Equilibrium patterns of genetic diversity shuffled by migration and recombination

Verónica Miró Pina

Sorbonne Université

École doctorale de sciences mathématiques de Paris centre

Laboratoire de Probabilités, Statistique et Modélisation

Thèse de doctorat

Discipline : Mathématiques

présentée par

Verónica Miró Pina

Equilibrium patterns of genetic diversity shuffled by migration and recombination

---

dirigée par Amaury Lambert et Emmanuel Schertzer

Présentée et soutenue publiquement le 7 septembre 2018

Devant le jury composé de :

M. Bastien Fernandez	Directeur de recherches CNRS	Examinateur
M. Amaury Lambert	Professeur Sorbonne Université	Directeur
M. Emmanuel Schertzer	Maître de conférences Sorbonne Université	Directeur
M. Viet Chi Tran	Maître de conférences Univ Lille	Examinateur
Mme Amandine Véber	Chargée de recherches CNRS	Rapporteuse
Mme Aleksandra Walczak	Directrice de recherches CNRS	Examinatrice
M. Carsten Wiuf	Professeur University of Copenhagen	Rapporteur

Sorbonne Université - LPSM
Campus Pierre et Marie Curie
Case courrier 158
4 place Jussieu
75 252 Paris Cedex 05

Sorbonne Université
École doctorale de sciences
mathématiques de Paris centre
Boîte courrier 290
4 place Jussieu
75 252 Paris Cedex 05

A mi yayo Pepe por haber sembrado en mi el germen de la curiosidad. A mi abuelo Antonio por haber intentado sembrar el de la paciencia. A Janito, el perro más bueno.

Remerciements

Je voulais tout d'abord remercier chaleureusement mes rapporteurs, Amandine Véber et Carsten Wiuf ainsi que les membres de mon jury, Bastien Fernandez, Viet Chi Tran et Aleksandra Walczak d'avoir consacré du temps et de l'énergie à l'évaluation de cette thèse.

Merci à mes deux directeurs de thèse pour leur encadrement exceptionnel. Merci de m'avoir laissé autant de liberté, de m'avoir permis de prendre le temps de choisir mon chemin, de me tromper, de changer de direction, tout en me guidant et en étant toujours là quand j'en avais besoin. Merci pour votre patience. Et surtout merci de m'avoir permis de découvrir deux personnes extraordinaires. Merci Emmanuel pour ta créativité sans fin, tes idées incroyables. Merci de m'avoir poussé à faire toujours mieux. Merci pour toutes les fois où tu es venu me demander si j'avais besoin d'aide. Merci de m'avoir supporté dans les périodes difficiles et d'avoir fait semblant d'être serein quand j'étais stréssée. Merci pour ton optimisme qui a essayé de compenser mon pessimisme. Merci Amaury pour tes conseils avisés, sur le plan scientifique mais aussi sur le plan personnel. Merci de m'avoir aidé à reprendre confiance en moi et de m'avoir rappelé à plein d'occasions que le plus important dans la recherche c'est de prendre du plaisir. Merci pour toutes les conversations passionnantes sur tout et n'importe quoi. Merci de faire toujours attention à l'ambiance dans l'équipe et à ce que chacun trouve sa place. Je suis très contente que Félix puisse prendre la relève et j'espère qu'il saura profiter encore plus que moi de ce binome d'encadrants de choc. Et surtout, j'espère que nous aurons l'occasion de travailler ensemble tous les trois dans le futur.

Je voudrais aussi remercier tous ceux et celles qui ont contribué à ma formation avant la thèse. Tout d'abord, le Lycée Francais de Murcia, dans lequel j'ai passé 15 années merveilleuses. Merci, non seulement aux professeurs, mais aussi à tout le personnel, des CPE aux surveillantes, de m'avoir supporté et de m'avoir aidé à grandir. Porque el liceo, es el liceo ! Et merci à Mme Gohin, M. Prost et M. Etienne de m'avoir donné le goût de la science et de m'avoir encouragé à continuer mes études en France. Merci à Farouk Boucekkine, qui m'a fait voir la beauté des mathématiques et qui a été un grand soutien pendant les périodes de doute en prépa. Merci Mme Gazeau d'avoir cru en moi. Merci aussi à un autre professeur de prépa, qui a juré un jour que jamais je ne rentrerai à l'ENS

et jamais je ne ferai de thèse. Vous m'avez dit que c'était impossible donc je l'ai fait. Merci à l'ENS de m'avoir offert 4 années de liberté, 4 années où j'ai pu profiter de cours passionnants, découvrir le monde merveilleux de la Recherche et être entourée de personnes exceptionnelles. Merci aussi à Luis Almeida et au reste de l'équipe enseignante du M2 math-bio de Jussieu pour leur patience quand j'étais une petite biologiste perdue dans le monde des mathématiciens. Enfin, je voulais remercier la Danse, de m'avoir inculqué très jeune la discipline dont j'ai eu besoin plus tard et d'avoir été là pour me permettre de m'évader.

Je voudrais remercier l'équipe SMILE de m'avoir accompagné dans cette aventure et m'avoir permis de travailler dans la meilleure ambiance possible. Merci d'avoir partagé des cafés, des gâteaux, des repas à l'AURA, des mots sur le tableau mais aussi des apéros sur la terrasse, des voyages au ski, des randonnées sans eau, des raclettes, des escape game, des manifs, de la musique stochastique, des fonctions test et des uniformes, des descentes de l'infini avec de la poussière, du coloriage et du découpage de chromosomes... Merci Marc, Fanny, Miraine, François Bienvenu, Jean-Jil, Félix, Élise, Florence, Marguerite, Kader, Thomas, Pascal, Julie, Guillaume, François Blanquart, Peter, Luca, Cécile, Airam, Todd, Odile, Tristan, Salma, Mélanie, Charles, Sophie, Eliott et tous les autres smileurs qui sont passés par l'équipe. Merci aussi aux collègues de l'èquipe Modélisation aléatoire du vivant et aux organisateurs et membres assidus du Groupe de Travail des Thésards du LPSM en particulier Yann, Michel, Paul, Clément, Laure... Merci aux professeurs et autres chargés de TD avec qui j'ai pu travailler pendant ces années, ainsi qu'aux élèves enthousiastes qui m'ont fait découvrir le plaisir d'enseigner. Merci aussi à tous les gens avec qui j'ai passé de bons moments lors des conférences et écoles d'été: Raphael, Paul, Delphin, Lucas, Amélie, Luisa, Sebastian and Sebastian, Sandra, Philippe, Joseba... Gracias Adrián González Casanova pour ser mi mejor amigo de conferencias y gracias también a Arno Siri-Jégousse por haber confiado en mi y haberme dado la oportunidad de continuar esta aventura del otro lado del charco.

Je voulais aussi remercier les amis qui m'ont accompagné dans cette aventure. Merci Mariane d'être toujours la voix de la sagesse. Merci pour cette capacité que tu as à me faire retrouver le chemin de la raison et merci pour ton soutien indefectible. Merci aussi à Spin, Totoro et BB-8 (mais pas à Cyrille, sinon il va être jaloux d'avoir moins de mots que toi). Merci Antoine Petit d'être la personne sur qui je peux toujours compter et qui me connait mieux que je ne me connais moi même. Merci d'être toujours là pour me faire voir la part des choses et m'empêcher de me prendre trop aux sérieux. Merci de m'avoir accueilli au retour de Princeton et d'être venu à mon secours à chaque fois que j'en ai eu besoin. Merci Laura pour les réunions du Club des Ppetits. Merci Olivier de m'avoir soutenu quand j'avais cassé ma thèse et quand j'étais stréssée. Merci Émilie, Corentin et Claire pour toutes les soirées à Alésia. Merci Nath et JackieJack, mes colocs pendant la

première année de thèse avec qui j'ai partagé beaucoup de pâtes au pesto, des soirées et de conversations philosophiques. Merci aussi Julia et Mathieu d'avoir fait partie de la famille de Villejuif. Merci Anne, Valentine et Justine, mes stars de la K-Pop, pour les safaris antropologiques et les soirées films niais. Merci Clément et tes poissons ponctuels, pour les conseils avisés pour survivre dans le monde des probas. Merci Judith, Noémie et Aurore, la team pompom toujours. Merci Xavier, Raph, Alexis et Mathieu, Laure et Jaime, et tous les autres amis de l'ENS. Merci Lola pour les soirées salsa et les débats politiques toujours passionants. Obrigada David, muchas gracias por confiar siempre en mi y por los mensajes de ánimo.

Y gracias al Pisi por ser como una familia. Gracias Ana y Alba por haberme dejado descubriros. Gracias por haberme acompañado en este viaje, por haber hecho que los momentos difíciles lo fueran algo menos. Gracias Ana, por ser un ejemplo de que quien la sigue la consigue. Y por todos los momentos divertidos y absurdos. Gracias Alba por acompañarme en la misión de hermana mayor, por nuestros debates, por ser tan distinta a mi y ayudarme a ver las cosas desde otro punto de vista. Creo que hemos aprendido mucho la una de la otra. Por los viajes, por nuestras cenas, por las noches interminables y absurdas, por nuestras locuras: gracias grinchitas ! Y gracias a todos los habitantes temporales del pisi (Giorgio, Gloria, Mari...) por tantos momentos especiales.

También quería darle las gracias a esas personas que me han apoyado desde la distancia. Gracias a mis amigas de siempre (en ese "amigas" se reconocerá también algún que otro amigo). Algunas lleváis 24 años aguantandome, otras algo menos, hemos vivido tantas cosas juntas, hemos tenido nuestros enfados y nuestras peleas, como hermanas que somos. Pero la unión que seguimos teniendo después de 9 años viviendo en ciudades y países distintos es el tesoro más grande que tengo. Gracias vuestros whatsapps, vuestros audios infinitos y vuestros mensajes por instagram, sin los cuales la redacción de esta tésis se me hubiese hecho interminable. Ojalá nunca se acaben los aperitivazos de nochebuena, los amigos invisibles, los entierros de la sardina, las noches de verano... Que sigamos apoyándonos siempre desde la distancia. Y que estoy muy orgullosa de teneros como amigas.

Muchas gracias a toda la pandilla de Lopa, por haber hecho que los veranos siempre sean especiales (y no solo por la rareza anual de Rafa). Por hacer que los cursos siempre empezaran con las pilas cargadas de migas, barbacoas, bocadillos tropicales y helados de la jijonenca (y sin ningún kilo de menos). Porque podré viajar lejos, conocer un millón de sitios nuevos, pero siempre querré volver a la llana, y el verano no será verano sin ir en bici por las salinas. Gracias a mis amigos frikis, mis medulpis, por estar desde siempre y entendernos como nadie.

Por último, quería dar las gracias a mi familia, por haberme apoyado siempre y haber

estado tan presentes desde la distancia. Gracias a mi tía Mari Angeles y a mis primas Aitana y Alba, sin las cuales habría perdido aún más aviones de los que ya perdí. Gracias Aitana por estar siempre ahí. Gracias al gen Pina por haberme dado una familia tan absurda y divertida. Gracias a mi abuela Cari y a mis tíos Ricardo y Antonio, por haberme acompañado tantas veces desde el otro lado del teléfono del laboratorio a casa y por haber confiado en mi siempre. Gracias a mi hermana pequeña que a veces es mi hermana mayor, a la que no ha dudado en echarme la bronca cuando me lo merecía para que nunca tirara la toalla. Confía en ti pequeña, que vas a llegar muy lejos. No sabes lo que me gustaría haberte tenido más cerca. Gracias a mi otra hermana, mi polo opuesto, la que me hace reir, la que me recuerda que nunca hay que tomarse las cosas demasiado en serio, que hay que dejarse fluir, la responsable de que diga tantas tonterías. Gracias por ser tan inútil, pero ser mi inútil. Gracias a mis perris por ser tan bebuchis. Y gracias a mis padres por confiar en mi, por haberme dejado elegir mi camino con total libertad, caerme, levantarme y hacerme fuerte. Vuestro apoyo y vuestro consejo nunca me han faltado y no dudasteis en hacer 2000 km de la noche a la mañana para salvarme en el momento más difícil de estos 9 años. Y sin eso, probablemente hoy no estaría donde estoy. Gracias por estar siempre al otro lado del teléfono, cuando necesito desahogarme o reirme. Gracias por haberme enseñado a esforzarme y a luchar, pero sobre todo por haberme enseñado a ser la persona que soy hoy.

Contents

Introduction

Population genetics is the field that studies how different factors influence genetic variability within and between populations. It is also devoted to the study of ancestry relationships between genes by the means of "genealogical trees". As we shall see, these two questions are related.

In most organisms, genetic information is carried by molecules of deoxyribonucleic acid (DNA), which is a chain of nucleotides that can be formed of one of four bases: cytosine (C), guanine (G), adenine (A) or thymine (T). Individuals may carry one or several DNA molecules or *chromosomes*. In *haploid* species each individual carries one copy of each chromosome, whereas in *diploid* species, each individual carries two copies of each chromosome. For example, humans have 23 pairs of chromosomes. The DNA sequence of an individual is inherited from its parents and constitutes its *genotype*. The *phenotype* of an individual is the set of all its observable characteristics or traits and is encoded by its genotype. A *gene* is a portion of DNA that codes for a given protein. Genes, but also regulatory sequences and other non-coding sequences determine the phenotype of an individual. We will often use the term *locus* to refer to a region of the chromosome, without specifying if it is a gene, a portion of a gene, a regulatory sequence, a single base... An *allele* is a version of a locus. Genetic variability is the fact that individuals have different genotypes i.e. carry different alleles. As we shall see, the linear arrangement of the different loci in a chromosome has an important effect in the way they are transmitted from one generation to another, and therefore in genetic variability.

As already mentioned, we are going to study how different factors influence genetic variability within and between populations. Among these "evolutionary forces" are:

— **Mutations**, that are changes in the DNA sequence of an individual that occur randomly. Mutation creates new alleles and is the ultimate source of genetic variability.

— **Genetic recombination** is the mechanism by which an individual can inherit a chromosome that is a mosaic of two parental chromosomes. Genes that are close to one another often share the same evolutionary history and are in linkage disequilibrium (i.e. the frequency of association of their different alleles is higher or lower than

what would be expected if loci were independent). Recombination breaks up linkage disequilibrium.

— **Population structure**, which, from a geneticist's point of view, is the fact that there are mating restrictions. For example, in a geographically structured population, individuals are more likely to reproduce with individuals that live close to them, which promotes genetic differentiation.

— **Migration**, that allows gene flow between individuals from different areas and weakens population structure.

— **Natural selection**, that is the fact that some individuals are better adapted to the environment, so they have more chances to survive and produce offspring.

— **Competition** between individuals for resources (and other biological interactions such as predation, parasitism, etc...).

— **Demographic variation**: the size of a population has an important effect on genetic variability: the smaller the population size, the more likely it is that two randomly chosen individuals have a recent common ancestor.

During my PhD I have studied the effect of three of these forces: recombination, population structure and migration. I have used two different models to study how recombination and migration shuffle genetic diversity.

In the first model, recombination is the only evolutionary force and we look at its effect on the chromosome of a randomly sampled individual. We consider a model in which, at time 0 each individual has her unique chromosome painted in a distinct color. By the blending effect of recombination, the genomes of descending individuals look like mosaics of colors, where each segment of the same color is called an identical-by-descent (IBD) segment. The goal of my first project was to characterize this mosaic at equilibrium. For example, if the leftmost locus is red, we have been able to characterize the distribution of the amount of red in the mosaic and of the positions of the red segments. The results of this project are presented in Chapter I and in an article to be submitted .

In Chapter II, I present an application of the results of Chapter I, in which we used the distribution of IBD block lengths to study hybrid populations.

In the second project, I have studied the effects of geographic structure, migration, mutation and recombination in the genetic composition of a metapopulation. The metapopulation is modelled as a graph where vertices correspond to subpopulations and edges are associated to migration rates. The idea behind this project was to study speciation: when two subpopulations accumulate enough genetic differences they may become separate species. We have been able to characterize the distribution of the genetic distances between

subpopulations in a low mutation - low migration regime, depending on the geographic structure, and to show that some geographic configurations can promote speciation. The results of this project are presented in Chapter III and in an article that is under revision for *Stochastic Processes and their Applications* [MPS17].

1 A brief history of population genetics

Mathematical models have played an important role in the study of genetic variation since the 1930s. John B.S. Haldane, Ronald A. Fisher and Sewall Wright are considered as the "founding fathers" of population genetics. They developed the first models describing the evolution of the genetic composition of a population, opening the way for the development of a fruitful branch of mathematics devoted to the study of refinements and generalizations of their models. In this section, we explain the historical context in which population genetics emerged.

1.1 Darwin's legacy and the study of natural selection

By the end of the 19^{th} century, the concept of evolution was largely accepted by the scientific community, but the mechanisms of evolution and the supports of heredity remained controversial. In fact Charles Darwin had published his book *On the origin of species* in 1859, suggesting that populations evolve over the course of generations through a process of natural selection. His theory was based on three concepts: variation between individuals, adaptation to the environment and heredity of traits. But he did not propose a mechanism for species formation and his theory on inheritance, pangenesis, did not meet any success [GHA97].

Among Darwin's successors one of the most famous was his half-cousin Francis Galton, who was the first to develop a statistical theory of heredity. He was particularly interested in describing variation in human populations and identifying which human abilities were hereditary. In 1877, in an article called *Typical laws of inheritance*, Galton described how, when crossing peas that produce large seeds, the offspring produce seeds whose size is closer to the population mean. Galton called this phenomenon "reversion" (although later he changed the name into *regression*). He made similar observations in human height and published the results in an article *Regression toward mediocrity in hereditary stature*. Galton would interpret this as meaning that the small variation by which natural selection was supposed to act according to Darwin could not work because small changes would be neutralized by regression toward the mean. In other words, evolution had to proceed via discontinuous steps [Gil01].

Galton is considered as one of the founders of modern statistics and introduced important concepts and tools such as correlation studies, linear regression, standard deviation and the Gaussian law of error. On the other hand he is also considered as the founder

of eugenics: he believed that the human species could help direct its future by selectively breeding individuals who have "desired" traits. This ideology had terrible consequences on European and American politics in the first half of the 19^{th} century, but this goes beyond the aims of this introduction.

Galton is also the founder of the "biometric" school, which was devoted to the mathematical description of the effects of natural selection. Galton's work was continued by his student Karl Pearson, who also made major contributions to the field of statistics (correlation coefficient, hypothesis testing, p-value, χ^2 test, principal component analysis...). Pearson and his colleague Raphael Weldon expanded statistical reasoning to the study of inheritance and natural and sexual selection. Unlike Galton, Pearson and Weldon developed a continuous theory of evolution in which natural selection was supposed to act by gradual variation. As we will see, this gave rise to an intense debate between biometricians and geneticists.

1.2 Mendel and the birth of genetics

The history of genetics begins in 1865, with the work of Gregor Mendel. His breeding experiments on peas allowed him to show that the patterns of inheritance obey simple statistical rules, with some traits being dominant and others recessive [Mik08]. But his work was not given any attention by the scientific community. There is evidence that Darwin was aware of Mendel's results, but without the concept of mutation, his laws seemed to imply that traits remain fixed. This could be the reason why Darwin did not pay much attention to his work. In 1900, De Vries, Correns and von Tschermak rediscovered Mendel's laws. In addition, De Vries introduced the concept of mutation, after observing some rare but brutal changes from one generation to the next. In 1906 William Bateson introduced the term "genetics" to describe the study of inheritance. The term "gene" was introduced three years later by Wilhelm Johannsen to describe the units of hereditary information.

Chromosomes had been observed under the microscope by W. Fleming at the end of the 19^{th} century, but it was not until the 1900s that Theodor Bovery linked chromosomes and heredity. Walter Sutton was the first to suggest that chromosomes constitute the physical basis of the Mendelian law of heredity [OM08]. Thomas H. Morgan and his colleagues demonstrated the chromosomal theory experimentally. They introduced the idea that a gene corresponds to a specific region in the chromosome. They proposed the idea that genetic linkage was related to the distance between genes in a chromosome and suggested a process of crossing over to explain recombination [LS08].

However it took several decades to discover the molecular basis of heredity and, at the time Wright, Fisher and Haldane wrote their theories, the role of DNA had not been established yet. In fact it was not until the 1940s that the experiments by Oswald Avery and his colleagues, together with the work of Alfred Hershey and Martha Chase on bacterial

phages, allowed to identify DNA as the hereditary material. In 1944, Erwin Chargaff noted that the nucleotide composition of DNA varied across species, but the proportion of A was always the same as the proportion of T and the proportion of G was equal to the proportion of C. This realization, together with some important X-ray cristallography work by Rosalind Franklin and Maurice Wilkins, allowed James Watson and Francis Crick to discover the double helix structure of DNA in 1953. Some years later, they established the central dogma of molecular biology, which explains the flow of genetic information, from DNA to proteins (via RNA) and which allowed to establish that phenotypic variation arises from changes in the DNA sequence [O'C08, Pra08].

1.3 The birth of population genetics

At the beginning of the 20^{th} century there was an intense debate between "Mendelians" (Bateson, De Vries ...) and "Darwinians" (Pearson, Weldon, ...). Part of the controversy was about the mechanisms of evolution: while the biometricians claimed that species evolve through the action of natural selection (that acts via small, gradual changes), the geneticists believed that discontinuous, brutal changes (mutations) were responsible for evolutionary change and did not believe in natural selection [GHA97]. But there was also a disagreement on methodology. While Pearson and Weldon wanted to make predictions, Bateson and the geneticists were more focused on describing the mechanisms of heredity. Pearson criticized the biologists for not being able to use mathematical techniques, stating that "before we can accept any cause of a progressive change as a factor we must have not only shown its plausibility but if possible have demonstrated its quantitative ability". For the geneticists, the work of the biometricians was "almost metaphysical speculation as to the causes of heredity".

The Mendelian and the biometrician models were eventually reconciled with the development of population genetics, thanks to the work of Fisher, Haldane and Wright. Fisher, who belonged to the same school as Galton and Pearson and who made major contributions in the field of statistics, e.g. the Monte Carlo method and the maximum likelihood estimation, is the man who allowed to conciliate the two different points of view on methodology. For Fisher, values obtained in experiments were no longer considered for what they were but as representations of a set of possibilities with probabilities attached. Because experimental results fluctuate, they have to be analyzed by probabilistic methods. This methodology at hand, Fisher and Haldane developed stochastic models that assumed Mendelian inheritance and where the combined effect of mutation and selection produced genetic variation and evolutionary change. Haldane applied statistical analysis to the study of real examples of natural selection such as the peppered moth. While Fisher and Haldane studied large populations, Wright was more interested in studying genetic drift, which is the phenomenon according to which, in a finite population, gene frequencies can evolve by the randomness of births and deaths. One of his major contributions

is the shifting-balance theory to explain species formation by population subdivision (see 4.4). The theoretical work of these three authors was a critical step towards developing a unified theory of evolution. The models they developed and their extensions are still used nowadays and presented in the next section.

2 Classical models in population genetics

2.1 The Wright-Fisher model

This model was developed by Fisher (1930) and Wright (1931). The hypothesis of the model are the following:

— The population size, N, is constant.

— Individuals are haploid (i.e. each individual carries a single copy of each gene).

— Mating is random.

— The population is panmictic i.e. all individuals are potential partners and there are no mating restrictions.

— Generations are non-overlapping, i.e. all individuals reproduce and die at the same time.

Definition 2.1. *In the neutral Wright-Fisher model, the individuals from generation $t+1$ choose their unique parent uniformly (and independently) at random from generation t.*

Consider a single gene with two alleles a and A. If, at generation t there are k individuals carrying allele A, we denote by X_t^N the proportion of individuals of type A in the population. If $X_t^N = k/N$, NX_{t+1}^N follows a binomial distribution of parameters N and k/N. $(X_t^N)_{t\in\mathbb{N}}$ is a Markov chain, valued in $\{0, 1/N, \dots, 1\}$ and that has two absorbing states, 0 and 1. We say that an allele is fixed when its frequency reaches 1 (otherwise we say that it is extinct). Let us call τ_N the absorption time time, i.e.

$$\tau_N = \min\{t, X_t^N = 0 \text{ or } X_t^N = 1\}.$$

In this model, all individuals of generation 0 have the same probability of being the common ancestor of all the individuals of generation τ_N, so we have

$$\mathbb{P}(X_{\tau_N}^N = 1 | X_0^N = \frac{i}{N}) \;=\; \frac{i}{N}.$$

In words, the probability of fixation of a neutral allele is its initial frequency.

This model is neutral, in the sense that all individuals have the same probability of being parents of an individual in the next generation, independently of their genotype.

Variation in allele frequency is only due to random sampling. This phenomenon is called *genetic drift*.

We can also consider the case where natural selection confers an advantage to the individuals that carry allele A with respect to those carrying allele a. In the Wright-Fisher model with selection, when an individual from generation $t+1$ chooses her parent in such a way that each individual of type A from generation t has a probability $(1+s)/(N(1+sX_t))$ of being chosen and each individual of type a a probability $1/(N(1+sX_t^N))$. Then, given X_t^N, the number of individuals carrying allele A at generation $t+1$ follows a binomial distribution with parameters N and $\frac{(1+s)X_t^N}{(1+sX_t^N)}$. The ratio between the mean number of offspring of an individual of type A and of an individual of type a is given by s, which is called the relative fitness. The concept of fitness was introduced by Haldane, and it represents the marginal ability to survive and reproduce in a given environment.

When the population size is large, the changes in the genetic frequencies from one generation to another are small, so it is quite natural to approximate $(X_t^N)_{t\in\mathbb{N}}$ by a diffusion process. This concept already appeared in an article by William Feller in 1951 [Fel51]. As X_{t+1}^N follows a binomial distribution, using Taylor expansions, we have

$$\begin{aligned}\mathbb{E}(X_{t+1}^N - X_t^N | X_t^N = x) &= sx(1-x) + o(s)\\ \mathbb{E}((X_{t+1}^N - X_t^N)^2 | X_t^N = x) &= \frac{1}{N}x(1-x) + o(s).\end{aligned} \tag{1}$$

It can be shown that, if we consider the series of Markov processes $(X_{\lfloor Nt\rfloor}^N, N\in\mathbb{N})$, where X^N is the frequency of allele A in a Wright-Fisher model with fitness rate $s \equiv s_N$, that scales with the population size in such a way that

$$Ns_N \underset{N\to\infty}{\longrightarrow} s \tag{2}$$

and if $\forall N \in \mathbb{N},\ X_0^N = x_0$, then

Proposition 2.2. *For all $T > 0$, $(X_{\lfloor Nt\rfloor}^N)_{N\in\mathbb{N}}$ converges in distribution in the Skorokhod topology $D([0,T],\mathbb{R})$ to the solution of:*

$$\begin{cases} dX_t &= sX_t(1-X_t)dt + \sqrt{X_t(1-X_t)}dB_t,\\ X_0 &= x_0, \end{cases} \tag{3}$$

where B is a standard Brownian motion.

See for example [EK05] (Theorem 2.2, Chapter 10) for a proof of this result.

Remark 2.3. *To obtain the diffusion approximation we had to renormalize time (i.e. to consider $X_{\lfloor Nt\rfloor}^N$ instead of X_t^N. We say that in the Wright-Fisher diffusion time is measured in units of N generations.*

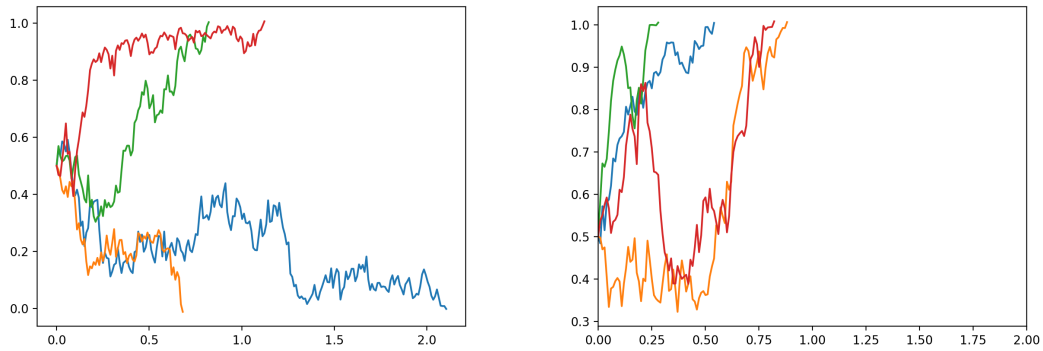


Figure 1 – The Wright-Fisher diffusion. The curves represent the frequency of allele A as a function of time. In the left-hand side $s = 0$ and in the right-hand side $s = 2$. We simulated 4 trajectories, with $X_0 = 0.5$ and stopped the simulations when absorption was reached.

Remark 2.4. *In the stochastic differential equation* (3)*, the drift term corresponds to natural selection while the diffusion term corresponds to genetic drift.*

From (1) *one can see that the mean variation in allelic frequency from one generation to another only depends on the selection coefficient* s *while the variance of* $(X^N_{t+1} - X^N_t)$ *depends on the inverse of the population size. Hypothesis 2 guarantees that there is a balance between natural selection and genetic drift. But small populations are more sensitive to genetic drift. If* $Ns_N \to 0$*, the magnitude of genetic drift can overwhelm the effect of selection resulting in non adaptive evolution (i.e. fixation of alleles is totally random and does not depend on the fitness value). On the contrary, when the population size is large and* $Ns_N \to \infty$*, the Wright-Fisher model converges to a deterministic model i.e.* $(X^N_{\lfloor Nt \rfloor}, N \in \mathbb{N})$ *converges to the solution of*

$$X'(t) = sX(t)(1 - X(t)).$$

The Wright-Fisher diffusion was used for example by Motoo Kimura [KO69] to compute the time to fixation of an allele. For the sake of simplicity, we will only present this result for the neutral case (i.e. when $s = 0$). Let $(X_t, t \geq 0)$ solution of (3) and Q its infinitesimal generator. Let

$$\tau = \min\{t, X_t = 0 \text{ or } X_t = 1\}$$

For $x \in [0, 1]$, define $g(x) = \mathbb{E}(\tau | X_0 = x)$. It is not hard to prove that

$$Qg(x) = -1 \text{ and } g(0) = 0. \tag{4}$$

By solving this differential equation, it can be shown that

$$g(x) = -2(x\log(x) + (1-x)\log(1-x)). \tag{5}$$

In addition, as a corollary of Proposition 2.2, it can be shown that

$$\lim_{N\to\infty} \mathbb{E}(\tau_N) = \mathbb{E}(\tau),$$

see [EK05], Corollary 2.4, Chapter 10, for a proof of this result. Consider a population of size $N \gg 1$ where all individuals carry allele a, except one mutant of type A. We have $X_0^N = 1/N \ll 1$. Taking into account the fact that time is measured in units of N generations, replacing into (5), we get

$$\mathbb{E}(\tau^N) \simeq \frac{2}{N} \times N = 2.$$

Kimura and Ohta [KO69] showed that, when the process is conditioned to the fixation of A (which happens with probability $1/N$), then

$$\mathbb{E}(\tau^N | X_{\tau^N} = 1) \simeq 2 \times \frac{1}{1/N} = 2N.$$

In words, the time it takes for a neutral allele to be fixed in the population is of the order of the population size.

Remark 2.5. *To model a population of N diploid individuals, where each individual carries two copies of each chromosome, one can consider a Wright-Fisher model with population size $2N$.*

2.2 The Moran model

This model was proposed by Patrick A.P. Moran in 1958. It is a continuous time analogous to the Wright-Fisher model. The hypothesis are the same as in the Wright-Fisher model except that the generations are overlapping.

Definition 2.6. *In the neutral Moran model, each individual reproduces at rate 1. She produces an offspring, which is a copy of herself, that replaces a randomly chosen individual in the population (who dies simultaneously).*

Again, consider a single gene with two alleles, A and a. Let Y_t^N the fraction of the population carrying allele A at time t. The reproduction events between individuals of the same type do not change the genetic composition of the population, so we have the

following transition rates for (Y_t^N)

$$\begin{cases} \frac{i}{N} & \to \frac{i+1}{N} \quad \text{at rate } N\frac{i}{N}(1-\frac{i}{N}) \\ \frac{i}{N} & \to \frac{i-1}{N} \quad \text{at rate } N\frac{i}{N}(1-\frac{i}{N}). \end{cases} \tag{6}$$

Remark 2.7. *In the Moran model, the total reproduction rate is N, but it takes at least N reproduction events to replace the whole population. So one time unit corresponds to one generation in the Wright-Fisher model.*

It is also possible to add selection into this model. Assume allele A is favoured by natural selection and its relative fitness is $s > 0$. Then we assume that individuals of type A reproduce at rate $1 + 2s$ instead of 1. Then the transition rates become

$$\begin{cases} \frac{i}{N} & \to \frac{i+1}{N} \quad \text{at rate } (1+2s)N\frac{i}{N}(1-\frac{i}{N}) \\ \frac{i}{N} & \to \frac{i-1}{N} \quad \text{at rate } N\frac{i}{N}(1-\frac{i}{N}) \end{cases} \tag{7}$$

Time is accelerated by $N/2$ and we let Q^N be the infinitesimal generator of the process $(Y^N_{Nt/2}; t \geq 0)$. We have

$$\begin{aligned} Q^N f\left(\frac{i}{N}\right) &= (1+2s)\frac{N}{2}N\frac{i}{N}(1-\frac{i}{N})\left(f\left(\frac{i+1}{N}\right)-f\left(\frac{i}{N}\right)\right) \\ &\quad +\frac{N}{2}N\frac{i}{N}(1-\frac{i}{N})\left(f\left(\frac{i-1}{N}\right)-f\left(\frac{i}{N}\right)\right) \end{aligned}$$

Again, assume that in the Moran model with population size N, the fitness rate is s_N, such that $Ns_N \underset{N\to\infty}{\longrightarrow} s$. For every function f that is at least twice differentiable in $[0,1]$, using Taylor expansions, we have:

$$\begin{aligned} Q^N f(x) &= (1+2s_N)\frac{N^2}{2}x(1-x)\left(\frac{1}{N}f'(x)+\frac{1}{2N^2}f''(x)+o(1/N^2)\right) \\ &\quad +\frac{N^2}{2}x(1-x)\left(\frac{-1}{N}f'(x)+\frac{1}{2N^2}f''(x)+o(1/N^2)\right) \\ &\longrightarrow Qf(x)\ f'(x)sx(1-x)+\frac{1}{2}f''(x)x(1-x) \end{aligned}$$

which is the generator of the diffusion process that corresponds to the solution of (3). As in the case of the Wright-Fisher model, it can be proved properly that, if $Y^N = x_0$, for $T > 0$, the sequence of processes $(Y^N)_{N\geq 1}$ converges in distribution, in the Skorokhod topology $D([0,T],\mathbb{R})$ to the solution of (3).

