is no 'left' and 'right' ordering of the child nodes of an internal node. The leaves are labelled by the sampled species, or by $\{1, \dots, n\}$ for simplicity. The parameter $\beta > -2$ modulates the shape and balance of the tree produced by this model by determining the split distribution of a node subtending m leaves. More precisely, Aldous' recursive construction involves a fixed n, the number of leaves representing the extant species in a tree with at least two leaves and $\{q_n^\beta(i) : i = 1, 2, \dots, n-1\}$, a symmetric probability distribution (i.e., $q_n^\beta(i) = q_n^\beta(n-i)$) which specifies the numbers i and $n-i$ of descendants along the two branches emanating from the root node of the tree. Once this split $(i, n-i)$ is fixed, the construction carries on recursively in the two subtrees pending from the root, with respective numbers of leaf nodes i and $n-i$, and stops when all subtrees considered have only one leaf. In the Beta-splitting model with $\beta > -2$, the split distribution q_n^β takes the form

$$q_n^\beta(i) = \frac{1}{a_n} \binom{n}{i} \int_0^1 x^{i+\beta}(1-x)^{n-i+\beta} dx \tag{3.1}$$

for $1 \leq i \leq n-1$, where a_n is a normalizing factor given by

$$a_n = \int_0^1 \big(1 - x^n - (1-x)^n\big) x^\beta (1-x)^\beta dx.$$

This *Markov branching model* has now become a reference in the literature [15, 90], in particular because it provides a family of random tree topologies indexed by a single parameter, which contains the most commonly used Kingman-Yule tree ($\beta = 0$) and Proportional to Distinguishable Arrangements (or PDA) model in which every cladogram is equally likely ($\beta = -3/2$). The parameter β tunes the balance of the tree, since '$\beta = -2$' corresponds to the totally unbalanced tree or comb, whereas the generated trees become more and more balanced as β tends to infinity. In [3], Aldous also proposes a measure of the balance of a tree which has the advantage of being independent of the tree size, at least for large n's: the median of the split distribution q_n^β. This measure is used to perform maximum likelihood estimation of β or to compare the global balance of several trees [3, 15].

Unfortunately, there seems to be no underlying evolutionary process whose outcome is the random cladogram obtained for a given value of β (unless $\beta = 0$, corresponding to Yule's pure birth process). Indeed, since the number of leaves has to be known before recursive splitting begins, Aldous' Beta-splitting model is not based on an incremental construction from one ancestral to n present species/individuals, nor is it defined jointly on the product space of tree topologies and branch-lengths for every value of $\beta > -2$. It thus lacks evolutionary interpretability.

To overcome this problem, in [15], Blum and François introduce an evolutionary Beta-splitting model based on ideas of Kirkpatrick and Slatkin [68] and Aldous [2]. The idea is that the 'speciation potential' is shared between the two offspring species in a random way, as may occur e.g. in the case where speciation is influenced by available niche or geographical space that is shared between the two new species. In this model, a (rooted binary non-planar) tree is constructed incrementally by starting from a single node (the root) with speciation rate (or 'potential') 1. When this first species branches, a parameter p_1 is sampled in $[0, 1]$ according to a Beta$(\beta+1, \beta+1)$ distribution. Then the first offspring species is given the speciation rate p_1, and the second the speciation rate $1 - p_1$. The next species to split is thus the first one with probability p_1, or the second one with probability $1 - p_1$. Carrying on the construction, upon the split of a species with speciation rate λ, a new parameter p_i is sampled independently of the previous ones according to the same Beta$(\beta + 1, \beta + 1)$ distribution, and the two sister species receive the speciation rates λp_i and $\lambda(1 - p_i)$. Then, each species is the next one to branch with a probability equal to its speciation rate/potential.

Though the Blum-François and the Aldous Beta-splitting models coincide for $\beta = 0$, in general they do not yield the same distribution on cladograms. See the Supplementary Material of [15] for a discussion of the relations between the two families of processes. Nevertheless, the principles behind the two constructions are similar and the Blum-François model offers an approximate evolutionary construction of Aldous' Beta-splitting model, with a slightly restricted range of parameters ($\beta > -1$ instead of $\beta > -2$). In fact, the range of topologies covered by the Blum-François model is quite wide as well, since '$\beta = -1$' corresponds to the totally unbalanced trees while '$\beta = \infty$' corresponds to highly balanced trees.

In [SV16], we extend the Blum-François model by allowing asymmetric Beta-distributions for the split distribution. That is, the fraction of 'speciation potential' allocated to the first offspring species is now distributed according to a Beta$(\alpha + 1, \beta + 1)$ distribution, for some $\alpha > -1$ and $\beta > -1$. Of course this lack of symmetry makes sense only if we distinguish a first and second (or later 'left' and 'right') offspring species. That is, instead of cladograms we now work with planar trees. In order to include some information on relative speciation times without keeping track of the full set of speciation times, we also rank the split events in the trees. Because various tree shape statistics are functions of the unranked and/or non-planar *lumping* of such trees, we consider four types of (rooted binary) trees:

— **Ranked planar trees:** In this case, we distinguish the left and right child nodes of an internal node, and every internal node is labelled by an integer keeping track of the ordering in which the splits occur during the construction of the tree. Since a binary tree with n leaves has $n-1$ internal nodes, the labels run from 1 (the root) to $n-1$ (the last split).

- **Unranked planar trees:** Left and right child nodes are distinguished, but the internal nodes are not labelled (so that the order of the splits is not recorded).
- **Ranked non-planar trees:** In this case, the internal nodes are ranked and labelled according to the splitting order, but left and right child nodes play equivalent roles.
- **Trees:** Unranked and non-planar trees. Aldous' cladograms are such trees whose leaves are further labelled by the n taxa.

We can give explicit expressions for the probability of any tree at the resolutions of ranked planar and unranked planar trees for any $\alpha, \beta > -1$, and for the probability of any tree at the resolutions of ranked or unranked non-planar trees for any $\alpha = \beta > -1$. For example, at the resolution of the ranked planar trees we have the following expressions. For a given ranked planar tree and an internal node labelled by i, let us write n_i^L (resp., n_i^R) for the number of internal nodes in the left (resp., right) subtree below node i. In particular, if node i subtends two leaves, then $n_i^L = 0 = n_i^R$.

Theorem 3.5. (Th. 3.1 in [SV16]). *For any ranked planar binary tree τ with n leaves, we have*

$$\begin{aligned}\mathbb{P}(\tau) &= \prod_{i=1}^{n-1}\left\{\frac{1}{B(\alpha+1,\beta+1)}\int_0^1 b_i^{n_i^L+\alpha}(1-b_i)^{n_i^R+\beta}\mathrm{d}b_i\right\}\\ &= \prod_{i=1}^{n-1}\frac{B(n_i^L+\alpha+1, n_i^R+\beta+1)}{B(\alpha+1,\beta+1)}, \qquad (3.2)\end{aligned}$$

where

$$B(\alpha,\beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x.$$

The probabilities for the other tree resolutions are obtained by computing how many ranked planar trees are lumped into the tree of interest. In the case of the unranked planar tree, it only boils down to multiplying (3.2) by a combinatorial factor for any α, β. When we consider non-planar trees, on the other hand, unless $\alpha = \beta$ there are no simplifications (due to the fact that $B(n_i^L+\alpha+1, n_i^R+\beta+1) \neq B(n_i^R+\alpha+1, n_i^L+\beta+1)$ in general) and the probability of a given non-planar tree is the sum of all probabilities of corresponding planar trees. All these probabilities can be found in Sections 3 and 4 of [SV16].

Our generalisation of the Beta-splitting model can be constructed via an evolutionary process, in which it is then easy to add death or *freezing* of lines (although computing probabilities of trees with freezing becomes tricky once the labels have been erased.) We present here the simpler construction without death. Let us fix $\alpha, \beta > -1$. Let $(B_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of Beta$(\alpha+1, \beta+1)$ random variables, and $(U_i)_{i\in\mathbb{N}}$ be an independent i.i.d. sequence of uniform random variables on $[0,1]$. The form of the parameters of the Beta distribution is chosen so that its density (proportional to $x^\alpha(1-x)^\beta$) matches that of Aldous' and Blum and François's Beta-splitting models. Once these sequences are realised, we proceed incrementally from the root and for $n-1$ steps, where n is the number of leaves we want to reach (at the end of step i, the tree has $i+1$ leaves). We start with a single root node, labelled by the interval $[0,1]$.

- **Step 1:** Split the root into a left leaf labelled by $[0, b_1]$ and a right leaf labelled by $[b_1, 1]$. Change the label of the root to the integer 1.

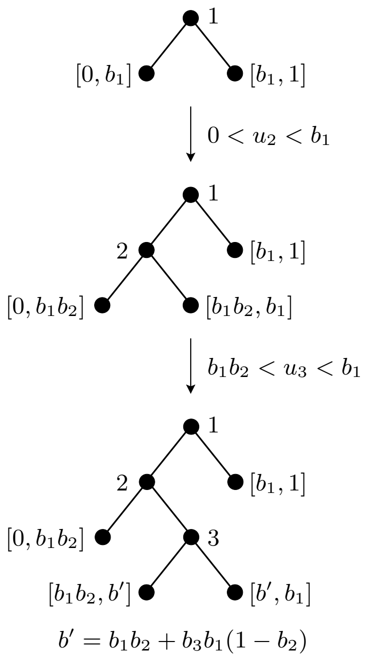


Figure 3.4 – An example of construction for $n = 4$.

- **Step 2:** If $u_2 \in [0, b_1]$, split the left child node of the root into a left leaf and a right leaf respectively labelled by $[0, b_1b_2]$ and $[b_1b_2, b_1]$. If $u_2 \in [b_1, 1]$, then instead split the right child node of the root into left and right leaves with respective labels $[b_1, b_1 + (1-b_1)b_2]$, $[b_1 + (1 - b_1)b_2, 1]$. Label the former leaf that is split during this step by 2.
- **Step i:** Find the leaf whose interval label $[a, b]$ contains u_i. Change its label to the integer i and split it into a left leaf with label $[a, a + (b - a)b_i]$ and a right leaf with label $[a + (b - a)b_i, b]$.
- Stop at the end of Step $n - 1$.

An example is provided in Figure 3.4. We see that as long as we have not erased the interval labels of the leaves, we can resume the construction and add another m leaves in the tree. Furthermore, once a realisation of the sequence $((B_i, U_i))_{i \in \mathbb{N}}$ is fixed, the construction of the tree is entirely deterministic, and so the sequence characterises the tree. To add the freezing of leaves which prevents them from being split at some later stage, we fix some freezing probability $\delta \in [0, 1)$ and we augment the vector (B_i, U_i) driving the ith step by adding two coordinates (V_i, D_i) which are both uniformly distributed on $[0, 1]$ (and independent of every other random variable). These new components are then used to decide whether the ith move is a split or a freezing, and which leave is chosen to freeze if needed.

To give a few examples of the kind of topologies we can obtain, it is not difficult to see from Theorem 3.5 that $\alpha = \beta = 0$ corresponds to the Kingman-Yule tree, $\alpha = \beta \to -1$ gives rise to comb trees, and $\alpha = \beta \to +\infty$ yields very balanced trees (see Section 3.2 in [SV16]). Because this family of random tree topologies contains a large range of tree balances and allows to give asymmetric roles to the two child nodes of a given node, while being parametrised by two real-valued parameters only, we believe that it could be a relevant statistical model to compare and classify different ancestral, evolutionary, or transmission trees without imposing mechanistic rules explaining their shapes. The code developped for this model is publicly shared at `https://cloud.sagemath.com/projects/2c5f7f68-e689-4c70-a4b4-5b5d4dc4f93f/files/2015-10-27-082849.sagews`.

3.4 Perspectives

The different models explored in the previous sections show a high dependence of the form of the pedigrees on the biology of the organisms. Of course gene genealogies are embedded in the physical ancestry of the individuals of the population, and so *a priori* the whole pedigree constrains the genetic diversity observed in the population. Yet, in Section 2.6.3 we have seen that these constraints may not be significant at the genetic level and, in many cases, statistically the genealogy/diversity at a typical locus depends only on a few summary statistics of the evolution of the population (see for example [111] and Section 3.2). However, when considering large recombining stretches of genome with linkage, or loci on sex chromosomes and autosomes, the shape of the physical ancestry is likely to have a much larger impact and, in particular, to explain much of the correlations across loci. For example, if males and females in a sexual population do not behave and reproduce in the same way during their lifetimes, we expect the correlation between the genetic diversities at two loci situated on an autosomal and a sex chromosome to have a specific form. Indeed, they are both transmitted through the same pedigree but the first locus can be inherited from both parents, while a gene on the Y-chromosome (in mammals), or mitochondrial DNA (maternally inherited in general) can come from one of the two parents only. This problematic is at the intersection between population genetics and behavioural ecology and further studies of the relationship between pedigrees and gene genealogies could shed some light on questions like why sex differences in dispersal seem to be promoted in many species of birds and mammals [29, 53], or what is the effect of monogamy, territoriality, dispersal distances, etc. on the genetic diversity of the population (work in progress with Raazesh Sainudiin).

Chapter 4

A new approach for the inference of demographic parameters

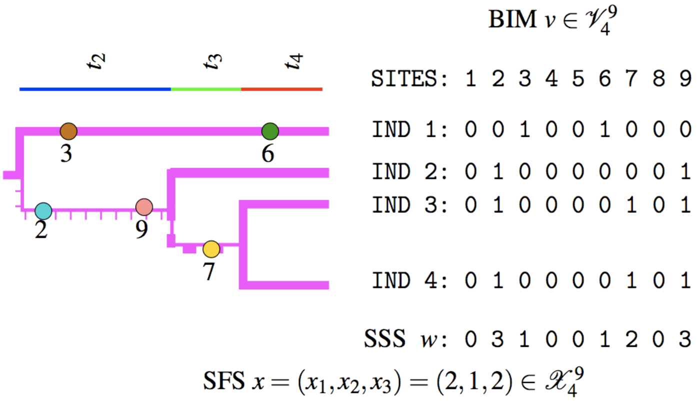
4.1 Motivations

Population genetics methods based on genealogies can be used to infer demographic parameters such as effective population sizes, rates of exponential growth or decline, bottleneck times and strengths, geographical structure, etc. Indeed, each of these phenomena affects the genealogies at all loci in the same way. Assuming that we can consider L loci as evolving independently, the allelic distributions observed at these loci in a sample of individuals can thus be seen as L realisations of the same genealogical and mutational random processes, and we can therefore hope to extract some information on the demographic parameters of the population from this data. In the same vein, we can also try to detect outlier loci, in particular those subject to natural selection. Different types of data can be used to infer the parameters of interest. Those we shall discuss further below are the *site frequency spectrum* (or SFS) and *sequence alignments*.

Let us consider a single non-recombining gene. Assuming the corresponding locus is long and the per-base mutation rate is small, we can make the approximation that every mutation falls on a different site (or base pair). This is the *infinitely many sites model* of mutation. In this case, for a sample of n individuals the most precise data we can observe is the set of n locus-specific DNA sequences. This is what we call a sequence alignment. In general only the segregating sites, where a mutation occurred, are represented, in the form of a *binary incidence matrix* of 0's (ancestral or reference bases) and 1's (derived bases). See Figure 4.1 for an example in which some non segregating sites are also shown. We may also restrict our attention to the less detailed site frequency spectrum $\mathbf{S} = (S_1, \ldots, S_{n-1})$, where

$$S_k = \# \text{ mutations carried by exactly } k \text{ individuals} \tag{4.1}$$

counts the number of segregating sites at which the mutation is carried by k individuals of the sample. In the example of Figure 4.1, we have $n = 4$ and $\mathbf{S} = (2, 1, 2)$. Note that this definition supposes that we know which base is ancestral and which is derived at every site, which may be possible if we can compare the sequences to an *outgroup*, i.e. an homologous sequence sampled in a closely related species. When this identification is not possible, we instead consider the *folded* site frequency spectrum defined for every $k \in \{1, \ldots, \lfloor n/2 \rfloor\}$ as the number of sites at



BIM $v \in \mathscr{V}_4^9$

```
SITES: 1 2 3 4 5 6 7 8 9
IND 1: 0 0 1 0 0 1 0 0 0
IND 2: 0 1 0 0 0 0 0 0 1
IND 3: 0 1 0 0 0 0 1 0 1
IND 4: 0 1 0 0 0 0 1 0 1
```

SSS w: 0 3 1 0 0 1 2 0 3

SFS $x = (x_1, x_2, x_3) = (2, 1, 2) \in \mathscr{X}_4^9$

Figure 4.1 – A genealogical tree with mutations for a sample of 4 individuals and the corresponding binary incidence matrix and site frequency spectrum. The coloured dots represent the mutations, and the associated numbers indicate the site affected by each of them. Here there are five segregating sites, namely 2, 3, 6, 7, 9.

which k individuals carry one base and $n-k$ carry another base. That is:

$$\tilde{S}_k = \frac{S_k + S_{n-k}}{1 + \delta_{\{k=n-k\}}}.$$

On top of the mutation model, we must also choose a model for the genealogical tree on which the mutations that we observe in the sample appeared at some point in the past. Recall Kingman's coalescent introduced in Section 2.1 as a continuous-time Markov process on the set of all partitions of $\{1, \dots, n\}$. The whole trajectory until it reaches the absorbing state $\{\{1, \dots, n\}\}$ can be represented more graphically as a genealogical tree topology whose leaves are labelled by $1, \dots, n$, accompanied by a time vector $(T_2, \dots, T_n)$ where for every $i \in \{2, \dots, n\}$,

$$T_i = \text{duration of the epoch during which the sample has } i \text{ ancestors.} \tag{4.2}$$

See again Figure 2.2. In the case of Kingman's coalescent, T_i is an exponential random variable with parameter $i(i-1)/2$, since we suppose that each of the $i(i-1)/2$ pairs of blocks tries to merge at rate 1. Because of its simple evolution rule, many computations can be done to describe its topology and the lengths of some portions of the tree of interest. For example, conditionally on the total length L_k of all the edges of the tree subtending k individuals in the sample, the number S_k of mutations carried by k individuals follows a Poisson distribution with parameter μL_k, where μ is the per-locus mutation rate. The average number of such mutations is thus proportional to the expectation of L_k.

Based on the intuition that the larger the population, the longer it takes to two lineages to reach a common ancestor, a natural (and now classical) generalisation of Kingman's coalescent to the case of a panmictic population with fluctuating census sizes is to suppose that t units of time in the past, each pair of lineages tries to coalesce at the instantaneous rate $1/N_t$ (see, e.g., Chapters 4.2 to 4.4 of [34] and references therein). In other words, for any sampling size n, the probability that none of the $n(n-1)/2$ initial pairs of lineages has coalesced by time t in the past ($t = 0$ corresponding to the present) is given by

$$\mathbb{E}\left[\exp\left\{-\int_0^t \frac{n(n-1)}{2N_s}\,\mathrm{d}s\right\}\right]. \tag{4.3}$$

To fully specify the model, we need a description of the backward process $(N_t)_{t\geq 0}$ of (effective) population sizes. As an example, if we want to model an exponential growth at rate g from the past towards the present, we can take

$$N_t = N_0 e^{-gt},$$

where N_0 is the current population size. To model a bottleneck occurring between a and b units of time in the past and during which the population is reduced by a factor ε, we can take

$$N_t = \begin{cases} N_0 & \text{if } t \in [0,a), \\ \varepsilon N_0 & \text{if } t \in [a,b), \\ N_0 & \text{if } t \geq b. \end{cases}$$

These are examples of parametric models, but we may also consider nonparametric models. For instance, the work presented in Section 4.3 makes the assumption that $(\ln(N_t))_{t\geq 0}$ is a Gaussian process whose correlation function needs to be estimated (see also [88]).

A large panel of methods already exist to infer demographic parameters from mutation data. To name a few:

- **The *Poisson Random Field* approach** [56, 87, 99] considers a series of independent SNPs in a sample of size n (SNP stands for 'Single Nucleotide Polymorphism', or 'segregating site' in the terminology introduced earlier). Assuming the infinitely many sites mutation model with a very low mutation rate, the distribution of the number of mutations carried by $k \in \{1,\dots,n-1\}$ individuals is approximated by a Poisson distribution. In general the parameter of the Poisson distribution corresponding to each k must be estimated by means of simulations, but we can then derive an approximate likelihood expression for $(N_t)_{t\geq 0}$.
- Methods based on the **probability generating functions of the branch lengths in the genealogy** are available, see for example [19]. These methods are untractable for $n \geq 5$ (at best).
- **Skyline plots** form another family of inference methods for demographic history, as reviewed in [60]. These nonparametric methods rest on the assumption that there is not much variability in the data-compatible reconstructed tree on which the estimation of the local harmonic means of effective population size is based (c.f., [92]). However, this will typically not be the case when the per-locus mutation rate is low and the data contains only a few mutations.
- Methods based on the **Sequential Markov Coalescent** [77, 81] have been set up to relax the strong assumption of independent loci, and to exploit the information provided by recombination and partial linkage. This tree-valued random process is a Markovian approximation for the sequence of genealogical trees relating a sample of n individuals at each locus, as we go along the genome. More precisely, we assume that the tree at each locus follows the law of Kingman's coalescent, but as we scan the genome, we encounter breakpoints at which the genealogy of the sample changes due to a recombination event. The true sequence of trees that we obtain in this way is not Markovian because the lineage which splits off from the current genealogy during a recombination event could correspond to an ancestor present in a genealogy seen previously. For a few example of such studies, see [58, 88, 103].
- To overcome the difficulty of computing analytical (or even approximate) likelihoods in potentially complex population models, a simple simulation-intensive approach known as **Approximate Bayesian Computation** or ABC [10] is now routinely used in a

wealth of studies. These methods consist in simulating a large number of trees with mutations for each parameter in a given subset of the parameter space, and to compute approximate likelihoods by keeping only the realisations which are close, in some appropriate sense, to the data. See for instance [16, 89] for recent works in this direction.

The approach we propose in this section is not fundamentally different from those reviewed above. It originated from the observation made by Raazesh Sainudiin that many summary statistics of mutation data do not actually depend on the full description of Kingman's coalescent. A caricature of this point is the following. Suppose we want to reconstruct the past fluctuations in population size based on the number of segregating sites (of course noone would do that but for very simple models, since the likelihood surface is likely to be very flat). In the infinitely many sites model, this statistics corresponds to the number of mutations which sit on the genealogical tree of the sample. Let μ be the per-locus mutation rate, n be the sample size and recall the notation T_k for the amount of time during which the sample has exactly k ancestors. Let also $(|\pi_t|)_{t\geq 0}$ be the process counting the number of blocks in Kingman's coalescent starting at $\{\{1\},\dots,\{n\}\}$. Then the total length of the tree is

$$L^{(n)} = \sum_{k=2}^{n} kT_k = \int_0^{T_n+\cdots+T_2} |\pi_t|\mathrm{d}t.$$

Conditionally on $L^{(n)}$, the number of segregating sites has a Poisson distribution with parameter $\mu L^{(n)}$. As a consequence, we see that recording the labels of the individuals of the sample and who merges with whom during each coalescence event is useless, and the law of the data depends only on the process $(|\pi_t|)_{t\geq 0}$ counting the number of blocks in the coalescent. In particular, the state space of this Markovian *lumping* of Kingman's coalescent is much smaller than the number of partitions of $\{1,\dots,n\}$, even for relatively small sample sizes. As we shall see below the same holds true for data made of site frequency spectra, or even sequence alignments, except that the optimal *resolution* (or lumping) of Kingman's coalescent is obtained by erasing less information than in our caricatural example. Although this observation does not look very deep, it does not seem to be used in practice and could drastically enhance the exploration of the set of possible topologies and the computation of (approximate) likelihoods.

To illustrate the last point, suppose we have a parametric family of models, indexed by some parameter space Θ. In general, computing the probability of the data directly from the law $\mathbb{P}_\theta$ is not feasible. However, the law of the data knowing the hidden genealogical tree is much easier to obtain (under the infinitely many sites model in particular) and in many examples the law of the tree itself under $\mathbb{P}_\theta$ is accessible. Thus, writing D for the data and $(C,\mathbf{T})$ for the discrete topology and the vector of epoch times (4.2) of the genealogical tree, the likelihood of $\theta\in\Theta$ is computed as the integral

$$\mathbb{P}_\theta(D) = \sum_{c\in\mathcal{C}_n} \int_{\mathbb{R}_+^{n-1}} \mathbb{P}_\theta(D\,|\,(c,\mathbf{t}))\mathbb{P}_\theta(C=c,\mathbf{T}\in\mathrm{d}\mathbf{t}). \tag{4.4}$$

As discussed in Section 4.2, already for $n=10$ the state space $\mathcal{C}_n$ of all possible topologies with n labelled leaves is huge, and the computation of the sum in (4.4) needs to be done through an exploration of $\mathcal{C}_n$ by Monte Carlo or importance sampling methods. Replacing the sum over $\mathcal{C}_n$ by a sum over a much smaller state space would enable us to either exhaustively explore this new space and compute the true likelihood, or at least improve the exploration of the set of all possible topologies to obtain an approximate likelihood at a much lower computational cost.

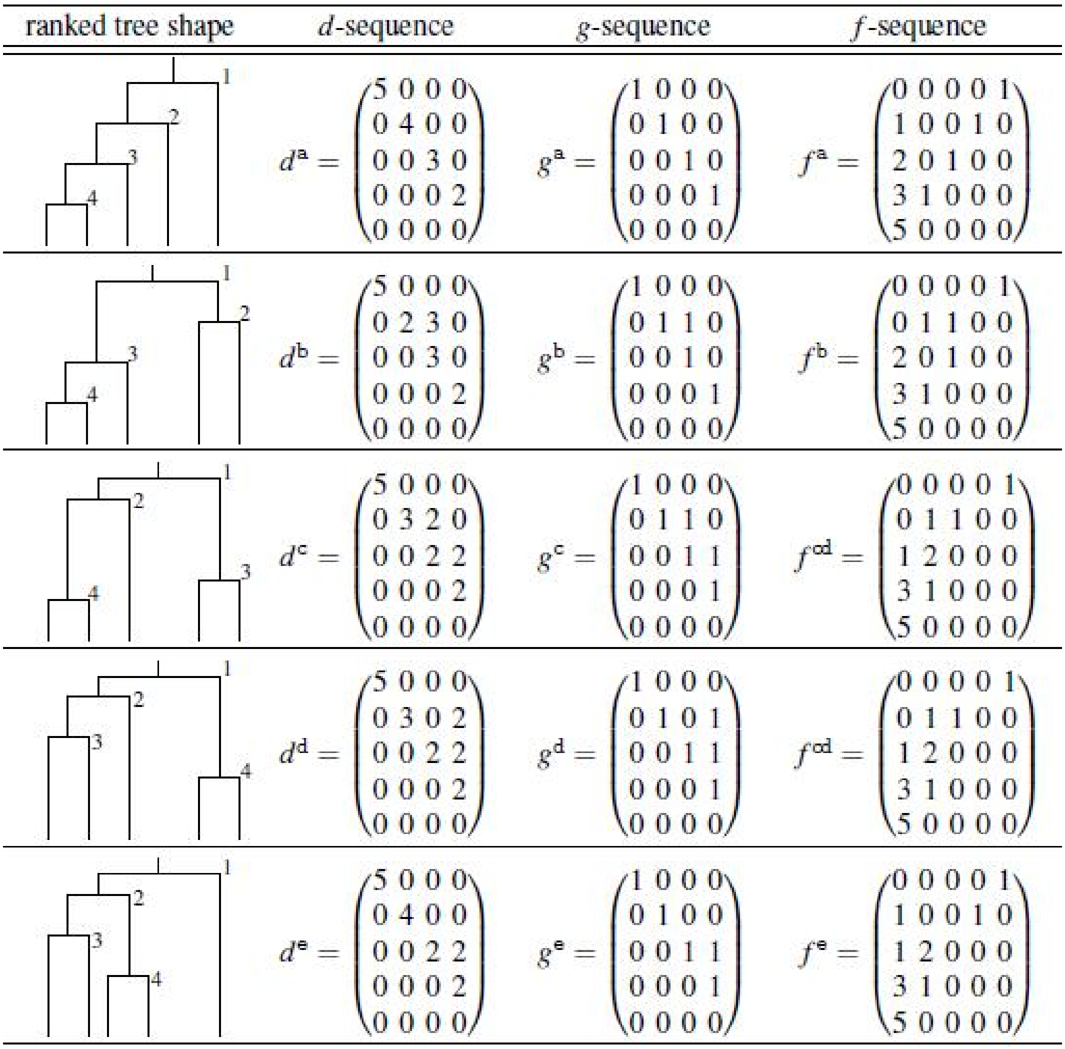
In Section 4.2, we list the different Markovian resolutions of Kingman's coalescent and give a few examples of types of data for which coarser resolutions than the standard Kingman labelled coalescent are optimal. In Section 4.3, we outline the results of a work currently in progress in which we use Tajima's *ranked tree shapes* instead of Kingman's labelled topologies to reconstruct $(N_t)_{t\geq 0}$ in a nonparametric way. In Section 4.4, we present the construction of an importance sampler which produces only *unlabelled and sized coalescents* which are compatible with a given site frequency spectrum. In this work, the law of the tree shape is given by Aldous' Beta-splitting model (see Section 3.3). This enables us to test whether the tree underlying the site frequency spectrum departs significantly from Kingman's coalescent, which corresponds to the case where the balance parameter β is equal to 0.