Remark 2.8. *In the Wright-Fisher model we had to accelerate time by a factor N to obtain the convergence to the Wright-Fisher diffusion. In the Moran model we had to accelerate*

time by $N/2$. From a biological point of view, this means that differences in the breeding structure of a population can lead to differences in the timescale of changes in the population (see [Wak16], Chapter 3).

2.3 The Kingman coalescent

Another important line of research in population genetics consists in tracing backwards in time the ancestry of a population. Instead of looking at how the genetic composition of the population evolves, we look at how individuals are related to one another, i.e. for each pair of individuals, we want to know how many generation ago lived their last common ancestor. The Kingman coalescent was introduced by John F.C. Kingman in 1982 [Kin82] to describe the genealogy of a panmictic population, of constant size, made of haploid individuals, where mating is uniformly random.

We start by considering a population of size N. The N-Kingman coalescent is the process valued in $\mathcal{P}_N$, the set of partitions of $\{1, \dots, N\}$, such that two lineages are in the same block at time t if they share a common ancestor at time t. See Figure 2) for a representation of the N-Kingman coalescent.

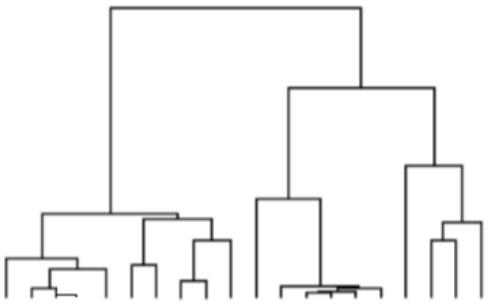


Figure 2 – Example of realization of the N-Kingman coalescent ($N = 20$). Time goes from bottom to top and each line represents a lineage.

Definition 2.9. *(i) The N-Kingman coalescent is the process valued in $\mathcal{P}_N$ such that Π_0 is the partition made of singletons, and each pair of blocks merges (or coalesces) at rate 1.*

(ii) The standard Kingman coalescent is the process valued in the partitions of $\mathbb{N}$ such that for each N its restriction to $\{1, \dots, N\}$ is a N-Kingman coalescent.

Let us now recall two important properties of the Kingman coalescent:

— *(Exchangeability).* The distribution of the Kingman coalescent is invariant under finite permutation.

— *(Consistency).* For $L > N$, a L-Kingman coalescent restricted to $\mathcal{P}_N$ is a N-Kingman coalescent. In particular, if one has a sample of L individuals, and restricts the genealogical tree relating them to a subsample of N individuals, the genealogical tree that is obtained is the same in distribution as if one had taken the smaller sample since the beginning.

The Kingman coalescent and the neutral Wright-Fisher diffusion (i.e. when $s = 0$) are linked, in the sense that the Wright-Fisher diffusion describes the evolution of a population forwards in time and the Kingman coalescent describes backwards in time the ancestry relations between the individuals of the same population. Mathematically, we can formalize this fact by means of a duality relation. Duality is a powerful mathematical tool to obtain information about one process by studying another process, its dual.

Definition 2.10. *Two Markov processes X and Y, with laws $\mathcal{X}$ and $\mathcal{Y}$ that take values in E and F respectively are said to be dual with respect to a bounded measurable function function h on $E \times F$ if for all $x \in E$, $y \in F$, $t \geq 0$,*

$$\mathbb{E}_{\mathcal{X}}(h(X_t, y)|X_0 = x) \;=\; \mathbb{E}_{\mathcal{Y}}(h(x, Y_t)|Y_0 = y).$$

Let N_t be the block counting process of the Kingman coalescent. N_t is a pure death process, where, the transition rate from i to $i-1$ is given by $\binom{i}{2}$. Let $(X_t;\ t \geq 0)$ be the solution of

$$\sqrt{x_t(1-x_t)}dB_t.$$

Proposition 2.11 (Duality between the Wright-Fisher diffusion and the Kingman coalescent)**.**

$$\mathbb{E}(h(X_t, k)|X_0 = x) \;=\; \mathbb{E}(h(x, N_t)|N_0 = k), \quad \textit{where} \;\; h(x, k) = x^k.$$

In the LHS $\mathbb{E}$ denotes the expectation with respect to the distribution of the Wright-Fisher diffusion and in the RHS $\mathbb{E}$ denotes the expectation with respect to the distribution of the Kingman coalescent.

Proof. We consider Q, the infinitesimal generator of the Wright-Fisher diffusion (see (8)). We have

$$\begin{aligned} Qh(x,k) &= \frac{1}{2}x(1-x)\frac{\partial^2 h}{\partial x^2}(x,k) \\ &= \frac{1}{2}(1-x)k(k-1)x^{k-1} \\ &= \binom{n}{2}\left(h(x,k-1) - h(x,k)\right) \\ &= \bar{Q}h(x,k), \end{aligned}$$

where $\bar{Q}$ is the infinitesimal generator of the Kingman coalescent. In the first line Q acts

on h seen as a function of the first variable (x), whereas in the last line, $\bar{Q}$ acts on h seen as a function of the second variable (k). We consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ in which, under $\mathbb{P}$, $(N_t, t \geq 0)$ and $(X_t; t \geq 0)$ are independent. The previous computation implies that

$$\frac{d}{ds}\mathbb{E}(h(X_s, N_{t-s})) \;=\; \mathbb{E}\left(Qh(X_s, N_{t-s}) - \bar{Q}h(X_s, Y_{t-s})\right) \;=\; 0,$$

and the conclusion follows by integrating between 0 and t. □

Now consider the Moran model for finite population of size N and the N-Kingman coalescent.

Proposition 2.12 (Duality between the Moran and the N-Kingman coalescent)**.** *For $i \geq k$, define $H_N(i,k) = \binom{i}{k} / \binom{N}{k}$*

$$\mathbb{E}(H_N(NX^N_{Nt/2}, k)|X_0 = x) \;=\; \mathbb{E}(H_N(Nx, N_t)|N_0 = k).$$

where in the LHS, $\mathbb{E}$ denotes the expectation with respect to the distribution of the Moran model and in the RHS, $\mathbb{E}$ denotes the expectation with respect to the distribution of the N-Kingman coalescent (backwards in time).

This result can be interpreted as follows: if, at time t, one samples k individuals from a population of size N, the probability that they are all of type A is the same as the probability that their ancestors at time 0 were of type A. The proofs of these results can be found in [M01, M99].

3 Multi-locus models and Recombination

3.1 Why are they important?

The models presented above consider the evolution of one single locus. However, what are transmitted from one generation to another are chromosomes, or blocks of chromosomes, so the evolutionary histories of the different loci carried by an individual are not independent. Generalizing these models from one to 2 (or n) loci is not trivial, as one needs to take into account genetic recombination.

In fact, during meiosis (which is one of the steps of sexual reproduction), homologous chromosomes are paired and can exchange fragments by a mechanism called chromosomal crossover (see Figure 4). Thanks to this recombination mechanism, the offspring can inherit chromosomes that are either copies of one of the parental chromosomes or mosaics of the two parental chromosomes. Recombination is a widespread mechanism that is not

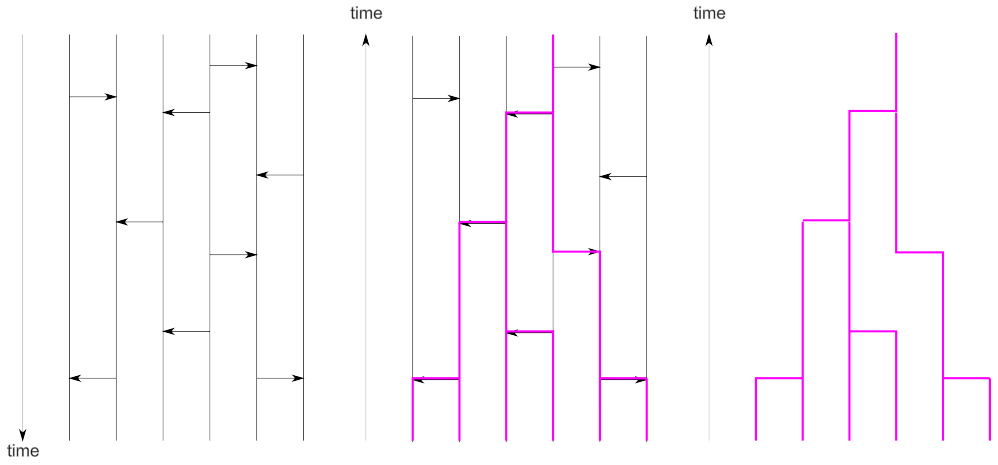


Figure 3 – Duality between the Moran model and the N-Kingman coalescent ($N = 6$). The left-hand side represents a realization of the Moran model. Time goes from top to bottom and an arrow represents a birth event in which the individual that is at the base of the arrow replaces the individual at the tip of the arrow. In the second panel we show how the N-Kingman coalescent can be obtained by reversing time. We start from the individuals in the present population and we follow their ancestral lineage: each time it finds the tip of an arrow, the lineage jumps to the base of the arrow. Arrows indicate coalescence events. The right-hand side panel represents the N-Kingman coalescent.

only present in sexual organisms. For example, bacteria have their own mechanisms of gene transfer and homologous recombination that allow them to exchange portions of their genomes.

All these mechanisms are complex and costly. Trying to explain how the recombination mechanisms emerged and why they are maintained has been an important line of research in evolutionary biology [Mul64, Fel74, OG06, BC98, Bar10]. If asked why sex and recombination are such widespread mechanisms, most biologists would say that it increases genetic variability and hence allows evolution to proceed faster. However, recombination does not allow to produce new alleles, so the link between genetic variance and recombination is not clear. In addition, recombination can break up favorable associations between alleles that have been accumulated by selection (this is known as the "recombination load").

One of the hypotheses that has been favored to explain the maintenance of recombination is finite population size and genetic drift [OG06, Bar10]. In fact, as it was already pointed out by Fisher [Fis30] and Muller [Mul64], mutation is rare, so different favorable mutations will tend to arise in different individuals. In asexual populations, favorable mutations have to be fixed one by one (see Figure 7), whereas in sexual populations, recombination can bring them together, so several favorable mutations can be fixed at a time. In addition, in the absence of recombination, a mutation that arises in an individual

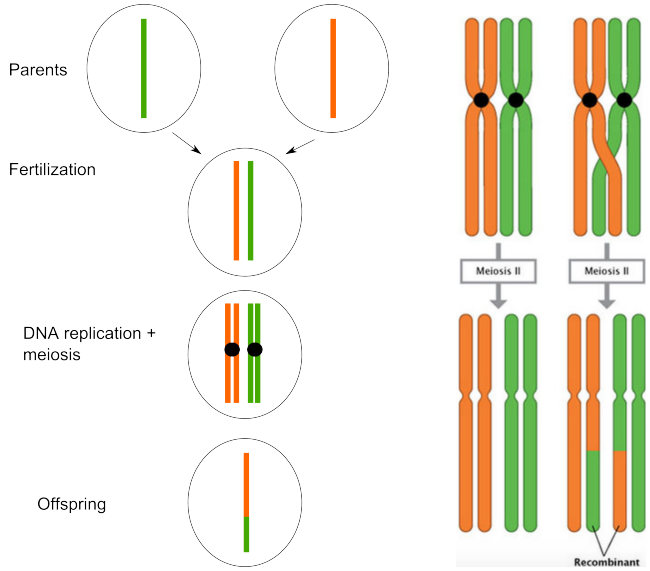


Figure 4 – The left-hand side represents an example of a haploid life cycle. The right-hand side comes from [LS08] and represents the mechanism of a crossing-over.

that carries deleterious mutations at other loci will tend to be lost. In contrast, recombination can allow to bring favorable mutations into good genetic backgrounds and therefore increase the rate of evolution. Finally, in asexual populations there is a substantial probability that all fittest individuals will eventually acquire a slightly deleterious mutation and therefore go extinct, so that only "second fittest" individuals survive. The population can therefore accumulate several deleterious mutations resulting in a global reduction in fitness. This mechanism is known as Muller's ratchet [Mul64, Bar09].

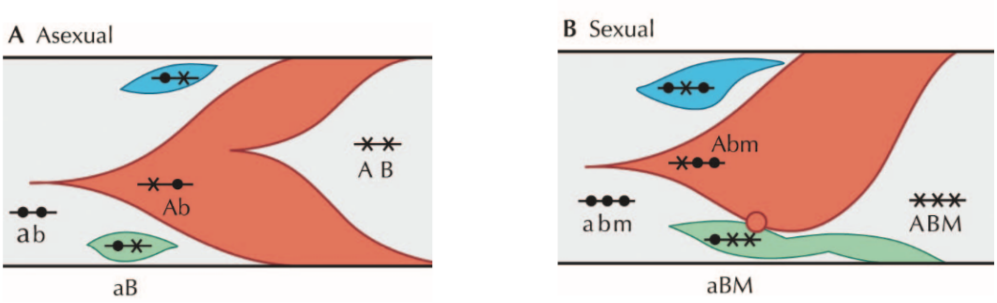


Figure 5 – In an asexual population (A), Favorable mutations must be established sequentially. If allele A is going to be fixed in the population, then any favorable allele that occur at other loci (B, for instance) can only be fixed if they occur within a genome that already carries A. (B) Recombination allows favorable mutations at different loci to be combined: a favorable allele B that occurs with the unfavorable allele a can be fixed if it can recombine into association with A (red circle). From [Bar09].

But recombination is not only important form an evolutionary point of view. From a modeler's point of view, models that take into account recombination are complex but can be very powerful. In fact, nowadays with the advent of new generation sequencing techniques it has become usual to have access to whole genomes. Considering that loci have evolved independently results in an important loss of statistical power and can lead to incorrect inferences. Multi-locus models that take into account linkage disequilibrium have become of particular interest to analyse data. As we shall see in Section 3.5, analysing recombination patterns can be useful for the detection of selection [SRH$^+$02], to study recent demography [RCB17], in candidate gene studies [Cla04] or to analyse data from experimental evolution [TEPB17]. In Chapter II we use recombination patterns to study hybrid populations.

3.2 Two-locus models

In this section we are going to define a multi-locus version of the Wright-Fisher and the Moran model described in 2.1 and 2.2. We will start by considering the case of two loci. Locus 1 has two alleles, A_1 and A_2, and locus 2 has two alleles B_1 and B_2. The population size is N, the recombination rate between the two loci is ρ_N and all the alleles are neutral, so $s = 0$ for any genotype. We denote by

$$\begin{aligned} X_{11}^N &\text{ the frequency of genotype } A_1B_1 \\ X_{12}^N &\text{ the frequency of genotype } A_1B_2 \\ X_{21}^N &\text{ the frequency of genotype } A_2B_1 \\ X_{22}^N &\text{ the frequency of genotype } A_2B_2 \end{aligned}$$

Recall, that, at each time t we have

$$X_{11}^N(t) + X_{12}^N(t) + X_{21}^N(t) + X_{22}^N(t) = 1.$$

Finally, we denote by X^N the vector $(X_{11}^N, X_{12}^N, X_{21}^N, X_{22}^N)^T$.

Definition 3.1. *In the Wright-Fisher model with recombination, each individual from generation $t+1$ chooses two parents uniformly (and independently) at random from generation t. With probability $1-\rho_N$, she inherits the alleles at both loci from one of the parents (chosen at random). With probability ρ_N, she inherits allele at locus 1 from one parent (chosen at random) and allele at locus 2 from the other parent.*

Fix $i, j \in \{1, 2\}$. Considering the different ways an individual A_iB_j can be formed, we

have

$$\begin{aligned}\mathbb{E}(X_{ij}^N(t+1) - X_{ij}^N)(t) \mid X^N(t) = (x_{ii}, x_{ij}, x_{ji}, x_{jj})^T) \\ = (1-\rho_N)x_{ij} + \rho_N(x_{ij}^2 + x_{ii}x_{jj} + x_{ii}x_{ij} + x_{ij}x_{jj}) - x_{ij} \\ = \rho_N(x_{ii}x_{jj} - x_{ij}x_{ji}).\end{aligned}$$

In addition, conditional on $X^N(t) = (x_{ii}, x_{ij}, x_{ji}, x_{jj})$, $X^N(t+1)$ follows a multinomial distribution of parameters N and $(x_{ii}, x_{ij}, x_{ji}, x_{jj})$, which gives

$$\mathrm{Var}(X_{ij}^N(t+1) \mid X^N(t) = (x_{ii}, x_{ij}, x_{ji}, x_{jj})^T) = \frac{1}{N}x_{ij}(1-x_{ij})$$
$$\forall (i',j') \neq (i,j),\ \mathrm{Cov}(X_{ij}^N(t+1), X_{i'j'}^N(t+1) |\ X^N(t) = (x_{ii}, x_{ij}, x_{ji}, x_{jj})^T) = -\frac{1}{N}x_{ij}x_{i'j'}$$

Assume that $X_0^N = x_0$ and

$$N\rho_N \underset{N\to\infty}{\longrightarrow} \rho.$$

Proposition 3.2. *For any $T > 0$, $(X_{\lfloor Nt \rfloor}^N)_{N\in\mathbb{N}}$ converges in distribution (in the Skorokhod topology $D([0,T], \mathbb{R}^4)$ to the solution of*

$$\begin{cases} dX_t &= \rho(-1,1,1,-1)^T D(X(t))dt + \sigma(X(t))dB_t, \\ X_0 &= x_0. \end{cases} \tag{8}$$

where B is a standard Brownian motion in $\mathbb{R}^4$. and

$$D(X) = X_1X_4 - X_2X_3,$$

and $\sigma(X)\sigma(X)^T = M(X)$ where

$$\begin{aligned}\forall i,j \in \{1,2,3,4\}, i \neq j, \quad M(X)_{i,i} &= X_i(1-X_i) \\ M(X)_{i,j} &= -X_iX_j.\end{aligned}$$

We let the reader refer to [EN89] for a formal proof or to [Dur08] for a modern exposition of this result.

Remark 3.3. *D is called the "linkage disequilibrium". In fact, if X_{1-}^N is the total frequency of allele A_1 ($X_{1-}^N = X_{11}^N + X_{12}^N$) and X_{-1}^N is the total frequency of allele B_1 ($X_{-1}^N = X_{11}^N + X_{21}^N$), we have*

$$\begin{aligned}D(X^N) &= X_{11}^N - (X_{11}^N + X_{12}^N)(X_{11}^N + X_{21}^N) \\ &= X_{11}^N - X_{1-}^N X_{-1}^N.\end{aligned}$$

Linkage disequilibrium is a measure of the non-random association between alleles A_1 and B_1. From (8), we have

$$D(X^N(t)) \underset{t\to\infty}{\longrightarrow} 0,$$

but the rate of convergence depends on ρ. Therefore, in a neutral setting, linkage disequilibrium should decrease with time and with distance in the chromosome. However patterns of linkage disequilibrium can be affected by many factors such as population subdivision, demographic bottlenecks or natural selection (see [Sla08] or Section 3.5).

Similarly, we can define a Moran model with recombination in the following way

Definition 3.4. *In the Moran model with recombination each individual reproduces at rate 1. She chooses a random partner in the population.*

- *With probability $1-\rho_N$, the chromosome of the offspring is a copy of one of the two parents (chosen at random).*
- *With probability ρ_N, there is a crossing over between these two loci. The offspring copies the allele at one locus from one parent and the allele at the other locus from the other parent.*

Define $e_{11} = (1,0,0,0)^T$, $e_{12} = (0,1,0,0)^T$, $e_{21} = (0,0,1,0)^T$, $e_{22} = (0,0,0,1)^T$. Again, we accelerate time by $N/2$ and we assume that

$$\frac{N}{2}\rho_N \underset{N\to\infty}{\longrightarrow} \rho.$$

Let Q^N be the infinitesimal generator of the Markov process $(X^N(Nt/2);\ t\geq 0)$. For any function f at least twice differentiable, if $x=(x_{ij}))_{i,j\in\{1,2\}}$,

$$\begin{aligned}
Q^N f(x) = {} & \frac{N^2}{2}(1-\rho_N)\sum_{i,j\in\{1,2\}}\sum_{k,p\in\{1,2\}} x_{ij}x_{kp}\left(f(x+\frac{1}{N}e_{ij}-\frac{1}{N}e_{kp})-f(x)\right)\\
& + \rho_N\frac{N^2}{2}\sum_{i,j\in\{1,2\}}\sum_{k,p\in\{1,2\}} x_{ij}x_{kp}\left(f(x+\frac{1}{N}e_{ip}-\frac{1}{N}e_{kj})-f(x)\right)\\
\underset{N\to\infty}{\longrightarrow} {} & \rho\sum_{i,j\in\{1,2\}}\sum_{k\neq i,p\neq j}(x_{ip}x_{jk}-x_{ij}x_{kp})\frac{\partial f}{\partial x_{ij}}(x)\\
& + \frac{1}{2}\sum_{(i,j),(k,p)\in\{1,2\}} x_{ij}(\mathbb{1}_{i=k,j=p}-x_{kp})\frac{\partial^2 f}{\partial x_{ij}\partial x_{kp}}(x).
\end{aligned}$$

where the last equality is obtained by means of Taylor expansions (as in the one-locus Moran model). The last line corresponds to the generator of the diffusion process defined in (8). As in the case of the simple Moran model, it can be shown that the sequence $(Y^N, N\geq 1)$ is tight in $D([0,T],\mathbb{R}^4))$ converges in distribution, in the Skorokhod topology $D([0,T],\mathbb{R})$ to the solution of (8).

3.3 The case of n loci

These two models can be extended to the case of n loci. In this work, we will only consider single crossing over recombination which means that at each reproduction event, if recombination takes place, the chromosome of each parent is partitioned into two segments. The offspring inherits the genetic material to the right of the cutpoint from one parent and the genetic material from the left of the cutpoint from the other parent. The position of the cut point is uniformly distributed in the chromosome, so the probability that the crossing over occurs between two given loci depends on the distance between these two loci on the chromosome.

The n loci are identified to their positions in a chromosome, which are given by $z = (z_1, \dots, z_n) \in [0,1]$ such that $z_1 = 0 < z_2 < \dots < z_n = 1$.

Definition 3.5. *In the Wright-Fisher model with recombination, each individual from generation $t+1$ chooses two parents uniformly (and independently) at random from generation t.*

- *With probability $1 - \rho_N$ there is no recombination and she inherits all the loci from one of the parents (chosen uniformly at random).*
- *For $i \in \{2, \dots, n\}$, with probability $\rho_N(z_i - z_{i-1})$ the offspring inherits loci $z_1, \dots, z_{i-1}$ from one parent and $z_i, \dots, z_n$ from the other one.*

Definition 3.6. *In the Moran model with recombination each individual reproduces at rate 1 and chooses a random partner*

- *With probability $1 - \rho_N$ there is no recombination and the offspring inherits all the loci from one of the parents (chosen at random).*
- *For $i \in \{2, \dots, n\}$, with probability $\rho_N(z_i - z_{i-1})$ the offspring inherits loci $z_1, \dots, z_{i-1}$ from one parent and $z_i, \dots, z_n$ from the other one.*

The offspring replaces a randomly chosen individual in the population, who simultaneously dies.

As in the case of two loci, when time is rescaled by N and the recombination rate scales with the population size in such a way that $\rho = \lim_{N\to\infty} \rho_N N$ the Wright-Fisher model with recombination has a diffusive limit. For the Moran model with recombination this diffusive limit arises when time is rescaled by $N/2$ and $\rho = \lim_{N\to\infty} \rho_N N/2$.

We follow closely [GJL16] and we assume that each locus i has k possible alleles $\{i_1, \dots, i_k\}$. The set of al possible genotypes is $E = \prod_{i=1}^{n} \{i_1, \dots, i_k\}$. For each $e \in E$, we denote by x_e the frequency of genotype e in the population. Let $\mathcal{S}$ be the set of non-empty subsets of $\{1, \dots, n\}$. For $S \in \mathcal{S}$, we denote by x_e^S, the marginal frequency of the alleles

e into S. In particular, for $\ell \in \{1,\dots,n\}$, we denote by $x_e^{\leq l}$ (resp. $x_e^{>l}$) the marginal frequency of the alleles e into $\{1,\dots,\ell\}$ (resp. $\{\ell+1,\dots,n\}$

$$x_e^{\leq l} = \sum_{j\in E,\ j|_{E_{\leq \ell}}=e} x_j, \quad x_e^{\leq l} = \sum_{j\in E,\ j|_{E_{>\ell}}=e} x_j$$

where $E_{\leq \ell} = \prod_{i\leq \ell}\{i_1,\dots,i_k\}$ and $E_{>\ell} = \prod_{i>\ell}\{i_1,\dots,i_k\}$. The generator of the Wright-Fisher diffusion for n recombining loci with k alleles per loci is given by

$$L = \sum_{e\in E}\left(\sum_{\ell=1}^{n-1}\rho(z_{\ell+1}-z_\ell)(x_e^{\leq l}-x_e^{\leq l})\frac{\partial}{\partial x_e} + \sum_{j\in E} x_e(\mathbb{1}_{e=k}-x_j)\frac{\partial^2}{\partial x_e \partial x_j}\right). \tag{9}$$

Remark 3.7. *In Chapter III, we consider a slightly different version of the Moran model with recombination, which is more general because we do not assume single crossing-over recombination. We assume that each individual carries a chromosome of length* 1*. The positions of the cutpoints are given by a Poisson Point Process of intensity* λdx*. The chromosomes of each of the parents are cut into fragments at these positions . The offspring inherits each fragment of the chromosome from one of the two parents, chosen uniformly at random, so the probability of observing a crossing-over at a given cutpoint is* $1/2$*. The probability that a crossing-over occurs between* z_i *and* z_j *is then given by*

$$r_{i,j} \ = \ \frac{1}{2}(1-\exp(-\lambda|z_i-z_j|).$$

This model is known as the Haldane model. For the purpose of Chapter I, this model is too complex to handle, and we use the single crossing-over model, which is a good approximation of this one, if one looks at a portion of chromosome that is small enough so that the probability of observing more than one crossing over at each reproduction event is negligible.

Remark 3.8. *The parameter* ρ_N *corresponds to a recombination rate. If we rescale the positions of the loci in such a way that* $\bar{z}_i = \rho_N z_i$*, this corresponds to measuring the chromosome in units of recombination, or "morgans" (this unit was named in honor of Thomas H. Morgan).*

3.4 The Ancestral Recombination Graph and the partitioning process

In the previous section we showed how the genealogy for a single locus can be described using the Kingman coalescent. In this section, we consider the Ancestral Recombination Graph (ARG) which follows backwards in time the ancestry of several recombining loci. The ARG was introduced by Hudson [Hud83] and Griffiths [Gri81, Gri91]. The idea is to sample n loci from an individual in the present population and to follow backwards in time the lineages carrying each of these loci. The dynamics of these lineages are controlled by splitting and coalescence events.

We start with considering the Moran model with recombination for a population of size N and recombination rate ρ_N. We follow the ancestry of n loci whose positions in are given by $z = (z_1, \dots, z_n) \subset [0,1],\ z_1 < \dots < z_n$. The N-ARG for the set of loci z has the following transition rates:

- **Coalescence:** Each pair of lineages coalesces at rate $(1-\rho_N)^2 \frac{2}{N} + O(\rho_N/N)$. Forwards in time, it corresponds to a birth event in which there is no recombination and individual i replaces individual j, which happens with probability $1/N$ (or a birth event in which individual j replaces individual i, which happens with probability $1/N$) .
- **Splitting:** The lineage carrying $z_{i_1}, \dots z_{i_k}$ is split into $z_{i_1}, \dots, z_{i_j}$ and $z_{i_{j+1}}, \dots z_{i_k}$ at rate $\rho_N(z_{i_{j+1}} - z_{i_j})$. Forwards in time this corresponds to a reproduction event in which the offspring inherits loci $z_{i_1}, \dots . z_{i_j}$ from one parent and loci $z_{i_{j+1}}, \dots z_{i_k}$ from the other parent.
- Events in which two blocks coalesce and the resulting block is split simultaneously happen at rate $O(\rho_N/N)$ (and when N is large enough, we can neglect them).

It can readily be seen that, if $\lim_{N\to\infty} \rho_N N/2 = \rho$ and time is rescaled by $N/2$, when $N \to \infty$ this process converges to a process, known as the ARG, in which pairs of lineages coalesce at rate 1 and a block is split between z_i and z_j at rate $\rho|z_i - z_j|$.

Remark 3.9. *Similarly, in the Wright- Fisher model with recombination, if we follow backwards in time the ancestry of a sample of n loci, we have the following transition probabilities*

- *At each generation, each pair of lineages coalesces with probability* $(1-\rho_N)^2/N$. *This corresponds to individuals i and j choosing the same parent (with happens with probability* $1/N^2$ *and there are N possible parents)*
- *At each generation, the lineage carrying* $z_{i_1}, \dots z_{i_k}$ *is split into* $z_{i_1}, \dots, z_{i_j}$ *and* $z_{i_{j+1}}, \dots, z_{i_k}$ *with probability* $\rho_N(z_{i_{j+1}} - z_{i_j})$.
- *At each generation, the probability that there is more than one coalescence or recombination event is* $O(\rho_N/N)$

If time is rescaled by N and $\lim_{N\to\infty} \rho_N N = \rho$, *this process also converges to the ARG.*

In a finite population, it can readily be seen from the graphical representation (Figure 6) that the N-ARG is dual to the Moran model with recombination. The ARG (in an infinite population) is dual to the Wright-Fisher diffusion with recombination. Griffiths et al. [GJL16] showed that, for $S \in \mathcal{S}$, if n^S is the number of lineages that carry genetic material

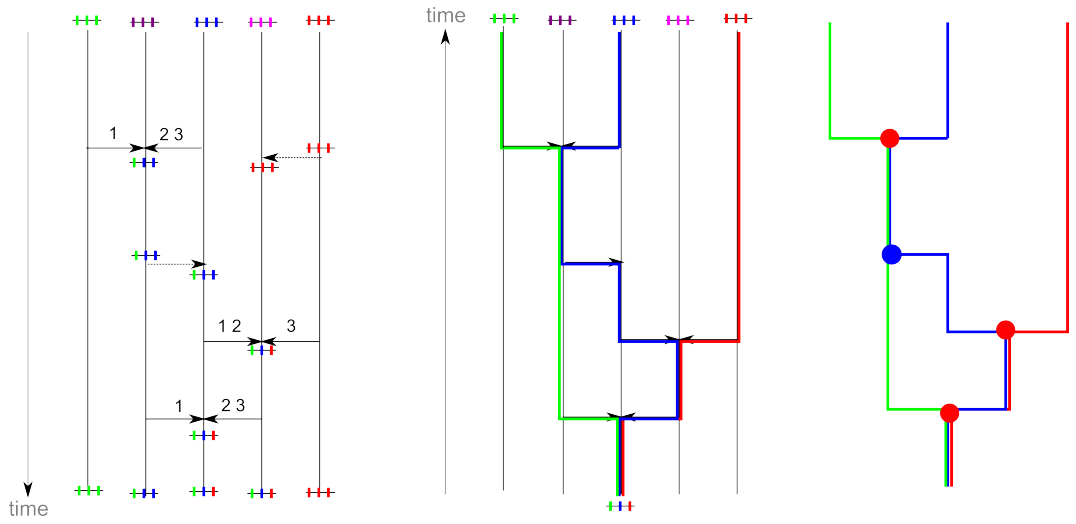


Figure 6 – Duality between the Moran model and the N- ARG ($N = 5,\ n = 3$). The left panel represents a realization of the Moran model with recombination. On the top we represented the chromosomes of the different individuals at generation 0, assuming that each one has her loci painted in a different color (which allows us to distinguish them). Arrows represent reproduction events. For pure resampling events, the base of the arrow represents the parent from which the genetic material is inherited, and the arrow points at the individual that is replaced. For reproduction events with recombination (represented by two arrows), we indicated on top of each arrow the loci that are inherited from each parent. In the second panel we show how the N-ARG can be obtained by reversing time. We sample the 3 loci in an individual in the present population and we follow the ancestral lineages corresponding to each of her loci. Each time an ancestral lineage finds the tip of an arrow, it jumps to the base of the arrow. If there are two arrows, the lineage corresponding to locus i jumps to the base of the arrow labelled i. The right-hand side panel represents the resulting N-ARG. Red dots represent splitting events and blue dots represent coalescence events.

that is ancestral to the loci in S of a randomly sampled individual in the population, the diffusion process whose generator is defined in (9) is dual to the ARG with duality function

$$F(x,n) = \prod_{S\in\mathcal{S}} (x_j^S)^{n^S}.$$

In Chapter I, we extend the ARG to the case of a continuous, possibly infinite chromosome. We call this process the $\mathbb{R}$-partitioning process. Let $\mathcal{P}^{loc}$ be the set of partitions of $\mathbb{R}$ that are locally finite and right continuous, i.e. such that the blocks of the partition are disjoint unions of left-closed (right-open) intervals and such that in any compact subset of $\mathbb{R}$ there is only a finite number of these intervals. For any z finite subset of $\mathbb{R}$, for any $\pi \in \mathcal{P}^{loc}$, the trace of π on z, t_z, is the partition of z induced by π. The $\mathbb{R}$-partitioning process $(\Pi_t^\rho;\ t \geq 0)$ is the only càdlàg process valued in $\mathcal{P}^{loc}$ such that for any z finite subset of $\mathbb{R}$, $(t_z(\Pi_t^\rho);\ t \geq 0)$ is a partitioning process at rate ρ for the set of loci z (with the transition rates described above). In Chapter I, we will study the partitioning process in the limit $\rho \to \infty$. This corresponds to observing a frame of the chromosome of the order of $1/(2N)$ morgans and letting the size of the frame tend to infinity. A similar model was studied by Wiuf and Hein [WH97] to address the question of how many genetic ancestors there are to a contemporary human chromosome.