4.2 How to make good resolutions

The results presented in this section correspond to the publication [SSV15]. They constitute a preliminary step in the program consisting in developping inference tools based on appropriate tree resolutions. In fact, Kingman's coalescent is only an example of random models of genealogies. Because this model is used quite extensively, we found it useful to provide the transition probabilities of all Markovian resolutions to have them gathered in one place. We expect that these derivations may be extended to more general models, as in Section 3.3.

Recall that what we call a *resolution* is a process which is a function of the coalescent. In what follows, we focus on the tree topologies without times, corresponding to the embedded chains of the six continuous-time resolutions of Kingman's coalescent which are Markovian. These discrete-time resolutions are:

- The **vintaged and labelled coalescent**: it is Kingman's coalescent, except that to each block of the partition is associated a number called its *vintage*. This number records the epoch in which the block was created by the merger of two blocks. That is, a block has vintage $k \in \{2, \ldots, n\}$ if it was created at the end of the epoch during which the coalescent has k blocks. The state space $\mathbb{B}_n$ of this resolution is an augmentation of the set $\mathbb{C}_n$ of partitions of $\{1, \ldots, n\}$ with vintage tags. But since there is one and only one way to label the blocks of a $\mathbb{C}_n$-valued coalescent, there is a one-to-one correspondence between the full $\mathbb{B}_n$- and $\mathbb{C}_n$-valued sequences. The interesting feature of the vintaged and labelled coalescent is that, at any stage, we know in which order the blocks which constitute the current state of the process were created.
- The **unvintaged and labelled coalescent**, corresponding to Kingman's (discrete) coalescent. It takes its values in $\mathbb{C}_n$.
- The **vintaged and sized coalescent**, obtained from the vintaged and labelled coalescent by keeping track only of the vintage and the size of each block of the partition, and dropping the integer label $1, \ldots, n$. Its state space $\mathbb{D}_n$ is the space of all ordered integer partitions.
- The **vintaged and shaped coalescent**, obtained from the vintaged and sized coalescent by keeping track only of the vintages of the blocks present at each time step, and throwing away the sizes of these blocks. In other words, this coalescent records only the presence (1) or absence (0) of a block with vintage k in each epoch and for every $k \in \{2, \ldots, n\}$. Its state space $\mathbb{G}_n$ is a subset of $\{0, 1\}^{n-1}$. The sequence of states visited by this process gives Tajima's evolutionary relationships (see [105], Figures 1-4), which are ranked, rooted binary tree shapes. Since by knowing the states through which this process goes until the k-th transition enables us to reconstruct the size of each block in



ranked tree shape	d-sequence	g-sequence	f-sequence
	$d^{a} = \begin{pmatrix} 5&0&0&0 \\ 0&4&0&0 \\ 0&0&3&0 \\ 0&0&0&2 \\ 0&0&0&0 \end{pmatrix}$	$g^{a} = \begin{pmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \\ 0&0&0&0 \end{pmatrix}$	$f^{a} = \begin{pmatrix} 0&0&0&0&1 \\ 1&0&0&1&0 \\ 2&0&1&0&0 \\ 3&1&0&0&0 \\ 5&0&0&0&0 \end{pmatrix}$
	$d^{b} = \begin{pmatrix} 5&0&0&0 \\ 0&2&3&0 \\ 0&0&3&0 \\ 0&0&0&2 \\ 0&0&0&0 \end{pmatrix}$	$g^{b} = \begin{pmatrix} 1&0&0&0 \\ 0&1&1&0 \\ 0&0&1&0 \\ 0&0&0&1 \\ 0&0&0&0 \end{pmatrix}$	$f^{b} = \begin{pmatrix} 0&0&0&0&1 \\ 0&1&1&0&0 \\ 2&0&1&0&0 \\ 3&1&0&0&0 \\ 5&0&0&0&0 \end{pmatrix}$
	$d^{c} = \begin{pmatrix} 5&0&0&0 \\ 0&3&2&0 \\ 0&0&2&2 \\ 0&0&0&2 \\ 0&0&0&0 \end{pmatrix}$	$g^{c} = \begin{pmatrix} 1&0&0&0 \\ 0&1&1&0 \\ 0&0&1&1 \\ 0&0&0&1 \\ 0&0&0&0 \end{pmatrix}$	$f^{cd} = \begin{pmatrix} 0&0&0&0&1 \\ 0&1&1&0&0 \\ 1&2&0&0&0 \\ 3&1&0&0&0 \\ 5&0&0&0&0 \end{pmatrix}$
	$d^{d} = \begin{pmatrix} 5&0&0&0 \\ 0&3&0&2 \\ 0&0&2&2 \\ 0&0&0&2 \\ 0&0&0&0 \end{pmatrix}$	$g^{d} = \begin{pmatrix} 1&0&0&0 \\ 0&1&0&1 \\ 0&0&1&1 \\ 0&0&0&1 \\ 0&0&0&0 \end{pmatrix}$	$f^{cd} = \begin{pmatrix} 0&0&0&0&1 \\ 0&1&1&0&0 \\ 1&2&0&0&0 \\ 3&1&0&0&0 \\ 5&0&0&0&0 \end{pmatrix}$
	$d^{e} = \begin{pmatrix} 5&0&0&0 \\ 0&4&0&0 \\ 0&0&2&2 \\ 0&0&0&2 \\ 0&0&0&0 \end{pmatrix}$	$g^{e} = \begin{pmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&1 \\ 0&0&0&1 \\ 0&0&0&0 \end{pmatrix}$	$f^{e} = \begin{pmatrix} 0&0&0&0&1 \\ 1&0&0&1&0 \\ 1&2&0&0&0 \\ 3&1&0&0&0 \\ 5&0&0&0&0 \end{pmatrix}$

Figure 4.2 – The different ranked tree shapes and the corresponding lower resolutions for a sample of size 5. The superscripts $a, \dots, e$ are simply meant to give a name to each example.

the k-th state of the Markov chain, there is a one-to-one correspondence between the full $\mathbb{D}_n$- and $\mathbb{G}_n$-valued sequences.

- The **unvintaged and sized coalescent** is obtained from the unvintaged and labelled coalescent by keeping track only of how many blocks there are of each size. Its state space $\mathbb{F}_n$ is the set of all integer partitions of n.
- The **block-counting process**, which simply records the number of blocks in the coalescent. The discrete block counting process of Kingman's coalescent is always $(n, n-1, \dots, 1)$ (its continuous-time version is slightly more exciting).

See Figure 4.2 for an example of different resolutions with $n = 5$. Table 4.1 compares the cardinalities of the state spaces of the most interesting resolutions. We see that the reduction in size from the state space of Kingman's coalescent to that of lower resolutions is significant.

Computing the law of each of these resolutions of Kingman's coalescent is rather straightforward and we refer to [SSV15] for the corresponding probabilities. Let us now give a few examples of statistics of the tree shape or of the observed mutation pattern which depend on

n	4	10	30	60	90	120
$\|\mathbb{C}_n\|$	15	1.2×10^5	8.5×10^{23}	9.8×10^{59}	1.4×10^{101}	5.1×10^{145}
$\|\mathbb{G}_n\|$	5	88	1.3×10^6	2.5×10^{12}	4.7×10^{18}	8.7×10^{24}
$\|\mathbb{F}_n\|$	5	42	5.6×10^3	9.7×10^5	5.7×10^7	1.8×10^9

Table 4.1 – Cardinalities of the state spaces $\mathbb{C}_n$, $\mathbb{G}_n$ and $\mathbb{F}_n$.

particular coalescent resolutions.

Mutation statistics

We have already seen that in the infinitely many sites mutation model, the total number of segregating sites depends only of the block-counting process.

In Section 4.4, we shall see that the law of the site frequency spectrum (4.1) depends only on the unvintaged and sized coalescent resolution. Indeed the mutations carried by k individuals in the sample are those which fall on the branches of the tree subtending k leaves, or equivalently those hitting a block of size k in the partition. For every $k \in \{1, \dots, n-1\}$, the total length L_k of all these edges can be computed from the knowledge of the vector of epoch times and the number of blocks of size k in each epoch.

Finally, in Section 4.3 we shall see that the optimal resolution to characterise the law of the pattern of mutations seen in a sequence alignment (also encoded by the binary incidence matrix, or BIM) is that of the vintaged and sized coalescent. Indeed, the additional information that the BIM contains compared to the site frequency spectrum is which mutations are carried by the same individuals. Thus, to compute the probability of the data knowing the genealogical tree, we need to be able to say that a particular block/ancestor (whose lineage may have experienced a mutation event already) is the one that takes part to the next merger to create a larger block (on which another mutation may appear). This is the role of the vintages. Observe that singletons do not have vintages as they have not been created by a coalescence event. But these blocks are interchangeable, since there is no history before their creation.

Tree shape statistics

In a slightly different perspective, we may want to understand the properties a given ancestral or evolutionary tree reconstructed from the observed mutation pattern. This question is more common in phylogenetics, in which the precise realisation of the tree matters (we want to know the evolutionary relationships between a given set of species). In population genetics, in general we do not care about the idiosyncratic genealogy of a sample of individuals taken at random in the population, what matters is the *law* of this tree which tells us something about the way in which the population evolves.

Many statistics of a tree shape have been introduced. In Sections 4.2 and 4.3 of [SSV15], we argue that the most famous ones are fully described by a coarser resolution than the classical unvintaged and labelled coalescent. For example, *Sackin's index* [96] is the sum of the number of leaves subtended by each internal node. But the number of leaves subtended by an internal node is simply the size of the corresponding block, and so Sackin's index can be computed from the unvintaged and sized coalescent. *Colless' index* [23] is defined as the sum of the absolute values of the differences between the number of leaves subtended by the two branches bifurcating from each internal node (up to a constant factor). This statistics also depends only on block sizes, scanned in any order, and so on the unvintaged and sized coalescent resolution. See [SSV15] for more examples.

4.3 BESTT: an inference methodology based on Tajima's trees

This section describes the work in progress [PVWR17]. In this project, we use the vintaged and sized coalescent to reconstruct the variations in effective population size $(N_t)_{t\geq 0}$ under the assumption that Kingman's coalescent with fluctuating population size is an appropriate model for the genealogies.

In [88], the authors already use this resolution to infer $(N_t)_{t\geq 0}$ from the knowledge of the sequence of local genealogies of a sample that we encounter as we go along the genome. They assume that this sequence of trees (including the epoch times) is observed and moreover that it can be modelled by a sequential Markov coalescent. As they consider long stretches of DNA, the state space of all sequences of labelled trees is huge and they argue that considering this optimal resolution drastically reduces the computational cost. Then, they assume as a prior distribution that $(\ln(N_t))_{t\geq 0}$ follows a Gaussian process, and they compute the posterior distribution of the process thanks to an MCMC sampling algorithm on the state space of the suitably discretized process (joint with the space of parameters of interest).

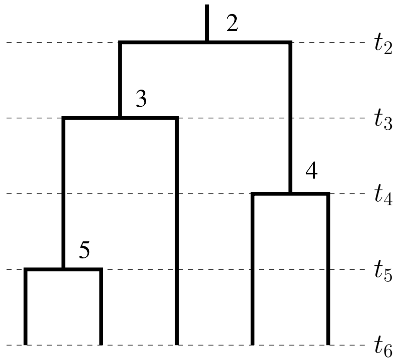
Observe that up to now we have not mentioned mutations. The implicit assumption in the previous paragraph is that the pattern of mutations observed in the sample enabled us to propose a sequence of underlying hidden trees on which to perform the techniques expounded above. In [PVWR17], we make a second step towards the inference of $(N_t)_{t\geq 0}$ based on whole sequence polymorphism data, which we present here. In what follows, we consider a stretch of DNA which has not recombined between the time of the most recent common ancestor to the sample at this locus and the present. A binary incidence matrix describes the mutations observed in the sample, and again we assume the infinitely many sites model of mutation. Our aim is to compute the probability a given BIM knowing the trajectory $(N_t)_{t\geq 0}$. This probability is at the basis of the calculation of the likelihood of a population size trajectory, again under the assumption that it is log-Gaussian.

As in [88], we encode a vintaged and sized (discrete) coalescent by an $n \times n$ matrix D. To do so, let us call t_i the time at which the number of blocks or ancestors decreases from i to $i-1$. Recalling the epoch times T_i introduced in (4.2), we thus have $t_i = T_n + T_{n-1} + \ldots + T_i$ for every $i \in \{2, \ldots, n\}$. By convention, we set $t_{n+1} = 0$. Now for every $2 \leq j \leq i \leq n$, let us define $D(i,j)$ by

$$D(i,j) = \#\ \text{lineages/blocks which do not coalesce in } [t_{i+1}, t_j). \tag{4.5}$$

The first row is filled with zeros just for completion. Proposition 2 in [88] states that there is a one-to-one correspondence between vintaged and sized coalescents (or ranked tree shapes) and D-matrices. The set of all such matrices can be characterised by a small number of conditions ensuring that a given integer-valued matrix indeed encodes a tree. We do not detail these conditions here, see instead Figure 4.3 for an example. Our approach rests on the likelihood decomposition (4.4), in which we replace the sum over $\mathcal{C}_n$ by a sum over the much smaller space $\mathcal{D}_n$ of vintaged and sized coalescents. Of course we need to compute each conditional probability at this coarser resolution. The 'probability' of the tree topology and epoch times knowing $(N_t)_{t\geq 0}$ can be written as the product of the probability of the topology, given in [SSV15], and of the density of the epoch times, which can be obtained from p.g.f.'s of the form (4.3). The last step is to compute the probability of the observed BIM knowing the underlying tree topology and time vector.

To compute the latter probability in the infinitely many sites model, we can use the tree structure to proceed incrementally. We sort the observed mutations into groups of mutations carried by exactly the same individuals. We then have to place them on the tree and to



$$\begin{bmatrix} 0 & & & & \\ 0 & 2 & & & \\ 0 & 1 & 3 & & \\ 0 & 0 & 2 & 4 & \\ 0 & 0 & 1 & 3 & 5 \end{bmatrix}$$

Figure 4.3 – A ranked tree shape with five leaves and the corresponding matrix D counting the number of lineages which do not coalesce in the time interval $[t_{i+1}, t_j)$.

compute the probability of observing such an arrangement of mutations (using the Poissonian structure of the cloud of mutations). To this end, we first consider the mutations carried by the largest number of individuals in the sample. Let us write k for this number. For any allocation of edges in the tree (or vintages) subtending k leaves to each of these groups of mutations, we can then compute the probability of seeing these mutations on these edges, no mutations on edges in the tree subtending more leaves, and call recursively our likelihood function on the subtrees pending from the k-edges, using the sub-pattern of other mutations carried by a subset of their leaves. This enables us to calculate the likelihood of a BIM given the vintaged and sized coalescent with epoch times.

Now the sum in (4.4) is still over a very large set, and so we devise an MCMC method to explore the space of D-matrices in an efficient way. All these techniques can then be extended to heterochronous coalescents, in which all individuals are not sampled at the same time. This should enable us to add ancient DNA polymorphism to present-time samples to improve the accuracy of the reconstruction. The last step will then consist in merging the techniques of [88] and [PVWR17] to infer population size trajectories based on whole sequence polymorphism data.

4.4 An importance sampler of trees compatible with a site frequency spectrum

Another way of improving the exploration of the space of possible tree topologies to compute the likelihood of a given population trajectory is to consider only tree topologies which are compatible with the data. Indeed, for a given pattern of mutations that we observe, we can find many trees that cannot explain this pattern because they have no edges subtending a given number k of leaves on which we may place an observed mutation carried by k individuals. These trees contribute nothing to the decomposition (4.4) of the likelihood, and therefore it would be (much) more efficient to restrict our attention to data-compatible trees. This is the basis of the work in progress [SV17] which we describe in this section. Here we shall work with site frequency spectrum instead of BIM data.

In [97], the authors set up a controlled Markov chain method to produce unvintaged and sized Kingman's coalescents which are always compatible with a given site frequency spectrum. The idea is that each positive coordinate of the SFS yields a constraint on the topology of the tree, since there needs to be a block of size k somewhere in the path from $\{\{1\}, \ldots, \{n\}\}$ to $\{\{1, \ldots, n\}\}$ to explain a mutation carried by k individuals in the sample. As we have already mentioned, the unvintaged and sized coalescent resolution is the optimal one for this problem: knowing the epoch times T_i during which there are i blocks in the ancestral process (or i edges in the corresponding layer of the tree), and the sizes of the blocks present in each epoch, we can compute the total length L_k of the edges in the genealogy which subtend k leaves. Then if μ is the per-locus mutation rate, conditional on $(L_1, \ldots, L_{n-1})$ the coordinates of the SFS are independent Poisson random variables with parameters μL_i. The method developped in [97] is specific to Kingman's coalescent, and we extend it in several ways in [SV17].

First, we consider more general a priori distributions on tree topologies. More precisely, we use Aldous' Beta-splitting model with parameter $\beta > -2$, of which Kingman's topology is a particular case ($\beta = 0$, see Section 3.3 for the definition of this model). Second, as a prior distribution on the vector of epoch times, we use a vector of $n-1$ independent exponential distributions whose parameters $A(2), \ldots, A(n)$ are the inverse of the average epoch times in the model for which we want to do likelihood calculations (independently of the observed data, which means that these times can either be obtained analytically or by a quick round of simulations). Our aim is to inform the sampler with a minimal number of parameters describing the general shape and length of the tree, and see how this information is turned into a posterior distribution once we add the constraints given by the SFS.

Thus, the input of our sampler consists of a vector (A, β, n, S, μ), where A is the vector of *a priori* rates for the $n-1$ exponentially distributed epoch times, β is the parameter of Aldous' Beta-splitting model, n is the sample size, S is the observed site frequency spectrum and μ is the per-locus mutation rate. Its output is a vector (F, M, T, w_F, w_M, w_T), where:

— F is a matrix encoding an unvintaged and sized coalescent as follows. For every $i \in \{2, \ldots, n\}$ and $j \in \{1, \ldots, n-1\}$,

$$F(i,j) = \#\text{ edges in epoch } i \text{ subtending } j \text{ leaves.}$$

— M is a matrix of mutation placements on the topology. For every $i \in \{2, \ldots, n\}$ and $j \in \{1, \ldots, n-1\}$,

$$M(i,j) = \#\text{ mutations placed on one of the } F(i,j)\ j\text{-edges in epoch } i.$$

— T is a vector of $n-1$ epoch times.

— w_F, w_M and w_T are the importance weights associated respectively to the topology F, the mutation placement M and the time vector T.

The construction of a particle (F, M, T, w_F, w_M, w_T) is incremental and is based on the same idea as in Section 4.3 of using first the mutations carried by the largest numbers of individuals. Initially all the coefficients of the matrices F and M are set to 0 and the vector T is initialised by sampling each coordinate from an exponential distribution with parameter $A(i)$, $2 \leq i \leq n$. We start with the mutations carried by $n-1$ individuals. If $S_{n-1} > 0$, such mutations are observed indeed and this imposes that the first split in the tree (seen from the root to the leaves) should separate the sample into a subsample of size $n-1$ and a subsample of size 1, so that the edge in epoch 2 which subtends $n-1$ leaves may carry the S_{n-1} mutations. If $S_{n-1} = 0$, we have no constraints on the first split and we can use the Beta-splitting model to decide at random whether we create an $(n-1)$-edge in epoch 2 or not. If it is created then $F(2, n-1) \leftarrow 1$ and $F(2,1) \leftarrow 1$, if not we do not update the F matrix. Since epoch 2 is the only layer in the tree where we can see an $(n-1)$-edge, the updating of F stops here. We update M by placing the S_{n-1} mutations carried by $n-1$ individuals on the $(n-1)$-edge (which necessarily exists if $S_{n-1} > 0$) by setting $M(2, n-1) \leftarrow S_{n-1}$. Finally, if $S_{n-1} > 0$, we resample T_2 according to the distribution of an exponential r.v. with parameter $A(2)$ conditioned on $\text{Poisson}(\mu T_2) = S_{n-1}$, which turns T_2 into a $\text{Gamma}(1 + S_{n-1}, A(2) + \mu)$ r.v. All these moves have probabilities (or densities for T) that we use to update the importance weights too.

Let us just do another step and consider the mutations carried by $n-2$ individuals in the sample. Again, we proceed from the top of the tree, epoch 2, down to the bottom (knowing that an $(n-2)$-edge can be found only in epochs 2 and 3). If an $(n-1)$-edge was created in the previous step, then we go directly to epoch 3 and force the presence there of an $(n-2)$-edge if $S_{n-2} > 0$, or use the Beta-splitting distribution to decide whether such an edge is created or not if $S_{n-2} = 0$. If it is created, then $F(3, n-2) \leftarrow 1$, $F(3, 1) \leftarrow 2$ (recall that the first split already created a 1-edge). If $F(2, n-1) = 0$, then we decide or not to create an $(n-2)$-edge in epoch 2 using the same rules. If it is created, we choose at random with a probability depending on the block sizes in epoch 2 whether this edge still exists in epoch 3 or is split into two edges subtending less individuals. In the first case, F is updated by setting $F(3, n-2) \leftarrow 1$ and necessarily $F(3, 1) \leftarrow 2$ (the 2-edge in epoch 2 is split into 2 1-edges). In the second case, we only update $F(3, 1) \leftarrow 1$ since we do not know the sizes of the edges created by the split yet. Epoch 4 cannot have $(n-2)$ edges and so the updating of F stops there. We then distribute the S_{n-2} mutations carried by $n-2$ individuals by using the fact that conditionally on their number, the mutations are independently and uniformly distributed over the total length $F(2, n-2)T_2 + F(3, n-2)T_3$ of $(n-2)$-edges in the tree. Consequently, the number of these mutations which are placed in epoch 2 follow a Binomial distribution with parameters S_{n-2} and $F(2, n-2)T_2/(F(2, n-2)T_2 + F(3, n-2)T_3)$. Finally, for each epoch $i \in \{2, 3\}$ in which we could have placed some mutations (i.e., such that $F(i, n-2) > 0$), we update T_i by sampling a new time according to its distribution $\text{Gamma}(1 + M(i, n-1), A(i) + \mu F(i, n-1)$ in the previous step conditioned on $\text{Poisson}(\mu F(i, n-2)T_i) = M(i, n-2)$. This turns the distribution of T_i into a Gamma distribution with parameters $1 + M(i, n-1) + M(i, n-2)$ and $A(i) + \mu(F(i, n-1) + F(i, n-2))$ (this stability of the Gamma distribution with respect to the Poissonian conditioning was one of the main motivations for the choice of this form of prior distributions). Again, the importance weights are updated at the same time.

We carry on updating F, M, T, w_F, w_M and w_T as described above, by considering the mutations carried by k individuals for $k = n-3$ down to $k = 1$. Each time, we take into account the information on these objects brought by the previous steps to build a tree topology and a vector of epoch times which are necessarily compatible with the observed site frequency spectrum. The matrix M is only here to relate the constructions of F and T, but its weight needs to be taken into account in the approximate likelihood calculations for which we want to use the sampler.

The code for the sampler and the likelihood procedure is publicly shared at
`https://cloud.sagemath.com/projects/ac7f397f-eab9-45fc-9278-f486af09ca55/files/FullLikelihoodInferenceSFS.sagews`
It is more proof of concept than made to be efficient, and still we can compute approximate likelihoods of basic scenarii for a reasonable number (~ 20) of parameter values, assuming the data is made of site frequency spectra at up to 1000 independent loci and producing about 100 particles for each locus, in a few hours. The code would need to be optimised to consider larger subsets of the parameter space.

4.5 Perspectives

The idea of using the appropriate resolution of the genealogical trees in simulations and likelihood computations deserves to be developped into practical tools, at least for its computational efficiency. In particular, the sampler of Section 4.4 considers only one locus and the associated likelihood computations assume that we can find sufficiently many independent loci to reconstruct the parameter values from a series of data correlated only through the population size history. It could be generalised to several loci with recombination, in order to take into account the additional information brought by linkage, in the same spirit as what we want to develop with BIM data from the results of Section 4.3. We expect this generalisation to be difficult, as a recombination point modifies the genealogies only locally and the mutation placements on the genealogies are much less constrained when we know only the site frequency spectrum. However, if we could set up such a multi-locus sampler building sequences of unvintaged and sized coalescents always compatible with a series of SFSs, the significant reduction of the size of the state space would surely compensate for the loss of information due to summarizing a sequence alignment into a set of site frequency spectra.

Chapter 5

Communication networks with logarithmic weights

The final chapter of this thesis deals with the modelling of a network of interacting queues, in which the sharing of common resources (servers) follows a particular policy with logarithmic weights. Indeed, it is a desirable property of such systems that the fraction of the servers' capacity received by a queue should be an increasing function of the number of pending requests in this queue. In this way, larger queues are served more often and this discipline should regulate the global number of requests in the network.