3.5 Applications: linkage disequilibrium, haplotype blocks and inference

Recombination patterns are of particular interest for analysing data from whole genome scans. Nowadays it has become feasible to have access to whole DNA sequences of individuals and models that take into account correlations between loci are needed to analyse these data. The sequence of a chromosome is called a *haplotype*. When comparing two haplotypes from two individuals of a population it is possible to detect blocks of loci that have been inherited from the same common ancestor. These blocks are called IBD blocks ("identical-by-descent"). The distribution of the IBD blocks in a chromosome, i.e. the lengths and the positions of the different segments of the chromosome that are IBD (to a locus of interest) can be studied using the partitioning process. This is the goal of Chapter I. This model assumes that the population size is constant and all the loci are neutral. However, changes in the population size, or natural selection at some loci can alter the distribution of the sizes of the IBD blocks.

Delimiting IBD blocks from real data is not straightforward. Unlike mutation, recombination events do not always leave a footprint on the DNA sequence. A recombination event can only be observed if it occurs between two loci where the parents carried different alleles (so it is possible to determine which fragment has been inherited from which parent). A single-nucleotide polymorphism (SNP) is a variation in a single nucleotide, at a

specific locus. The higher the density of SNPs the most accurately we can infer IBD blocks (see Chapter II). Different methods have been developed to infer IBD blocks. Algorithms such as fastIBD [BB11] or IBD_Haplo [BGZT12] identify long haplotype segments that are shared between two individuals by a combination of likelihood methods and Hidden Markov Models (HMM).

The distribution of IBD block lengths can be used to infer the recent demographic history of populations. Classical methods use mutation patterns to infer past demographic variation (see e.g. [PHN10] for a review on this topic). But mutation rates are usually too low to be used to detect fast demographic changes. Ralph and Coop [RC13] and Ringbauer et al. [RCB17] used IBD blocks to infer recent migration patterns in European human populations.

Recombination patterns can also be used to detect loci under selection. The idea behind these methods is that loci that are under selection tend to be fixed rapidly in a population. During a selective sweep, loci that are close to the locus under selection tend to be hitchhiked (i.e. the alleles that are in the same haplotype where the beneficial mutation arose also tend to be fixed, because recombination does not have time to break up the linkage). Therefore alleles that are under positive selection tend to be located within long IBD blocks. Some examples of these type of methods are Extended Haplotype Homozygosity [SRH$^+$02] or Runs of Homozygosity [MLAR$^+$08].

Janzen et al. [JNT18] used the IBD block length distribution to infer the time since admixture in hybrid populations. The idea is that by comparing the genotypes of the hybrids to those of the ancestral populations, one can infer which haplotype blocks have been inherited from each of the ancestral populations. As the blocks are split by recombination, their size tend to decrease with time (until fixation is reached), so the sizes of the blocks are informative about the admixture time. But the quality of the inference depends on the density and the positions of the SNPs (or markers) that segregate between the two ancestral populations. In [JNT18] the authors had assumed that the markers were regularly spaced and derived a formula to infer the admixture time from the number of junctions between blocks. In Chapter II I present some work in collaboration with Thijs Janzen in which I derived a formula to infer the admixture time using markers that are randomly distributed across the genome.

Recombination patterns can also be useful to analyse data from experimental evolution. For example, in the experiment by Teotonio et al. [TEPB17], individuals from different subpopulations of *C. elegans* have been crossed for several generations. By sequencing the offspring and comparing their haplotypes to those of the ancestral individuals, it is possible to infer which haplotype blocks have been inherited from each of the ancestral

subpopulations. One of the motivations behind I is to derive a neutral model of the IBD-block distribution to analyse this type of experiment.

4 Geographic structure and speciation

In *On the origin of species*, Darwin called species formation the "mystery of mysteries". He was perplexed by the clustering of individuals into discrete species and the absence of "transitional forms". When listing the drawbacks of his theory he wrote: "Why is not all nature in confusion instead of the species being, as we see them, well defined?".

It is not surprising that since then, the process of speciation has received an enormous amount of attention from evolutionary biologists. The speciation process is complex and difficult to understand from a theoretical point of view because there are many factors controlling the dynamics of speciation (mutation, geographic structure, migration, recombination, natural selection, sexual selection...). Chapter III is devoted to the study of a particular model of speciation that can be used to understand how the geographic structure of a population can promote species formation. In this section we will give some biological background and review some important models of speciation that will allow us to justify some of the hypotheses made in that chapter.

4.1 Geographic structure and genetic differentiation

Geographic structure is one of the main drivers of within species genetic variability: if the geographical range is larger than the typical dispersal rate of its individuals, a species can be structured into different local subpopulations with limited contact. On the contrary, migration allows the different subpopulations (or *demes*) to exchange genes and has an homogenising effect.

One of the first models that was proposed to explain how the geographic structure promotes genetic variability was Wright's stepping stone model [Wri43], which was later improved by Kimura [Kim53]. In this model, a population is divided into several demes which can exchange migrants with their nearest neighbours in $\mathbb{Z}$ or $\mathbb{Z}^2$. Wright proposed a statistical theory on how population differentiation should vary as a function of the migration rates between demes, which was called "Isolation by distance".

Malécot studied the case of a population in continuous space where individual dispersal is assumed to be normally distributed [Mal48] . He proposed a formula for the probability P that two individuals sampled at distance r on the real line have the same allele at a given locus

$$\frac{P(r)}{P(0)} \;=\; \exp(-r\sqrt{(2\mu)/\sigma}),$$

where μ is the mutation rate and σ is the dispersal coefficient (i.e. the standard deviation of the dispersal distribution). This formula is known as Malécot's formula and can be extended to $\mathbb{R}^2$ or $\mathbb{R}^3$.

Samuel Karlin analysed how migration patterns can influence genetic variability in a metapopulation [Kar82]. A metapopulation is a population that is formed by several demes connected by migration. The geographic structure of a metapopulation can be modelled by a graph in which the vertices represent the subpopulations. Two vertices are connected if the corresponding demes can exchange migrants and each edge is associated to a migration rate. Using a deterministic model, Karlin studied which geographic configurations promote genetic variability and speciation: for example he showed that, in the presence of selection, some geographic structures can promote speciation. This is for example the case of a metapopulation graph that is clustered, in the sense that it is formed from the union of several (almost) complete graphs connected by a limited number of edges. In Chapter III, we showed, using a stochastic model, that, even if the absence of selection, a clustered geographic structure promotes genetic differentiation which can lead to speciation.

4.2 The biological species concept and reproductive barriers

It is difficult and beyond the scope of this thesis to give a universal definition of a species. For sexual organisms, one of the most commonly used definitions was given by Mayr in 1942 [May42]:

"Species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups."

This defines the *biological species concept*. A more general definition of this concept, based on evolutionary considerations, was given by de Queiroz in 1998 [dQ98]:

"Species are separately evolving metapopulation lineages ; they form an independent gene pool and reproductive community that evolves together."

But there are many other ways to define a species. For example:

- A *phenotypic* species is a morphologically distinguishable group of individuals.
- A *phylogenetic* species is a monophyletic group of individuals i.e. a group that consists of all the descendants of a common ancestor.
- An *ecological* species is a group of individuals that occupy the same niche, i.e. that are adapted to a particular set of resources in the environment.

Although all these definitions do not always lead to the same classification, they all have some advantages and are used in different contexts ([dQ05]). In this work we are going to consider sexual organisms and adopt the biological species concept, which is one of the most commonly used in evolutionary biology.

The biological species concept focuses on the capacity of individuals to interbreed, which means that they can mate and produce viable and fertile offspring. Speciation can be seen as the emergence of mechanisms that prevent individuals from different groups to interbreed. These mechanisms are called *reproductive barriers*. Among these mechanisms we can distinguish between:

- *Prezygotic barriers*, that are mechanisms that prevent fertilization. They include ecological mechanisms and habitat or behavioural differences that prevent mating. For example the American toad and the Fowler's toad are closely related species that live in the same areas of North America but that are unable to reproduce because their mating season is different [Bla41]. They also include anatomical differences and gametic incompatibility. For example, in the *Drosophila* genus, the differences in the shape of the genital organs prevent mating between individuals from different species [Mas12].
- *Postzygotic barriers*, that are mechanisms that prevent the development of hybrids. They include embryo inviability, sterility and reduced fitness of hybrids. Mules, that are hybrids between a donkey and a horse, are a classic example of hybrid sterility.

These mechanisms are controlled genetically and different models have been proposed to study how these barriers emerge and are maintained.

• The Dobzhansky-Muller model

This model was proposed independently by Dobzhansky [Dob37] and Muller [Mul42]. It involves two loci in diploid individuals.

The model assumes that, in an ancestral population, all individuals carry the same genotype $AABB$. The population is split into two subpopulations and gene flow is interrupted (for example, by a geographic barrier). In the first one, a mutation is fixed in the first locus and the genotype of the population becomes $aaBB$. In the second one, a mutation is fixed in the second locus and the the genotype of the population becomes $AAbb$. Fixation is possible because $AABb$ and $AaBB$ heterozygotes are viable and their fitness is similar to the fitness of the homozygotes. In a cross between two parents from different subpopulations, the genotype of the offspring would be $aAbB$. If these hybrids are non viable, a postzygotic reproductive barrier has emerged.

In this model, speciation may be adaptive or not. a and b can be neutral or each one may confer a selective advantage in the environment where it emerges. In the second case, fixation should be faster.

Different biological mechanisms have been proposed for these Dobzhansky-Muller incompatibilities. For example, they can involve a maternal and a paternal gene, both

involved in reproduction. One can code for a sperm protein and the other one for an egg protein, that interact during fertilization. The proteins produced from the derived alleles a and b may lack matching, which prevents fertilization (see e.g. [VS11] for some example in marine invertebrates). It has also been proposed that Dobzhansky-Muller incompatibilities can arise by gene duplication. Initially both loci (A and B) code for the same trait but one may lose its function. If each locus loses its function in one sub-population, hybrids will lack this trait and may not be viable. Also, this kind of incompatibilities can be caused by negative epistatic interactions. Epistatic interactions arise when the effect of a gene depends on the genetic context in which it acts.

However, very few examples of real Dobzhansky-Muller incompatibilities, involving only two loci have been found in nature and are well characterized. One occurs between two closely related species of fish, the platyfish, *X. maculatus* and the swordtail *X. helleri* (see e.g. [Gor31, CO04, PMN10]). The first one has spots on the dorsal fin whereas the second one lacks spots. The spots are produced by a X -linked gene which expression is controlled by a second locus. While the platyfish has both genes, the swordtail lacks both. Hybrids, which have the spot producing gene but not the repressor, have large spots which develop into tumors. Therefore the fitness of the hybrids is reduced.

- **The Nei model**

Another model for the evolution of reproductive isolation was developed by Nei and his collaborators [NMW83]. It is a one-locus multi-allelic model of post-zygotic incompatibility. A locus A which is involved in hybrid sterility or inviability has a series of different alleles $(A_i)_{i\geq 1}$. Homozygotes A_iA_i and heterozygotes A_iA_{i+1} and A_iA_{i-1} are viable and have the same fitness. But heterozygotes A_iA_{i+2} and A_iA_{i-2} are lethal or sterile.

In their model, the different alleles are neutral, in the sense that all viable individuals have the same fitness. The model assumes that in an ancestral population allele A_i is fixed. The population is split into two subpopulations. By genetic drift, allele A_{i+1} can be fixed in one subpopulation and allele A_{i-1} in the other one. The genotype of the hybrids is $A_{i+1}A_{i-1}$, so they not viable and a reproductive barrier is built.

Models of incompatibilities involving a large number of loci

Some multi-locus generalizations of these models have been studied. Orr [Orr95] and Gavrilets and Gravner [GG97] studied how the probability of reproductive isolation depends on the number of substitutions between two sub-populations. Assume that the ancestral genotype is *abcde* which is replaced in one subpopulation by *AbcdE* and by *aBcde* in the second one. The allele E can be incompatible with alleles a and B. More generally, the k^{th} substitution can be incompatible with $k-1$ alleles. If each new derived allele has a probability p of being incompatible with each locus, the expected number of incompatibilities between two populations differing at k loci was $\frac{pk^2}{2}$. They predicted a

"snowballing effect": if the number of substitutions increases linearly with time, the number of incompatibilities, increases faster than linearly with time (and therefore the probability that two individuals are able to interbreed decreases faster than linearly with time). He also concluded that speciation should occur more easily if many loci are involved in reproductive barriers: in a multi-locus systems there are more possible combinations of loci that can give rise to incompatibilities than in a classical two locus model.

The model we use in Chapter III is a simplified version of this model, proposed by Yamaguchi and Iwasa in [YI13]. We will consider a set of ℓ incompatibility controlling loci. The *genetic distance* between two individuals is defined as the number of these loci that differ between the two. Inspired by this "snowballing effect", we will assume that there is a threshold of speciation s and reproductive incompatibility emerges when the genetic distance becomes higher than the threshold. One of the advantages of this type of approach is that it takes into account the fact that speciation takes time, and two populations that are not completely isolated can produce hybrids. Many examples of hybridization between lineages that had been thought to be separate species have been found in nature, for example in cichlid fishes [KDS$^+$07, KWG$^+$13], warblers [BBI11], fruit flies [SMSBM05], butterflies [MSB$^+$06, CDRM15] and sculpins [NFST05]. Chapter II is devoted to the study of hybrid populations. The idea is to use recombination patterns to infer the time since admixture.

4.3 Geography and speciation

As we have seen, geographic structure promotes genetic differentiation between subpopulations. The accumulation of genetic differences between populations may give rise to reproductive incompatibilities and therefore to the formation of new species. Geographic structure can play a crucial role in speciation and models of speciation have been classified depending on the geographical setting.

- *Sympatric speciation* is the emergence of two or more species from a single ancestral population, in the same geographic location, without any spatial isolation. One of the most studied examples of sympatric speciation is that of the clichids, a very diversified family that lives in the Rift Valley lakes. The plausibility and generality of sympatric speciation has been a source of controversy amongst evolutionary biologists. In the 2000s Gavrilets [Gav04, Gav05] derived general conditions for sympatric speciation. It requires disruptive selection, which occurs when extreme phenotypes have a fitness advantage over more intermediate phenotypes, and assortative mating, which means that mating between individuals with similar genotypes is promoted.
- *Allopatric speciation* is the formation of new species from subpopulations which are geographically isolated. In the absence of gene flow, each subpopulation accumulates

mutations independently, which promotes the emergence of reproductive barriers. Local adaptation can accelerate the divergence between the different subpopulations. The models of reproductive incompatibilities presented above were introduced in the context of allopatric speciation.

- *Peripatric speciation* is a form of allopatric speciation in which a new species is formed in an isolated, small peripheral subpopulation that is isolated from the main population.
- *Parapatric speciation* is an intermediary form of speciation, where subpopulations are partially isolated but there exists some gene flow between them.

4.4 Fitness landscapes

To visualize the link between fitness, adaptation and speciation one can use the metaphor of fitness landscapes. The idea was introduced by Wright in 1932 [Wri32]. Fitness landscapes represent individual fitness as a function on the genotype space, which is a multidimensional space representing all possible genotypes. An individual is represented as a point and a population corresponds to a cloud of points. Wright suggested the idea that adaptive landscapes were "rugged", with peaks, corresponding to the different species and valleys corresponding to unfit hybrids. Speciation can be seen as a population moving from one peak to another. Wright suggested that a small population can move across an adaptive valley to the basis of another, maybe higher peak. Then, natural selection will move the population up to the new peak. Finally, this newly adapted population can expand its range. This mechanism of speciation is known as the shifting-balance theory. However, Wright's argument was only verbal and further theoretical analyses of this model ([CBT00]) have shown that, although the mechanisms underlying this theory can in principle work, the conditions are very strict (for example, a really small population size is required).

Within this context, Gavrilets [Gav97] suggested the concept of "holey" adaptive landscapes, where two reproductively incompatible genotypes can be connected by a chain of intermediary fit genotypes, forming a "ridge". This is for example the case in the Dobzhansky-Muller model where $AAbb$ and $aaBB$ are incompatible and are connected by a chain of fit genotypes ($AABb$, $AABB$, $AaBB$). Each mutation is neutral, so two reproductively isolated species can be formed without going through a valley of maladaptation. The resulting landscape is flat, with a "hole" corresponding to the unfit hybrids $aAbB$ (and all the fitness values are 0 or 1). This idea can be extended to higher dimensional genotype spaces. Gavrilets and Gravner [GG97] showed, using percolation theory that, when many loci are involved, typically fit genotypes will be connected by evolutionary ridges. Speciation is seen as a population diffusing across a ridge by several neutral mutation steps until it stands at the other side of a hole.

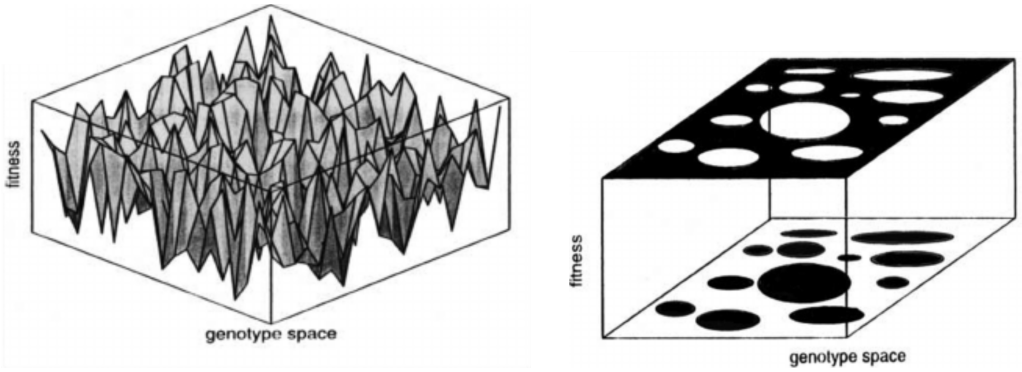


Figure 7 – The left panel represents a rugged adaptive landscape and the right panel represents a holey adaptive landscape [GG97].

In Chapter III we will consider a model of parapatric speciation, in which a metapopulation is divided into several subpopulations. The geographic structure of the metapopulation is modelled by a weighted oriented graph, where each vertex corresponds to a subpopulation. Each directed edge is associated to a migration rate in each direction. This model is a generalization of the classical isolation-by-distance model, in the sense that the metapopulation is subdivided into discrete demes, but we can allow migration between any pair of subpopulations (and not only the nearest neighbours). In addition, we will consider a set of ℓ incompatibility controlling loci and assume that neutral mutations for this set of loci are rarer than in the typical population genetics context. This rare mutation hypothesis is commonly used when studying speciation, see for example [GAG00] or [YI13]. The idea behind this hypothesis is that the incompatibility controlling loci are potentially involved in pre or post-zygotic reproductive barriers so they must participate in reproduction or in development and interact with other genes. For this type of loci, random mutations are very likely to be deleterious, so they will be washed away by selection at the micro-evolutionary timescale. This can be visualized using a holey adaptive landscape where neutral mutations along the evolutionary ridge connect the different fit genotypes, but most mutations make the resulting genotype fall into the "hole".

Chapter I

Chromosome Painting

Abstract

We consider a Moran model with recombination in a haploid population of size N. At each birth event, with probability $1 - \rho_N$ the offspring copies one parent's chromosome, and with probability ρ_N she inherits a chromosome that is a mosaic of both parental chromosomes. We assume that at time 0 each individual has her chromosome painted in a different color and we study the color partition of the chromosome that is asymptotically fixed in a large population, when we look at a portion of the chromosome such that $\rho := \lim_{N\to\infty} \frac{\rho_N N}{2} \to \infty$. To do so, we follow backwards in time the ancestry of the chromosome of a randomly sampled individual. This yields a Markov process valued in the color partitions of the half-line, that was introduced by [EPB16], in which blocks can merge and split, called the partitioning process. Its stationary distribution is closely related to the fixed chromosome in our Moran model with recombination. We are able to provide an approximation of this stationary distribution when $\rho \gg 1$ and an error bound. This allows us to show that the distribution of the (renormalised) length of the leftmost block of the partition (i.e. the region of the chromosome that carries the same color as 0) converges to an exponential distribution. In addition, the geometry of this block can be described in terms of a Poisson point process with an explicit intensity measure.

1 Introduction

1.1 Motivation: a Moran model with recombination

Genetic recombination is the mechanism by which, in species that reproduce sexually, an individual can inherit a chromosome that is a mosaic of two parental chromosomes. Many classical population genetics models ignore recombination and only focus on a single locus, i.e. a location on the chromosome with a unique evolutionary history. In this setting, many analytical results are known. For example the time to fixation (i.e. the first time at

which all individuals carry the same allele) or the fixation probabilities (see for example [Eth11]). However, understanding the joint evolution of different loci is well known to be mathematically challenging, as one needs to take into account non-trivial correlations between loci along the chromosome. For instance, loci that are close to one another are difficult to recombine, so they often inherit their genetic material from the same parent and as a consequence, often share a similar evolutionary history. On the contrary, loci that are far from one another will tend to have different, but not independent, evolutionary histories.

To visualize the questions that will be addressed in this work, let us imagine that in the ancestral population, each individual carries a single continuous chromosome painted in a distinct color. By the blending effect of recombination, after a few generations, the chromosome of each individual looks like a mosaic of colors, each color corresponding to the genetic material inherited from a single ancestral individual. Some natural questions arise: How does the mosaic of colors that is fixed in the population look like? How many colors are there? If the leftmost locus is red (i.e. is inherited from the individual with red chromosome in the ancestral population), what is the amount of red in the mosaic and where are the red loci located? These questions are interesting from a biological point of view: for example, the number of colors in the mosaic corresponds to the number of ancestors that have contributed to an extant chromosome. Loci that are of the same color (i.e., that have been inherited from the same individual in the ancestral population) are called identical-by-descent (IBD).

It is known that changes in the population size or natural selection can alter the sizes of the IBD segments: for example genes that are under selection tend to be located within large IBD segments. This prediction can guide the detection of genes that are under selection (see for example the methods developed by [SRH$^+$02] or [MLAR$^+$08]). The aim of this article is to characterize the distribution of the IBD blocks along a chromosome in the absence of selection or demography. Our results may then be used as predictions under the null hypothesis, that can serve as a standard to which compare real data, e.g., to infer selection or demography.

Also, our results may be relevant to the analysis of data obtained in experimental evolution. For example, in the experiment carried out by Teotónio et al. [TEPB17], the authors intercrossed individuals from 16 different subpopulations of the worm *C. elegans* and let the population evolve for several generations at controlled population size. Then, each individual is genotyped, each of its variants is mapped to one of the 16 ancestor subpopulations, so as to get a represention of each DNA sequence of each individual as a partition of the sequence into 16 colors. Again, our model (or an extension to our model accomodating for the finite number of colors), might be used as a null model whose predictions can be compared to these real color mosaics.

Sampling the chromosome, seen as a continuous, single-ended strand modelled by the

positive half-line, of an individual in the present population and tracing backwards in time the ancestry of every locus yields a process valued in the partitions of $\mathbb{R}^+$, called the $\mathbb{R}$-partitioning process. Here, $x \geq 0$ and $y \geq 0$ belong to the same block of the $\mathbb{R}$-partitioning process at time t if the loci at positions x and y on the sampled chromosome shared the same ancestor t units of time ago.

The $\mathbb{R}$-partitioning process is the continuum analog of the celebrated Ancestral Recombination Graph [Hud83, Gri91, GM97]. Before giving a formal description of this object, we start by showing how it arises naturally from a multi-locus Moran model. The population size is N and each haploid individual carries a single linear chromosome of length R. At time 0, each individual has her (unique) chromosome painted in a distinct color (see Figure 1.1). Each individual reproduces at rate 1, and upon reproduction, the individual chooses a random partner in the population. Let $\rho_N \in (0, 1)$,

— With probability $1 - \rho_N R$, the offspring copies one parent's chromosome (chosen uniformly at random).

— With probability $\rho_N R$, a recombination event occurs. We assume single-crossover recombination which means that each parental chromosome is cut into two fragments. The position of the cutpoint (i.e. the crossover) is uniformly distributed along the chromosome (see Figure 1.1). The offspring copies the genetic material to the left of this point from one parent and the genetic material to the right of this point from the other parent.

The offspring then replaces a randomly chosen individual in the population. Because of recombination, at time t each chromosome is a mosaic of colors, each color corresponding to the genetic material inherited from one individual in the founding generation. (In other words, loci sharing the same color are IBD.)

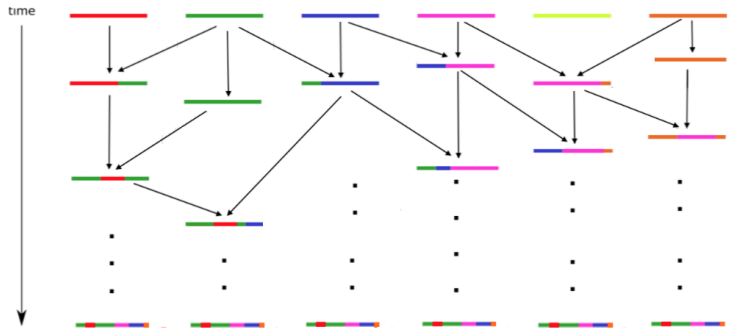


Figure I.1 – Moran model model with recombination.

Let us now consider $z = (z_0, z_1, \ldots, z_n) \in [0, R]$ corresponding to the locations of $n+1$ loci along the chromosome (with $z_0 < z_1 < \ldots < z_n$). Forward in time, the evolution of the genetic composition of the population can be described in terms of a $(n+1)$-locus

Moran model with recombination as described in [Dur08, BWK10]. Backward in time, the genealogy of those loci (sampled from the same individual) is described in terms of the discrete partitioning process, introduced by [EPB16], which traces the history of the $n+1$ loci under consideration (see Figure 1.1). More precisely, the discrete partitioning process (associated to z) is a Markov process valued in the partitions of $z = \{z_0, z_1, \dots, z_n\}$ such that z_i and z_j are in the same block at time t if and only if they inherit their respective genetic material from the same individual t units of time ago. (In other words, z_i and z_j are IBD if we look t units of time in the past). In a population of size N, it can be seen that the dynamics of the discrete partitioning process are controlled by the following transitions.

— Each pair of blocks coalesces at rate $2/N + O(\rho_N/N)$.

— Each block $b = \{z_{i_1}, \dots, z_{i_k}\}$ is fragmented into $\{z_{i_1}, \dots, z_{i_j}\}$ and $\{z_{i_{j+1}}, \dots, z_{i_k}\}$ at rate $\rho_N(z_{i_{j+1}} - z_{i_j})$.

— Simultaneous splitting and coalescence events happen at rate $O(\rho_N/N)$.

The interesting scaling for this process is when time is accelerated by $N/2$ and the recombination probability scales with N in such a way that

$$\lim_{N\to\infty} \rho_N N/2 = \rho, \tag{I.1}$$

for some $\rho > 0$. It can readily be seen that the discrete partitioning process in a population of size N converges in distribution (in the Skorokhod topology) to a process $(\Gamma_t^{\rho,z}; t \geq 0)$, which is the Markov process with the following transition rates:

— Each pair of blocks coalesces at rate 1.

— Each block $b = \{z_{i_1}, \dots, z_{i_k}\}$ is fragmented into $\{z_{i_1}, \dots, z_{i_j}\}$ and $\{z_{i_{j+1}}, \dots, z_{i_k}\}$ at rate $\rho(z_{i_{j+1}} - z_{i_j})$.

In the literature, $\Gamma^{\rho,z}$ is also referred to as the Ancestral Recombination Graph (ARG) [Hud83, Gri91, GM97] associated to z (with recombination rate ρ). The following scaling property can easily be deduced from the description of the transition rates. We assume that at time 0 all loci are sampled in the same individual, i.e. we consider the ARG started from the coarsest partition. Then

$$\forall R > 0, \quad \Gamma^{R,z} = \Gamma^{1,Rz} \text{ in distribution.} \tag{I.2}$$

In the following, we are going to consider a high recombination regime, i.e. that ρ is large. This relation states that this is equivalent to considering that the distances between loci of interest are large.

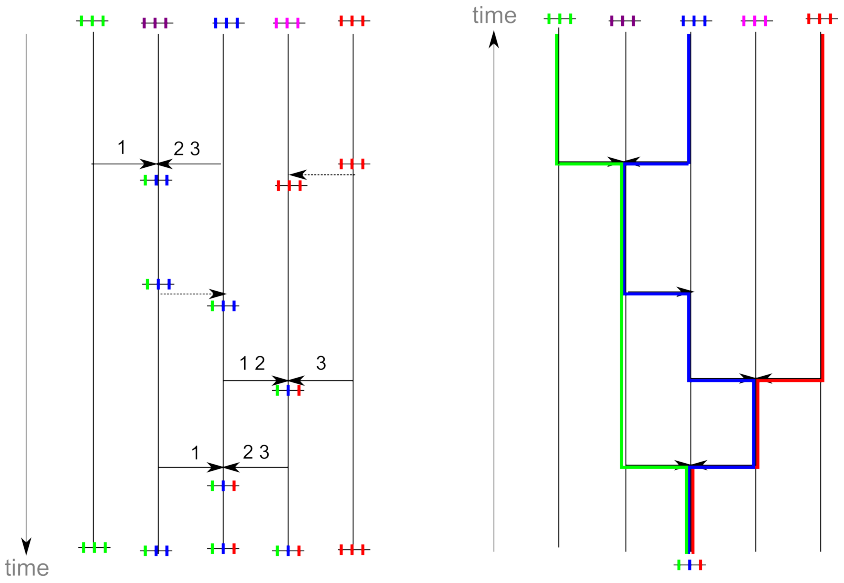


Figure I.2 – Duality between the Moran model and the discrete partitioning process ($N = 5$, $n = 3$). The left panel represents a realization of the Moran model with recombination. On the top we represented the chromosomes of the different individuals at generation 0, assuming that each one is painted in a different color. Arrows represent reproduction events. For reproduction events without recombination, the base of the arrow represents the parent from which the genetic material is inherited, and the arrow points at the individual that is replaced. For reproduction events with recombination (represented by two arrows), we indicate on top of each arrow the loci that are inherited from each parent. In the second panel we show how the discrete partitioning process can be obtained by reverting time. We sample the 3 loci in an individual in the present population and we follow the ancestral lineages corresponding to each of her loci. Each time an ancestral lineage finds the tip of an arrow, it jumps to the base of the arrow. If there are two arrows, the lineage corresponding to locus i jumps to the base of the arrow labelled i.

1.2 The $\mathbb{R}$-partitioning process

As the goal of this article is to characterize the distribution of the IBD blocks in a continuous chromosome in an infinite population, we extend the ARG $(\Gamma_t^{\rho,z}; t \geq 0)$ to the whole positive real line. To do so, we will consider partitions of $\mathbb{R}^+$. We call a segment a maximal set of connected points belonging to the same block of the partition. A partition of $\mathbb{R}^+$ is right-continuous if the segments of the partition are left-closed (right-open) intervals and the blocks correspond to disjoint unions of such intervals. Let $\mathcal{P}^{loc}$ be the set of partitions of $\mathbb{R}^+$ that are right-continuous and locally finite, i.e. such that each compact subset of $\mathbb{R}^+$ contains only a finite number of segments. For any z finite subset of $\mathbb{R}^+$, $\mathcal{P}_z$ is the set of partitions of z and t_z is the trace of z, i.e. the function $\mathcal{P}^{loc} \to \mathcal{P}_z$ such that for any $\pi \in \mathcal{P}^{loc}$, $t_z(\pi)$ is the partition of z induced by π. We define $\mathcal{F}$ as the σ-field on $\mathcal{P}^{loc}$ generated by

$$\mathcal{C} = \{\{\omega \in \mathcal{P}^{loc}, t_z(\omega) = \pi\},\ n \in \mathbb{N},\ z = (z_0, \dots, z_n) \subset \mathbb{R}^+,\ \pi \in \mathcal{P}_z\}.$$

Finally, for any measure μ on a measured space $(\Omega, \mathcal{A})$ and any $\mathcal{A}$-measurable function f, we will denote by $f \star \mu$ the pushforward of μ i.e. the measure such that $\forall B \in \mathcal{A}$, $f \star \mu(B) = \mu(f^{-1}(B))$.

Theorem 1.1. *Let μ_0 be a probability measure on $(\mathcal{P}^{loc}, \mathcal{F})$. The $\mathbb{R}$-partitioning process $(\Pi_t^\rho; t \geq 0)$ started at μ_0 is the unique càdlàg stochastic process valued in $(\mathcal{P}^{loc}, \mathcal{F})$ such that for any z, finite subset of $\mathbb{R}$, $(t_z(\Pi_t^\rho);\ t \geq 0)$ is the ARG at rate ρ for the set of loci z started at $t_z \star \mu_0$.*

The proof of this theorem can be found in Section 2. The goal of this paper is to study some properties of the invariant measure of the $\mathbb{R}$-partitioning process.

Theorem 1.2. *The $\mathbb{R}$-partitioning process $(\Pi_t^\rho; t \geq 0)$ has a unique invariant probability measure μ^ρ in $(\mathcal{P}^{loc}, \mathcal{F})$. In addition, for any finite subset z of $\mathbb{R}^+$,*

$$t_z \star \mu^\rho \ = \ \mu^{\rho,z}$$

where $\mu^{\rho,z}$ is the unique invariant measure of $\Gamma^{\rho,z}$.

We let the reader refer to Section 3 for proof of this result.

1.3 Approximation of the stationary distribution of the ARG

The ARG with more than two loci is a complex process and some authors have considered that characterizing its distribution is "computationally not tractable" (see [BWK10]). In [GJL16] and [EPB16], the authors provided methods to compute the stationary distribution that fail when considering a large number of loci. Our goal is to provide an

approximation of the stationary distribution of the ARG that is relatively easy to handle, even when we consider a large number of loci.

One of the main contributions of this paper is an explicit approximation (and an error bound for it) of the stationary distribution of the ARG $(\Gamma_t^{\rho,z}; t \geq 0)$ when the typical distance between the z_i's is large (or equivalently when the rate of recombination ρ is large).

We fix $z = (z_0, \ldots, z_n) \subset \mathbb{R}$. We define

$$\alpha = \min_{i \neq j} |z_i - z_j|$$

and we assume that $\alpha > 0$ (or equivalently that the coordinates of z are pairwise distinct). Let $r \in \{0, \ldots, n\}$, $\mathcal{P}_z^r$ is the set of partitions of z containing $n+1-r$ blocks. In particular, the only partition in $\mathcal{P}_z^0$ is π_0, the partition made of singletons. We define a "coalescence scenario of order r" as a sequence of partitions $(s_k)_{0 \leq k \leq r}$ such that s_0 is the partition made of singletons and for $1 \leq k \leq r$, s_k is a partition of order k that can be obtained from s_{k-1} by a single coagulation event. For any partition in $\pi \in \mathcal{P}_z^r$, $\mathcal{S}(\pi)$ is the set of coalescence scenarios of order r such that $s_r = \pi$.

For $\pi \in \mathcal{P}_z^r$, let $b_1, \ldots, b_{n+1-r}$ be the blocks of π. We denote by $C(\pi)$ the cover length of π defined as:

$$C(\pi) \ := \ \sum_i \max_{x,y \in b_i} |x - y|.$$

In particular, the cover length of π_0 is equal to 0.

Let $s = (s_k)_{0 \leq k \leq r}$ be a scenario of coalescence of order r, with $1 \leq r \leq n$. We define the energy of s, $E(s)$ as

$$E(s) := \prod_{i=1}^{r} C(s_i).$$

where $C(s_i)$ is the total rate of fragmentation at state s_i. Finally, define

$$\forall \pi \in \mathcal{P}_z \setminus \mathcal{P}_z^0, \quad F(\pi) := \sum_{S \in \mathcal{S}(\pi)} \frac{1}{E(S)}. \tag{I.3}$$

Theorem 1.3. *There exists a function*

$$f^n : \mathbb{R}_*^+ \to \mathbb{R}^+, \quad \lim_{x \to \infty} f^n(x) = 0,$$

independent of the choice of $z = (z_0, \ldots, z_n)$ and ρ, such that

$$\forall \rho > 0, \ \forall k \in [n], \ \forall \pi_k \in \mathcal{P}_z^k, \quad \left| \mu^{\rho,z}(\pi_k) - \frac{1}{\rho^k} F(\pi_k) \right| \leq f^n(\alpha\rho) \frac{1}{\rho^k} F(\pi_k).$$

Recall that the RHS goes to 0 either when $\rho \to \infty$ or $\alpha \to \infty$. As already mentioned (see (I.2)), these two scaling limits are equivalent. We let the reader refer to Section 4 for a proof of this result.

1.4 Characterization of the leftmost block of the $\mathbb{R}$-partitioning process

As an application of our approximation of $\mu^{\rho,z}$, we characterize the geometry of the leftmost block on a large scale. Motivated by the Moran model and the scaling relation (I.2), without loss of generality, we study the $\mathbb{R}$-partitioning process at rate 1 restricted to $[0, R]$.

For any partition π, $x \sim_\pi y$ means x and y are in the same block of π. Let Π_{eq} be the random partition with law μ^1. Let $\mathcal{L}_R(0)$ be the length of the block containing 0, rescaled by $\log(R)$. More precisely,

$$\mathcal{L}_R(0) = \frac{1}{\log(R)} \int_{[0,R]} \mathbb{1}_{\{x \sim_{\Pi_{eq}} 0\}} dx.$$

We define the random measure $\vartheta^R[a, b]$ such that

$$\forall a, b \in [0, 1],\ a \leq b, \quad \vartheta^R[a, b] = \frac{1}{\log(R)} \int_{R^a}^{R^b} \mathbb{1}_{\{x \sim_{\Pi_{eq}} 0\}} dx$$

so that ϑ^R encapsulates the whole information about the positions of the loci that are IBD to 0 in the *logarithmic scale* (which will be seen to be the natural scaling for the partitioning process at equilibrium). In the following ϑ^R will be considered as a random variable valued in $\mathcal{M}([0, 1])$, the space of locally finite measures of $[0, 1]$ equipped with the weak topology (i.e. the coarsest topology making $m \to \langle m, f\rangle$ continuous for every function f bounded and continuous). In the following, $\Longrightarrow$ denotes the convergence in distribution.

Theorem 1.4. *Consider a Poisson point process $\mathcal{P}^\infty$ on $[0, 1] \times \mathbb{R}^+$ with intensity measure*

$$\lambda(x, y) = \frac{1}{x^2} \exp(-y/x) dx dy$$

and define the random measure on $\mathcal{M}([0, 1])$

$$\vartheta^\infty := \sum_{(x_i, y_i) \in \mathcal{P}^\infty} y_i\ \delta_{x_i}.$$

Then

1. *$\vartheta^R \underset{R\to\infty}{\Longrightarrow} \vartheta^\infty$ in the weak topology.*
2. *In particular, $\mathcal{L}_R(0) \underset{R\to\infty}{\Longrightarrow} \varepsilon(1)$ where $\varepsilon(1)$ denotes the exponential distribution of parameter 1.*

This result can be interpreted as follows. As $R \to \infty$, there are distinct regions of genetic material that is IBD to 0, and at the limit, those regions are clustered into points. The locations of those regions are encapsulated by the x_i's (in the logarithmic scale) – in other words, at R^{x_i}, there is a cluster of genetic material IBD to 0 – and the coordinate y_i corresponds the amount of genetic material that is IBD to 0 present in this cluster (see Figure 1.4). Note that the positions of the segments (in the logarithmic scale) are given by the Poisson process of intensity $(1/x)dx$, which is known as "the scale invariant Poisson Process" (see for example [Arr98]).

We performed some numerical simulations of the partitioning process to illustrate the second part of Theorem 1.4. Figure 1.4 shows how the length of the cluster covering 0 is exponentially distributed.

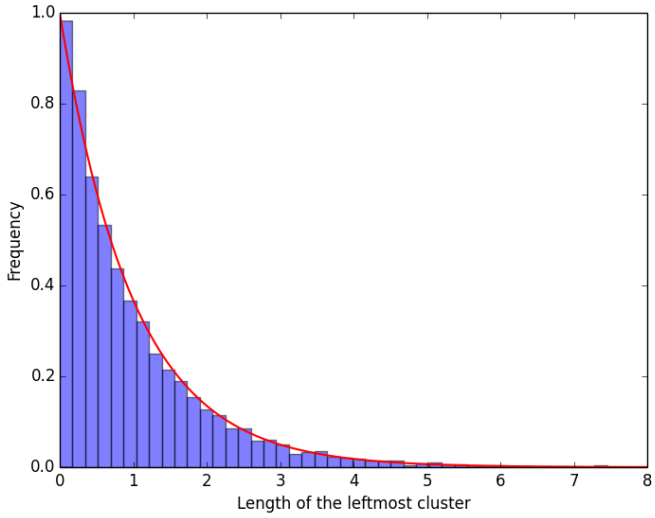


Figure I.3 – **Distribution of the length of the leftmost block (R = 5000).** The blue histogram represents the empirical distribution, that was obtained by simulating the partitioning process, for a chromosome of length $R = 5000$. The number of replicates is 10000. The red curve is the probability density function of an exponential distribution of parameter 1. We compared the empirical distribution to an exponential distribution using a Kolmogorov-Smirnov test, which was positive, with a p-value of 10^{-4}.

1.5 Biological relevance

Recall that a "morgan" is a unit used to measure genetic distance. The distance between two loci is 1 morgan if the average number of crossovers is 1 per reproduction event. In other words, in a population of size N, if we consider the discrete partitioning process at rate 1, two loci z_i and z_j are at distance $\frac{2}{N}|z_i - z_j|$ morgans.

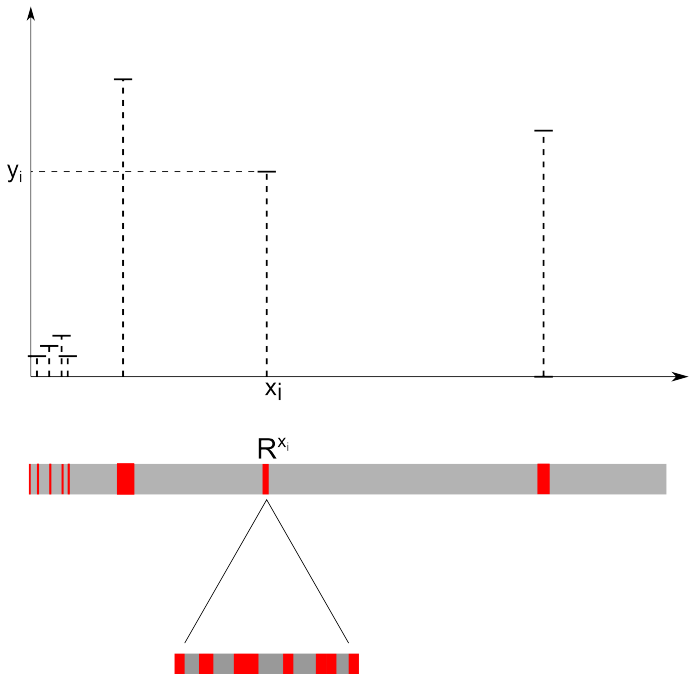


Figure I.4 – Example of a realization of ϑ^∞ and its interpretation. Regions of the chromosome (in the log-scale) that are IBD to 0 are represented in red. In the limit, those regions are clustered into points which can have a complex geometry on a finer scale (see lower figure). y_i is the amount of genetic material IBD to 0 in the region located at R^{x_i}.

We studied the $\mathbb{R}$-partitioning process at rate 1 restricted to $[0, R]$, which should correspond to a portion (or frame) of the chromosome that is of size R/N morgans (and small enough so that the single crossing-over approximation is valid). Then, we first let the population size N tend to infinity (in order to get the partitioning process from the underlying finite population model), and then the size of the frame go to ∞ (in the $\mathbb{R}$-partitioning process). Note that since we take successive limits (first $N \to \infty$ and then $R \to \infty$), this gives no clue on how the population size and the size of the observation frame should scale with one another to ensure that the approximation is correct.

[WH97] used the same hypothesis. They explained that this approximation should be valid in human populations, where for example, the size of chromosome 1 is 2.93 morgans and the effective population size is $N = 20000$. If one looks at a frame of this chromosome of length 1 morgan (1/3 of the chromosome), then $R = 20000$.

1.6 Outline

This paper is organized as follows. In Section 2 we propose a construction of the $\mathbb{R}$-partitioning process and we prove Theorem 1.1. In Section 3 we show the existence and uniqueness of a stationary distribution for this process (Theorem 1.2). Finally, Sections 4 and 5 are devoted to the proofs of Theorems 1.3 and 1.4 respectively.

2 The $\mathbb{R}$-partitioning process

2.1 Some preliminary definitions

We start by recalling some definitions that are useful for the rest of the paper and by clarifying some notation. In the following, we consider partitions of $\mathbb{R}^+$ or of subsets of $\mathbb{R}^+$. We call a *segment* a maximal set of adjacent points belonging to the same block of the partition. For $E = \mathbb{R}^+$ or $E \subset \mathbb{R}^+$, we say that a partition ω of E is *locally finite* if for any compact subset K of E such that $K \cap E \neq \varnothing$, $t_{K\cap E}(\omega)$ contains a finite number of segments. We say that a partition is *right continuous* if the segments of the partition are left-closed (right-open) intervals (and the blocks correspond to disjoint unions of such intervals). Note that for a partition that is right continuous, infinite sequences of small intervals can only accumulate to the left of a point. For $a < b \in \mathbb{R}^+$, we denote by $\mathcal{P}^{loc}_{[a,b]}$ (resp. $\mathcal{P}^{loc}$) the set of the partitions of $[a, b]$ (resp. $\mathbb{R}^+$) that are right continuous and finite (resp. right continuous and locally finite). We define the σ-field $\mathcal{F}$ on $\mathcal{P}^{loc}$ generated by

$$\mathcal{C} = \{\{\omega \in \mathcal{P}^{loc}, t_z(\omega) = \pi\},\ n \in \mathbb{N},\ z = (z_0, \ldots, z_n) \subset \mathbb{R}^+,\ \pi \in \mathcal{P}_z\}.$$

We also need to define a distance d on $\mathcal{P}^{loc}$. To do so, we start by identifying each partition in $\mathcal{P}^{loc}$ to a function from $\mathbb{R}^+$ to itself. More precisely, we define a map ϕ :

$\mathcal{P}^{loc} \to D(\mathbb{R}^+, \mathbb{R}^+)$ such that, for $\pi \in \mathcal{P}^{loc}$, $\phi(\pi)$ is constructed as follows. For each block B of π and for each $x \in B$, we set $\phi(\pi)(x) := \min(B)$. Note that ϕ is injective and $\forall x \in \mathbb{R}^+,\ \phi(\pi)(x) \leq x$. Also, as $\pi \in \mathcal{P}^{loc}$, ϕ is càdlàg and has a finite number of jumps in any compact set of $\mathbb{R}^+$. Now, for any $\pi_1, \pi_2 \in \mathcal{P}^{loc}$, define

$$d(\pi_1, \pi_2) \ := \ \int_0^{+\infty} |\phi(\pi_1)(x) - \phi(\pi_2)(x)| \exp(-x) dx.$$

It can easily be checked that d defines a distance on $\mathcal{P}^{loc}$. For $T > 0$, we will denote by $D([0,T], \mathcal{P}^{loc})$ the Skorokhod space associated to $(\mathcal{P}^{loc}, d)$ equipped with the standard Skorokhod topology. For each partition $\pi \in \mathcal{P}^{loc}$ we define a natural ordering on its blocks. We denote by $b^0, b^1, \ldots, b^i, \ldots$ the blocks of π indexed in such a way that $\min(b^0) < \min(b^1) < \ldots$

The space $\mathcal{P}^{loc}$ is separable under d. Indeed, for $n \in \mathbb{N}^*$, let $\mathcal{S}_n$ be the set of partitions in $\pi \in \mathcal{P}^{loc}$ such that in $\pi|_{[0,n[}$ each block is a finite union of segments whose endpoints are in $[0, n[\cap \mathbb{Q}$ and $[n, +\infty[$ is included in a block of π. $S = \cup_n \mathcal{S}_n$ is countable and using standard methods, it can be shown that given $\pi \in \mathcal{P}^{loc}$ and $\epsilon > 0$, there exists a partition $\pi' \in S$ such that $d(\pi, \pi') < \epsilon$. The space $\mathcal{P}^{loc}$ is not complete but we define its completion $\bar{\mathcal{P}}^{loc}$.

In the following, we will also consider partitions of $\mathbb{Q}^+$. We define $\mathcal{P}_{\mathbb{Q}}^{loc}$ as the set of locally finite partitions of $\mathbb{Q}^+$ that are right continuous (in the sense that if $\mathbb{Q}^+ \ni x_n \downarrow x \in \mathbb{Q}^+$ then x_n is in the same segment as x for n large enough) and $\mathcal{F}_{\mathbb{Q}}$ the σ-field generated by

$$\mathcal{C}_{\mathbb{Q}} = \{\{\omega \in \mathcal{P}_{\mathbb{Q}}, t_z(\omega) = \pi\}, n \in \mathbb{N},\ z = (z_0, \ldots, z_n) \subset \mathbb{Q}^+,\ \pi \in \mathcal{P}_z\}.$$

2.2 Definition of the $\mathbb{R}$-partitioning process

We start by defining ARG for a finite set of loci [Hud83, Gri91, GM97]. Note that here we only consider the case of single-crossover recombination. We consider a finite set of loci, whose positions in the chromosome are given by $z \subset \mathbb{R}^+,\ z = \{z_0, \ldots, z_n\}$ (with $z_0 < z_1 < \ldots < z_n$). Let $\rho > 0$. Let $(\Gamma_t^{\rho,z}; t \geq 0)$ be the Markov process on $(\mathcal{P}_z, \mathcal{F}_z)$, with the following transition rates:

— **Coagulation:** Consider $\pi_1 \in \mathcal{P}_z$ and a and b two blocks of π_1. Let $c = a \cup b$ and π_2, the partition obtained by coalescing the blocks a and b into c and letting all the other blocks unchanged. A transition from π_1 to π_2 occurs at rate:

$$q(\pi_1, \pi_2) = 1.$$

— **Fragmentation:** Now take $\pi_1 \in \mathcal{P}_z$ and a a block of π_1 containing k elements $z_{i_1}, \ldots, z_{i_k}$ such that $z_{i_1} < \ldots < z_{i_k}$. Let $j < k$. Let $b = \{z_{i_1}, \ldots, z_{i_j}\}$ and $c =$

$\{z_{i_{j+1}}, \dots, z_{i_k}\}$ and π_2, the partition obtained by fragmenting a into b and c. A transition from π_1 to π_2 occurs at rate:

$$q(\pi_1, \pi_2) = \rho(z_{i_{j+1}} - z_{i_j}).$$

— All these events are independent and all other events have rate 0.

This process is called the ARG at recombination rate ρ, for the set of particles (or loci) z. It is easily seen that $\Gamma_t^{\rho,z}$ has a finite state-space and is irreducible. We call $\mu^{\rho,z}$ its unique invariant probability measure, that will be characterized in Section 4.

We now want to define a process on $(\mathcal{P}^{loc}, \mathcal{F})$, called the $\mathbb{R}$-*partitioning process* so that for any z finite subset of $\mathbb{R}^+$, the trace on z is distributed as the ARG $\Gamma^{\rho,z}$.

We set $\Pi_0^{\rho,L} = \pi_0,\ \pi_0 \in \mathcal{P}^{loc}_{[0,L[}$. We assume that the blocks of this partition are indexed with the natural order defined on the previous section. The partitioning process on $\mathcal{P}^{loc}_{[0,L[}$ is generated by a sequence of independent Poisson point processes as follows:

— For all $i, j \in \mathbb{N}$, $Y^{i,j}$ is a Poisson point process of intensity 1. For $t \in Y^{i,j}$, at time t^- there is a **coagulation** event: blocks b^i and b^j are replaced by $b^i \cup b^j$. If i or j does not correspond to the index of any block, nothing happens.

— For all $i \in \mathbb{N}$, X^i is a Poisson point process on $\mathbb{R}^+ \times [0, L[$ with intensity $\rho\ dt \otimes dx$. The atoms of X^i correspond to **fragmentation** events. For $(t, x) \in X^i$, if at time t^-, $\Pi_{t^-}^{\rho,L} = \pi$, if b^i is a block of π and $x \in\ ]\min(b^i), \sup(b^i)[$, b^i is fragmented into two blocks $b^{i,-}$ and $b^{i,+}$ such that $b^{i,-} = b^i \cap [0, x[$ and $b^{i,+} = b^i \cap [x, L]$. Then $\Pi_t^{\rho,L}$ is equal to the partition obtained by replacing b^i by $b^{i,-}$ and $b^{i,+}$. If $x \notin\ ]\min(b^i), \sup(b^i)[$, nothing happens.

After each event, blocks are relabelled in such a way that they remain ordered, in the sense specified above. Recall that, with this construction, the partitions that are formed are always right continuous. Also the number of blocks of $\Pi^\rho|_{[0,L[}$ is stochastically dominated by a birth-death process which jumps from n to $n+1$ at rate $\rho L n$ and from n to $n-1$ at rate $n(n-1)/2$ with initial condition the number pf blocks in π_0, which is known to remain locally bounded (and even to have $+\infty$ as entrance boundary, see [Lam05]). There is the same stochastic domination between the two processes for the numbers of jump events on any fixed time interval. This shows that the number of blocks in $\Pi_t^\rho|_{[0,L[}$ is a.s. locally bounded and since the number of segments jumps at most by $+1$ at each event, the number of segments is also a.s. locally bounded. So a.s. for all t, $\Pi_t^{\rho,L} \in \mathcal{P}^{loc}_{[0,L[}$.

Finally, we define the partitioning process in $\mathbb{R}$, as the projective limit of $(\Pi_t^{\rho,L};\ t \geq 0)_{L \in \mathbb{R}^+}$ as $L \to \infty$. In fact, by construction, $\forall L' > L,\ \forall t \geq 0,\ \Pi_t^{\rho,L'}|_{[0,L[} = \Pi_t^{\rho,L}$, where $\Pi^{\rho,L'}|_{[0,L[}$ is the natural restriction of $\Pi_t^{\rho,L'}$ to $[0, L[$.

Proposition 2.1. *The $\mathbb{R}$-partitioning process, $(\Pi_t^\rho;\ t \geq 0)$ with initial measure π_0 is the*

unique càdlàg stochastic process valued in $(\mathcal{P}^{loc}, \mathcal{F})$ *such that*

$$\forall L \geq 0,\ (\Pi_t^\rho \cap [0, L];\ t \geq 0) = (\Pi_t^{\rho,L};\ t \geq 0)$$

with $\Pi_0 = \pi_0$. *Further for any finite subset* z *in* $\mathbb{R}^+$, $t_z(\Pi^\rho)$ *is distributed as* $\Gamma^{\rho,z}$, *the ARG with initial condition* $t_z(\pi_0)$.

Proof. We need to check that, for any $T > 0$, $(\Pi_t^\rho;\ 0 \leq t \leq T) \in D([0,T], \mathcal{P}^{loc})$ almost surely. To do so, we need to prove that with probability 1, for every $t \in [0,T]$, for every $\epsilon > 0$, one can find $s > 0$ such that $d(\Pi_t^\rho, \Pi_{t+s}^\rho) < \epsilon$. Fix $\epsilon > 0$ and pick $L > 0$ such that $2\exp(-L)(L+1) < \epsilon$. From the Poissonian construction, for any $T > 0$, the process $\Pi^\rho|_{[0,L[}$ has a finite number of jumps in $[0,T]$, which happen at times $t_1, \ldots, t_n$. We choose $s > 0$ such that $|t - s| < \min_i |t_{i+1} - t_i|$. Then $\Pi_t|_{[0,L[} = \Pi_{t+s}|_{[0,L[}$. As $\phi(\Pi_t^\rho|_{[0,L[}) = \phi(\Pi_t^\rho)|_{[0,L[}$, for any $x \in [0, L[$, $\phi(\Pi_t^\rho)(x) = \phi(\Pi_{t+s}^\rho)(x)$, so

$$\begin{aligned} d(\Pi_t^\rho, \Pi_{t+s}^\rho) &= 0 + \int_L^{+\infty} |\phi(\Pi_t^\rho)(x) - \phi(\Pi_{t+s}^\rho)(x)| e^{-x} dx \\ &\leq \int_L^{+\infty} 2x \exp(-x) dx \ = \ 2\exp(-L)(L+1) \ < \ \epsilon, \end{aligned}$$

and similarly for left-hand limits. So $(\Pi_t^\rho;\ 0 \leq t \leq T) \in D([0,T], \mathcal{P}^{loc})$. The fact that $t_z(\Pi^\rho)$ is distributed as $\Gamma^{\rho,z}$, the ARG with initial condition $t_z(\pi_0)$ can be readily seen from the definition. □

In addition, the following proposition can easily be deduced. Note that the second equality is just a trivial consequence of the first one.

Proposition 2.2 (Consistency). *For all* z *and* y, *finite subsets of* $\mathbb{R}^+$ *such that* $y \subset z$,

$$\Gamma^{\rho,y} = \Gamma^{\rho,z}|_y,$$

where $\Gamma^{\rho,z}|_y$ *denotes the restriction of* $\Gamma^{\rho,z}$ *to* $\mathcal{P}_y$, *and*

$$\mu^{\rho,y} = t_y \star \mu^{\rho,z}.$$

We now turn to the proof of the main result of this section, i.e. Theorem 1.1.

Proof of Theorem 1.1. Let $(\Pi_t; t \geq 0)$ be a càdlàg process in $(\mathcal{P}^{loc}, \mathcal{F})$ such that for any z, finite subset of $\mathbb{R}$,

$$(t_z(\Pi_t);\ t \geq 0) \stackrel{d}{=} (\Gamma_t^{\rho,z};\ t \geq 0) \stackrel{d}{=} (t_z(\Pi_t^\rho);\ t \geq 0).$$

We denote by $\{z_i\}_{i\in\mathbb{N}}$ an enumeration of the rational numbers and for all $n \in \mathbb{N}$, we define $z^n := \{z_0, \dots z_n\}$. For every $n > 1$, we have

$$t_{z^n}(\Pi^\rho) = t_{\mathbb{Q}}(\Pi^\rho)|_{z^n} \quad \text{and} \quad t_{z^n}(\Pi) = t_{\mathbb{Q}}(\Pi)|_{z^n},$$

so we have

$$(t_{\mathbb{Q}}(\Pi^\rho_t); t \geq 0) \stackrel{d}{=} (t_{\mathbb{Q}}(\Pi_t); t \geq 0).$$

In particular

$$\forall (t_1, \dots, t_n) \subset \mathbb{R}^+ \quad (t_{\mathbb{Q}}(\Pi^\rho_{t_1}), \dots, t_{\mathbb{Q}}(\Pi^\rho_{t_n})) \stackrel{d}{=} (t_{\mathbb{Q}}(\Pi_{t_1}), \dots, t_{\mathbb{Q}}(\Pi_{t_n})). \tag{I.4}$$

Similarly as done p.55, the number of blocks of and number of events undergone by $(t_z(\Pi_t);\ t \in [0,T])$ are stochastically dominated, uniformly in $z \subset \mathbb{Q} \cap [0, L]$, by those of a birth-death process which jumps from n to $n+1$ at rate $\rho L n$ and from n to $n-1$ at rate $n(n-1)/2$ with initial condition the number of blocks in Π_0. This shows that a.s. for all $t \geq 0$, $t_{\mathbb{Q}}(\Pi_t) \in \mathcal{P}^{loc}_{\mathbb{Q}}$ (and of course $t_{\mathbb{Q}}(\Pi^\rho_t) \in \mathcal{P}^{loc}_{\mathbb{Q}}$). Since the partitions in $\mathcal{P}^{loc}$ (resp. $\mathcal{P}^{loc}_{\mathbb{Q}}$) are right-continuous and since $\mathbb{Q}$ is dense in $\mathbb{R}$, for every $\bar{\pi} \in \mathcal{P}^{loc}_{\mathbb{Q}}$ there exists a unique $\pi \in \mathcal{P}^{loc}$ such that $t_{\mathbb{Q}}(\pi) = \bar{\pi}$. In other words, the projection map

$$t_{\mathbb{Q}} : (\mathcal{P}^{loc}, \mathcal{F}) \to (\mathcal{P}^{loc}_{\mathbb{Q}}, \mathcal{F}_{\mathbb{Q}})$$

is bijective. With a little bit of extra work, one can show that $t_{\mathbb{Q}}^{-1}$ is measurable so, from (I.4),

$$\forall (t_1, \dots, t_n) \subset \mathbb{R}^+, \quad (\Pi^\rho_{t_1}, \dots, \Pi^\rho_{t_n}) \stackrel{d}{=} (\Pi_{t_1}, \dots, \Pi_{t_n}).$$

This implies that $\forall T > 0$, $(\Pi_t;\ 0 \leq t \leq T) \stackrel{d}{=} (\Pi^\rho_t;\ 0 \leq t \leq T)$, in the Skorokhod topology $D([0,T], \bar{\mathcal{P}}^{loc})$ (see [Bil68], Theorem 16.6). So $(\Pi^\rho_t;\ t \geq 0)$ is the unique process in $D([0,T], \bar{\mathcal{P}}^{loc})$ such that for any z, finite subset of $\mathbb{R}$, $(t_z(\Pi^\rho_t); t \geq 0)$ is distributed as $(\Gamma^{\rho,z}_t; t \geq 0)$. As $\mathcal{P}^{loc} \subset \bar{\mathcal{P}}^{loc}$, the theorem is proved. □

3 Stationary measure for the $\mathbb{R}$-partitioning process

The goal of this section is to prove Theorem 1.2. The idea of the proof is to consider the stationary measure of the partitioning process on finite sets of rational numbers. Using Kolmogorov's extension theorem we define its unique projective limit in $\mathcal{P}^{loc}_{\mathbb{Q}}$. Then, using continuity arguments, we prove that there is a unique extension of this measure to the partitions of $\mathbb{R}$. Let us now go into more details. We decompose the proof into several lemmas.

Lemma 3.1. *A measure ν is invariant for $(\Pi^\rho_t;\ t \geq 0)$ iff for any finite subset z of $\mathbb{R}^+$, $\nu \circ t_z^{-1}$ is invariant for $(t_z(\Pi^\rho_t);\ t \geq 0)$.*

Proof. We obviously only prove the "if" part. We consider a probability measure ν and for each finite $z \subset \mathbb{R}^+$, we define $\nu_z := \nu \circ t_z^{-1}$. We assume that for any subset $z \in \mathbb{R}$, ν_z is invariant for $(t_z(\Pi_t^\rho))$. We assume that $\Pi_0^\rho = \pi_0$ is distributed according to ν. We want to prove that

$$\forall B \in \mathcal{F}, \quad \forall t \in \mathbb{R}^+, \ \mathbb{P}(\Pi_t^\rho \in B) = \mathbb{P}(\Pi_0^\rho \in B).$$

As $\mathcal{F}$ is the σ-field generated by $\mathcal{C}$, and $\mathcal{C}$ is closed under finite intersection, we only need to prove that for any z finite subset of $\mathbb{R}^+$,

$$\forall \pi \in \mathcal{P}_z, \ \forall t \in \mathbb{R}^+, \ \mathbb{P}(t_z(\Pi_t^\rho) = \pi) = \mathbb{P}(t_z(\Pi_0^\rho) = \pi).$$

As ν_z is invariant for $t_z(\Pi^\rho)$,

$$\forall \pi \in \mathcal{P}_z, \ \mathbb{P}(t_z(\Pi_t^\rho) = \pi) = \mathbb{P}(t_z(\Pi_0^\rho) = \pi) = \nu_z(\pi),$$

which completes the proof of Lemma 3.1. □

Lemma 3.2. *There exists a unique probability measure $\bar{\mu}^\rho$ on $(\mathcal{P}_\mathbb{Q}, \mathcal{F}_\mathbb{Q})$ charging right continuous partitions such that, for every finite $z \subset \mathbb{Q}^+$,*

$$t_z \star \bar{\mu}^\rho = \mu^{\rho,z}.$$

Furthermore, $\bar{\mu}^\rho$ only charges locally finite partitions of $\mathbb{Q}^+$ and for every $x \in \mathbb{Q}^+$,

$$\bar{\mu}^\rho(x \text{ is the extremity of a segment}) = 0.$$

Proof. From Proposition 2.2, the family $(\mu^{\rho,z};\ z \subset \mathbb{Q}^+)$ is consistent in the sense that for two finite subsets $z \subset z'$ then

$$t_z \star \mu^{\rho,z'} \ = \ \mu^{\rho,z}.$$

By an application of the Kolmogorov extension theorem, there exists a unique measure $\bar{\mu}^\rho$ defined on $(\mathcal{P}_\mathbb{Q}, \mathcal{F}_\mathbb{Q})$ such that for every finite subset z in $\mathbb{Q}$ we have

$$t_z \star \bar{\mu}^\rho \ = \ \mu^{\rho,z}.$$

(To see how one can apply Kolmogorov theorem in the context of consistent random partitions, we refer the reader to [Ber09], Proposition 2.1.)

We now need to prove that $\bar{\mu}^\rho$ only charges locally finite partitions of $\mathbb{Q}^+$. To do so, we follow closely [WH97]. We fix $a, b \in \mathbb{N},\ a < b$. We want to prove that, if π is a partition of $\mathbb{Q}$ distributed as $\bar{\mu}^\rho$, then $S_{[a,b]}$, the number of segments in $\pi|_{[a,b]\cap\mathbb{Q}}$ is finite almost surely.

To do so, we define

$$\begin{aligned}\forall n \in \mathbb{N}^*,\ \epsilon_n &:= 2^{-n},\\ X_{in} &:= \mathbb{1}_{((a+(i-1)\epsilon_n) \not\sim (a+i\epsilon_n))}\\ z_{in} &:= (a+(i-1)\epsilon_n, a+\epsilon_n) \in \mathbb{R}^2.\end{aligned}$$

In words, $X_{in} = 1$ if $(i-1)\epsilon_n$ and $i\epsilon_n$ belong to different segments. Let us compute the expectation of $S_{[a,b]}$. Using the monotone convergence theorem we have

$$\begin{aligned}\mathbb{E}(S_{[a,b]}) &= 1 + \mathbb{E}(\lim_{n\to\infty} \sum_{i=1}^{\lfloor 2^n(b-a)\rfloor} X_{in}) = 1 + \lim_{n\to\infty} \sum_{i=1}^{\lfloor 2^n(b-a)\rfloor} \mathbb{E}(X_{in})\\ &= 1 + \lim_{n\to\infty} \sum_{i=1}^{\lfloor 2^n(b-a)\rfloor} \mu^{\rho,z_{in}}(\{a+(i-1)\epsilon_n\}, \{a+i\epsilon_n\}).\end{aligned}$$

The ARG at rate ρ for the set of loci z_{in} has only two types of transitions: coagulation at rate 1 and fragmentation at rate $\rho\epsilon_n$, so

$$\mu^{\rho,z_{in}}(\{a+(i-1)\epsilon_n\}, a+\{i\epsilon_n\}) = \frac{\rho\epsilon_n}{1+\rho\epsilon_n}$$

which gives

$$\mathbb{E}(S_{[a,b]}) = 1 + \lim_{n\to\infty} \sum_{i=1}^{\lfloor 2^n(b-a)\rfloor} \frac{\rho 2^{-n}}{1+\rho\epsilon_n} = 1 + \rho(b-a).$$

Then $S_{[a,b]}$ is finite almost surely, which implies that $\bar{\mu}^\rho$ only charges locally finite partitions of $\mathbb{Q}$.