To illustrate why logarithmic policies may be of interest, let us suppose that J queues share a single server. Let us write $L_i(t)$ for the size of the ith queue at time t. We assume that new requests arrive in queue i at rate $\lambda_i > 0$ and that the instantaneous rate at time t at which queue i is served is given by

$$\mu_i \, \frac{f(L_i(t))}{\sum_{j=1}^{J} f(L_j(t))}, \tag{5.1}$$

where $\mu_i > 0$ and $f : \mathbb{N} \to (0, \infty)$ is an increasing function. Arrivals and service are supposed to be independent between the queues. Such processes can be obtained as limits of rather classical models of wireless networks, as the size of the quantum of information transmitted during one service time and the amount of time between two transmissions from the same non-empty queue both tend to zero. See [1, 83, 107] for historical references. The first examples which come to mind are $f(l) = l$, or more generally $f(l) = l^{\alpha}$ for some $\alpha > 0$. To derive some stability properties of the system when the global number of pending requests is large, a standard approach is to look for a *fluid limit*. That is, we suppose that each queue starts with an initial number of jobs of the form Nl_i, where $l_i \geq 0$, and that the parameters λ_i^N and μ_i^N are of the form $N\tilde{\lambda}_i$ and $N\tilde{\mu}_i$ respectively. This corresponds to looking at a regime of high turnover, which in the light of the different scalings done in the previous chapters could also be seen as looking at the behaviour of the system with parameters $\tilde{\lambda}_i$, $\tilde{\mu}_i$, but on the timescale $(Nt, t \geq 0)$. To describe the queues by quantities which remain of order 1 as N tends to infinity, for every $i \in \{1, \ldots, J\}$ we consider the evolution of $(L_i^N(t)/N)_{t \geq 0}$. If we take $f(l) = l^\alpha$ with $\alpha > 0$ in (5.1), we obtain that for every $i \in \{1, \dots, J\}$,

$$\begin{aligned}\bar{L}_i^N(t) := \frac{L_i^N(t)}{N} &= l_i + \frac{N\tilde{\lambda}_i t}{N} - \frac{N\tilde{\mu}_i}{N}\int_0^t \frac{(L_i^N(s))^\alpha}{\sum_{j=1}^J (L_j^N(s))^\alpha}\,\mathrm{d}s + \frac{M_i^N(t)}{N}\\ &= l_i + \tilde{\lambda}_i t - \tilde{\mu}_i \int_0^t \frac{(\bar{L}_i^N(s))^\alpha}{\sum_{j=1}^J (\bar{L}_j^N(s))^\alpha}\,\mathrm{d}s + \frac{M_i^N(t)}{N},\end{aligned} \tag{5.2}$$

where M_i^N is a martingale with a quadratic variation

$$\langle M_i^N\rangle(t) = N\tilde{\lambda}_i t + N\tilde{\mu}_i \int_0^t \frac{(L_i^N(s))^\alpha}{\sum_{j=1}^J (L_j^N(s))^\alpha}\,\mathrm{d}s.$$

Hence, as N tends to infinity, $M_i^N(t)/N$ tends to 0 for any $i \in \{1, \dots, J\}$ and any $t \geq 0$, and $((\bar{L}_1^N(t), \dots, \bar{L}_J^N(t)))_{t\geq 0}$ converges in distribution to the solution to the system of ordinary differential equations

$$\frac{\mathrm{d}\ell_i(t)}{\mathrm{d}t} = \tilde{\lambda}_i - \tilde{\mu}_i \frac{(\ell_i(s))^\alpha}{\sum_{j=1}^J (\ell_j(s))^\alpha}, \qquad \ell_i(0) = l_i, \qquad 1 \leq i \leq J. \tag{5.3}$$

Summing these equations over $i \in \{1, \dots, J\}$, we obtain that the weighted sum $L(t) = (\tilde{\mu_1})^{-1}\ell_1(t) + \cdots + (\tilde{\mu}_J)^{-1}\ell_J(t)$ satisfies

$$\frac{\mathrm{d}L(t)}{\mathrm{d}t} = \sum_{i=1}^J \frac{\tilde{\lambda}_i}{\tilde{\mu}_i} - 1.$$

Thus, writing $\rho_i = \tilde{\lambda}_i/\tilde{\mu}_i$ for the *load* of queue i, a necessary and sufficient condition for the total 'number' of requests to come back to 0 is to impose that

$$\rho_1 + \cdots + \rho_J < 1. \tag{5.4}$$

This is the stability condition of the system. However, we also see from (5.3) that if queue 1 starts at $\ell_1 = 0$, corresponding to an initial state for L_1^N which is negligible compared to N, then at least for some amount of time the service term in (5.3) will be very small and the other queues will monopolise the server. Consequently, instead of being emptied rapidly as we would expect from its small size, queue 1 keeps on increasing until it reaches some macroscopic size which enables it to compete with the other queues.

One way to overcome this problem of monopoly by the largest queues is to use a function f which is still increasing in the number of requests, but much more slowly than polynomially. In all that follows we choose $f(l) = \ln(1 + l)$, but we could conduct analogous analyses with other slowly increasing functions. Related algorithms based on log policies have been considered in the context of wireless networks, see [102] and references therein. However, to our knowledge no detailed mathematical analysis had been carried out before those presented here.

Below we consider two different networks of queues. This underlying structure encodes the interference between the different users of the same servers. In Section 5.1, we consider the case of a single server which can be used by only one queue (or node of the network) at a time. In this case, all the nodes interfere since they cannot be served at the same time. In Section 5.2, we consider a star network, in which the central node interferes with all the peripheral nodes. Peripheral nodes can be served at the same time. This difference of network topologies will appear in the service rates of the queues which differ between the two models. To study the response of these systems to the appearance of a very large queue, we focus on the particular case where one queue starts at some large value N and the others start at 0. We look for the behaviour of each queue on the timescale $(Nt,\, t \geq 0)$ over which the largest queue varies macroscopically.

Observe that an approach like (5.2) does not work when $f(l) = \ln(1+l)$, since it does not lead to a system of autonomous equations in $(\ln(1+L_i))_{1\leq i\leq J}$. To obtain a fluid limit, we have to study first what happens on timescales of the form N^γ, $\gamma \in (0,1)$. Indeed, it is on this timescale that the initially empty queues start increasing until they reach some equilibrium values that will dictate the behaviour of the largest queue on the timescale $(Nt,\, t \geq 0)$. These equilibria and the fluid limit of the system depend on the network considered.

5.1 Interacting queues on the complete graph

The work presented here corresponds to the publication [RV15]. As mentioned earlier, we suppose that all queues interfere, and that queue i receives a weight proportional to $\ln(1+L_i)$ when the idle server chooses a next queue to serve. If furthermore the service time of a client in this queue follows an exponential distribution with parameter μ_i, we can model the service of clients in queue i as occurring at rate

$$\mu_i \frac{\ln(1+L_i)}{\sum_{j=1}^J \ln(1+L_j)},$$

as in (5.1). We again assume that new clients arrive in queue i at rate $\lambda_i > 0$. Finally, we also assume that $L_J(0) = N$ and $L_i(0) = 0$ for every $i \in \{1, \ldots, J-1\}$, and we add a superscript N to each L_i to recall the dependency on N. The load of queue i is defined as $\rho_i = \lambda_i/\mu_i$.

Suppose as a start that $J = 2$. That is, there are only two queues and the first one is initially empty, while the second one starts with N clients. The main result in [RV15] is that on the fluid timescale, the system (L_1^N, L_2^N) behaves as in Figure 5.1. To give a precise statement and some elements of proof, we decompose the analysis of the long term behaviour of the pair of queues into several steps. In what follows we consider large queue sizes, and so we shorten the notation by writing $\ln(l)$ instead of $\ln(1+l)$.

First phase: the timescale $t \mapsto N^t$, $t < 1$.

On this timescale, L_2^N remains of the order of $N + \mathcal{O}(N^t) \approx N$. Hence, L_1^N is well approximated by the Markov process $(X^N(t))_{t\geq 0}$ such that

$$\begin{aligned} X^N &\to X^N + 1 \quad \text{at rate } \lambda_1, \\ X^N &\to X^N - 1 \quad \text{at rate } \mu_1 \frac{\ln X^N}{\ln X^N + \ln N}. \end{aligned}$$

The infinitesimal drift of X^N is

$$\Delta X^N = \lambda_1 - \mu_1 \frac{\ln X^N}{\ln X^N + \ln N} = \mu_1 \bigg(\rho_1 - \frac{(\ln X^N)/\ln N}{1 + (\ln X^N)/\ln N} \bigg),$$

which is initially positive since X^N starts at 0 (or 1, for the service rate to make sense) and decreases until it reaches 0 when

$$\frac{(\ln X^N)/\ln N}{1 + (\ln X^N)/\ln N} \approx \rho_1 \quad \Leftrightarrow \frac{\ln X^N}{\ln N} \approx \frac{\rho_1}{1-\rho_1} =: \alpha_1^*. \tag{5.5}$$

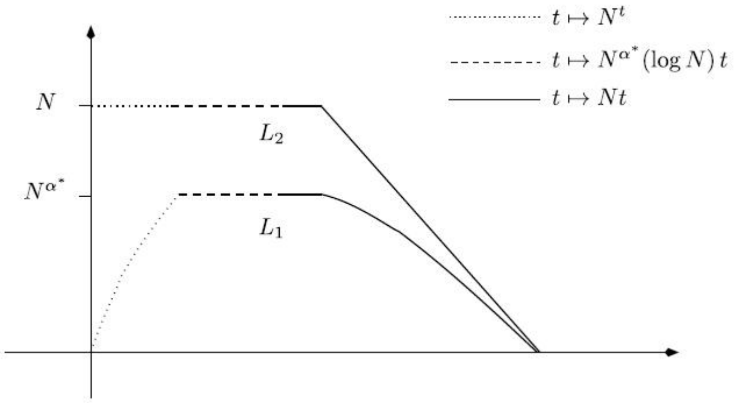


Figure 5.1 – A first order picture of the network with $\rho_1 + \rho_2 < 1$, $\rho_1 < 1/2$ and $(L_1^N(0), L_2^N(0)) = (0, N)$. During a first phase L_1^N grows proportionally to N^t, during a second phase it behaves like an Ornstein-Uhlenbeck process around the equilibrium value $N^{\alpha_1^*}$, then finally the system (L_1^N, L_2^N) converges to zero in a way described by Theorem 5.1.

In other words, the drift remains positive as long as $X^N < N^{\alpha_1^*}$ and so initially we have $X^N(N^t) \propto N^t$. Using the analogue of (5.2) in our case and changing variable in the integral corresponding to the service term, we obtain that

$$\frac{X^N(N^t)}{N^t} \approx \lambda_1 - \frac{\mu_1}{N^t}\int_0^t \frac{s\ln N}{(s+1)\ln N}(\ln N)N^s \mathrm{d}s + \frac{M^N(N^t)}{N^t}.$$

As before, the scaled martingale term vanishes in the limit as $N \to \infty$. Furthermore, changing variable again in the integral by setting $u = t - s$, we obtain that the integral is equal to

$$\mu_1 \int_0^t \frac{t-u}{1+t-u}(\ln N)e^{-u\ln N}\mathrm{d}u.$$

But $(\ln N)e^{-u\ln N}$ is the density of an exponential random variable with parameter $\ln N$, which converges in distribution towards the Dirac mass at 0 as $N \to \infty$. As a consequence, the integral converges to $\mu_1 t/(1+t)$. Combining all this, we obtain that

$$\left(\frac{X^N(N^t)}{N^t}\right)_{0<t<\alpha_1^*\wedge 1} \xrightarrow{(d)} \left(\lambda_1 - \mu_1\frac{t}{1+t}\right)_{0<t<\alpha_1^*\wedge 1},$$

and this convergence is in fact uniform over compact subintervals of $(0, \alpha_1^* \wedge 1)$. Observe that the limiting process cancels at $t = \alpha_1^*$, and so nothing tells us that X^N actually reaches the value $N^{\alpha_1^*}$. In fact it does, but in a time slightly larger than $N^{\alpha_1^*}$.

Proposition 5.1. (Prop. 3 in [RV15]). *For every $a > 0$, define H_a^N by*

$$H_a^N := \inf\{t > 0 : X^N(t) \geq a\}.$$

(a) If $\alpha_1^* < 1$, then for every $\delta \in (0,1)$ there exists $C > 0$ such that for every $N \geq 1$,

$$\mathbb{E}\left[H^N_{\delta N^{\alpha_1^*}}\right] \leq \frac{C\delta}{\ln(1/\delta)} N^{\alpha_1^*}\ln N.$$

(b) If $\alpha_1^* > 1$, for every small $\delta > 0$ (so that L_2^N does not vary too much over the time interval considered) we have

$$\limsup_{N\to\infty} \frac{1}{N}\,\mathbb{E}\big[H_{\delta N}^N\big] \leq \frac{\delta}{\lambda_1 - \mu_1/2}.$$

Second phase: the timescale $t \mapsto N^{\alpha_1^*}(\ln N)t,\ t \geq 0$.

Let us suppose that $\alpha_1^* < 1$ (or equivalently that $\rho_1 < 1/2$), and that $(L_1^N(0), L_2^N(0)) = (\delta N^{\alpha_1^*}, N)$ for some $\delta \in (0,1]$.

Using the same type of arguments as for the first phase, and writing $L_1^N(N^{\alpha_1^*}(\ln N)t) = h^N(t)N^{\alpha_1^*}$, we obtain that the infinitesimal drift of L_1^N on this timescale is given by

$$\begin{aligned}\Delta L_1^N(N^{\alpha_1^*}(\ln N)t) &= (N^{\alpha_1^*}\ln N)\bigg(\lambda_1 - \mu_1\,\frac{\ln h^N(t) + \alpha_1^*\ln N}{\ln h^N(t) + (1+\alpha_1^*)\ln N}\bigg)\\ &\approx -\frac{\mu_1}{(1+\alpha_1^*)^2}\,N^{\alpha_1^*}(\ln N)\,\frac{\ln h^N(t)}{\ln N},\end{aligned}$$

where the approximation uses the fact that $\alpha_1^*/(1+\alpha_1^*) = \rho_1$ and a Taylor expansion in $(\ln h^N(t))/(\ln N)$. Consequently, we expect that when N is large, we should have that h^N satisfies $(h^N)'(t) = -(\mu_1/(1+\alpha_1^*)^2)\ln h^N(t)$. These heuristics are good:

Proposition 5.2. (Prop. 4 in [RV15]). *If $\rho_1 < 1/2$ and if $(L_1^N(0), L_2^N(0)) = (\delta N^{\alpha_1^*}, N)$ for some $\delta \in (0,1]$, then the sequence of processes*

$$\left(\frac{L_1^N(N^{\alpha_1^*}(\ln N)t)}{N^{\alpha_1^*}}\right)_{t\geq 0}$$

converges in distribution to $(h(t))_{t\geq 0}$ defined by $h(t) \equiv 1$ if $\delta = 1$, and if $\delta \neq 1$ by

$$\int_\delta^{h(t)} \frac{\mathrm{d}u}{\ln u} = -\frac{\mu_1 t}{(1+\alpha_1^*)^2}.$$

Notice that this result is not completely intuitive, since from the Poisson process formulation we would expect the fluctuations of L_1^N to be of the order of $N^{\alpha_1^*}\ln N$ on this timescale. Note also that $h(t)$ tends to 1 as $t \to \infty$. A natural next question is to look for a central limit theorem which would give some hindsight on the behaviour L_1^N close to the equilibrium value $N^{\alpha_1^*}$. That is, we would like to obtain a sort of Central Limit Theorem describing the asymptotic behaviour of

$$\left(\frac{L_1^N(N^{\alpha_1^*}(\ln N)t) - h(t)N^{\alpha_1^*}}{\sqrt{N^{\alpha_1^*}\ln N}}\right)_{t\geq 0}.$$

However, we can show that the convergence of h^N to h is too slow and we have to replace $h(t)$ in the above by some function $\mathfrak{h}^N$ which converges to h at rate $1/\ln N$.