For the last statement let $x \in \mathbb{Q}^+$. By the previous argument,

$$\bar{\mu}^\rho(x \text{ is the extremity of a segment}) = \lim_{\epsilon\downarrow 0} \bar{\mu}^\rho(x-\epsilon \not\sim x+\epsilon) = 0,$$

which completes the proof. □

Lemma 3.3. *There exists a unique measure μ^ρ on $(\mathcal{P}^{loc}, \mathcal{F})$ such that*

$$t_{\mathbb{Q}} \star \mu^\rho \ = \ \bar{\mu}^\rho,$$

where $\bar{\mu}^\rho$ is the measure defined in Lemma 3.2.

Proof. Let $\tilde{\mathcal{P}}^{loc}_{\mathbb{Q}}$ the set of locally finite partitions of $\mathbb{Q}$ such that for all $x \in \mathbb{Q}^+$, x is not an extremity of a segment of π. Note that here we do not assume that the partitions of $\mathbb{Q}$ are right continuous. From the previous Lemma, $\bar{\mu}^\rho(\tilde{\mathcal{P}}^{loc}_{\mathbb{Q}}) = 1$. Similarly, let $\tilde{\mathcal{P}}^{loc}$ be the

set of elements π of $\mathcal{P}^{loc}$ such that for all $x \in \mathbb{Q}^+$, x is not an extremity of a segment of π. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, it is easy to see that for every $\bar{\pi} \in \tilde{\mathcal{P}}^{loc}_{\mathbb{Q}}$ there exists a unique $\tilde{\pi} \in \tilde{\mathcal{P}}^{loc}$ such that $t_{\mathbb{Q}}(\pi) = \bar{\pi}$. In other words, the projection map

$$t_{\mathbb{Q}} : (\tilde{\mathcal{P}}^{loc}, \mathcal{F}) \to (\tilde{\mathcal{P}}^{loc}_{\mathbb{Q}}, \mathcal{F}_{\mathbb{Q}})$$

is bijective. (Note that the condition that there are no rational extremities for the latter statement to hold, can be understood with the following counterexample. Let $\bar{\pi}$ be the partition of $\mathbb{Q}^+$ consisting of the two blocks $[0,1] \cap \mathbb{Q}$ and $]1, +\infty) \cap \mathbb{Q}$. Then there is no right-continuous partition $\pi \in \mathcal{P}^{loc}$ such that $t_{\mathbb{Q}}(\pi) = \bar{\pi}$.) With a little bit of extra work, one can show that $t_{\mathbb{Q}}^{-1}$ is measurable. As already mentioned in the proof of Theorem 1.1, the projection map

$$t_{\mathbb{Q}} : (\tilde{\mathcal{P}}^{loc}, \mathcal{F}) \to (\tilde{\mathcal{P}}^{loc}_{\mathbb{Q}}, \mathcal{F}_{\mathbb{Q}})$$

is bijective and measurable, so the measure μ^ρ defined by

$$\mu^\rho = t_{\mathbb{Q}}^{-1} \star \left[\bar{\mu}^\rho(\cdot \cap \tilde{\mathcal{P}}^{loc}_{\mathbb{Q}}) \right]$$

has mass 1 and satisfies

$$t_{\mathbb{Q}} \star \mu^\rho \ = \ \bar{\mu}^\rho.$$

To prove uniqueness, let μ on $(\mathcal{P}^{loc}, \mathcal{F})$ such that $t_{\mathbb{Q}} \star \mu \ = \ \bar{\mu}^\rho$. Because $\bar{\mu}^\rho$ only charges $\tilde{\mathcal{P}}^{loc}_{\mathbb{Q}}$,

$$t_{\mathbb{Q}} \star \mu = \bar{\mu}^\rho(\cdot \cap \tilde{\mathcal{P}}^{loc}_{\mathbb{Q}}).$$

Because μ only charges right continuous partitions, μ only charges $\tilde{\mathcal{P}}^{loc}$ (i.e., elements with no rational extrmities). Taking the pushforward of the two members of the previous equality by $t_{\mathbb{Q}}^{-1}$, we get

$$\mu(\cdot \cap \tilde{\mathcal{P}}^{loc}) = t_{\mathbb{Q}}^{-1} \star (t_{\mathbb{Q}} \star \mu) = t_{\mathbb{Q}}^{-1} \star \left[\bar{\mu}^\rho(\cdot \cap \tilde{\mathcal{P}}^{loc}_{\mathbb{Q}}) \right] = \mu^\rho.$$

Since μ only charges $\tilde{\mathcal{P}}^{loc}$, $\mu = \mu^\rho$. □

Proof of Theorem 1.2. We have proved that there exists a unique probability measure μ^ρ on $(\mathcal{P}^{loc}, \mathcal{F})$ such that, for any finite subset z of $\mathbb{Q}^+$, $t_z \star \mu^\rho$ is invariant for $(t_z(\Pi^\rho_t); t \geq 0)$ (by combining Lemmas 3.2 and 3.3). Using Lemma 3.1, we still need to prove that the same property holds for any finite subset $z \subset \mathbb{R}^+$. This will be shown by a continuity argument.

We fix $\rho > 0$. We denote by $\mathbb{P}^\rho$ the law of the process $(\Pi^\rho_t;\ t \geq 0)$, with initial condition Π^ρ_0 with law μ^ρ. We also fix $z = (z_1, \ldots, z_n) \subset \mathbb{R}^+$. For each $z^* = (z^*_1, \ldots, z^*_n) \subset \mathbb{Q}^+$, we define a function $g^* : \mathcal{P}_{z^*} \to \mathcal{P}_z$ such that, if π is a partition of z^*, $g^*(\pi)$ is the partition of z such that for every $i, j \in [n]$, $z_i \sim_{g^*(\pi)} z_j$ iff $z^*_i \sim_\pi z^*_j$. For every $t > 0$, we define the

event

$$A(z^*,t) = \{\forall s \in [0,t],\ t_z(\Pi_s^\rho) \ = \ g^*(t_{z^*}(\Pi_s^\rho))\}.$$

We want to prove that for every $t > 0$ and for $\mathcal{F}_z$-measurable bounded function f on $\mathcal{P}_z$,

$$\mathbb{E}^\rho(f(t_z(\Pi_t^\rho))) \ = \ \mathbb{E}^\rho(f(t_z(\Pi_0^\rho)).$$

As μ^ρ is a measure on $\mathcal{P}^{loc}$, for every $\epsilon > 0$ one can find $z^* = (z_1^*, \ldots, z_n^*) \subset \mathbb{Q}^+$ such that

$$\mathbb{P}^\rho(A(z^*,t)^\complement)||f||_\infty < \epsilon/2 \ \text{ and } \ |\mathbb{E}^\rho(f(g^*(t_{z^*}(\Pi_0^\rho)), A(z^*,t)^\complement)| < \epsilon/2.$$

Then

$$\begin{aligned}\mathbb{E}^\rho(f(t_z(\Pi_t^\rho))) \ &= \ \mathbb{E}^\rho(f(t_z(\Pi_t^\rho)), A(z^*,t)) + \mathbb{E}^\rho(f(t_z(\Pi_t^\rho)), A(z^*,t)^\complement) \\ &= \ \mathbb{E}^\rho(f(g^*(t_{z^*}(\Pi_t^\rho)), A(z^*,t)) + \mathbb{E}^\rho(f(t_z(\Pi_t^\rho)), A(z^*,t)^\complement).\end{aligned}$$

As $z^* \subset \mathbb{Q}^+$, $\mu^\rho \circ t_{z^*}^{-1}$ is invariant for $t_{z^*}(\Pi_t^\rho))$,

$$\mathbb{E}^\rho(f(g^*(t_{z^*}(\Pi_t^\rho)), A(z^*,t)) = \mathbb{E}^\rho(f(g^*(t_{z^*}(\Pi_0^\rho))) - \mathbb{E}^\rho(f(g^*(t_{z^*}(\Pi_0^\rho)), A(z^*,t)^\complement).$$

Then,

$$\begin{aligned}|\mathbb{E}^\rho(f(t_z(\Pi_t^\rho))) - \mathbb{E}^\rho(f(g^*(t_{z^*}(\Pi_0^\rho)))| \ &\leq \mathbb{P}^\rho(A(z^*,t)^\complement)||f||_\infty \\ &+|\mathbb{E}^\rho(f(g^*(t_{z^*}(\Pi_0^\rho)), A(z^*,t)^\complement)|\end{aligned}$$

so

$$|\mathbb{E}^\rho(f(t_z(\Pi_t^\rho))) - \mathbb{E}^\rho(f(t_z(\Pi_0^\rho))| \ < \ \epsilon,$$

and the conclusion follows by letting $\epsilon \to 0$. □

To conclude this section, we state an important property of μ^ρ.

Proposition 3.4 (Scaling)**.** *Fix $\rho > 0$. For every $\lambda \in> 0$, define $h_\lambda : \mathbb{R} \to \mathbb{R}$ such that $\forall x \in \mathbb{R},\ h_\lambda(x) = \lambda x$. Then*

$$h_\lambda \star \mu^\rho \ = \ \mu^{\lambda\rho}$$

Similarly, for any $z \subset R$,

$$h_\lambda \star \mu^{\rho,z} \ = \ \mu^{\lambda\rho,z}$$

Proof. This proposition can easily be deduced from the definition of the ARG and the scaling (I.2) and the construction of the $\mathbb{R}$-partitioning process given in the previous section. □

Without loss of generality, in Section 5, we will consider the partitioning process with recombination rate $\rho = 1$.

4 Proof of Theorem 1.3

Theorem 1.3, provides an approximation of the stationary measure of the discrete partitioning process when $\rho \to \infty$ or $\alpha \to \infty$, i.e. when recombination is much more frequent than coalescence. This approximation of $\mu^{\rho,z}$ is easy to handle, and that will be used in the proof of Theorem 1.4. We start by clarifying some notation that were already defined in the introduction and by introducing some new notation. In the following, we fix $z = (z_0, \dots, z_n)$ a finite subset of $\mathbb{R}$, and we define

$$\alpha = \min_{i \neq j} |z_i - z_j|$$

and we assume that $\alpha > 0$ (or equivalently that the coordinates of z are pairwise distinct).

Definition 4.1. *We consider the ARG ρ for the set of loci z, $\Gamma^{\rho,z}$. We say that a partition $\pi \in \mathcal{P}_z$ is of order r if it can be obtained from the finest partition ($\pi_0 := \{z_0\}, \dots, \{z_n\}$) by r successive coagulation events. We denote by $\mathcal{P}_z^k$ the subset of $\mathcal{P}_z$ containing all the partitions of order k.*

For example, for $i, j, k, l \in \{0, \dots, n\}$:

— $\pi_0 = \{z_0\}, \dots, \{z_n\}$ is the only partition of order 0.

— $\{z_0\}, \dots, \{z_i, z_j\}, \dots, \{z_n\}$ is of order 1.

— $\{z_0\}, \dots, \{z_i, z_j, z_k\}, \dots, \{z_n\}$ is of order 2.

— $\{z_0\}, \dots, \{z_i, z_j\}, \{z_k, z_l\}, \dots, \{z_n\}$ is also of order 2.

— $\{z_0, z_1, \dots, z_n\}$ is the only partition of order n.

Note that as the number of blocks decreases by 1 at each coalescence event, in a partition of order k, there are always $n + 1 - k$ blocks, so this definition is equivalent to the one given in the Introduction.

Definition 4.2. *Let $(s_k)_{0 \leq k \leq r}$ be a sequence of r elements of $\mathcal{P}_z$. The sequence (s_k) is called a "(coalescence) scenario of order r" if:*

— *s_0 is the finest partition.*

— *For $1 \leq k \leq r$, s_k is a partition of order k that can be obtained from s_{k-1} by a single coagulation event.*

If π is a partition of order r, we denote by $\mathcal{S}(\pi)$ the set of coalescence scenarios of order r, such that $s_r = \pi$.

For example, the partition $\{z_0\}, \dots, \{z_i, z_j, z_k\}, \dots, \{z_n\}$ can be obtained from the

finest partition with three different scenarios:

$$\begin{aligned}
\{z_i\}\{z_j\}\{z_k\}\dots &\to \{z_i,z_j\}\{z_k\}\ \dots &&\to \{z_i,z_j,z_k\}\\
\{z_i\}\{z_j\}\{z_k\}\dots &\to \{z_i,z_k\}\{z_j\}\ \dots &&\to \{z_i,z_j,z_k\}\\
\{z_i\}\{z_j\}\{z_k\}\dots &\to \{z_k,z_j\}\{z_i\}\ \dots &&\to \{z_i,z_j,z_k\}\dots
\end{aligned}$$

For $\pi \in \mathcal{P}_z$, let $b_1, b_2, \dots, b_k, \dots$ be the blocks of π. We denote by $C(\pi)$ the cover length of π defined as:

$$C(\pi) \ := \ \sum_i \max_{x,y\in b_i} |x-y|.$$

In particular, the cover length of π_0 is equal to 0.

If π_1 and π_2 are two partitions in $\mathcal{P}_z$, we define $\theta(\pi_1,\pi_2)$ as the transition rate from π_1 to π_2 in the finite partitioning process $\Gamma^{1,z}$ with recombination rate $\rho = 1$ (and we set $\theta(\pi_1,\pi_2) = 0$ if the transition is not possible). By definition, in the ARG $\Gamma^{\rho,z}$ (with recombination rate ρ), the transition rate from π_1 to π_2 is $\theta(\pi_1,\pi_2)$ if the transition corresponds to a coagulation event and $\rho\theta(\pi_1,\pi_2)$ if it is a fragmentation. It can readily be seen that,

$$\forall \pi \in \mathcal{P}_z^r, \quad \sum_{\omega\in\mathcal{P}_z^{r-1}} \theta(\pi,\omega) \ = \ C(\pi).$$

In words, when $\rho = 1$, the total fragmentation rate corresponds to the cover length. For general values of ρ, the fragmentation rate is the cover length multiplied by ρ.

Also, the total coalescence rate from a partition of order k only depends on n and k (and not in the values of $z_0, \dots, z_n$ and ρ) and is given by

$$\sum_{\omega\in\mathcal{P}_z^{r+1}} \theta(\pi,\omega) = \gamma_k := \frac{(n-k)(n-k+1)}{2},$$

where γ_k corresponds to the number of pairs of blocks in a partition of order k.

Definition 4.3. *Let $s = (s_k)_{0\le k\le r}$ be a scenario of coalescence of order r, with $1 \le r \le n$. We define the energy of s, $E(s)$ as:*

$$E(s) := \prod_{i=1}^{r} C(s_i) = \prod_{i=1}^{r} \sum_{\pi\in\mathcal{P}_z^{i-1}} \theta(s_i,\pi).$$

In words, the energy of a scenario corresponds to the product of the successive cover lengths at each step.

Now we can state the main result of this section, that gives an approximation of $\mu^{\rho,z}$, when ρ or α is large. The idea behind this theorem is that, when $\rho \gg 1$ or $\alpha \gg 1$, fragmentation events occur much more often than coalescence events. This implies that

the partition made of singletons is the most likely configuration and the probability of a partition decreases with its order. Define

$$\forall z \subset \mathbb{R}^+,\ \forall \pi \in \mathcal{P}_z \setminus \{\pi_0\}, \quad F(\pi) := \sum_{S \in \mathcal{S}(\pi)} \frac{1}{E(S)}. \tag{I.5}$$

We recall the statement of Theorem 1.3.

Theorem. *There exists a function*

$$f^n : \mathbb{R}_*^+ \to \mathbb{R}^+, \quad \lim_{x \to \infty} f^n(x) = 0,$$

independent of the choice of $z = (z_0, \dots, z_n)$ *and* ρ*, such that*

$$\forall \rho > 0,\ \forall k \in [n],\ \forall \pi_k \in \mathcal{P}_z^k, \quad \left| \mu^{\rho,z}(\pi_k) - \frac{1}{\rho^k} F(\pi_k) \right| \leq f^n(\alpha\rho) \frac{1}{\rho^k} F(\pi_k).$$

Before proving Theorem 1.3 we need to prove some technical results. But to give the reader some intuition on this result, we will start by giving a brief sketch of the proof. Until further notice, we are going to fix $\rho > 0,\ k \in [n],\ \pi \in \mathcal{P}_z^k$ a partition of order $k \geq 1$. We will start by defining some notation.

- $t_0^+ = \inf\{t > 0,\ \Gamma_t^{\rho,z} \neq \pi_0\}$.
- $\mathcal{T}_\pi = \inf\{t > 0,\ \Gamma_t^{\rho,z} = \pi\}$, $\mathcal{T}_0 = \inf\{t > t_0^+,\ \Gamma_t^{\rho,z} = \pi_0\}$.
- $\mathbb{P}_\pi$ (resp $\mathbb{P}_0$) denotes the law of $\Gamma^{\rho,z}$ conditioned on the initial condition $\Gamma_0^{\rho,z} = \pi$ (resp $\Gamma_0^{\rho,z} = \pi_0$).

Recall that the variables defined above depend on z and ρ, but for the sake of clarity this dependency is not made explicit.

The idea behind the proof of Theorem 1.3 is to use excursion theory and a well known extension of Blackwell's renewal theorem [Bla48] that states that

$$\mu^{\rho,z}(\pi) = \frac{\mathbb{E}_0(Y_1^\pi)}{\mathbb{E}_0(\Delta_0)}, \tag{I.6}$$

where Δ_0 is the time between two renewals at π_0 and Y_1^π is the time spent in π during an excursion out of π_0. (More precise definitions of these variables will be given in the proof of Theorem 1.3).

As we consider that $\alpha \gg 1$ or $\rho \gg 1$, fragmentation occurs much more often than coalescence so π_0 is the most likely configuration and $\Gamma^{\rho,z}$ spends most of the time at π_0. Then $\mathbb{E}_0(\Delta_0)$ can be approximated by the expectation of the holding time at π_0 which is $1/\gamma_0$. Also, in this regime, most excursions out of π_0 will only visit π at most one time, so

$\mathbb{E}_0(Y_1^\pi)$ can be approximated by

$$\mathbb{P}_0(\mathcal{T}_\pi < \mathcal{T}_0)\frac{1}{\rho C(\pi)},$$

where $\mathbb{P}_0(\mathcal{T}_\pi < \mathcal{T}_0)$ is the probability that π is reached during the excursion out of π_0 and $\frac{1}{\rho C(\pi)}$ is approximately the expectation of the holding time at π when $\rho C(\pi) \gg \gamma_k$ (i.e. when recombination occurs much more often than coalescence).

The core of the proof is to compute $\mathbb{P}_0(\mathcal{T}_\pi < \mathcal{T}_0)$. (This will be done in Corollary 4.6.)To do so, we will consider $\bar{\Gamma}^{\rho,z}$, the embedded chain of the ARG $\Gamma^{\rho,z}$, conditioned on the initial condition $\bar{\Gamma}_0^{\rho,z} = \pi_0$. We call a "direct path" a trajectory that goes from π_0 to π in only k coalescence steps (without recombination events). Indirect paths are trajectories that are longer and that contain at least a recombination event. As we consider a high recombination regime, where coalescence occurs much more often than recombination, direct paths will be much more likely that indirect paths. (This will be formalized in Lemma 4.5.) So we can approximate $\mathbb{P}_0(\mathcal{T}_\pi < \mathcal{T}_0)$ by the sum of the probabilities of the direct paths. Then the conclusion will follow by realizing that a direct path corresponds to a scenario of coalescence and showing that $\mathbb{P}_0(\mathcal{T}_\pi < \mathcal{T}_0)$ can be approximated by $\frac{C(\pi)}{\rho^{k-1}\gamma_0}F(\pi)$. (This will be formalized in Corollary 4.6.) Finally, replacing in (I.6), we find that $\mu^{\rho,z}(\pi)$ can be approximated by $\frac{F(\pi)}{\rho^k}$.

Before turning to the formal proof of Theorem 1.3, we start by proving some technical results. We consider $\bar{\Gamma}^{\rho,z}$, the embedded chain of the ARG $\Gamma^{\rho,z}$. Let P_0 denote the law of $\bar{\Gamma}^{\rho,z}$ conditioned on $\bar{\Gamma}_0^{\rho,z} = \pi_0$ and $\forall \pi' \in \mathcal{P}_z$, $P_{\pi'}$ denotes the law of $\bar{\Gamma}^{\rho,z}$ conditioned on $\bar{\Gamma}_0^{\rho,z} = \pi'$. We will consider paths that go from π_0 to π. A path is defined as follows.

Definition 4.4. *For $j \in \mathbb{N}^*, \pi', \pi'' \in \mathcal{P}_z$, we define:*

$$\begin{aligned} G(j, \pi' \to \pi'') \;=\; & \{(\pi^{(0)} = \pi', \pi^{(1)}, \dots, \pi^{(j-1)}, \pi^{(j)} = \pi''), \\ & \pi^{(1)}, \dots, \pi^{(j-1)} \in \mathcal{P}_z \setminus \{\pi', \pi''\} \textit{ such that} \\ & \theta(\pi^{(i)}, \pi^{(i+1)}) > 0 \quad \forall i \in \{0, \cdots, j-1\}\}. \end{aligned}$$

In words, $G(j, \pi' \to \pi'')$ contains every possible path (admissible for the partitioning process) that connects π' to π'' in j steps.

We are going to consider paths p that go from π_0 to π, which have at least k steps (as π is of order k).

- p is a *direct* path if $p \in G(k, \pi_0 \to \pi)$, i.e., p can only be composed of coalescence events.
- p is an *indirect* path if $p \in G(k + N, \pi_0 \to \pi)$, $N \in \mathbb{N}^*$. Indirect paths contain at least one recombination event. Note that the parity of the process implies that $G(k + 2N + 1, \pi_0 \to \pi)$ is empty.

Lemma 4.5. *Fix $N \in \mathbb{N}^*$ and a path p in $G(k+2N, \pi_0 \to \pi)$. There exists a path $\tilde{p} \in G(k, \pi_0 \to \pi)$ such that*

$$\frac{P_0(p)}{P_0(\tilde{p})} \leq \left(\frac{(1+\frac{\gamma_1}{\rho\alpha})^k}{\alpha\rho} \right)^N .$$

Proof of Lemma 4.5. We fix $N \in \mathbb{N}^*$ and we start with proving that

$$\forall p \in G(k+2N, \pi_0 \to \pi),\ \exists\ \tilde{p} \in G(k+2(N-1), \pi_0 \to \pi),\quad \frac{P_0(p)}{P_0(\tilde{p})} \leq \frac{(1+\frac{\gamma_1}{\rho\alpha})^k}{\alpha\rho}. \tag{I.7}$$

We consider a path $p \in G(k+2N, \pi_0 \to \pi)$ such that

$$p = (\pi_0, \bar{\pi}_1, \dots, \bar{\pi}_j, \tilde{\pi}_{j-1}, \pi_{i_1}, \pi_{i_2}, \dots, \pi).$$

where the indices of the $\tilde{\pi}, \bar{\pi}$'s coincide with the order of the partition (for instance, in the transition $\bar{\pi}_j \to \tilde{\pi}_{j-1}$, the order of the partition decreases by one unit, which corresponds to a fragmentation event). We do not specify the order of $\pi_{i_1}, \pi_{i_2}, \dots$ As $N \geq 1$ there is at least one recombination event $(\bar{\pi}_j \to \tilde{\pi}_{j-1})$. The path p can be decomposed into p_1 and p_2 such that:

$$\begin{aligned} p_1 \in G((j-1)+2, \pi_0 \to \tilde{\pi}_{j-1}),\ \ p_1 = (\pi_0, \bar{\pi}_1, \dots, \bar{\pi}_{j-1}, \bar{\pi}_j, \tilde{\pi}_{j-1}) \\ p_2 \in G(k+2N-(j-1)-2, \pi_{j-1} \to \pi),\ \ p_2 = (\tilde{\pi}_{j-1}, \pi_{i_1}, \pi_{i_2}, \dots, \pi). \end{aligned}$$

In words, we decompose p into two paths, p_1 that goes from π_0 until the first recombination event and p_2 that contains the rest of the path.

The idea now is to find a direct path

$$\tilde{p}_1 \in G(j-1, \pi_0 \to \tilde{\pi}_{j-1}),\ \ \tilde{p}_1 = (\pi_0, \tilde{\pi}_1, \dots, \tilde{\pi}_{j-1})$$

such that

$$\frac{P_0(p_1)}{P_0(\tilde{p}_1)} \leq \frac{1}{\alpha\rho}\left(1+\frac{\gamma_1}{\rho\alpha}\right)^{j-1} \leq \frac{1}{\alpha\rho}\left(1+\frac{\gamma_1}{\rho\alpha}\right)^{k}.$$

To do so, consider the fragmentation event that occurs between step j and step $j+1$ in p (when transitioning from $\bar{\pi}_j$ to $\tilde{\pi}_{j-1}$). $\bar{\pi}_j$ contains $n+1-j$ blocks and let $(b_1, \dots b_{n-j}, b^*)$ be the blocks of $\bar{\pi}_j$ such that b^* is the block of $\bar{\pi}_j$ that is fragmented during this fragmentation event and $z_a < z_b$ the two elements of b^* such that b^* is fragmented between z_a and z_b (i.e. such that b^* is fragmented into b_a^* and b_b^* where z_a is the rightmost element in b_a^* and z_b

the leftmost element in b_b^*). We have

$$C(\bar{\pi}_j) = C(\tilde{\pi}_{j-1}) + z_b - z_a. \tag{I.8}$$

Let $i^* \leq j$ be the first step of p such that z_a and z_b are in the same block, i.e

$$i^* = \min_{i \in [j]} \{i,\ z_a \sim_{\bar{\pi}_i} z_b\}.$$

We will construct a direct path $\tilde{p}_1 = (\tilde{\pi}_0, \cdots, \tilde{\pi}_{j-1})$ in such a way that

$$\begin{aligned} \forall\ 1 \leq i < i^*, \quad & C(\tilde{\pi}_i) \ \leq \ C(\bar{\pi}_i) \\ \text{if } i^* < j-1,\ \forall\ i^* < i \leq j, \quad & C(\tilde{\pi}_{i-1}) \ \leq \ C(\bar{\pi}_i), \end{aligned} \tag{I.9}$$

(Note that the terminal value of $\tilde{p}_1$ coincides with the terminal value of p_1 and its length is $j-1$ instead of $j+1$).) See Figure 4 for a concrete example. In words, we skip step i^*, and rearrange the path in such a way that $\tilde{p}_1$ is admissible, ends at $\tilde{\pi}_{j-1}$ and the inequalities (I.9) are satisfied along the way. Formally, the path $\tilde{p}_1$ is constructed as follows :

— If $i^* < j-1$, for $i \in \{i^*+1, \ldots, j-1\}$, let $(b_1^i, \ldots b_{n-i}^i, b_*^i)$ be the blocks of $\bar{\pi}_i$, where b_*^i is the one that contains z_a and z_b. The blocks of $\tilde{\pi}_{i-1}$ are $(b_1^i, \ldots, b_{n-i}^i, b_{n-i+1}^i, b_{n-i+2}^i)$ such that:

- if $z \in b_*^i$ and $z \leq z_a$, $z \in b_{n-i+1}^i$.
- if $z \in b_*^i$ and $z \geq z_b$, $z \in b_{n-i+2}^i$.

If $i^* = j-1$ we skip the present step in the construction of $\tilde{p}$.

— If in $\bar{\pi}_{i^*-1}$, z_a is the rightmost element in its block and z_b the leftmost element in its block, then we define $(\tilde{\pi}_1, \ldots, \tilde{\pi}_{i^*-1}) = (\bar{\pi}_1, \ldots, \bar{\pi}_{i^*-1})$. With this construction $\tilde{\pi}_{i^*}$ can be obtained from $\tilde{\pi}_{i^*-1}$ by a coalescence event, so the path $\tilde{p}$ is admissible for $\Gamma^{\rho,z}$.

— Else, $(\tilde{\pi}_1, \ldots, \tilde{\pi}_{i^*-1})$ are constructed from $(\bar{\pi}_1, \ldots, \bar{\pi}_{i^*-1})$ in the following way. Let us denote by b_a and b_b the blocks of $\bar{\pi}_{i^*-1}$ that contain z_a and z_b respectively. For $1 \leq i \leq i^* - 1$,

- If the coalescence event between $\bar{\pi}_{i-1}$ and $\bar{\pi}_i$ involves two blocks b_c and b_d such that in π_{i^*-1}, $b_c, b_d \subset b_a$ (resp. $b_c, b_d \subset b_b$) and if b_c contains an element that is smaller than z_a and b_d contains an element is larger than z_b, then in the coalescence step between $\tilde{\pi}_{i-1}$ and $\tilde{\pi}_i$, b_c (resp. b_d) coalesces with the block containing z_a (resp. z_b). (And nothing happens to b_d - resp. b_c).
- Otherwise the same coalescence event occurs between $\bar{\pi}_{i-1}$ and $\bar{\pi}_i$ and between $\tilde{\pi}_{i-1}$ and $\tilde{\pi}_i$.

With this construction $\tilde{\pi}_{i^*}$ can be obtained from $\tilde{\pi}_{i^*-1}$ by a coalescence event, and as a consequence the path $\tilde{p}$ is admissible, in the sense that $\theta(\pi_i, \pi_{i+1}) > 0$ (see Figure 4 for an example).

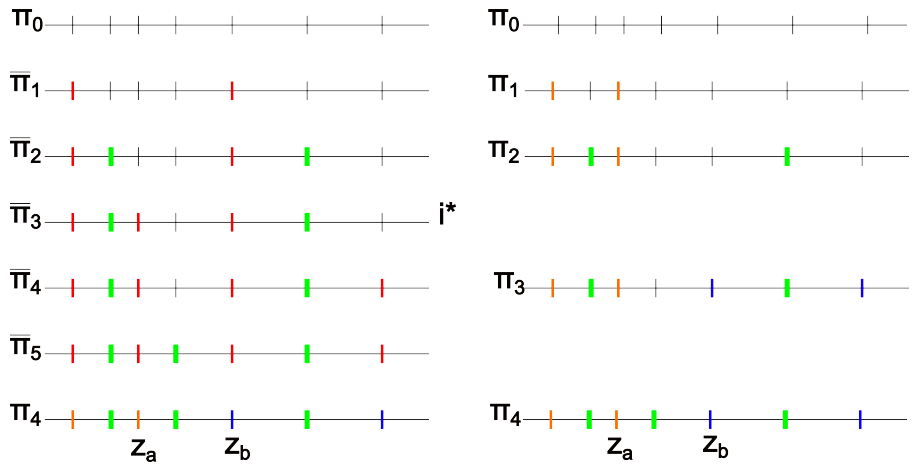


Figure I.5 – Example of two paths that go from π_0 to π_4, for $n = 7$. Loci in the same block are of the same color and the black loci corresponds to to loci that are in singleton blocks. The path on the right corresponds $p \in G(\pi_0 \to \pi_4, 4+2)$ and the path on the left is $\tilde{p} \in G(\pi_0 \to \pi_4, 4)$ constructed from π with the method presented above.

First,

$$P_0(\tilde{p}_1) = \frac{1}{\gamma_0 \rho^{j-2}} \prod_{i=1}^{j-1} \frac{1}{C(\tilde{\pi}_i) + \gamma_i/\rho}$$
$$P_0(p_1) = \frac{1}{\gamma_0 \rho^{j-1}} \prod_{i=1}^{j} \frac{1}{C(\bar{\pi}_i) + \gamma_i/\rho} \frac{\rho(z_b - z_a)}{\rho C(\bar{\pi}_j) + \gamma_j},$$

From (I.9), if $i^* < j-1$

$$\begin{aligned}
\frac{P_0(p_1)}{P_0(\tilde{p}_1)} &= \frac{1}{\rho} \frac{1}{C(\bar{\pi}_{i^*}) + \gamma_{i^*}} \prod_{i=1}^{i^*-1} \frac{C(\tilde{\pi}_i) + \gamma_i/\rho}{C(\bar{\pi}_i) + \gamma_i/\rho} \prod_{i=i^*+1}^{j} \frac{C(\tilde{\pi}_{i-1}) + \gamma_{i-1}/\rho}{C(\bar{\pi}_i) + \gamma_i/\rho} \frac{\rho(z_b - z_a)}{\rho C(\bar{\pi}_j) + \gamma_j} \\
&\leq \frac{1}{\alpha\rho} \prod_{i=i^*+1}^{j} \frac{C(\bar{\pi}_i) + \gamma_{i-1}/\rho}{C(\bar{\pi}_i) + \gamma_i/\rho} \leq \frac{1}{\alpha\rho} \prod_{i=i^*+1}^{j} \frac{1 + \frac{\gamma_{i-1}}{\rho C(\bar{\pi}_i)}}{1 + \frac{\gamma_i}{\rho C(\bar{\pi}_i)}} \\
&\leq \frac{1}{\alpha\rho} \prod_{i=i^*+1}^{j} (1 + \frac{\gamma_{i-1}}{\rho C(\bar{\pi}_i)}) \leq \frac{1}{\alpha\rho} \left(1 + \frac{\gamma_1}{\rho\alpha}\right)^k.
\end{aligned}$$

where the second inequality is a consequence of (I.8) and (I.9). The case $i^* = j-1$ follows along the same lines.

Let us define

$$\tilde{p} \in G(k+2, \pi_0 \to \pi), \;\; \tilde{p} = (\pi_0, \tilde{\pi}_1, \ldots, \tilde{\pi}_{j-1}, \pi_{i_1}, \pi_{i_2} \ldots, \pi).$$

Since

$$P_0(p) \;=\; P_0(p_1)\; P_{\tilde{\pi}_{j-1}}(p_2), \quad P_0(\tilde{p}) \;=\; P_0(\tilde{p}_1)\; P_{\tilde{\pi}_{j-1}}(p_2)$$

we have

$$\frac{P_0(p)}{P_0(\tilde{p})} = \frac{P_0(p_1)}{P_0(\tilde{p}_1)} \leq \; \frac{\left(1+\frac{\gamma_1}{\rho\alpha}\right)^k}{(\alpha\rho)},$$

which completes the proof of (I.7). Lemma 4.5 then follows by a simple induction on N using (I.7). □

Corollary 4.6. *There exists a function*

$$u^n : \mathbb{R}_*^+ \to \mathbb{R}^+, \quad \lim_{x\to\infty} u^n(x) = 0,$$

independent of the choice of z, π and ρ, such that

$$\left|\mathbb{P}_0(\mathcal{T}_\pi < \mathcal{T}_0) \;-\; \frac{C(\pi)}{\rho^{k-1}\gamma_0} F(\pi)\right| \leq u^n(\alpha R)\frac{C(\pi)}{\rho^{k-1}\gamma_0} F(\pi).$$

Proof. We have

$$\mathbb{P}_0(\mathcal{T}_\pi < \mathcal{T}_0) = \sum_{N\geq k} \sum_{p\in G(N,\pi_0\to\pi)} P_0(p).$$

As π is of order k, a path from π_0 to π has at least k steps. In addition, as the order of the partition can only increase or decrease by 1 at each step, a path from π_0 to π can only have $k+2N$ steps, with $N \geq 0$, so

$$\mathbb{P}_0(\mathcal{T}_\pi < \mathcal{T}_0) = \sum_{p\in G(k,\pi_0\to\pi)} P_0(p) + \sum_{N\geq 1} \sum_{p\in G(k+2N,\pi_0\to\pi)} P_0(p). \tag{I.10}$$

We start by considering the first term in the right hand side. We consider a path p such that

$$p \in G(k, \pi_0 \to \pi), \;\; p = (\pi_0, \pi_1, \ldots, \pi).$$

We have:

$$P_0(p) = \frac{1}{\gamma_0} \prod_{i=1}^{k-1} \frac{1}{\rho C(\pi_i) + \gamma_i}.$$

Recall that

$$F(\pi) \ = \ \sum_{s \in \mathcal{S}(\pi)} \prod_{i=1}^{k} \frac{1}{C(s_i)}.$$

Further, paths that have k steps are only composed of coalescence events, and therefore $G(k, \pi_0 \to \pi) = \mathcal{S}(\pi)$. It follows that

$$\sum_{p \in G(k,\pi_0 \to \pi)} P_0(p) - \ \frac{C(\pi)}{\rho^{k-1}\gamma_0} F(\pi) \ = \ \frac{1}{\rho^{k-1}\gamma_0} \ \sum_{s \in \mathcal{S}(\pi)} \left(\prod_{i=1}^{k-1} \frac{1}{C(s_i) + \gamma_i/\rho} - \prod_{i=1}^{k-1} \frac{1}{C(s_i)} \right)$$

and using the fact that $\gamma_i \leq \gamma_0$,

$$\begin{aligned} \sum_{s \in \mathcal{S}(\pi)} \left| \prod_{i=1}^{k-1} \frac{1}{C(s_i) + \gamma_i/\rho} - \prod_{i=1}^{k-1} \frac{1}{C(s_i)} \right| &= \sum_{s \in \mathcal{S}(\pi)} \prod_{i=1}^{k-1} \frac{1}{C(s_i)} \left(1 - \prod_{i=1}^{k-1} \frac{1}{1 + \gamma_i/(\rho C(s_i))} \right) \\ &\leq \sum_{s \in \mathcal{S}(\pi)} \prod_{i=1}^{k-1} \frac{1}{C(s_i)} \left(1 - \left(\frac{1}{1 + \gamma_0/(\rho\alpha)} \right)^{k-1} \right) \\ &\leq C(\pi) F(\pi) \left(1 - \frac{1}{(1 + \gamma_0/(\rho\alpha))^{k-1}} \right) \end{aligned}$$

so

$$\left| \sum_{p \in G(k,\pi_0 \to \pi)} P_0(p) - \ \frac{C(\pi)}{\rho^{k-1}\gamma_0} F(\pi) \right| \ \leq \ \frac{C(\pi)}{\rho^{k-1}\gamma_0} F(\pi) \left(1 - \frac{1}{(1 + \gamma_0/(\rho\alpha))^{k-1}} \right). \qquad \text{(I.11)}$$

To prove Proposition 4.6, we still need to consider the second term in the right hand side of (I.10). Using Lemma 4.5, we have

$$\begin{aligned} &\sum_{N \geq 1} \sum_{p \in G(k+2N, \pi_0 \to \pi)} P_0(p) \\ &\leq \sum_{\tilde{p} \in G(k, \pi_0 \to \pi)} P_0(\tilde{p}) \ \left(\sum_{N \geq 1} |G(k + 2N, \pi_0 \to \pi)| \left(\frac{(1 + \frac{\gamma_1}{\rho\alpha})^k}{\alpha\rho} \right)^N \right). \end{aligned} \qquad \text{(I.12)}$$

To compute $|G(k + 2N, \pi_0 \to \pi)|$, let us recall that, at each step in a path:

— If it corresponds to a coalescence event from a partition of order j there are γ_j possibilities, and $\forall j \in \{0, \ldots, n\}$ $\gamma_j \leq n(n+1)$.

— If it corresponds to a fragmentation event, there are at most $(n+1)$ blocks in the

partition and each one contains at most $(n+1)$ elements, so that each block can be fragmented in n different ways.

From there, it can easily be seen that

$$|G(k+2N, \pi_0 \to \pi)| \leq (n(n+1))^{k+2N}.$$

Combining this with (I.12), we have:

$$\begin{aligned}
&\sum_{N\geq 1} \sum_{p\in G(k+2N,\pi_0\to\pi)} P_0(p) \\
&\leq \left(\sum_{p\in G(k,\pi_0\to\pi)} P_0(p) \right) (n(n+1))^k \sum_{N\geq 1} \left(\frac{n^2(n+1)^2(1+\frac{\gamma_1}{\rho\alpha})^k}{\alpha\rho} \right)^N,
\end{aligned}$$

which combined with (I.10) and (I.11), gives:

$$\left| \mathbb{P}_0(\mathcal{T}_\pi < \mathcal{T}_0) - \frac{C(\pi)}{\rho^{k-1}\gamma_0} F(\pi) \right| \leq \frac{C(\pi)}{\rho^{k-1}\gamma_0} F(\pi) u^{n,k}(\alpha\rho)$$

where $u^{n,k}$ is a function independent of z and ρ and vanishing at ∞. The conclusion follows by setting $u^n(\alpha\rho) = \max_{k\in[n]}(u^{n,k}(\alpha\rho))$. □

Before stating the last technical result that is needed in the proof of Theorem 1.3, we need to introduce some notation:

- $t_\pi^+ = \inf\{t > 0,\ \Gamma_t^{\rho,z} \neq \pi\}$
- $T_\pi = \inf\{t > t_\pi^+,\ \Gamma_t^{\rho,z} = \pi\}$, $T_0 = \inf\{t > t_\pi^+,\ \Gamma_t^{\rho,z} = \pi_0\}$.

Lemma 4.7. *For any $n \in \mathbb{N}$, there exist two functions g^n and h^n such that*

$$\lim_{x\to\infty} g^n(x) = 0,\ \lim_{x\to\infty} h^n(x) = 1$$

independent on the choice of z, π and ρ such that

$$\begin{aligned}
&(i)\ \ \mathbb{E}_0(\mathcal{T}_0 - t_0^+) \leq g^n(\alpha\rho) \\
&(ii)\ \forall k > 0, \forall \pi \in \mathcal{P}_z^k,\ \mathbb{P}_\pi(T_0 < T_\pi) \geq h^n(\alpha\rho).
\end{aligned}$$

Proof. We fix $\rho > 0, n \in \mathbb{N}, z = (z_0, \ldots, z_n),\ k \in [n], \pi \in \mathcal{P}_z^k$.

The idea of the proof is to consider the stochastic process $(X_t^{\rho,z}; t \geq 0)$ valued in $\{0, \ldots, n\}$ and such that $\forall t \geq 0$, $X_t^{\rho,z}$ is the order of the partition $\Gamma_t^{\rho,z}$. This process is not Markovian, but it can easily be compared to a Markov process $(W_t^{\rho,z}, t \geq 0)$ in such a way that the excursions out of 0 of $W^{\rho,z}$ are longer than those of $X^{\rho,z}$.

More precisely, let $W^{\rho,z}$ be the birth-death process in $\{0,\cdots,n\}$ where all the death rates are equal to $\rho\alpha$ and the birth rate at state k is γ_k (note that $\gamma_n = 0$).

With these transition rates, for any $\pi_k \in \mathcal{P}_z^k$, the total coalescence rate from π_k for the process $\Gamma_t^{\rho,z}$ is the same as the birth rate from k for $W_t^{\rho,z}$. On the other hand, the total fragmentation rate for $\Gamma_t^{\rho,z}$ when $\Gamma_t^{\rho,z} = \pi_k$ is equal to $\rho C(\pi_k)$ and is always higher than the death rate at k for $W_t^{\rho,z}$. We can find a coupling between $W^{\rho,z}$ and $X^{\rho,z}$ such that the holding times at 0 of the two process are the same (as the birth rate in 0 for $W^{\rho,z}$ is the same as the coagulation rate from 0 for $\Gamma^{\rho,z}$). In addition, during an excursion out of 0, the holding time at $k > 0$ for $X^{\rho,z}$ is shorter than the holding time at k for $W^{\rho,z}$ and the embedded chain of $X^{\rho,z}$ jumps more easily to the right than the embedded chain of $W^{\rho,z}$.

Let us denote by $\bar{\mathbb{E}}_0$ the probability with respect to the distribution of $W^{\rho,z}$, conditional to $W_0^{\rho,z} = 0$, and define

$$\begin{aligned} \bar{t}_0^+ &= \inf\{t > 0, W^{\rho,z} \neq 0\} \\ \bar{\mathcal{T}}_0 &= \inf\{t > \bar{t}_0^+,\ W_t^{\rho,z} = 0\}. \end{aligned}$$

By construction, we have

$$\mathbb{E}_0(\mathcal{T}_0 - t_0^+) \ \leq\ \bar{\mathbb{E}}_0(\bar{\mathcal{T}}_0 - \bar{t}_0^+).$$

Finally, $\bar{\mathbb{E}}_0(\bar{\mathcal{T}}_0 - \bar{t}_0^+)$ only depends on $\rho\alpha$ and n and it can be checked that

$$\bar{\mathbb{E}}_0(\bar{\mathcal{T}}_0 - \bar{t}_0^+) \underset{\alpha\rho\to\infty}{\longrightarrow} 0$$

so (i) is verified.

(ii) can be handled by similar methods. Namely, let $\bar{W}^{\rho,z}$ denote the embedded chain of $W^{\rho,z}$

$$\begin{aligned} \mathbb{P}_\pi(T_0 < T_\pi) &\geq \mathbb{P}(\bar{W}_0^{\rho,z} = k, \bar{W}_1^{\rho,z} = k-1, \ldots, \bar{W}_k^{\rho,z} = 0) \\ &= \prod_{i=1}^{k} \frac{\rho\alpha}{\rho\alpha + \gamma_i} \geq \prod_{i=1}^{n} \frac{\rho\alpha}{\rho\alpha + \gamma_i} \underset{\rho\alpha\to\infty}{\longrightarrow} 1 \end{aligned}$$

where the first inequality is obtained by the same argument as in (i). This completes the proof of Lemma 4.7. □

We are now ready to prove the main results of this section.

Proof of Theorem 1.3. We will consider excursions of $(\Gamma_t^{\rho,z})$ out of π_0. Let us consider $(J_i)_{i\in\mathbb{N}}$ the renewal times at π_0 i.e. the successive jump times of $(\Gamma_t^{\rho,z})$ such that $\Gamma_{J_i}^{\rho,z} = \pi_0$

(and $\Gamma^{\rho,z}_{J_i-} \neq \pi_0$). For $i \in \mathbb{N}^*$, let us define

$$\Delta_0^i := J_i - J_{i-1}$$

the time between two renewals at π_0. The $(\Delta_0^i)_{i\in\mathbb{N}}$ are independent and identically distributed random variables. Also, for $i \in \mathbb{N}^*$, consider

$$Y_i^\pi := \int_{J_{i-1}}^{J_i} \mathbb{1}_{\Gamma_t^{\rho,z}=\pi} dt,$$

which corresponds to the time spent by $\Gamma^{\rho,z}$ in π during the i^{th} excursion out of π_0. By standard excursion theory, the Y_i^π's are independent and identically distributed random variables.

From the ergodic theorem we have

$$\begin{aligned}\mu^{\rho,z}(\pi) &= \lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbb{1}_{\{\Gamma_s^{\rho,z}=\pi\}} ds \quad \text{a.s.} \\ &= \lim_{n\to\infty} \frac{1}{J_n}\sum_{k=1}^{n}\int_{J_{n-1}}^{J_n} \mathbb{1}_{\{\Gamma_s^{\rho,z}=\pi\}} ds \quad \text{a.s..}\end{aligned}$$

Since the excursions are independent from one another, using Blackwell's renewal theorem [Bla48], and the law of large numbers,

$$\lim_{n\to\infty}\frac{n}{J_n} = \frac{1}{\mathbb{E}_0(\Delta_0^1)} \quad \text{and} \quad \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\int_{J_{n-1}}^{J_n} \mathbb{1}_{\{\Gamma_s^{\rho,z}=\pi\}} ds = \mathbb{E}_0(Y_1^\pi) \quad \text{a.s.},$$

which easily gives

$$\mu^{\rho,z}(\pi) = \frac{\mathbb{E}_0(Y_1^\pi)}{\mathbb{E}_0(\Delta_0^1)}.$$

Let H_0 be the holding time at π_0 and H the holding time at π. H_0 follows an exponential distribution of parameter γ_0. H follows an exponential distribution of parameter $(C(\pi) + \gamma_k)$. By standard excursion theory,

$$\begin{aligned}\mathbb{E}(Y_1^\pi) &= \mathbb{P}_0(\mathcal{T}_\pi < \mathcal{T}_0)\sum_{k\geq 1} k\mathbb{E}(H)\mathbb{P}_\pi(T_\pi < T_0)^{k-1}\mathbb{P}_\pi(T_0 < T_\pi) \\ &= \mathbb{P}_0(\mathcal{T}_\pi < \mathcal{T}_0)\frac{1}{(\rho C(\pi) + \gamma_k)}\frac{1}{\mathbb{P}_\pi(T_0 < T_\pi)},\end{aligned}$$

and

$$\mathbb{E}(\Delta_0^1) = \mathbb{E}(H_0) + \mathbb{E}_0(\mathcal{T}_0 - t_0^+) \text{ where } \mathbb{E}(H_0) = \frac{1}{\gamma_0}.$$

Combining this with Corollary 4.6, we have

$$\begin{aligned}
&\left| \mu^{\rho,z}(\pi) \;-\; \frac{C(\pi)F(\pi)}{\gamma_0\rho^{k-1}} \frac{1}{(\rho C(\pi)+\gamma_k)} \frac{1}{\mathbb{P}_\pi(T_0<T_\pi)} \;\frac{1}{1/\gamma_0+\mathbb{E}_0(\mathcal{T}_0-t_0^+)} \right| \\
&\quad\leq\; u^n(\alpha\rho)\frac{C(\pi)F(\pi)}{\gamma_0\rho^{k-1}} \frac{1}{(\rho C(\pi)+\gamma_k)} \frac{1}{\mathbb{P}_\pi(T_0<T_\pi)} \;\frac{1}{1/\gamma_0+\mathbb{E}_0(\mathcal{T}_0-t_0^+)} \\
&\quad\leq\; u^n(\alpha\rho)\; \frac{F(\pi)}{\rho^k} \;\frac{1}{1+\frac{\gamma_k}{\rho C(\pi)}} \;\frac{1}{\mathbb{P}_\pi(T_0<T_\pi)} \;\frac{1}{1+\mathbb{E}_0(\mathcal{T}_0-t_0^+)\gamma_0}
\end{aligned}$$

so

$$\begin{aligned}
&\left| \mu^{\rho,z}(\pi) \;-\; \frac{F(\pi)}{\rho^k} \right| \\
&\leq\; u^n(\alpha\rho)\; \frac{F(\pi)}{\rho^k} \;\frac{1}{1+\frac{\gamma_k}{\rho C(\pi)}} \;\frac{1}{\mathbb{P}_\pi(T_0<T_\pi)} \;\frac{1}{1+\mathbb{E}_0(\mathcal{T}_0-t_0^+)\gamma_0} \\
&+\; \left| \frac{C(\pi)F(\pi)}{\gamma_0\rho^{k-1}} \frac{1}{(\rho C(\pi)+\gamma_k)} \frac{1}{\mathbb{P}_\pi(T_0<T_\pi)} \;\frac{1}{1/\gamma_0+\mathbb{E}_0(\mathcal{T}_0-t_0^+)} \;-\; \frac{F(\pi)}{\rho^k} \right| \\
&\leq \frac{F(\pi)}{\rho^k} \left(u^n(\alpha\rho)\; \frac{1}{\mathbb{P}_\pi(T_0<T_\pi)} \;\frac{1}{1+\mathbb{E}_0(\mathcal{T}_0-t_0^+)\gamma_0} \right. \\
&+\; \left. \left| 1 \;-\; \frac{1}{1+\frac{\gamma_k}{\rho C(\pi)}} \;\frac{1}{\mathbb{P}_\pi(T_0<T_\pi)} \;\frac{1}{1+\mathbb{E}_0(\mathcal{T}_0-t_0^+)\gamma_0} \right| \right)
\end{aligned}$$

and using Lemma 4.7 (and the fact that $\rho C(\pi) \geq \rho\alpha$), the term between parentheses can be bounded by $f^n(\alpha\rho)$, where f^n is independent on the choice of z and ρ and is such that

$$\lim_{x\to+\infty} f^n(x) = 0,$$

which completes the proof of Theorem 1.3. □

5 Proof of Theorem 1.4

Thanks to Proposition 3.4 (scaling), in this section we will assume without loss of generality that $\rho = 1$ and consider the $\mathbb{R}$-partitioning process restricted to $[0, R]$. The strategy of the proof is based on the following lemma. (Note that the second point will allow us to rephrase the convergence of ϑ^R in the weak topology in terms of a moment problem).

Lemma 5.1. *(i) For every k-tuple of disjoint intervals $\{[a_i, b_i]\}_{i=1}^k$ in $[0,1]$ and any k-tuple of integers $\{n_i\}_{i=1}^k$*

$$\mathbb{E}\left(\prod_{i=1}^k \vartheta^\infty([a_i,b_i])^{n_i}\right) \;=\; \prod_{i=1}^k n_i!\; b_i^{n_i-1}(b_i-a_i).$$

(*ii*) *Let* $\{\nu^R\}_{R\geq 0}$ *be a sequence of random variables in* $\mathcal{M}([0,1])$ *that have no atoms and such that and for every* k*-tuple of disjoint intervals* $\{[a_i,b_i]\}_{i=1}^k$ *in* $[0,1]$ *and any* k*-tuple of integers* $\{n_i\}_{i=1}^k$

$$\lim_{R\to\infty} \mathbb{E}\left(\prod_{i=1}^k \nu^R([a_i,b_i])^{n_i}\right) \;=\; \prod_{i=1}^k n_i!\; b_i^{n_i-1}(b_i-a_i).$$

Then $\nu^R \underset{R\to\infty}{\Longrightarrow} \vartheta^\infty$ *in the weak topology.*

Proof of Lemma 5.1. We start by proving (i). We fix $k=1$. We fix $a,b\in[0,1], a\leq b$ and we compute M, the moment generating function of $\vartheta^\infty([a,b])$.

$$\begin{aligned} M(t) &= \mathbb{E}(\exp(t\vartheta^\infty[a,b])) \\ &= \mathbb{E}(\exp(t \sum_{(x_i,y_i)\in\mathcal{P}^\infty,\; x_i\in[a,b]} y_i)). \end{aligned}$$

$M(t)$ is the Laplace functional of $\mathcal{P}^\infty$ for $f(x,y)=-ty$, so it is well known that:

$$\begin{aligned} M(t) &= \exp\left(-\int_{[a,b]\times\mathbb{R}^+}(1-e^{ty})\lambda(x,y)dxdy\right) \\ &= \exp\left(-\int_{[a,b]}\frac{dx}{x}\int_{\mathbb{R}^+}(1-e^{ty})\;\frac{1}{x}e^{-y/x}dy\right) \\ &= \exp\left(\int_a^b \frac{tx}{1-tx}\frac{dx}{x}\right) \;=\; \exp\left(\log\left(\frac{1-ta}{1-tb}\right)\right) \\ &= \frac{1-ta}{1-tb}. \end{aligned}$$

Note that when $a=0$, $M(t)$ is the moment generating function of an exponential distribution of parameter $1/b$ and $M^{(n)}(0)=n!b^n$. When $a\neq 0$, we use a Taylor expansion of $M(t)$:

$$\begin{aligned} \frac{1-ta}{1-tb} &= (1-ta)\sum_{n=0}^\infty (tb)^n \\ &= \sum_{n=0}^\infty (tb)^n - \sum_{n=1}^\infty ab^{n-1}t^n \\ &= 1+\sum_{n=1}^\infty \frac{n!\; b^{n-1}(b-a)}{n!}t^n, \end{aligned}$$

so $M^{(n)}(0)=n!b^{n-1}(b-a)$ for $n\geq 1$, which implies (i). To prove this result for $k>1$, we use the fact that $\mathcal{P}^\infty$ is a Poisson point process so that, for any k-tuple of disjoint intervals

$B_1, \dots B_k$, $\vartheta^\infty(B_1), \dots, \vartheta^\infty(B_k)$ are mutually independent.

We now turn to the proof of (ii). Let $\{\nu^R\}_{R\geq 0}$ be a sequence of random variables in $\mathcal{M}([0,1])$. Note that for every $x \in [0,1]$, ϑ^∞ does not charge x almost surely. From [Kal02] (Theorem 16.16 page 316), it follows that proving

$$\nu^R \underset{R\to\infty}{\Longrightarrow} \vartheta^\infty \text{ in the weak topology}$$

boils down to proving that $\forall n \in \mathbb{N}$, for any k-tuple of intervals $B_1, \dots, B_k$

$$(\nu^R(B_1), \dots, \nu^R(B_k)) \underset{R\to\infty}{\Longrightarrow} (\vartheta^\infty(B_1), \dots, \vartheta^\infty(B_k)). \tag{I.13}$$

To prove (I.13), we use a method of moments. We will apply an extension of Carleman's condition for multi-dimensional random variables [KS13, ST50].

Fix $n, k \in \mathbb{N}$ and for a given k-tuple of disjoint intervals $\{[a_i, b_i]\}_{i=1}^k$, define

$$M_n^k = \sum_{i=1}^k \mathbb{E}(\vartheta^\infty([a_i,b_i])^n), \quad C = \sum_{n=1}^\infty (M_n^k)^{-\frac{1}{2n}}.$$

The condition states that, if $C = \infty$ (for any choice of k and $\{[a_i, b_i]\}_{i=1}^k$ that are not necessarily disjoint), proving (I.13) is equivalent to proving that for $k \in \mathbb{N}$, $n_1, \dots, n_k \in \mathbb{N}^k$

$$\mathbb{E}\left(\prod_{i=1}^k \nu^R([a_i,b_i])^{n_i}\right) \underset{R\to\infty}{\longrightarrow} \mathbb{E}\left(\prod_{i=1}^k \vartheta^\infty([a_i,b_i])^{n_i}\right). \tag{I.14}$$

From (i), we have

$$M_n^k \;=\; \sum_{i=1}^k n!\; b_i^{n-1}\; (b_i - a_i) \leq kn!$$

and since

$$\begin{aligned}\sum_{n=1}^\infty \frac{1}{(kn!)^{\frac{1}{2n}}} \;&\geq\; \frac{1}{k}\sum_{k=1}^\infty \frac{1}{(n!)^{\frac{1}{2n}}} \\ &\geq\; \frac{1}{k}\sum_{k=1}^\infty \frac{1}{n^{\frac{1}{2}}} = \infty\end{aligned}$$

we get $C = \infty$ and we can apply the extension of Carleman's condition. We use the fact that

$$\forall a \leq b \leq c, \quad \nu^R[a,c] = \nu^R[a,b] + \nu^R[b,c] \;\text{ and }\; \vartheta^\infty[a,c] = \vartheta^\infty[a,b] + \vartheta^\infty[b,c]$$

so that (I.14) reduces to the case where the intervals $\{[a_i, b_i]\}_{i=1}^k$ are pairwise disjoint. This

completes the proof of Lemma 5.1. □

Since ϑ^R is absolutely continuous with respect to the Lebesgue measure,

$$\forall a \leq b \leq c, \quad \vartheta^R[a,c] = \vartheta^R[a,b] + \vartheta^R[b,c],$$

so, from Lemma 5.1, the proof of Theorem 1.4 boils down to proving that for every k-tuple of disjoint intervals $\{[a_i,b_i]\}_{i=1}^k$ in $[0,1]$ and any k-tuple of integers $\{n_i\}_{i=1}^k$

$$\lim_{R\to\infty} \mathbb{E}\left(\prod_{i=1}^k \vartheta^R([a_i,b_i])^{n_i}\right) = \prod_{i=1}^k n_i!\, b_i^{n_i-1}(b_i-a_i). \tag{I.15}$$

The rest of this section will be dedicated to the proof of this asymptotical relation. We start by fixing $k \in \mathbb{N}$, $n_1, \dots n_k \in \mathbb{N}^k$, $n = n_1 + \ldots + n_k$ and $\{[a_i,b_i]\}_{i=1}^k$ a k-tuple of disjoint intervals. Without loss of generality we assume $a_1 < b_1 < a_2 \ldots < a_k < b_k$. For any $z = (z_0, z_1, \dots, z_n) \subset \mathbb{R}^+$ we define $c(z)$ as the coarsest partition of z.

We start by rewriting the right hand side of the equation (I.14):

$$\begin{aligned}
&\mathbb{E}\left(\prod_{i=1}^k \vartheta^R([a_i,b_i])^{n_i}\right) \\
&= \frac{1}{\log(R)^n}\, \mathbb{E}_{\mu^1}\left(\int_{[R^{a_1},R^{b_1}]^{n_1}\times\ldots\times[R^{a_k},R^{b_k}]^{n_k}} \mathbb{1}_{\{0\sim z_1\sim\ldots\sim z_n\}} dz_1\ldots dz_n\right) \\
&= \frac{1}{\log(R)^n} \int_{[R^{a_1},R^{b_1}]^{n_1}\times\ldots\times[R^{a_k},R^{b_k}]^{n_k}} \mu^{z,1}(c(z)) dz_1\ldots dz_n
\end{aligned} \tag{I.16}$$

where $\mathbb{E}_{\mu^1}$ denotes the expectation with respect to μ^1, $z_0 = 0$ and $\mu^{z,1}$ is defined as the invariant measure of the partitioning process for the set of loci $z = (z_0, \cdots, z_n)$ with a recombination rate equal to 1.

Let us now give some intuition for the rest of the section. Let V_R be the volume of the integration domain above. We have

$$\mathbb{E}\left(\prod_{i=1}^k \vartheta^R([a_i,b_i])^{n_i}\right) = \frac{V_R}{\log(R)^n}\mathbb{E}_Z\left(\mu^{(z_0,Z),1}(\{z_0,Z\})\right),$$

where $\mathbb{E}_Z$ denotes the expectation with respect to $Z = (Z_1, \dots, Z_n)$ distributed as a uniform random variable on $[R^{a_1},R^{b_1}]^{n_1} \times \ldots \times [R^{a_k},R^{b_k}]^{n_k}$ and where we recall that $\mu^{z,1} = t_z \star \mu$. When $R \gg 1$, for a "typical" configuration Z, the distances between the z_i's will be of order R. As $\rho = 1$, the fragmentation rates correspond to the distances between the z_i's and are of order $R \gg 1$, whereas the coalescence rate is always 1 for each pair of blocks. In this situation, fragmentation events occur much more often than

coalescence events, which is the framework of Theorem 1.3. The main idea behind (I.15) is to approximate the integrand using this theorem.

Let us now go into the details of the proof. We decompose the proof into four steps. In the following $\{[a_i, b_i]\}_{i=1}^n$ will denote a set of disjoint intervals listed in increasing order.

Step 1. Define

$$C_\beta^R := \{z_1, \dots, z_n \in \otimes_{i=1}^n [R^{a_i-1}, R^{b_i-1}]^{n_i} : \text{ s.t. if } z_0 := 0 \\ \forall i \neq j \in \{0, \dots, n\},\ |z_i - z_j| \geq \beta\}$$

(Note that in the rest of the proof, we will always set $z_0 = 0$.) The aim of this step is to prove the following proposition.

Proposition 5.2.

$$\forall \beta > 0, \quad \lim_{R\to\infty} \frac{1}{\log(R)^n} \int_{C_\beta^R} F(c(z)) dz_1 \dots dz_n = \prod_{i=1}^k n_i!\ b_i^{n_i-1}(b_i - a_i), \tag{I.17}$$

where, as defined in Section 4,

$$F(c(z)) = \sum_{s\in\mathcal{S}(c(z))} \frac{1}{E(s)},$$

and where $E(s)$ is the product of the successive cover lengths along the coalescence scenario s.

To see why this Proposition is useful for the proof of (I.15), we let the reader refer to Steps 2 and 3.

In the following we fix $\beta \geq 1$, and we assume that R is large enough so that $\forall i \in [k],\ R^{b_i-1} > \beta R^{-1}$. Let Σ_n be the set of permutations of $[n]$. For $\sigma \in \Sigma_n$ define

$$C_{\beta,\sigma}^R := \{z_1, \dots, z_n \in C_\beta^R,\ z_{\sigma(1)} < \dots . < z_{\sigma(n)}\}.$$

Recall that, as the intervals $\{[a_i, b_i]\}_{i=1}^n$ are disjoint, the z_i's belonging to $[a_j, b_j]$ are always smaller than those belonging to $[a_{j+1}, b_{j+1}]$. This means that there are only $n_1! \dots n_k!$ permutations for which $C_{\beta,\sigma}^R$ is non empty. Using the symmetry between the z_i's belonging to the same interval, we have

$$\begin{aligned} \int_{C_\beta^R} F(c(z))\ dz_1 \dots dz_n &= \sum_{\sigma\in\Sigma_n} \int_{C_{\beta,\sigma}^R} F(c(z))\ dz_1 \dots dz_n \\ &= n_1! \dots n_k! \int_{C_{\beta,Id}^R} F(c(z))\ dz_1 \dots dz_n, \end{aligned}$$

where Id is the identity permutation. To prove Proposition 5.2, it remains to show that:

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \int_{C^R_{\beta,Id}} F(c(z))dz_1 \dots dz_n = \prod_{i=1}^{k} b_i^{n_i-1}(b_i - a_i). \tag{I.18}$$

Recall that Finvolves over coalescence scenarios. The idea now is to consider separately two different types of scenarios of coalescence.

- $\mathcal{S}_C(c(z))$ corresponds to the set of the "contiguous scenarios" i.e. the scenarios where blocks only coalesce with their neighbouring blocks (i.e where at each step the block containing z_i can only coalesce with the blocks containing z_{i-1} or z_{i+1}). This is for example the case of scenarios S_1 and S_2 in Figure 5.
- $\bar{\mathcal{S}}_C(c(z))$ contains all the other scenarios (for example S_3 and S_4 in Figure 5).

$$\begin{aligned}\int_{C^R_{\beta,Id}} F(c(z))dz_1 \dots dz_n &= \int_{C^R_{\beta,Id}} \sum_{s\in\mathcal{S}_C(c(z))} \frac{1}{E(s)} dz_1 \dots dz_n \\ &\quad + \int_{C^R_{\beta,Id}} \sum_{s\in\bar{\mathcal{S}}_C(c(z))} \frac{1}{E(s)} dz_1 \dots dz_n. \end{aligned} \tag{I.19}$$

The rest of this step is devoted to the computation of each of the terms in the RHS of this equation.

Step 1.1. The aim of Step 1.1 is to prove the following lemma

Lemma 5.3.