Proposition 5.3. (Prop. 5 in [RV15]). *Suppose that for some $\delta \in (0,1]$, we have*

$$\frac{L_1^N(0) - \delta N^{\alpha_1^*}}{\sqrt{N^{\alpha_1^*}\ln N}} \longrightarrow y \in \mathbb{R} \qquad \text{as } N \to \infty.$$

Then

$$\left(\frac{L_1^N(N^{\alpha_1^*}(\ln N)t) - \mathfrak{h}^N(t)N^{\alpha_1^*}}{\sqrt{N^{\alpha_1^*}\ln N}}\right)_{t\geq 0} \xrightarrow{(d)} (R(t))_{t\geq 0},$$

where $\mathfrak{h}^N(0) = \delta$,

$$(\mathfrak{h}^N)'(t) = -\frac{\mu_1}{1+\alpha_1^*}\,\frac{\ln \mathfrak{h}^N(t)}{\alpha_1^* + 1 + (\ln \mathfrak{h}^N(t))/(\ln N)}$$

and $(R(t))_{t\geq 0}$ *is the solution to the stochastic differential equation*

$$\mathrm{d}R(t) = \sqrt{2\lambda_1}\mathrm{d}B_t - \frac{\mu_1}{(1+\alpha_1^*)^2}\,\frac{R(t)}{h(t)}\,\mathrm{d}t, \qquad R(0) = y.$$

When $\delta = 1$, we have $h \equiv 1$, $\mathfrak{h}^N \equiv 1$ and so $(N^{\alpha_1^*}\ln N)^{-1/2}(L_1^N((N^{\alpha_1^*}\ln N)\cdot) - N^{\alpha_1^*})$ converges to an Ornstein-Uhlenbeck process.

Third phase: the fluid timescale $t \mapsto Nt$, $t \geq 0$.

Suppose now that $(L_1^N(0), L_2^N(0)) = (N^{\alpha_1^*}, N)$. Initially, the infinitesimal drift of L_1^N is

$$\begin{aligned}\Delta L_1^N(0) &= \lambda_1 - \mu_1\,\frac{\ln(N^{\alpha_1^*})}{\ln(N^{\alpha_1^*}) + \ln N} = 0\\ &= \mu_1\left(\rho_1 - \frac{\ln((L_2^N(0))^{\alpha_1^*})}{\ln((L_2^N(0))^{\alpha_1^*}) + \ln(L_2^N(0))}\right).\end{aligned}$$

Besides, when $L_1^N(Nt)$ is larger than $L_2^N(Nt)^{\alpha_1^*}$, this drift is negative while it is positive when $L_1^N(Nt) < L_2^N(Nt)^{\alpha_1^*}$. Thus, L_1^N is kept in a neighbourhood of $(L_2^N)^{\alpha_1^*}$. Indeed, the Ornstein-Uhlenbeck-type force which brings back L_1^N to its equilibrium value (together with the fact that $L_1^N \ll N$ and so it can respond in much less time than L_2^N needs to change in a macroscopic way) counteracts the effect of the Poissonian fluctuations. Assuming that $L_1^N(Nt) \approx L_2^N(Nt)^{\alpha_1^*}$ at all times, we obtain that on the fluid scale, the infinitesimal drift of L_2^N/N is

$$\Delta\frac{L_2^N}{N} = \lambda_2 - \mu_2\,\frac{\ln(L_2^N)}{\ln((L_2^N)^{\alpha_1^*}) + \ln(L_2^N)} = \lambda_2 - \mu_2(1-\rho_1) = \mu_2(\rho_2 + \rho_1 - 1).$$

This gives us the following result.

Theorem 5.1. (Th. 3 in [RV15]). *Suppose that* $(L_1^N(0), L_2^N(0)) = (0, N)$, $\rho_1 < 1/2$ *and* $\rho_1 + \rho_2 < 1$. *Then as* $N \to \infty$,

$$\left(\frac{L_1^N(Nt)}{N^{\alpha_1^*}}, \frac{L_2^N(Nt)}{N}\right)_{0<t<t^*} \xrightarrow{(d)} \left(\gamma(t)^{\alpha_1^*}, \gamma(t)\right)_{0<t<t^*},$$

where $\gamma(t) = 1 - \mu_2(1-\rho_1-\rho_2)t$ *and* $t^* = 1/(\mu_2(1-\rho_1-\rho_2))$.

In particular, we see that if we had scaled L_1^N by N as in standard fluid limit results, the limit would be 0. Thus, despite the fact that the size of queue 1 is negligible compared to that of the other queue, the expression of $\gamma(t)$ shows that node 1 receives the fraction ρ_1 of the server's capacity. This enables it to remain at equilibrium by allowing all the pending requests in queue 1 to be treated. Only the fraction $1-\rho_1$ of the server's capacity is allocated to the largest queue, which consequently does not monopolise the server.

When $J > 2$, the first $J-1$ queues start with 0 clients and the J-th one with N clients, Figure 5.2 shows the asymptotic behaviour of the system on the timescale $t \to N^t$. Indeed, if $\rho_1 < \rho_2 < \cdots < \rho_J$ and $\rho_1 + \cdots + \rho_J < 1$, until time $t_1 := \rho_1/(1-(J-1)\rho_1)$ all initially empty queues grow in proportion of N^t. Then the first queue reaches its equilibrium value N^{t_1} and remains in a neighbourhood of it until the fluid timescale. This queue captures a

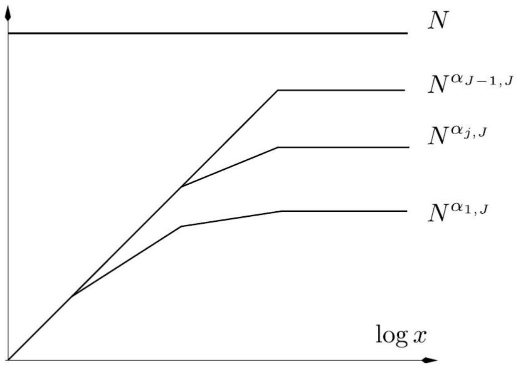


Figure 5.2 – The network with J nodes on the timescale $t \mapsto N^t$ (when all equilibria are reached before the fluid timescale).

fraction ρ_1 of the service capacity to maintain this equilibrium, and the remaining fraction $1-\rho_1$ is allocated to the other queues which carry on increasing proportionally to N^t, until time $t_2 := \rho_2/(1-\rho_1-(J-2)\rho_2)$ at which the second queue reaches its equilibrium, etc. The proof of this result is not fully given in [RV15], because it can be easily adapted from the proof of the more difficult results obtained in [RV16].

5.2 Star-shaped network of incompatibilities

In this second work, which corresponds to the preprint [RV16], we consider a star network of incompatibilities with $J+1$ nodes. This can be used for instance to model interfering communication channels in wireless networks, as depicted in Figure 5.3. We assume that the central node 0 cannot transmit at the same time as any of the peripheral nodes, which themselves can transmit at the same time. Let again L_i be the current number of pending messages at node i. In idle state, node 0 tries to transmit at rate $K\log(1+L_0)$, for some large constant K. The attempt is successful only if all the channels are free, i.e. if none of the nodes with index greater than or equal to 1 are currently transmitting at that time. When no communication is active, node 0 is therefore in competition with all the other nodes for transmission. Consequently, it succeeds at rate $K\log(1+L_0)$ or one of the other nodes starts transmitting at rate $K(\log(1+L_1)+\log(1+L_2)+\cdots+\log(1+L_J))$.

This situation will be represented as follows. Suppose the transmission times of requests at node i are exponentially distributed with rate μ_i and the state of the $J+1$ queues sitting at the nodes of the network is $L=(L_i, 0\le i\le J)$. Then in our model, any non-empty node with index greater than or equal to 1 receives the instantaneous capacity $W(L)$ to transmit and node 0 receives $1-W(L)$ (the total capacity of the channel is assumed to be 1), where

$$W(L) := \frac{\log(1+L_1)+\cdots+\log(1+L_J)}{\log(1+L_0)+\log(1+L_1)+\cdots+\log(1+L_J)}. \tag{5.6}$$

In particular, for every $i\in\{1,\dots,J\}$ (resp., $i=0$), node i completes a transmission at rate $\mu_i W(L)$ (resp. $\mu_0(1-W(L))$). This model assumes that K is sufficiently large so that the waiting times to try to access the channel are negligible.

As before, requests arrive at node $i\in\{0,\dots,J\}$ at rate λ_i, we write ρ_i for the load λ_i/μ_i of queue i and we define

$$\alpha_i^* = \frac{\rho_i}{1-\rho_i}.$$

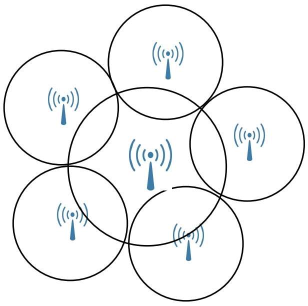


Figure 5.3 – Star Network with $J+1$=6.

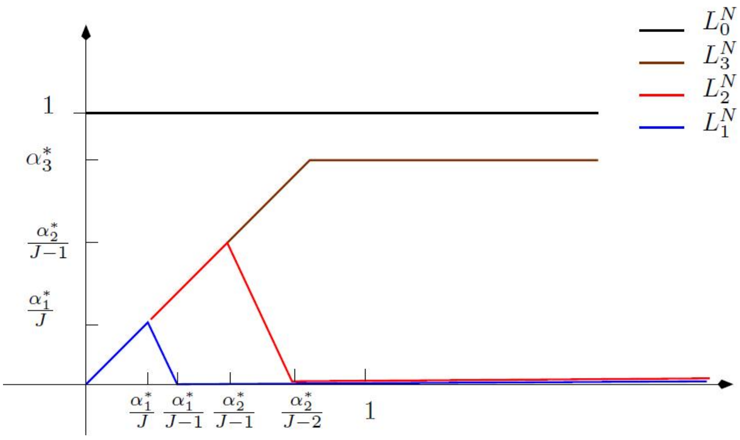


Figure 5.4 – Evolution of $\log(L_i^N(N^t))/\log N$, the exponent in N of L_i^N on the time scale $(N^t, t\in(0,1))$. Here J=3, $\rho_1<\rho_2<\rho_3$ and the initial state is $(N,0,0,0)$.

Without loss of generality, we assume that $\rho_1<\cdots<\rho_J$. We want to understand the fluid limit of this system when one of the queues, say the central one, becomes very big (the same arguments would enable us to analyse the easier case of the initially large queue being a peripheral one). Thus, let us suppose that $L_0^N(0)=N$, while $L_i^N(0)=0$ for $i\in\{1,\dots,J\}$. Our main result gives the large-N asymptotic behaviour of the queues which is illustrated by Figure 5.4. In fact, we obtain more than the convergence of the exponents in N of the size of each queue depicted in Figure 5.4. Theorem 5 in [RV16] describes the appropriate scalings of each L_i^N which give rise to a non-trivial limiting process for any value of J. In particular, it shows that the stability condition for the system of queues is now

$$\rho_0+\max\{\rho_i,\ 1\le i\le J\}<1.$$

For simplicity we consider only the case $J=2$ (hence with 3 nodes). The formal result we want to show is the following.

Theorem 5.2. (Th. 2 in [RV16]). *The following convergences of processes hold on the time interval* $(0,t_0)$.

1. *If* $\alpha_1^*/2>1$, *then* $t_0=3/(\mu_0-3\lambda_0)^+$ *and*

$$\lim_{N\to\infty}\left(\frac{L_0^N(Nt)}{N},\frac{L_1^N(Nt)}{N},\frac{L_2^N(Nt)}{N}\right)=\left(1+\mu_0\left(\rho_0-\frac{1}{3}\right)t,\mu_1\left(\rho_1-\frac{2}{3}\right)t,\mu_2\left(\rho_2-\frac{2}{3}\right)t\right).$$

2. *If $\alpha_1^*/2 < 1 < \alpha_1^*$, then $t_0 = 1/(1-\rho_0-\rho_1)^+$ and*

$$\lim_{N\to\infty}\left(\frac{L_0^N(Nt)}{N},\frac{L_1^N(Nt)}{N^{\alpha_1^*-1}},\frac{L_2^N(Nt)}{N}\right) = \left(1+\mu_0(\rho_0+\rho_1-1)t,\frac{1}{\mu_2(\rho_2-\rho_1)t},\mu_2(\rho_2-\rho_1)t\right).$$

3. *If $\alpha_1^* < 1 < \alpha_2^*$, then $t_0 = 1/(\mu_0(1/2-\rho_0))^+$ and*

$$\lim_{N\to\infty}\left(\frac{L_0^N(Nt)}{N},\frac{L_1^N(Nt)}{(\log N)^3},\frac{L_2^N(Nt)}{N}\right) = \left(1+\mu_0\Big(\rho_0-\frac{1}{2}\Big)t, 0, \mu_2\Big(\rho_2-\frac{1}{2}\Big)t\right).$$

4. *If $\alpha_2^* < 1$, then $t_0 = +\infty$ and*

$$\lim_{N\to\infty}\left(\frac{L_0^N(Nt)}{N},\frac{L_1^N(Nt)}{(\log N)^3},\frac{L_2^N(Nt)}{N^{\alpha_2^*}}\right) = ((\gamma(t),0,\gamma(t)^{\alpha_2^*})),$$

with $\gamma(t) = (1+\mu_0(\rho_0+\rho_2-1)t)^+$.