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \int_{C^R_{\beta,Id}} \sum_{s\in\mathcal{S}_C(c(z))} \frac{1}{E(s)} dz_1 \dots dz_n = \prod_{i=1}^{k} b_i^{n_i-1}(b_i - a_i).$$

For each $i \in [n]$, we define $u_i := z_i - z_{i-1}$. It is not hard to see that each scenario $s = (s_1, \dots, s_n) \in \mathcal{S}_C(c(z))$ is characterized by a unique permutation $\tau \in \Sigma_n$ which specifies the order of coalescence of the successive contiguous blocks in such a way that

$$\frac{1}{E(s)} = \prod_{i=1}^{n} \frac{1}{u_{\tau(1)} + \dots + u_{\tau(n)}}.$$

(see Figure 5 for some examples.) As a consequence, we can index each contiguous scenario by a permutation, and using the change of variables $u_i = z_i - z_{i-1}$, we get

$$\int_{C^R_{\beta,Id}} \sum_{s\in\mathcal{S}_C(c(z))} \frac{1}{E(s)}\, dz_1 \dots dz_n = \int_{U^R} \sum_{\tau\in\Sigma_n} \left(\prod_{i=1}^{n} \frac{du_{\tau(i)}}{u_{\tau(1)} + \dots + u_{\tau(i)}} \right), \tag{I.20}$$

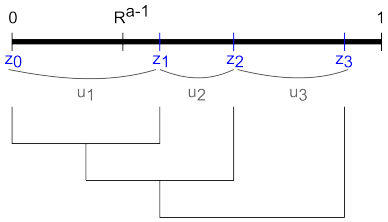
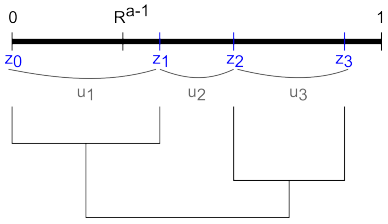
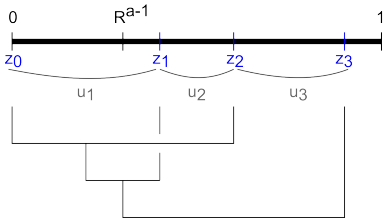
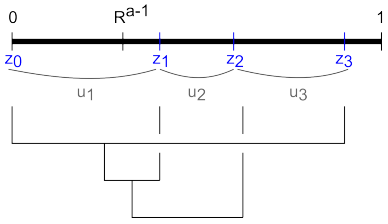
Type	Scenario of coalescence	Energy
$\mathcal{S}_1$	 S_1	$E(S_1) = (z_1 - z_0) \times (z_2 - z_0) \times (z_3 - z_0)$ $= u_1 \times (u_1 + u_2) \times (u_1 + u_2 + u_3)$ $(\tau(1) = 1,\ \tau(2) = 2,\ \tau(3) = 3)$
$\mathcal{S}_1$	 S_2	$E(S_2) = (z_1 - z_0) \times (z_1 - z_0 + z_3 - z_2)$ $\times (z_3 - z_0)$ $= u_1 \times (u_1 + u_3) \times (u_1 + u_3 + u_2)$ $(\tau(1) = 1,\ \tau(2) = 3,\ \tau(3) = 2)$
$\mathcal{S}_2$	 S_3	$E(S_3) = (z_2 - z_0) \times (z_2 - z_0) \times (z_3 - z_0)$ $= (u_1 + u_2) \times (u_1 + u_2)$ $\times (u_1 + u_2 + u_3)$
$\mathcal{S}_2$	 S_4	$E(S_4) = (z_3 - z_0) \times (z_3 - z_0) \times (z_3 - z_0)$ $= (u_1 + u_2 + u_3) \times (u_1 + u_2 + u_3)$ $\times (u_1 + u_2 + u_2)$

Figure I.6 – Some examples of coalescence scenarios and their energy. In these examples, $k = 1$, $b = 1$, $a_1 := a$.

where U^R is defined as follows. First, let us define (see also Figure I.7)

$$\begin{aligned} w_R(1) &:= \max(\beta R^{-1}, R^{a_1-1}) \\ W_R(1) &:= R^{b_1-1} \\ \forall 2 \le i \le k, w_R(i) &:= R^{a_i-1} - R^{b_{i-1}-1} \\ W_R(i) &:= R^{b_i-1} - R^{a_{i-1}-1} \\ \forall 1 \le i \le k, L_R(i) &:= R^{b_i-1} - R^{a_i-1}. \end{aligned}$$

Finally, we set $n_0 := 0$. Under the assumption that R is large enough so that $\forall i \in [k],\ R^{b_i-1} > \beta R^{-1}$

$$\begin{aligned} U^R :=\{\ u_1, \dots, u_n \in \otimes_{i=1}^k \left([w_R(i), W_R(i)] \times [\beta R^{-1}, L_R(i)]^{n_i-1}\right)\ : \\ \forall i \in [k], \sum_{j=n_{i-1}+2}^{n_i} u_j \le L_R(i) \ \text{ and } \sum_{j=1}^{n_1+\dots+n_i} u_j \le R^{b_i-1}\} \end{aligned}$$

In fact, by definition of $C^R_{\beta,Id}$, $\forall j \in [n]$, $\beta R^{-1} \le u_j = z_j - z_{j-1}$. In addition, the $L_R(i)$'s correspond to the lengths of the different intervals $[R^{a_i-1}, R^{b_i-1}]$, so when z_j and z_{j-1} belong to the same interval, $u_j = z_j - z_{j-1} \le L_R(i)$ (See Figure I.7). The $w_R(i)$'s correspond to the distance between two contiguous intervals and the $W_R(i)$'s to the maximal distance between two points of contiguous intervals. So, when z_j and z_{j-1} belong to different intervals then $u_j = z_j - z_{j-1} \in [w_R(i), W_R(i)]$ (See Figure I.7). Finally, the last inequalities come from the fact that for $j \in \{n_{i-1}+1, \dots, n_i\}$, the z_i's belong to the same interval $[a_i, b_i]$, so the sum of their distances cannot exceed the length of the interval. In addition, the distance between z_0 and z_{n_i} cannot exceed R^{b_i-1}.

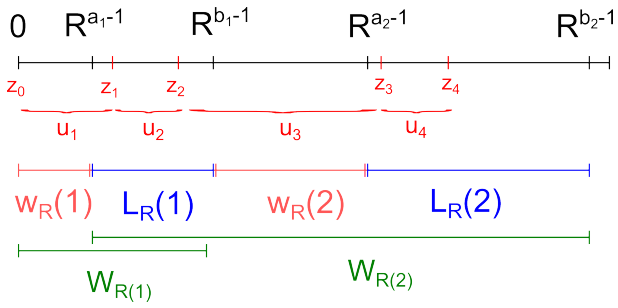


Figure I.7 – The set U^R.

To compute the RHS of (I.20), we start by proving the following Lemma.

Lemma 5.4. *For any $\tau \in \Sigma_n,\ \kappa > 1$, define*

$$\begin{aligned} K_\tau^{R,\kappa} &= \{u_1, \dots, u_n \in U^R, \forall i \in [n],\ u_{\tau(i)} > \kappa \sum_{j=1}^{i-1} u_{\tau(j)}\} \\ \bar{K}_\tau^{R,\kappa} &= U^R \setminus K_\tau^{R,\kappa} \end{aligned}$$

We have

$$\text{(i) } \forall \kappa > 1,\ \lim_{R\to\infty} \frac{1}{\log(R)^n} \sum_{\tau\in\Sigma_n} \int_{\bar{K}_\tau^{R,\kappa}} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(1)} + \ldots + u_{\tau(i)}} \right) = 0$$

$$\text{(ii) } \lim_{\kappa\to\infty} \lim_{R\to\infty} \frac{1}{\log(R)^n} \sum_{\tau\in\Sigma_n} \int_{K_\tau^{R,\kappa}} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(1)} + \ldots + u_{\tau(i)}} \right) = \prod_{i=1}^k b_i^{n_i - 1}(b_i - a_i).$$

Remark 5.5. *The proof of Lemma 5.4 is rather cumbersome, but the idea behind the proof is simple. In a nutshell, the idea is that, depending on the positions of the loci (the z_i's), one scenario is much more likely than the others. More precisely, for any configuration $z \in C^R_{\beta,Id}$, there exists a scenario $S_{min} \in \mathcal{S}_C(c(z))$ associated to permutation $\tau_{min} \in \Sigma_n$ such that*

$$u_{\tau_{min}(1)} \le u_{\tau_{min}(2)} \le \ldots \le u_{\tau_{min}(n)}.$$

By coalescing the u_i's in the increasing order, the successive cover lengths are minimised.

Proof. We start by proving (i). We fix $\tau \in \Sigma_n,\ \kappa > 1$. We make the following change of variables. Let us define Ψ^τ such that for $1 \le i \le n$, $(\Psi^\tau(u_1, \dots, u_n))_{\tau(i)} = u_{\tau(1)} + \ldots + u_{\tau(i)}$. We have

$$\int_{U^R} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(1)} + \ldots + u_{\tau(i)}} \right) = \int_{v\in\Psi^\tau(U^R)} \frac{dv_1 \ldots dv_n}{v_1 \ldots v_n}. \tag{I.21}$$

In particular, as

$$\forall i \in [n],\quad u_{\tau(i)} \le \kappa \sum_{j=1}^{i-1} u_{\tau(j)} \ \Leftrightarrow\ \forall i \in [n],\quad v_{\tau(i)} \le (1+\kappa) v_{\tau(i-1)},$$

we have

$$\int_{\bar{K}_\tau^{R,\kappa}} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(1)} + \ldots + u_{\tau(i)}} \right) = \int_{V_\tau^{R,\kappa}} \frac{dv_1 \ldots dv_n}{v_1 \ldots v_n}$$

where

$$V_\tau^{R,\kappa} = \{v \in \Psi^\tau(U^R), \exists i \in [n],\ v_{\tau(i)} \le (1+\kappa) v_{\tau(i-1)}\}.$$

For every $i \in [n]$, we define

$$V_\tau^{R,\kappa}(i) \ = \ \{v \in \Psi^\tau(U^R),\ v_{\tau(i)} \le (1+\kappa) v_{\tau(i-1)}\}.$$

We have

$$V_\tau^{R,\kappa} = \bigcup_{i=1}^{n} V_\tau^{R,\kappa}(i).$$

We fix $\tau \in \Sigma_n$ and $i \in [n]$. The $v_{\tau(j)}$'s are the successive cover lengths at each step of the scenario S associated to τ, so it can readily be seen that

$$\forall j \in [n], v_j \in [R^{-1}, 1] \text{ and } v_{\tau(1)} < \ldots < v_{\tau(n)},$$

which implies that

$$V_\tau^{R,\kappa}(i) \subset \{v \in [R^{-1}, 1]^n,\ v_{\tau(i-1)} < v_{\tau(i)} < (1+\kappa) v_{\tau(i-1)}\},$$

so

$$\begin{aligned}
\int_{V_\tau^{R,\kappa}(i)} \frac{dv_{\tau(1)} \ldots dv_{\tau(n)}}{v_{\tau(1)} \ldots v_{\tau(n)}} &\leq \left(\int_{R^{-1}}^{1} \frac{dv}{v}\right)^{n-2} \int_{R^{-1}}^{1} \frac{dv_{\tau(i-1)}}{v_{\tau(i-1)}} \int_{v_{\tau(i-1)}}^{(1+\kappa)v_{\tau(i-1)}} \frac{dv_{\tau(i)}}{v_{\tau(i)}} \\
&= \log(R)^{n-2} \int_{R^{-1}}^{1} \frac{dv_{\tau(i-1)}}{v_{\tau(i-1)}} \log(1+\kappa) \\
&= \log(R)^{n-1} \log(1+\kappa),
\end{aligned}$$

which completes the proof of (i).

We now turn to the proof of (ii). We decompose the proof into four steps.

Step a. Define

$$X^R := \otimes_{i=1}^{k} \left([w_R(i), W_R(i)] \times [\beta R^{-1}, L_R(i)]^{n_i - 1}\right).$$

Then

$$\begin{aligned}
\frac{1}{\log(R)^n} \int_{X^R} \left(\prod_{i=1}^{n} \frac{du_i}{u_i}\right) &= \frac{1}{\log(R)^n} \prod_{i=1}^{k} \int_{w_R(i)}^{W_R(i)} \frac{du}{u} \left(\int_{\beta R^{-1}}^{L_R(i)} \frac{du}{u}\right)^{n_i - 1} \\
&= \frac{1}{\log(R)^n} \prod_{i=1}^{k} \log\left(\frac{W_R(i)}{w_R(i)}\right) \log\left(\frac{L_R(i)}{\beta R^{-1}}\right)^{n_i - 1} \\
&\underset{R \to \infty}{\longrightarrow} \prod_{i=1}^{k} (b_i - a_i) b_i^{n_i - 1}. \qquad \text{(I.22)}
\end{aligned}$$

Step b. Next, for every $\kappa > 1$ and for every $\tau \in \Sigma_n$, we define

$$\begin{aligned}
X_\tau^R &:= \{\ u_1, \ldots, u_n \in X^R\ :\ u_{\tau(1)} \leq \ldots \leq u_{\tau(n)}\}, \\
A_\tau^\kappa &:= \{\ u_1, \ldots, u_n, \forall i \in [n],\ u_{\tau(i)} > \kappa \sum_{j=1}^{i-1} u_{\tau(j)}\}.
\end{aligned}$$

$$X_\tau^{R,\kappa} := X^R \cap A_\tau^\kappa.$$

By reasoning along the same lines as in the proof of (i), one can show that

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \left| \int_{X_\tau^R} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right) - \int_{X_\tau^{R,\kappa}} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right) \right| = 0.$$

From Step 1, we get that for every $\kappa > 1$

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \sum_{\tau\in\Sigma_n} \int_{X_\tau^{R,\kappa}} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right) = \prod_{i=1}^k (b_i - a_i) b_i^{n_i - 1} \tag{I.23}$$

Step c. We aim of this step is to prove that for every $\tau \in \Sigma_n$,

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \int_{K_\tau^{R,\kappa}} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right) = \lim_{R\to\infty} \frac{1}{\log(R)^n} \int_{X_\tau^{R,\kappa}} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right) \tag{I.24}$$

From the definition of $K_\tau^{R,\kappa}$, we have

$$K_\tau^{R,\kappa} = X_\tau^{R,\kappa} \cap (K_1^R \cap K_2^R)$$

where

$$K_1^R := \{\ u_1, \dots, u_n, \quad \forall i \in [k], \sum_{j=n_{i-1}+2}^{n_i} u_j \le L_R(i)\ \},$$
$$K_2^R := \{\ u_1, \dots, u_n, \ \ \forall i \in [k] \sum_{j=1}^{n_1+\dots+n_i} u_j \le R^{b_i-1}\ \}.$$

If

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \left| \int_{X_\tau^{R,\kappa}\cap K_1^R} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right) - \int_{X_\tau^{R,\kappa}} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right) \right| = 0 \tag{I.25}$$

and

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \left| \int_{X_\tau^{R,\kappa}\cap K_2^R} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right) - \int_{X_\tau^{R,\kappa}} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right) \right| = 0, \tag{I.26}$$

then

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \left| \int_{K_\tau^{R,\kappa}} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right) - \int_{X_\tau^{R,\kappa}} \left(\prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right) \right| = 0. \tag{I.27}$$

We will only prove (I.25), as (I.26) can be proved along the same lines. To do so, we

define

$$Y_\tau^{R,\kappa} := \{u_1,\dots,u_n \in \otimes_{i=1}^k \left([w_R(i), W_R(i)] \times \left[\beta R^{-1}, \frac{L_R(i)}{1+\frac{1}{\kappa}}\right]^{n_i-1}\right), \\ u_{\tau(1)} \le \dots \le u_{\tau(n)}\}.$$

so that $Y_\tau^{R,\kappa} \subset X_\tau^R$. By similar computations as those used in the proof of (i), it can be shown that

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \int_{X_\tau^R \setminus Y_\tau^{R,\kappa}} \prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} = 0,$$

so

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \left| \int_{X_\tau^{R,\kappa}} \prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} - \int_{Y_\tau^{R,\kappa}\cap A_\tau^\kappa} \prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right| = 0, \tag{I.28}$$

and

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \left| \int_{X_\tau^{R,\kappa}\cap K_1^R} \prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} - \int_{Y_\tau^{R,\kappa}\cap A_\tau^\kappa \cap K_1^R} \prod_{i=1}^n \frac{du_{\tau(i)}}{u_{\tau(i)}} \right| = 0 \tag{I.29}$$

Let us show that

$$Y_\tau^{R,\kappa} \cap A_\tau^\kappa \cap K_1^R = Y_\tau^{R,\kappa} \cap A_\tau^\kappa,$$

i.e. that

$$\forall (u_1,\dots,u_n) \in Y_\tau^{R,\kappa} \cap A_\tau^\kappa, \quad \forall i \in [k], \sum_{j=n_{i-1}+2}^{n_i} u_j \le L_R(i).$$

We fix $i \in [k]$ and we define

$$m_i := j \in \{n_{i-1}+2,\dots,n_i\},\ \tau^{-1}(j) = \max\{\tau^{-1}(n_{i-1}+2),\dots,\tau^{-1}(n_i)\}$$

As

$$u_{\tau(m_i)} > \kappa \sum_{j=1}^{m_i-1} u_{\tau(j)},$$

then

$$\sum_{j=n_{i-1}+2}^{n_i} u_j \le \sum_{j=1}^{m_i} u_{\tau(j)} = \left(1+\frac{1}{\kappa}\right) u_{\tau(m_i)} \le \left(1+\frac{1}{\kappa}\right) \frac{L_R(i)}{1+\frac{1}{\kappa}} = L_R(i).$$

Since $Y_\tau^{R,\kappa} \cap A_\tau^\kappa \cap K_1^R = Y_\tau^{R,\kappa} \cap A_\tau^\kappa$, combining (I.28) and (I.29), (I.25) is proved. Equation (I.26) can be proved along the same lines, so (I.27) is verified.

Step d. Finally, for any $\tau \in \Sigma_n$, for any $u_1,\dots,u_n \in K_\tau^{R,\kappa}$, we have

$$\forall i \in [n],\ u_{\tau(i)} \ \le\ u_{\tau(1)} + \dots + u_{\tau(i-1)} + \ u_{\tau(i)} \ \le\ \left(1+\frac{1}{\kappa}\right)\ u_{\tau(i)},$$

which implies that

$$\sum_{\tau\in\Sigma_n}\int_{K_\tau^{R,\kappa}}\left(\prod_{i=1}^n\frac{du_{\tau(i)}}{u_{\tau(i)}}\right)\le\sum_{\tau\in\Sigma_n}\int_{K_\tau^{R,\kappa}}\left(\prod_{i=1}^n\frac{du_{\tau(i)}}{u_{\tau(1)}+\ldots+u_{\tau(i)}}\right)$$

and

$$\sum_{\tau\in\Sigma_n}\int_{K_\tau^{R,\kappa}}\left(\prod_{i=1}^n\frac{du_{\tau(i)}}{u_{\tau(1)}+\ldots+u_{\tau(i)}}\right)\le\frac{1}{(1+\frac{1}{\kappa})^n}\sum_{\tau\in\Sigma_n}\int_{K_\tau^{R,\kappa}}\left(\prod_{i=1}^n\frac{du_{\tau(i)}}{u_{\tau(i)}}\right)$$

This, combined with Step b (see (I.23)) and Step c (see (I.24))

$$\begin{aligned}\prod_{i=1}^k(b_i-a_i)b_i^{n_i-1}&\le\lim_{R\to\infty}\frac{1}{\log(R)^n}\sum_{\tau\in\Sigma_n}\int_{K_\tau^{R,\kappa}}\left(\prod_{i=1}^n\frac{du_{\tau(i)}}{u_{\tau(1)}+\ldots+u_{\tau(i)}}\right)\\&\le\frac{\prod_{i=1}^k(b_i-a_i)b_i^{n_i-1}}{(1+\frac{1}{\kappa})^n},\end{aligned}$$

and the conclusion follows by taking $\kappa\to\infty$. □

Proof of Lemma 5.3. From (I.20), we have

$$\begin{aligned}&\lim_{R\to\infty}\frac{1}{\log(R)^n}\int_{C^R_{\beta,Id}}\sum_{s\in\mathcal{S}_C(c(z))}\frac{1}{E(s)}\,dz_1\ldots dz_n\\&=\lim_{R\to\infty}\frac{1}{\log(R)^n}\sum_{\tau\in\Sigma_k}\int_{U^R}\prod_{i=1}^n\frac{du_{\tau(i)}}{u_{\tau(1)}+\ldots+u_{\tau(i)}}\\&=\lim_{R\to\infty}\frac{1}{\log(R)^n}\sum_{\tau\in\Sigma_k}\left(\int_{K_\tau^{R,\kappa}}\prod_{i=1}^n\frac{du_{\tau(i)}}{u_{\tau(1)}+\ldots+u_{\tau(i)}}+\int_{\bar K_\tau^{R,\kappa}}\prod_{i=1}^n\frac{du_{\tau(i)}}{u_{\tau(1)}+\ldots+u_{\tau(i)}}\right).\end{aligned}$$

Using Lemma 5.4 and taking $\kappa\to\infty$ in the RHS, we have

$$\lim_{R\to\infty}\frac{1}{\log(R)^n}\int_{C^R_{\beta,Id}}\sum_{s\in\mathcal{S}_C(c(z))}\frac{1}{E(s)}\,dz_1\ldots dz_n\;=\;\prod_{i=1}^k b_i^{n_i-1}(b_i-a_i).\tag{I.30}$$

□

Step 1.2. The aim of this step is to compute the second term in the RHS of (I.19).

Lemma 5.6.

$$\lim_{R\to\infty}\frac{1}{\log(R)^n}\int_{C^R_{\beta,Id}}\sum_{S\in\bar{\mathcal{S}}_C(c(z))}\frac{1}{E(S)}dz_1\ldots dz_n\;=\;0.$$

Proof. We fix $z=z_0,z_1,\ldots,z_n\in C^R_{\beta,Id}$. We start by considering scenarios where blocks only coalesce with neighbouring blocks, except for one step. In other words, we start by

considering scenarios in $\bar{\mathcal{S}}'_C(c(z))$, the set of scenarios that contain one single coalescence event between two non-neighbouring blocks. For example, S_3 in Figure 5 is in $\bar{\mathcal{S}}'_C(c(z))$. We consider $S' = (s'_0, s'_1, \dots, s'_n) \in \bar{\mathcal{S}}'_C(c(z))$, a scenario of coalescence in which step $j > 1$, is the only coalescence event between two non neighbouring blocks. The idea is to compare S' with a scenario $S \in \mathcal{S}_C(c(z))$ and use this scenario to show that

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \int_{C^R_{\beta,Id}} \frac{1}{E(S')} dz_1 \dots dz_n = 0. \tag{I.31}$$

As already argued each scenario in $\mathcal{S}_C(c(z))$ is associated to a permutation τ such that, at each step i, the cover length increases by $u_{\tau(i)}$. The scenario S and the corresponding permutation τ are constructed as follows (and we let the reader refer to Figure I.8 for an example, where subfigures $(i), \dots, (iv)$ correspond to each step in the following construction).

(i) For $0 \le i < j$, we set $s_i = s'_i$. Before step j, there are only coalescence events between neighbouring blocks in S'. $\tau(1), \dots, \tau(j-1)$ are constructed in such a way that

$$\forall 1 \le i < j, \quad C(s_i) = C(s'_i) = \sum_{k=1}^{i} u_{\tau(k)}.$$

(ii) At step j, in scenario S' there is a coalescence event between two non neighbouring blocks, which means that there exists $i_1 < i_2 < \dots < i_\ell$ such that $C(s'_j) = C(s'_{j-1}) + u_{i_1} + \dots + u_{i_\ell}$. s_j is the partition of order j such that $C(s_j) = C(s_{j-1}) + u_{i_1}$. We set $\tau(j) := i_1$. We have

$$C(s_j) = \sum_{k=1}^{i} u_{\tau(k)} \le C(s'_j).$$

(iii) For $j < i \le j + \ell - 1$, $\tau(i) := i_{i-j}$, i.e., we add successively $u_{i_2}, \dots, u_{i_\ell}$. We have

$$\forall j < i \le j + \ell - 1, \ C(s_i) \le C(s'_i).$$

(iv) For $j + \ell \le i \le j$, the s_i's are constructed as follows. Let $u_{r_1}, \dots, u_{r_p}$ be the u_i's that have not been added yet to the cover length of S' (i.e. the u_i's that are not in $\{u_{\tau(1)}, \dots, u_{\tau(j+\ell)}\}$), indexed in such a way that in S', u_{r_1} coalesces before u_{r_2} etc …. Then we set $u_{\tau(j+\ell+i)} = u_{r_i}$. In other words, the u_{r_i}'s are added to the cover length in S in the same order as they are added in S' (see Figure I.8).

With this construction, we have

$$\frac{1}{E(S)} = \prod_{i=1}^{n} \frac{1}{u_{\tau(1)} + \dots + u_{\tau(i)}}$$

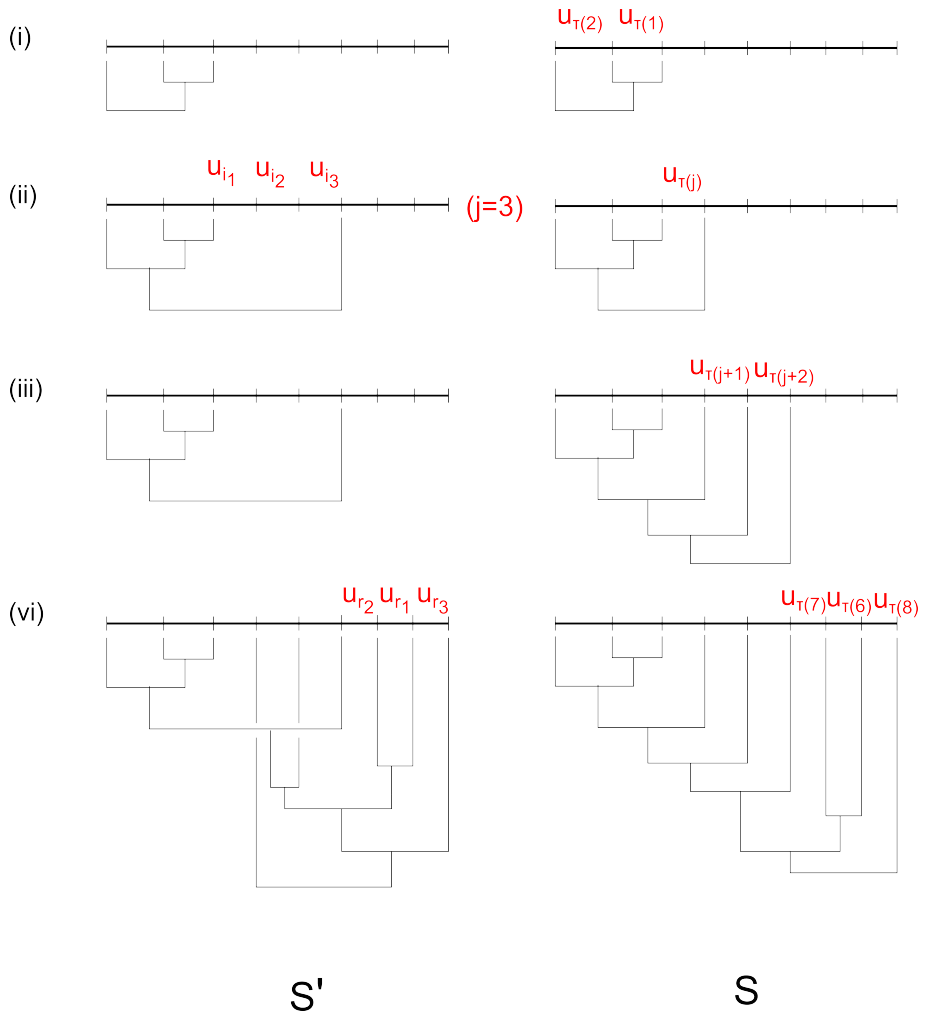


Figure I.8 – Example of the construction of a scenario $S \in \mathcal{S}_C(c(z))$ from a scenario $S' \in \bar{\mathcal{S}}'_C(c(z))$. The left-hand side corresponds to S' and the right-hand side to S. Steps (i) . . . (iv) correspond to the steps in construction of S from S'.

$$=\frac{1}{v_{\tau(1)}\dots v_{\tau(n)}}.$$

where for $i \in \{1,\dots,n\}$, $v_{\tau(i)} := \Psi^\tau(U^R)_{\tau(i)} = u_{\tau(1)} + \ldots + u_{\tau(i)}$. By construction, we have

$$\frac{1}{E(S')} \leq \left(\prod_{i=1}^{j-1}\frac{1}{C(s_i)}\right)\ \frac{1}{C(s'_j)}\ \left(\prod_{i=j+1}^{n}\frac{1}{C(s_i)}\right).$$

Using the fact that

$$\begin{aligned} C(s'_j) &= u_{\tau(1)} + \ldots + u_{\tau(j-1)} + u_{i_1} + \ldots + u_{i_\ell} \\ &= v_{\tau(j+\ell)}, \end{aligned}$$

we have

$$\begin{aligned} \frac{1}{E(S')} \leq &= \left(\prod_{i=1}^{j-1}\frac{1}{v_{\tau(i)}}\right)\frac{1}{v_{\tau(j+\ell)}}\left(\prod_{i=j+1}^{n}\frac{1}{v_{\tau(i)}}\right) \\ &\leq \left(\prod_{i=1}^{j-1}\frac{1}{v_{\tau(i)}}\right)\frac{1}{v_{\tau(j+1)}}\left(\prod_{i=j+1}^{n}\frac{1}{v_{\tau(i)}}\right) \end{aligned}$$

where the last inequality comes from the fact that $v_{\tau(j+1)} < \ldots < v_{\tau(j)+l}$. Using the same change of variables as in (I.21), we have

$$\int_{C^R_{\beta,Id}} \frac{1}{E(S')}dz_1\dots dz_n \quad = \quad \int_{\Psi^\tau(U^R)} \frac{dv_{\tau(1)}\dots dv_{\tau(n)}}{v_{\tau(1)}\dots v_{\tau(j-1)}v_{\tau(j+1)}v_{\tau(j+1)}\dots v_{\tau(n)}}$$

And, from the definition of Ψ^τ, it can easily be seen that for any $\tau \in \Sigma_n$

$$\Psi^\tau(U^R) \subset \Psi' = \{x_1,\dots,x_n \in [R^{-1},1]^n, x_1 \leq \ldots \leq x_n\}$$

so

$$\begin{aligned} \int_{C^R_{\beta,Id}} \frac{dz_1\dots dz_n}{E(S')} &\leq \int_{\Psi'} \frac{dx_1\dots dx_n}{x_1\dots x_{j-1}x_{j+1}x_{j+1}\dots x_n} \\ &= \int_{R^{-1}}^1 \frac{dx_n}{x_n}\dots\int_{R^{-1}}^{x_{j+2}}\frac{dx_{j+1}}{x_{j+1}^2}\int_{R^{-1}}^{x_{j+1}}dx_j\int_{R^{-1}}^{x_j}\frac{dx_{j-1}}{x_{j-1}}\dots\int_{R^{-1}}^{x_2}\frac{dx_1}{x_1} \\ &= \int_{R^{-1}}^1 \frac{dx_n}{x_n}\dots\int_{R^{-1}}^{x_{j+2}}\frac{dx_{j+1}}{x_{j+1}^2}\int_{R^{-1}}^{x_{j+1}}\frac{\log(Rx_j)^{j-1}}{(j-1)!}dx_j \\ &\leq \frac{\log(R)^{j-1}}{(j-1)!}\int_{R^{-1}}^1\frac{dx_n}{x_n}\dots\int_{R^{-1}}^{x_{j+2}}\frac{x_{j+1}-1/R}{x_{j+1}^2}\,dx_{j+1} \end{aligned}$$

$$\leq \log(R)^{j-1} \int_{R^{-1}}^{1} \frac{dx_n}{x_n} \dots \int_{R^{-1}}^{x_{j+2}} \frac{dx_{j+1}}{x_{j+1}} \;=\; \log(R)^{n-1},$$

which completes the proof of (I.31).

To complete the proof of Lemma 5.6, we are going show that, for every scenario $S^2 \in \bar{\mathcal{S}}_C(c(z))$ with more than one step of coalescence between non-contiguous scenarios, there exist a scenario $S^3 \in \bar{\mathcal{S}}'_C(c(z))$ such that $E(S^3) \leq E(S^2)$. We fix $S^2 = (s_0^2, s_1^2, \dots, s_n^2) \in \bar{\mathcal{S}}_C(c(z))$, and the idea is to construct $S^3 = (s_0^3, s_1^3, \dots, s_n^3) \in \bar{\mathcal{S}}'_C(c(z))$ along the same lines as in Step 1. Let j_1 the first step of coalescence between non contiguous blocks in S^3 and j_2 the second one.

— For $0 \leq i < j_2$, $s_i^3 := s_i^2$. In words, we copy all the steps, including j_1, the first step of coalescence between non neighbouring blocks.

— Steps $(s_{j_2}^3, \dots, s_n^3)$ are obtained from $(s_{j_2}^2, \dots, s_n^2)$ in the same way as S' was obtained from S in Step 1.