To sketch the proof of Theorem 5.2, we again decompose the evolution of the system of queues on the fluid timescale into several phases.

First phase: The timescale $t \mapsto N^t$, $t < \alpha_1^* \wedge 1$.

Using the same arguments as in the analysis of the first phase in Section 5.1, we see that L_0^N remains approximately equal to N whereas the other queues grow proportionally to N^t. That is:

Proposition 5.4. (Prop. 2 in [RV16]). *As $N \to \infty$, we have*

$$\left(\frac{L_1^N(N^t)}{N^t},\frac{L_2^N(N^t)}{N^t}\right)_{0<t<(\alpha_1^*/2)\wedge 1} \xrightarrow{(d)} \left(\lambda_1-\mu_1\frac{2t}{1+2t},\ \lambda_2-\mu_2\frac{2t}{1+2t}\right)_{0<t<(\alpha_1^*/2)\wedge 1},$$

uniformly over compact subintervals of $(0,\alpha_1^*/2\wedge 1)$.

Since $\rho_1 < \rho_2$ by assumption, the first coordinate of the limiting process is the first one to cancel. Assuming that $\alpha_1^*/2 < 1$, we now have to show that 'after' the timescale $N^{\alpha_1^*/2}$, L_2^N keeps on increasing while L_1^N decreases to 0. More precisely, we want to show that

Theorem 5.3. (Th. 1 in [RV16]). *Under the assumption that $\alpha_1^*/2 < 1$, as $N \to \infty$ we have*

$$\left(\frac{L_1^N(N^t)}{N^{\alpha_1^*-t}},\frac{L_2^N(N^t)}{N^t}\right)_{\alpha_1^*/2<t<\alpha_1^*\wedge 1} = \left(\frac{1}{\mu_2(\rho_2-\rho_1)},\mu_2(\rho_2-\rho_1)\right)_{\alpha_1^*/2<t<\alpha_1^*\wedge 1},$$

uniformly over compact subintervals of $(\alpha_1^*/2, \alpha_1^*\wedge 1)$.

This is where the new difficulty appears compared to the proof of Theorem 5.1. Indeed, instead of keeping a constant equilibrium value, L_1^N remains in equilibrium with L_2^N by adapting its value in such a way that the product $L_1^N L_2^N$ remains constant over time. The intrinsic reason why this convergence holds is again that whenever $\ln(L_1^N)+\ln(L_2^N) > \alpha_1^*\ln N$, the infinitesimal drift of L_1^N satisfies

$$\Delta L_1^N = \lambda_1 - \mu_1\frac{\ln(L_1^N)+\ln(L_2^N)}{\ln(L_1^N)+\ln(L_2^N)+\ln N} < \lambda_1 - \mu_1\frac{\alpha_1^*}{\alpha_1^*+1} = 0$$

(recall that $\alpha_1^*/(1+\alpha_1^*)=\rho_1$), and whenever $\ln(L_1^N)+\ln(L_2^N)<\alpha_1^*\ln N$, $\Delta L_1^N>0$. Now, just after time $N^{\alpha_1^*}$, the infinitesimal drift of L_2^N is still positive, which means that it keeps on increasing proportionally to N^t on the timescale N^t. Thus, on this timescale L_1^N becomes rapidly much smaller than L_2^N and the behaviour of the infinitesimal drift expounded just above implies that L_1^N can quickly adjust for the product $L_1^N L_2^N$ to come back to $N^{\alpha_1^*}$. But then the infinitesimal drift of L_2^N remains approximately equal to

$$\Delta L_2^N=\lambda_2-\mu_2\,\frac{\alpha_1^*}{\alpha_1^*+1}=\mu_2(\rho_2-\rho_1)>0,$$

and so $L_2^N(N^t)\approx\mu_2(\rho_2-\rho_1)N^t$ until L_1^N reaches a neighbourhood of 0.

The rigourous proof of Theorem 5.3 uses two main arguments. First, we prove the following extension of a result of Kingman [66] on subcritical birth and death processes.

Proposition 5.5. (Prop. 1 in [RV16]).

(a) If $(X(s))$ is a birth and death process on $\mathbb{Z}$ starting at 1 with birth rate λ and death rate $\mu>\lambda$, then for any integer $x\geq 0$,

$$\mathbb{P}\left(\sup_{s\geq 0}X(s)\geq x\right)\leq\left(\frac{\lambda}{\mu}\right)^x.$$

(b) If $(X_+(s))$ denotes the process with the same transitions as $(X(s))$ but with a reflection at 0, then for any $T>0$,

$$\mathbb{P}\left(\sup_{0\leq s\leq T}X_+(s)\geq x\right)\leq(\lambda T+1)\left(\frac{\lambda}{\mu}\right)^x$$

and

$$\mathbb{E}\left(\sup_{0\leq s\leq T}X_+(s)^2\right)\leq 2(\lambda T+1)\frac{\mu^2}{(\mu-\lambda)^2}.$$

These results enable us to control the excursions of L_1^N away from $N^{\alpha_1^*}/L_2^N$ over an interval of time during which L_2^N does not change much, and to show that

$$\left(\frac{\ln(L_1^N(N^t))}{\ln N}+\frac{(\ln L_2^N(N^t))}{\ln N}\right)\longrightarrow(\alpha_1^*)$$

as N tends to infinity, uniformly of compact time intervals. The second ingredient is a martingale argument in the spirit of the stochastic averaging results (see e.g. Chapter 1.7 in [44]). Indeed, if we define a function $F:\mathbb{N}^2\times\mathbb{R}_+\to\mathbb{R}$ by

$$F(l_1,l_2,t):=\frac{1}{2}\left(\frac{l_2}{N^t}-\mu_2(\rho_2-\rho_1)\right)^2-\frac{\mu_2}{\mu_1}\frac{l_1}{N^t}\left(\frac{l_2}{N^t}-\mu_2(\rho_2-\rho_1)\right),\tag{5.7}$$

with some courage we can compute that the generator G^N of the process $(L_1^N(N^t),L_2^N(N^t),t)$ (assuming that $L_0^N\equiv N$) applied to F can be written as

$$G^NF(l_1,l_2,t)=-(\ln N)\left(\frac{l_2}{N^t}-\mu_2(\rho_2-\rho_1)\right)^2+\mathcal{O}\left(\frac{l_1}{N^t}\right).$$

Using the associated martingale problem, a good control of the last term in the r.h.s. and finally Gronwall's inequality, we can show that $L_2^N(N^t)/N^t$ converges in $\mathbb{L}^2$ norm to $\mu_2(\rho_2-\rho_1)$ for any fixed $t \in (\alpha_1^*/2, \alpha_1^* \wedge 1)$. We then elaborate on this result to show the uniform pathwise convergence of $(L_2^N(N^t)/N^t)$, using again the martingale problem for $F(L_1^N(N^t), L_2^N(N^t), t)$.

Finally, another chain of arguments based on Proposition 5.5 enables us to conclude the proof of Theorem 5.3.

Second phase: the timescale $t \mapsto N^t$, $t \in (\alpha_1^*, 1+]$.

Assuming that $\alpha_1^* < 1$, at the end of the last phase L_1^N approaches 0. But its service rate remains identical to that of L_2^N when it is not empty. Since L_2^N is proportional to a power of N, all requests arriving at node 1 are treated very quickly and L_1^N remains of order at most $\mathcal{O}((\ln N)^2)$ for the rest of the evolution (cf. Proposition 3 in [RV16]). This implies that $\ln(L_1^N)$ remains negligible compared to $\ln L$ on this timescale and the system (L_0^N, L_2^N) is equivalent to the system of two queues studied in the previous section, which enables us to complete the proof of Theorem 5.2.

More than three nodes.

The above analysis remains valid when $J > 2$ and $\rho_1 < \rho_2 < \cdots < \rho_J$, except that the time t at which $L_1^N(N^t)/N^t \approx 0$ is now α_1^*/J and since all the other queues carry on increasing 'after' the timescale $N^{\alpha_1^*/J}$, we find that the time t at which $L_1^N(N^t) = \mathcal{O}((\ln N)^2)$ is now $\alpha_1^*/(J-1)$. Again, after this timescale, the first queue takes advantage of the fact that it is coupled with the very large queues $L_2^N, \ldots, L_J^N$ to receive a nonnegligible fraction of the service capacity and remain in a neighbourhood of 0. The remaining J queues (including the central one) then form a system of interacting queues of the same form as the initial $(J+1)$-system, but with one less queue. We can thus proceed by induction.

5.3 Perspectives

In the two studies presented above, we have explored two particular networks of service incompatibilities. We have seen that the resilience properties of the system depended on the precise form of the network. We could imagine more general graphs of interference representing the different clients in a wireless network using the same resources to transmit their messages. For example, the clients could be placed at the nodes of a finite subset of $\mathbb{Z}^2$, with interference between clients at distance less than some δ. Already in the case of a one-dimensional torus of odd size with nearest neighbour incompatibilities, the formulation of the service rate at queue i is not obvious. Indeed, considering a torus with 5 nodes, we see that client 1 can transmit at the same time as client 3 *or* client 4, but not both since clients 3 and 4 interfere. It is thus a first question to be able to write down a model for such a network. The next question is of course to explore the properties of this system, in particular when one of the clients is particularly demanding (i.e., the size of one of the queues is very large). Lastly, we could also ask the same questions when the clients move in space but the servers are fixed, as in a mobile phone network. For example, each client may use the closest server if it is not already transmitting, which would locally correspond to the system analysed in Section 5.1 until the client moves in space and switches server.

Bibliographie

[1] N. Abramson. The Aloha system : another alternative for computer communications. In *FJCC AFIPS Conf. Proc.*, pages 281–285. ACM, 1970.

[2] D. Aldous. Probability distributions on cladograms. In *Random discrete structures*, pages 1–18. Springer, 1996.

[3] D. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Stat. Sci.*, 16 :23–34, 2001.

[4] O. Angel, N. Berestycki, and V. Limic. Global divergence of spatial coalescents. *Probab. Theory Rel. Fields*, 152 :625–679, 2012.

[5] N.H. Barton. The effect of hitch-hiking on neutral genealogies. *Genet. Res.*, 72 :123–133, 1998.

[6] N.H. Barton, F. Depaulis, and A.M. Etheridge. Neutral evolution in spatially continuous populations. *Theor. Pop. Biol.*, 61 :31–48, 2002.

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

[8] N.H. Barton, J. Kelleher, and A.M. Etheridge. A new model for extinction and recolonisation in two dimensions : quantifying phylogeography. *Evolution*, 64 :2701–2715, 2010.

[9] N.H. Barton and I. Wilson. Genealogies and geography. *Philos. Trans. R. Soc. Lond. B*, 349 :49–59, 1995.

[10] M.A. Beaumont, W. Zhang, and D.J. Balding. Approximate Bayesian Computation in population genetics. *Genetics*, 162 :2025–2035, 2002.

[11] J. Berestycki, N. Berestycki, and J. Schweinsberg. The genealogy of branching Brownian motion with absorption. *Ann. Probab.*, 41 :527–618, 2013.

[12] N. Berestycki, A.M. Etheridge, and M. Hutzenthaler. Survival, extinction and ergodicity in a spatially continuous population model. *Markov Proc. Rel. Fields*, 15 :265–288, 2009.

[13] M. Birkner, J. Blath, M. Capaldo, A.M. Etheridge, M. Möhle, J. Schweinsberg, and A. Wakolbinger. Alpha-stable branching and Beta-coalescents. *Electron. J. Probab.*, 10 :303–325, 2005.

[14] M. Birkner, J. Blath, M. Möhle, M. Steinrücken, and J. Tams. A modified lookdown construction for the Ξ-Fleming-Viot process with mutation and populations with recurrent bottlenecks. *Alea*, 6 :25–61, 2009.

[15] M.G.B. Blum and O. Francois. Which random processes describe the tree of life ? A large-scale study of phylogenetic tree imbalance. *Syst. Biol.*, 55 :685–691, 2006.

[16] S. Boitard, W. Rodriguez, F. Jay, S. Mona, and F. Austerlitz. Inferring population size history from large samples of genome-wide molecular data - An Approximate Bayesian Computation approach. *PLoS Genetics*, 12(3) :e1005877, 2016.

[17] E. Bolthausen and A.-S. Sznitman. On Ruelle's probability cascades and an abstract cavity method. *Comm. Math. Phys.*, 197 :247–276, 1998.

[18] E. Brunet, B. Derrida, A.H. Mueller, and S. Munier. Noisy travelling waves : effect of selection on genealogies. *Europhys. Lett.*, 76 :1–7, 2006.

[19] L. Bunnefeld, L.A.F. Frantz, and K. Lohse. Inferring bottlenecks from genome-wide samples of short sequence blocks. *Genetics*, 201 :1157–1169, 2015.

[20] X. Cabré and J.-M. Roquejoffre. The influence of fractional diffusion in Fisher-KPP equations. *Communications in Mathematical Physics*, 320 :679–722, 2013.

[21] J. Chang. Recent common ancestors of all present-day individuals. *Adv. Applied Probab.*, 31 :1002–1038, 1999.

[22] B. Charlesworth, D. Charlesworth, and N.H. Barton. The effects of genetic and geographic structure on neutral variation. *Annu. Rev. Ecol. Evol. Syst.*, 34 :99–125, 2003.

[23] D.H. Colless. Review of phylogenetics : the theory and practice of phylogenetic systematics. *Syst. Zool.*, 31 :100–104, 1982.

[24] J.T. Cox. Coalescing random walks and voter model consensus times on the torus in $\mathbb{Z}^d$. *Ann. Probab.*, 17 :1333–1366, 1989.

[25] J.T. Cox and R. Durrett. The stepping stone model : new formulas expose old myths. *Ann. Appl. Probab.*, 12 :1348–1377, 2002.

[26] J.T. Cox and D. Griffeath. Diffusive clustering in the two-dimensional voter model. *Ann. Probab.*, 14 :347–370, 1986.

[27] J.T. Cox and D. Griffeath. Mean field asymptotics for the planar stepping stone model. *Proc. London Math. Soc.*, 61 :189–208, 1990.

[28] A. De and R. Durrett. Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum. *Genetics*, 176 :969–981, 2007.

[29] F.S. Dobson. The enduring question of sex-biased dispersal : Paul Greenwood's (1980) seminal contribution. *Anim. Behav.*, 85 :299–304, 2013.

[30] P. Donnelly, S.N. Evans, K. Fleischmann, T.G. Kurtz, and X. Zhou. Continuum-sites stepping-stone models, coalescing exchangeable partitions, and random trees. *Ann. Probab.*, 28 :1063–1110, 2000.

[31] P. Donnelly and T.G. Kurtz. A countable representation of the Fleming-Viot measure-valued diffusion. *Ann. Probab.*, 24 :698–742, 1996.

[32] P. Donnelly and T.G. Kurtz. Particle representations for measure-valued population models. *Ann. Probab.*, 27 :166–205, 1999.

[33] P. Donnelly, C. Wiuf, J. Hein, M. Slatkin, W. Ewens, and J.F.C. Kingman. Discussion : Recent common ancestors of all present-day individuals. *Adv. Applied Probab.*, 31 :1027–1035, 1999.

[34] R. Durrett. *Probability models for DNA sequence evolution.* Springer Science & Business Media, 2008.

[35] R. Durrett and J. Schweinsberg. Approximating selective sweeps. *Theor. Popul. Biol.*, 66 :129–138, 2004.

[36] B.K. Epperson. *Geographical genetics.* Princeton University Press, 2003.

[37] A. Etheridge. *Some mathematical models from population genetics : École d'été de probabilités de Saint-Flour XXXIX-2009*, volume 39. Springer Science & Business Media, 2011.

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

[39] A.M. Etheridge, N. Freeman, and S. Penington. Branching Brownian motion, mean curvature flow and the motion of hybrid zones. *arXiv preprint 1607.07563*, 2016.

[40] A.M. Etheridge, N. Freeman, S. Penington, and D. Straulino. Branching Brownian motion and selection in the spatial Lambda-Fleming-Viot process. *Ann. Applied Probab.*, (to appear), 2016.

[41] A.M. Etheridge, N. Freeman, and D. Straulino. The Brownian net and selection in the spatial Lambda-Fleming-Viot process. *arXiv preprint 1506.01158*, 2016.

[42] A.M. Etheridge and T.G. Kurtz. Genealogical constructions of population models. *arXiv preprint 1402.6724*, 2016.

[43] A.M. Etheridge, P. Pfaffelhuber, and A. Wakolbinger. An approximate sampling formula under genetic hitchhiking. *Ann. Applied Probab.*, 16 :685–729, 2006.

[44] S.N. Ethier and T.G. Kurtz. *Markov processes : characterization and convergence.* Wiley, 1986.

[45] S.N. Evans. Coalescing Markov labelled partitions and a continuous sites genetics model with infinitely many types. *Ann. Inst. H. Poincaré Probab. Statist.*, 33 :339–358, 1997.

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

[47] R. Fisher. *The genetical theory of natural selection.* Clarenson, Oxford, 1930.

[48] R.A. Fisher. The wave of advance of advantageous genes. *Ann. Eugenics*, 7 :353–369, 1937.

[49] R. Forien. Dispersal heterogeneity in the spatial Λ-Fleming-Viot process. *Preprint*, 2017.

[50] N. Freeman. The segregated Lambda-coalescent. *Ann. Probab.*, 43 :435–467, 2015.

[51] J.H. Gillespie. Genetic drift in an infinite population : the pseudohitchhiking model. *Genetics*, 155 :909–919, 2000.

[52] S. Gravel and M. Steel. The existence and abundance of ghost ancestors in biparental populations. *Theor. Popul. Biol.*, 101 :47–53, 2015.

[53] P.J. Greenwood. Mating systems, philopatry and dispersal in birds and mammals. *Anim. Behav.*, 28 :1140–1162, 1980.

[54] A. Greven, V. Limic, and A. Winter. Representation theorems for interacting Moran models, interacting Fisher-Wright diffusions and applications. *Electron. J. Probab.*, 10 :1286–1358, 2005.

[55] S. Guindon, H. Guo, and D. Welch. Demographic inference under the coalescent in a spatial continuum. *Theor. Popul. Biol.*, 111 :43–50, 2016.

[56] R.N. Gutenkunst, R.D. Hernandez, S.H. Williamson, and C.D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10) :e1000695, 2009.

[57] O. Hallatschek and D. Nelson. Life at the front of an expanding population. *Evolution*, 64 :193–206, 2010.

[58] K. Harris and R. Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, 9(6) :e1003521, 2013.

[59] B. Heuer and A. Sturm. On spatial coalescents with multiple mergers in two dimensions. *Theoret. Pop. Biol.*, 87 :90–104, 2013.

[60] S.Y.W. Ho and B. Shapiro. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Res.*, 11 :423–434, 2011.

[61] O. Kallenberg. *Foundations of modern probability.* Springer, New York, 2002.

[62] J. Kelleher, N.H. Barton, and A.M. Etheridge. Coalescent simulation in continuous space. *Bioinformatics*, 29 :955–956, 2013.

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

[64] M. Kimura. Stepping stone model of population. *Ann. Rep. Nat. Inst. Genet. Japan*, 3 :62–63, 1953.

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

[66] J.F.C. Kingman. Inequalities in the theory of queues. *J. Royal Stat. Soc. B*, 32 :102–110, 1970.

[67] J.F.C. Kingman. The coalescent. *Stochastic Process. Appl.*, 13 :235–248, 1982.

[68] M. Kirkpatrick and M. Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47 :1171–1181, 1993.

[69] A. Kolmogorov, I. Petrovsky, and N. Piscounov. Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. *Moscow Univ. Math. Bull.*, 1 :1–25, 1937.

[70] J.J. Koskela. *Consistency and intractable likelihood for jump diffusions and generalised coalescent processes.* Ph.D. thesis, University of Warwick, 2016.

[71] T.G. Kurtz and E.R. Rodrigues. Poisson representations of branching Markov and measure-valued branching processes. *Ann. Probab.*, 39 :939–984, 2011.

[72] J. Lachance. Inbreeding, pedigree size, and the most recent common ancestor of humanity. *J. Theor. Biol.*, 261 :238–247, 2009.

[73] Y.J. Li, Y. Satta, and N. Takahata. Paleo-demography of the *Drosophila melanogaster* subgroup : application of the maximum likelihood method. *Genes Genet. Syst.*, 74 :117–127, 1999.

[74] R.H. Liang. *Two continuum-sites stepping stone models in population genetics with delayed coalescence.* Ph.D. thesis, University of California, Berkeley, 2009.

[75] V. Limic and A. Sturm. The spatial Lambda-coalescent. *Electron. J. Probab.*, 11 :363–393, 2006.

[76] G. Malécot. *Les Mathématiques de l'hérédité.* Paris : Masson et Cie, 1948.

[77] P. Marjoram and J. Wall. Fast "coalescent" simulation. *BMC Genetics*, 7 :16, 2006.

[78] F.A. Matsen and S.N. Evans. To what extent does genealogical ancestry imply genetic ancestry ? *Theor. Popul. Biol.*, 74 :182–190, 2008.

[79] F.A. Matsen and J. Wakeley. Convergence to the island-model coalescent process in populations with restricted migration. *Genetics*, 172 :701–708, 2006.

[80] J. Maynard Smith and J. Haigh. The hitch-hiking effect of a favourable allele. *Genet. Res.*, 23 :23–35, 1974.

[81] G. McVean and N. Cardin. Approximating the coalescent with recombination. *Phil. Trans. Royal Soc. B*, 360 :1387–1393, 2005.

[82] G.A.T. McVean. A genealogical interpretation of linkage disequilibrium. *Genetics*, 162 :987–991, 2002.

[83] R.M. Metcalfe and D.R. Boggs. Ethernet : Distributed packet switching for local computer networks. *Communications of the ACM*, 19 :395–404, 1976.

[84] C. Mueller, L. Mytnik, and J. Quastel. Small noise asymptotics of traveling waves. *Markov Process. Related Fields*, 14 :333–342, 2008.

[85] C. Mueller, L. Mytnik, and J. Quastel. Effect of noise on front propagation in reaction-diffusion equations of KPP type. *Invent. Math.*, 184 :405–453, 2011.

[86] C. Mueller and R. Tribe. Finite width for a random stationary interface. *Electron. J. Probab.*, 2 :1–27, 1997.

[87] R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154 :931–942, 2000.

[88] J.A. Palacios, J. Wakeley, and S. Ramachandran. Bayesian nonparametric inference of population size changes from sequential genealogies. *Genetics*, 201 :281–304, 2015.

[89] B.M. Peter, D. Wegmann, and L. Excoffier. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Mol. Ecol.*, 19 :4648–4660, 2010.

[90] A.B. Phillimore and T.D. Price. Density-dependent cladogenesis in birds. *PLoS Biol.*, 6 :e71, 2008.

[91] J. Pitman. Coalescents with multiple collisions. *Ann. Probab.*, 27 :1870–1902, 1999.

[92] O.G. Pybus, A. Rambaut, and P.H. Harvey. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155 :1429–11437, 2000.

[93] P. Ralph and G. Coop. The geography of recent genetic ancestry across Europe. *PLoS Biol.*, 11 :e1001555, 2013.

[94] H. Ringbauer, G. Coop, and N.H. Barton. Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205 :1335–1351, 2017.

[95] Y. Rinott. On normal approximation rates for certain sums of dependent random variables. *J. Comp. Appl. Math.*, 55 :135–143, 1994.

[96] M.J. Sackin. “Good” and “bad” phenograms. *Syst. Zool.*, 21 :225–226, 1975.

[97] R. Sainudiin, K. Thornton, J. Harlow, J. Booth, M. Stillman, R. Yoshida, R.C. Griffiths, G. McVean, and P. Donnelly. Experiments with the site frequency spectrum. *Bulletin of Mathematical Biology*, 73 :829–872, 2011.

[98] R. Sainudiin and D. Welch. The transmission process : A combinatorial stochastic process for the evolution of transmission trees over networks. *J. Theor. Biol.*, 410 :137–170, 2016.

[99] S.A. Sawyer and D.L. Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132 :1161–1176, 1992.

[100] J. Schweinsberg. A necessary and sufficient condition for the Λ-coalescent to come down from infinity. *Electron. Comm. Probab.*, 5 :1–11, 2000.

[101] J. Schweinsberg and R. Durrett. Random partitions approximating the coalescence of lineages during a selective sweep. *Ann. Applied Probab.*, 15 :1591–1651, 2005.

[102] D. Shah and D. Wischik. Log-weight scheduling in switched networks. *Queuing Syst.*, 71 :97–136, 2012.

[103] M. Steinrücken, J.A. Kamm, and Y.S. Song. Inference of complex population histories using whole-genome sequences from multiple populations. *BioRxiv preprint*, 2016.

[104] W. Stephan, T.H. Wiehe, and M. Lenz. The effect of strongly selected substitutions on neutral polymorphism : analytical results based on diffusion theory. *Theor. Popul. Biol.*, 41 :237–254, 1992.

[105] F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105 :437–460, 1983.

[106] N. Takahata. Allelic genealogy and human evolution. *Mol. Biol. Evol.*, 10 :2–22, 1993.

[107] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Automat. Control*, 37 :1936–1948, 1992.

[108] J.E. Taylor and A. Véber. Coalescent processes in subdivided populations subject to recurrent mass extinctions. *Electron. J. Probab.*, 14 :242–288, 2009.

[109] A. Tenesa, P. Navarro, B.J. Hayes, D.L. Duffy, G.M. Clarke, M.E. Goddard, and P.M. Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome Res.*, 17 :520–526, 2007.

[110] J. Wakeley. *Coalescent Theory : An Introduction.* Greenwood Village : Roberts & Company Publishers, 2009.

[111] J. Wakeley, L. King, B.S. Low, and S. Ramachandran. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics*, 190 :1433–1445, 2012.

[112] J. Wakeley and S. Lessard. Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics*, 164 :1043–1053, 2003.

[113] J. Wakeley and T. Takahashi. The many-demes limit for selection and drift in a subdivided population. *Theoret. Pop. Biol.*, 66 :83–91, 2004.

[114] G.H. Weiss and M. Kimura. A mathematical analysis of the stepping stone model of genetic correlation. *J. Appl. Probab.*, 2 :129–149, 1965.

[115] D. White, J. Wolff, M. Pierson, and N. Gemmell. Revealing the hidden complexities of mtDNA inheritance. *Mol. Ecol.*, 17 :4925–4942, 2008.

[116] J.F. Wilkins. A separation of timescales approach to the coalescent in a continuous population. *Genetics*, 168 :2227–2244, 2004.

[117] J.F. Wilkins and J. Wakeley. The coalescent in a continuous, finite, linear population. *Genetics*, 161 :873–888, 2002.

[118] S. Wright. Evolution in Mendelian populations. *Genetics*, 16 :97–159, 1931.

[119] S. Wright. Isolation by distance. *Genetics*, 28 :114–138, 1943.

[120] S. Wright. The genetical structure of populations. *Ann. Eugen.*, 15 :323–354, 1951.

[121] G.U. Yule. A mathematical theory of evolution, based on the conclusions of Dr J.C. Willis. *Phil. Trans. R. Soc. Lond. B*, 213 :21–87, 1924.

[122] I. Zähle, J.T. Cox, and R. Durrett. The stepping stone model. II : Genealogies and the infinite sites model. *Ann. Applied Probab.*, 15 :671–699, 2005.