With this construction, $S^3 \in \bar{\mathcal{S}}'_C(c(z))$ (there is only one step of coalescence between non neighbouring blocks, which is j_1) and we have $\forall i \in [n],\ C(s_i^2) \leq C(s_i^3)$, so

$$\int_{C^R_{\beta,Id}} \frac{1}{E(S^2)} dz_1 \dots dz_n \leq \int_{C^R_{\beta,Id}} \frac{1}{E(S^3)} dz_1 \dots dz_n$$

As $S^3 \in \bar{\mathcal{S}}'_C(c(z))$, combining the previous equation with (I.31), for every scenario $S \in \bar{\mathcal{S}}_C(c(z))$,

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \int_{C^R_{\beta,Id}} \frac{1}{E(S)} dz_1 \dots dz_n = 0,$$

which completes the proof Lemma 5.6. □

Proof of Proposition 5.2. This is a direct consequence of Lemmas 5.3 and 5.6 and (I.19). □

Step 2. Define

$$\begin{aligned} D^R_\beta \;&:=\; \{z_1, \dots, z_n \in [R^{a_1}, R^{b_1}]^{n_1} \times \dots \times [R^{a_k}, R^{b_k}]^{n_k},\ \text{s.t. if } z_0 := 0 \\ &\qquad \forall i \neq j \in \{0, \dots, n\},\ |z_i - z_j| \geq \beta\} \\ I^R_\beta \;&:=\; \frac{1}{\log(R)^n} \int_{D^R_\beta} \mu^{z,1}(c(z)) dz_1 \dots dz_n, \end{aligned}$$

Using scaling (see Proposition 3.4) and a change of variables, we have:

$$I^R_\beta \;=\; \frac{R^n}{\log(R)^n} \int_{C^R_\beta} \mu^{z,R}(c(z)) dz_1 \dots dz_n. \tag{I.32}$$

Recall that $c(z)$ is a partition of order n, so from Theorem 1.3, we have:

$$\left| I_\beta^R - \frac{1}{\log(R)^n} \int_{C_\beta^R} F(c(z))\ dz_1 \dots dz_n \right| \ \leq\ \frac{f^n(\beta)}{\log(R)^n} \int_{C_\beta^R} F(c(z))\ dz_1 \dots dz_n$$

and $f^n(\beta) \underset{\beta\to\infty}{\longrightarrow} 0$. By taking successive limits, first $R \to \infty$ and then $\beta \to \infty$, using Proposition 5.2

$$\lim_{\beta\to\infty} \lim_{R\to\infty} I_\beta^R \ =\ \prod_{i=1}^k n_i!\ b_i^{n_i-1}(b_i - a_i). \tag{I.33}$$

Step 3. The aim of this step is to show that we can now approximate $\mathbb{E}\left(\prod_{i=1}^k \vartheta^R([a_i,b_i])^{n_i}\right)$ by I_β^R. In fact, I_β^R can be obtained from $\mathbb{E}\left(\prod_{i=1}^k \vartheta^R([a_i,b_i])^{n_i}\right)$ by removing a small fraction of the integration domain (see (I.16)). More precisely we will show that

Lemma 5.7.

$$\forall \beta \geq 1,\ \lim_{R\to\infty} \left(\mathbb{E}\left(\prod_{i=1}^k \vartheta^R([a_i,b_i])^{n_i} \right) - I_\beta^R \right) = 0.$$

Proof. We fix $k \in \mathbb{N},\ n_1, \dots, n_k \in \mathbb{N},\ n = n_1 + \ldots + n_k,\ a_1, \dots a_k \in [0,1],\ b_1, \dots, b_k \in [0,1],\ a_1 < b_1 < a_2 < b_2 \ldots < a_k < b_k,\ \beta \geq 1.$

Let us define

$$\dot{\Delta}_\beta^R := \{z_1, \dots, z_n \in [R^{a_1}, R^{b_1}]^{n_1} \times \ldots \times [R^{a_k}, R^{b_k}]^{n_k},\ \text{ such that if } z_0 := 0,\\ \exists\ i,j \in \{0, \dots, n\},\ |z_i - z_j| < \beta\}.$$

We have

$$\mathbb{E}\left(\prod_{i=1}^k \vartheta^R([a_i,b_i])^{n_i} \right) - I_\beta^R = \ \frac{1}{\log(R)^n} \int_{\dot{\Delta}_\beta^R} \mu^{z,1}(c(z))\ dz_1 \dots dz_n.$$

Lemma 5.7 can be reformulated as follows

$$\lim_{R\to\infty}\ \frac{1}{\log(R)^n} \int_{\dot{\Delta}_\beta^R} \mu^{z,1}(c(z))\ dz_1 \dots dz_n\ =\ 0.$$

By symmetry, proving this result reduces to proving that

$$\lim_{R\to\infty}\ \frac{1}{\log(R)^n} \int_{\Delta_\beta^R} \mu^{z,1}(c(z))\ dz_1 \dots dz_n\ =\ 0,$$

where

$$\Delta_\beta^R := \dot{\Delta}_\beta^R \bigcap \{z_1, \dots, z_n,\ z_0 := 0 < z_1 < \ldots < z_n\}.$$

Let $\mathbb{S}$ be the set of all subsets of $[n]$ containing at most $n-1$ elements. For $S \in \mathbb{S}$, define

$$\Delta^R_{\beta,S} = \Delta^R_\beta \bigcap \{z_1, \dots, z_n : \ \forall i \in S, \ |z_i - z_{i-1}| \geq \beta, \ \forall i \notin S, \ |z_i - z_{i-1}| < \beta\}.$$

in such a way that

$$\Delta^R_\beta = \bigcup_{S \in \mathbb{S}} \Delta^R_{\beta,S}.$$

It follows that

$$\begin{aligned} \int_{\Delta^R_\beta} \mu^{z,1}(c(z)) \ dz_1 \dots dz_n \ &= \ \sum_{S \in \mathbb{S}} \int_{\Delta^R_{\beta,S}} \mu^{z,1}(c(z)) \ dz_1 \dots dz_n \\ &\leq \ \sum_{S \in \mathbb{S}} \int_{\Delta^R_{\beta,S}} \mu^{z,1}(\pi_S) \ dz_1 \dots dz_n \end{aligned}$$

where $\forall S \in \mathbb{S}$, $\pi_S = \{\pi \in \mathcal{P}_z, \ \forall i,j \in S, \ \ z_i \sim_\pi z_j\}$ (and where the inequality follows from the fact that $c(z) \in \pi_S$). We define $z^S := \{z_i, \ i \in S\}$. Proposition 2.2 gives:

$$\mu^{z,1}(\pi_S) = \mu^{z^S,1}(c(z^S)).$$

Define

$$\bar{\Delta}^R_{\beta,S} := \{(z_i)_{i \in S} : \ \exists \ (z_j)_{j \in \{1,\dots,n\} \setminus S}, \ \ (z_1, \dots, z_n) \in \Delta^R_{\beta,S}\}.$$

We let the reader convince herself that, for any $S \in \mathbb{S}$ there exists $m_1, \dots, m_k \in \mathbb{N}$, $m_1 + \dots + m_k = |S|$, such that $\bar{\Delta}^R_{\beta,S}$ can be rewritten as

$$\begin{aligned} \bar{\Delta}^R_{\beta,S} \ = \{&\bar{z}_1, \dots \bar{z}_{|S|} \in [R^{a_1-1}, R^{b_1-1}]^{m_1} \times \dots \times [R^{a_k-1}, R^{b_k-1}]^{m_k} \ : \\ &\bar{z}_0 := 0 \leq \bar{z}_1 \leq \dots \leq \bar{z}_{|S|} \text{ and } \forall \ i,j \in S, \ i \neq j, \ |\bar{z}_i - \bar{z}_j| > \beta\}. \end{aligned}$$

This allows us to rewrite the previous inequality as

$$\begin{aligned} \int_{\Delta^R_\beta} \mu^{z,1}(c(z)) \ dz_1 \dots dz_n \ &\leq \ \sum_{S \in \mathbb{S}} \int_{\Delta^R_{\beta,S}} \mu^{z^S,1}(c(z^S)) \underbrace{dz_1 \dots dz_i \dots}_{i \in S} \ \underbrace{\dots dz_j \dots}_{j \notin S} \\ &= \ \sum_{S \in \mathbb{S}} \int_{\bar{\Delta}^R_{\beta,S}} \mu^{z^S,1}(c(\bar{z})) d\bar{z}_1 \dots d\bar{z}_{|S|} \left(\prod_{j \notin S} \int_{z_{j-1}}^{z_{j-1}+\beta} dz_j \right) \\ &= \ \beta^{n-|S|} \ \sum_{S \in \mathbb{S}} \int_{\bar{\Delta}^R_{\beta,S}} \mu^{z^S,1}(c(\bar{z})) d\bar{z}_1 \dots d\bar{z}_{|S|} \end{aligned} \tag{I.34}$$

where $\bar{z} = (\bar{z}_0, \dots, \bar{z}_{|S|})$ and $\bar{z}_0 = 0$. From (I.33), we have

$$\lim_{R\to\infty} \frac{1}{\log(R)^{|S|}} \int_{\bar{\Delta}^R_{\beta,S}} \mu^{z^S,1}(c(\bar{z}))d\bar{z}_1 \dots d\bar{z}_{|S|} \;=\; \prod_{i\in[k],\ m_i\neq 0} b_i^{m_i-1}(b_i - a_i).$$

As $|S| < n$,

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \sum_{S\in\mathbb{S}} \int_{\bar{\Delta}^R_{\beta,S}} \mu^{z^S,1}(c(\bar{z}))d\bar{z}_1 \dots d\bar{z}_{|S|} \;=\; 0,$$

which combined with (I.34) concludes the proof of the Lemma 5.7. □

Step 4. Conclusion. Combining (I.33) and with Lemma 5.7 (Step 3), we have proved that for every k-tuple of disjoint intervals $\{[a_i, b_i]\}_{i=1}^k$ and any k-tuple of integers $\{n_i\}_{i=1}^k$

$$\lim_{R\to\infty} \mathbb{E}\left(\prod_{i=1}^k \vartheta^R([a_i, b_i])^{n_i}\right) \;=\; \prod_{i=1}^k n_i!\; b_i^{n_i-1}(b_i - a_i).$$

So, using Lemma 5.1, ϑ^R converges to ϑ^∞ in distribution in the weak topology. In particular, we have

$$\mathcal{L}_R(0) = \vartheta^R[0,1] + \frac{1}{\log(R)} \int_{[0,1]} \mathbb{1}_{\{x\sim_\pi 0\}} dx$$

and

$$\frac{1}{\log(R)} \int_{[0,1]} \mathbb{1}_{\{x\sim_\pi 0\}} dx \;\leq\; \frac{1}{\log(R)} \underset{R\to\infty}{\longrightarrow} 0$$

so, using equation (I.14), we have

$$\forall n \in \mathbb{N},\ \lim_{R\to\infty} \mathbb{E}(\mathcal{L}_R(0)^n) = \lim_{R\to\infty} \mathbb{E}(\vartheta^R[0,1]^n) = n!$$

which are the moments of the exponential distribution of parameter 1. As in the proof of Lemma 5.1, using Carleman's condition (for $k = 1$), this implies that $\mathcal{L}_R(0)$ converges in distribution to an exponential distribution of parameter 1.

Chapter II

Deriving the expected number of detected haplotype junctions in hybrid populations

In this chapter, I present some work in progress, in collaboration with Thijs Janzen, from Carl von Ossietzky University (Oldenburg, Germany). It aims at using the Ancestral Recombination Graph to study hybridization.

The idea is to consider a hybrid population that emerged from a single hybridization event between two ancestral populations P and Q. We assume that, at time 0, the proportion of individuals from population P is p and the proportion of individuals from population Q is $q = 1 - p$. The evolution of the hybrid population can be modelled using a Wright-Fisher model with recombination, which is an analogous in discrete time to the Moran model with recombination used in Chapter I. Instead of assuming that, at time 0 each individual has her chromosome painted in a distinct color, we assume that there are only two colors, corresponding to each of the ancestral populations. By the blending effect of recombination, the chromosomes of the hybrids are mosaics of these two colors. For instance, in Figure II.1, the blue color indicates that the individual carrying the displayed chromosome inherited this portion of the chromosome from the blue ancestral subpopulation.

Inferring which fragments of the chromosome have been inherited from each ancestral subpopulation from genomic data is not straightforward. As the two ancestral populations are supposed to be closely related species, their genomes are quite similar. Therefore, one needs to use molecular markers: a marker is a locus that segregates between the two ancestral populations, i.e. such that each of the ancestral populations carries a different allele at this locus. If two consecutive markers carry different allelles (i.e. each one carries the allele associated to one of the subpopulations) we say that we observe a "junction".

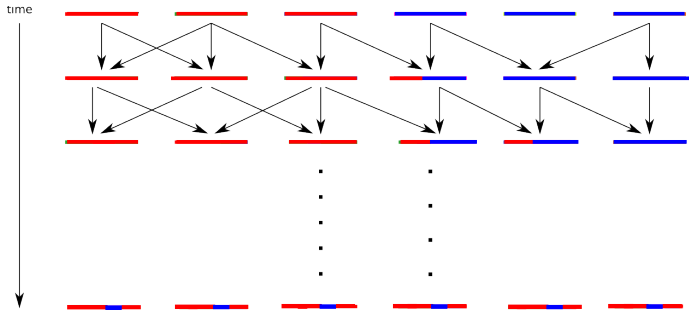


Figure II.1 – The Wright-Fisher model with recombination with two ancestral populations.

Note that the observed number of junctions between markers will depend on the density and the positions of the markers (see Figure II.2).

In [JNT18], Thijs Janzen and his collaborators derived a formula for the expected number of observed junctions using equidistant markers. However, in real data the positions of the markers are not regularly spaced, and their approach did not allow them to take this into account. Using the partitioning-process, Thijs Janzen and I extended this formula to the case of markers that are randomly distributed across the genome. In fact, if we have n markers, whose positions are given by $z = (z_1, \ldots, z_n)$, and we know that the recombination rate is R, we can follow their ancestry backwards in time, using the Ancestral Recombination Graph. If hybridization occured t generations ago, the probability of observing a junction between z_i and z_{i+1} corresponds to the probability that these two loci were carried by different lineages at time t and the two lineages correspond to individuals from different ancestral populations (which happens with probability $2pq$). We derived a formula for the expected number of junctions as a function of time and compared our results to simulations and to real data.

Figure II.2 – The number of observed junctions depends on the density of markers and their positions. Arrows represent markers, and their color indicates to which ancestral subpopulation they are associated. For the same realization of the Wright-Fisher model with recombination, the number of observed junctions is different. In the left-hand side, with 5 markers we observe 4 junctions whereas in the right-hand side, with 4 markers we only observe 2 junctions.

1 Introduction

The traditional view where species or lineages accumulate incompatibilities over time and become reproductively isolated from each other has led to insight into the processes generating and maintaining biodiversity [CO04]. This view has proven to be misleading however, and it has become apparent that lineages do not necessarily only branch, but that lineages can also come back together [AAA$^+$13]. In plants, it has been known for quite some time that hybridization between lineages can lead to not only viable offspring, but can also potentially lead to the formation of new lineages, and ultimately, species [Gra81]. It has long been debated whether this process could also happen in animals, but over the past few years numerous examples have appeared, including, but not limited to, butterflies [MSB$^+$06, CDRM15], cichlid fishes [KDS$^+$07, KWG$^+$13], warblers [BBI11], fruit flies [SMSBM05] and sculpins [NFST05].

Understanding the time-line of these hybridization events is paramount in obtaining a full understanding of the process and its impact. Often, hybridization processes occur fast, on a timescale that is too rapid to accumulate enough mutations. Instead, recombination processes are sufficiently rapid so as to leave a footprint in genomes undergoing hybridization. After admixture of two lineages, contiguous genomic blocks within the genome are broken down by recombination over time. The delineations between these blocks were termed "junctions" by Fisher [Fis49, Fis54], and inheritance of these junctions is similar to that of point-mutations. Further work on the theory of junctions has shown how junctions accumulate over time for sib-sib mating [Fis54], self-fertilization [Ben53], alternate parent-offspring mating [Fis59, Gal64], a randomly mating population [Sta80], and for sub-structured populations [CT02, CT03].

So far, applying the theory of junctions has shown to be difficult, as it requires extensive genotyping of the admixed lineage, but also of the parental lineages. With the current decrease in genotyping costs [MLL$^+$16], such analyses are currently coming well within reach, and frameworks are currently being developed that assist in inferring local ancestry, given molecular data of parental and admixed lineages (in order to identify the markers of the two populations). Nevertheless, molecular data always provide an imperfect picture of ancestry along the genome, and inferring the number of junctions along a genome remains limited by the number of diagnostic markers available. Currently, the theory of junctions does not take into account the effect of a limited number of genetic markers, and previous analyses taking into account the effect of a limited number of markers have had to resort to simulations [MHWS05, BR08]. Janzen et al. [JNT18] provided an analytical description of the effect of using evenly spaced genetic markers, and they find that predictions taking into account this marker effect differ substantially from more naive predictions not taking this marker effect into account. Molecular markers are rarely evenly spaced, and it is therefore warranted to extend the current theory of junctions by including the effect of

non-evenly spaced molecular markers on inferring the number of junctions in a genome, and inference of subsequent properties, such as the time since admixture. Here we present a new description of the expected number of junctions after admixture, given an arbitrary distribution of markers.

2 The expected number of detected junctions

We assume Wright-Fisher dynamics with non-overlapping generations, random mating and we assume that all individuals are hermaphrodite. Crossovers are assumed to be uniformly distributed along the chromosome. We only keep track of one pair of chromosomes, assuming that the accumulation of junctions between chromosomes is independent of each other. We assume that hybridization occurred at time 0 between two populations, P and Q. The proportion of individuals from population P at time 0 is p and the proportion of individuals of type Q is $q = 1 - p$.

We assume that the length of the chromosome is R morgans and that there are n molecular markers whose positions , scaled by R, are given by $(z_1, \dots z_n) \in [0, 1]$. We consider two markers at sites z_i and z_{i+1} and we define $d_i = d_i$. The question is how many junctions are interspersed between the two markers. If more than one junction is expected to be interspersed, the total number of junctions is expected to be underestimated. In fact if the real number of junctions between the tho markers is $2n + 1$ we will infer 1 junction and if it is $2n$ we will observe no junction. The number of junctions depends on Rd_i, which is the distance between the two markers in morgans.

We can solve this through the Ancestral Recombination Graph, which follows backwards in time the ancestry of the two sites. There are two possible states $(z_{i+1} \sim z_{i+1})$ (when both loci are carried by the same lineage) and state $(z_i \not\sim z_{i+1})$ (where each locus is carried by a different lineage). The transition probabilities for this process are the given by

$$\begin{aligned}
(z_i \sim z_{i+1}) &\to (z_i \not\sim z_{i+1}) \quad \text{with probability} \quad Rd_i \\
(z_i \sim z_{i+1}) &\to (z_i \sim z_{i+1}) \quad \text{with probability} \quad 1 - Rd_i \\
(z_i \not\sim z_{i+1}) &\to (z_i \sim z_{i+1}) \quad \text{with probability} \quad \frac{1}{2N} \\
(z_i \not\sim z_{i+1}) &\to (z_i \not\sim z_{i+1}) \quad \text{with probability} \quad 1 - \frac{1}{2N}.
\end{aligned}$$

Other events (such as simultaneous coalescence and recombination events) have probabilities that are negligible when N is large. This yields the following transition matrix:

$$\boldsymbol{M} = \begin{bmatrix} \mathbf{1 - Rd_i} & \boldsymbol{Rd_i} \\ \frac{\mathbf{1}}{\mathbf{2N}} & \mathbf{1 - \frac{1}{2N}} \end{bmatrix}.$$

Let P_t be the probability vector at time t for this Markov chain with two states. $(P_t)_1$ is the probability that $z_i \sim z_{i+1}$ at time t and $(P_t)_2$ the probability that $z_i \not\sim z_{i+1}$ at time t. We have $P_0 = (1, 0)$ (In the present we sample the two loci in the same individual) and

$$P_t = P_0 M^t.$$

We denote by $\mathbb{P}(J_t(z_i, z_{i+1}))$ the probability that a junction is observed between z_i and z_{i+1}, if the hybridization event happened t generations ago. We have

$$\mathbb{P}(J_t(z_i, z_{i+1})) \;=\; 2pq(P_t)_2,$$

which corresponds to the probability that the two loci were carried by different lineages t generations ago and the two lineages correspond to individuals from different ancestral subpopulations. Solving this gives:

$$\mathbb{P}(J_t(z_i, z_{i+1})) = 2pq\frac{2NR}{2NR + 1/d_i}\left(1 - \left(1 - Rd_i - \frac{1}{2N}\right)^t\right).$$

Let $\mathbb{E}(J_t)$ be the expected number of observed junctions, we have

$$\begin{aligned}\mathbb{E}(J_t) &= \sum_{i=1}^{n-1} \mathbb{P}(J_t(z_i, z_{i+1})) \\ &= \frac{4pqNR}{2NR + 1/d_i} \sum_{i=1}^{n-1}\left(1 - \left(1 - Rd_i - \frac{1}{2N}\right)^t\right). \end{aligned} \tag{II.1}$$

To infer the admixture time $\bar{T}$, given an observed number of junctions J_{obs}, we have to solve numerically solve this equation.

If we assume that the n molecular markers are uniformly spaced, i.e. that the distance between two consecutive markers is always $d := 1/n$, we get

$$\begin{aligned}\mathbb{E}(J_t) &= \sum_{i=1}^{n-1} \mathbb{P}(J_t(z_i, z_{i+1})) \\ &= \frac{4pqNR(n-1)}{2NR + n}\left(1 - \left(1 - \frac{R}{n} - \frac{1}{2N}\right)^t\right). \end{aligned} \tag{II.2}$$

In this case, we can easily solve this equation, and the estimated admixture time $\bar{T}$ given J_{obs} is

$$\bar{T} = \frac{\log(1 - \frac{J_{obs}(2NR+n)}{4pqRN(n-1)})}{\log(1 - \frac{R}{n} - \frac{1}{2N})}.$$

Finally, taking $t \to \infty$ in (II.2), gives

$$\mathbb{E}(J_\infty) = 2pqR(n-1)\frac{2}{2NR+n}$$

which is identical to the expectation of J_∞ in [JNT18], obtained from a model forwards in time.

3 Individual Based Simulations

To verify our findings, and validate their correctness, we compare results from individual based simulations with our analytical expectations. We perform individual based simulations as described in Janzen et al. [JNT18], using the package JUNCTIONS. Briefly, the simulations employ a Wright-Fisher model with a uniform recombination rate across the genome. We performed 1000 replicate simulations, and report the mean number of detected junctions across these replicates. Simulations were performed for two population sizes ($N = 100$, and $N = 1000$) and for three different marker densities: $n = 100, 1000$ and $n = 10000$. Across all parameter combinations we observe that the mean number of detected junctions in the simulations is close or identical to the predicted number following equation (II.1) (Figure II.3).

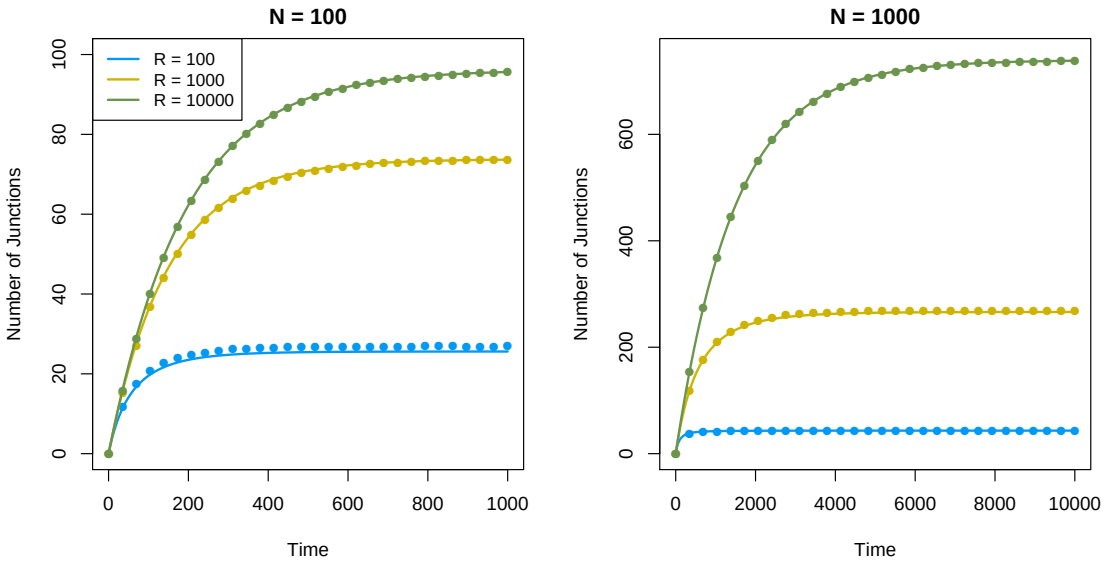


Figure II.3 – The mean number of detected junctions for either individual based simulations (dots) or as predicted by equation (II.1) (lines), for two different population sizes, and three different marker densities. Across all parameter combinations, results from individual based simulations are strongly congruent with the analytical predictions.

Alternatively, we can simulate the process of accumulating junctions over time for

a given number of generations, and then try to infer the time spent since the start of hybridization. This focuses solely on the final distribution after a given number of generations. Such a scenario more closely reflects the application of the extended theory of junctions. We simulate 100 replicate simulations with the following parameters: $N = 1000, R = 1, H_0 = 0.5$, and explore $n = [100, 300, 1000, 3000, 10000]$. We inferred the age of the hybrids using three different methods: firstly, we used equation (II.1) (by numerically solving for t). Secondly, we used the extended junctions framework (more specifically, the function `estimate time` in the junctions package), which assumes that markers are evenly distributed along the chromosome. Lastly, we ignored potentially confounding factors due to the used marker distribution, and assumed $n = \infty$. We find that with increasing marker number, all three approaches obtain more accurate results (Figure II.4 and Table 3). Assuming an even marker distribution tends to underestimate the age, especially for lower marker numbers. When the effect of markers is ignored, the underestimation is even worse. When using Eq. (II.1) , inaccuracies in the age estimate only arise for extremely low numbers of markers (< 300) (Table 3), which is also associated with an increase in the variance of the estimate (Figure II.4).

R	Random Markers	Even Markers	No Markers
100	284	108	64
300	202	151	117
1000	200	182	165
3000	198	191	185
10000	200	198	196

Table II.1 – Mean inferred age for 100 simulations ran for 200 generations. Used are three different methods to infer the age after obtaining the mean number of junctions at the end of the simulations using n randomly distributed markers: **Random markers** uses equation (II.1) to estimate the age, **Even Markers** uses the Extended theory of junctions frameworks and **No Markers** ignores the effect of markers by assuming that there are an infinite number of markers ($n = \infty$).

Discussion

We have extended the theory of junctions by including the effect of using only a limited, finite, number of molecular markers to detect the accumulated number of junctions over time after an admixture event.

Using individual based simulations, we have verified the accuracy of our method, and have shown how including the distribution of markers on the chromosome improves estimates of admixture time, even when the number of markers is very low (< 1000).

With this new extension to the theory of junctions, we hope to provide future research with valuable tools to accurately infer the timing of admixture. Code to perform individual

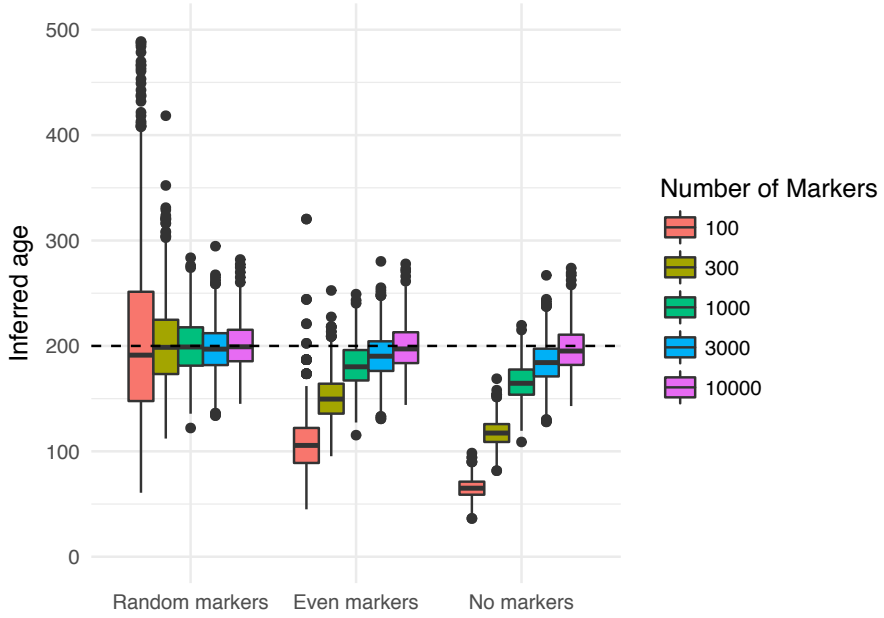


Figure II.4 – Distribution of inferred ages for 100 simulations ran for $t = 200$ generations. Shown are results for $n = [100, 300, 1000, 3000, 10.000]$, and three different inference methods. See text for explanation of the different methods.

based simulations, and functions to calculate the expected number of junctions, or to numerically estimate the expected time since admixture has been made available in the package JUNCTIONS.

Chapter III

How does geographical distance translate into genetic distance?

Joint work with Emmanuel Schertzer. Currently in revision in *Stochastic Processes and their Applications* [MPS17].

Abstract

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

1 Introduction

1.1 Genetic distances in structured populations. Speciation

In most species, the geographical range is much larger than the typical dispersal distance of its individuals. A species is usually structured into several local subpopulations

with limited genetic contact. Because migration only connects neighbouring populations, more often than not, populations can only exchange genes indirectly, by reproducing with one or several intermediary populations. As a consequence, the geographical structure tends to buffer the homogenising effect of migration, and as such, it is considered to be one of the main drivers for the persistence of genetic variability within species (see [Mal48] or [Kar82]).

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

Several authors have studied parapatric speciation, i.e. speciation in the presence of gene flow between subpopulations, for example [GLV98, GAG00, GLV00] and [YI13, YI15]. In their models, some loci on the chromosome are responsible for reproductive isolation. These loci may be involved in incompatibilities at any level of biological organisation (molecular, physiological, behavioural etc) and either prevent mating (pre-zygotic incompatibilities) or prevent the development of hybrids (post-zygotic incompatibilities). The number of segregating loci increases through the accumulation of mutations, and decreases after each migration event (creating the opportunity for some gene exchange between the migrants and the host population). When the number of segregating loci between two individuals reaches a certain threshold, they become reproductively incompatible. For example, Yamaguchi and Iwasa [YI13, YI15] studied the case of a metapopulation containing two homogeneous subpopulations. The authors studied how the genetic distance, defined as the number of loci differing between the two subpopulations, evolves through time, using a continuous-time model. When considering metapopulations with more than two subpopulations, this kind of dynamics may translate into complex patterns of speciation. One particularly intriguing example is the case of ring species [Noe97, GLV98], where two neighbouring subpopulations are too different to be able to reproduce with one another but can exchange genes indirectly, by reproducing with a series of intermediate subpopulations that form a geographic "ring". How these patterns emerge and are maintained is still poorly understood, and we hope that our analytical result might shed some new light on the subject.

1.2 Population divergence and fitness landscapes

To study speciation by accumulation of genetic differences, we model the evolution of some loci on the chromosome, that are potentially involved in reproductive incompatibilities. To visualise these evolutionary dynamics, Wright [Wri32] suggested the metaphor

of adaptive landscapes. Adaptive landscapes represent individual fitness as a function defined on the genotype space, which is a multi-dimensional space representing all possible genotypes. Wright emphasised the idea of 'rugged' adaptive landscapes, with peaks of fitness representing species and valleys representing unfit hybrids. Speciation, seen as a population moving from one peak to another, implies a temporary reduction in fitness, which is not very likely to occur in large populations, where genetic drift is not important enough to counterbalance the effect of selection (see [Gav97] for a more detailed discussion). However, Gavrilets [Gav97] suggested the idea of 'holey' adaptive landscapes, where local fitness maxima can be partitioned into connected sets (called evolutionary ridges). Speciation is therefore seen as a population diffusing across a ridge, by neutral mutation steps, until it stands at the other side of a hole. Theoretical models, such as [GG97], have shown, using percolation theory, that in high-dimensional genotype spaces, fit genotypes are typically connected by evolutionary ridges.

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

Further, we consider that the evolutionary dynamics along the ridge are slow (as random mutations are very likely to be deleterious, mutations along the evolutionary ridge are assumed to be rarer than in the typical population genetics framework), which is why we study our model in a low mutation - low migration regime (see Section 1.4 for more details). This assumption is commonly made when studying speciation, for example in [GAG00] or [YI13].

1.3 An individual based model (IBM)

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

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

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

— Each individual x reproduces at constant rate 1 and chooses a random partner y ($y \neq x$). Upon reproduction, their offspring replaces a randomly chosen individual in the population.

— The new individual inherits a chromosome which is a mixture of the parental chromosomes. Both parental chromosomes are cut into fragments in the following way: we assume a Poisson Point Process of intensity λ on $[0,1]$. Two loci belong to the same fragment iff there is no atom of the Poisson Point Process between them. For each fragment, the offspring inherits the fragment of one of the two parents chosen randomly.

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

— **Mutation** occurs at rate $b^{\gamma,\epsilon}$ per individual, per locus according to an infinite allele model.

— **Migration** from subpopulation i to subpopulation j occurs at rate m_{ij}^γ. At each migration event, one individual migrates from subpopulation i to j, and replaces one individual chosen uniformly at random in the resident population. (We set $\forall i \in E,\ m_{i,i}^\gamma = 0$.)

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

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

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

$$d^{\epsilon,\gamma}(i,j) = \left(\frac{1}{n_i^\epsilon}\sum_{x\in\{i_1,\cdots,i_{n_i^\epsilon}\}}\min_{y\in\{j_1,\cdots,j_{n_j^\epsilon}\}}\delta^{\gamma,\epsilon}(x,y)\right) \vee \left(\frac{1}{n_j^\epsilon}\sum_{y\in\{j_1,\cdots,j_{n_j^\epsilon}\}}\min_{x\in\{i_1,\cdots,i_{n_i^\epsilon}\}}\delta^{\gamma,\epsilon}(x,y)\right) \quad \text{(III.1)}$$

This corresponds to the so-called modified Haussdorff distance between subpopulations, as introduced by [DJ94]. (This distance has the advantage of averaging over the individuals in each subpopulation, so introducing a single mutant or migrant would produce a smooth variation in the genetic distances.)

1.4 Slow mutation–migration and large population - long chromosome regime.

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

It is well known that in the absence of mutation and migration, the neutral Moran model describing the dynamics at the local level reaches fixation in finite time: the average time to fixation for a single locus is of the order of the size of the subpopulation [KO69, Kim68] (In our multi-locus model, it will also depend on the number of loci and on the recombination rate λ.) Heuristically, if we assume a low mutation - low migration regime, i.e. that

$$\forall i,j \in E, \quad \frac{1}{b^{\gamma,\epsilon} n_i^\epsilon}, \ \frac{1}{m_{i,j}^\gamma} \ \gg \ n_j^\epsilon, \ l^\epsilon \gg 1, \tag{III.2}$$

the average time between two migration events ($1/m_{i,j}^\gamma$), and the average time between two successive mutations at a given locus ($1/(b^{\gamma,\epsilon} n_j^\epsilon)$) are much larger than the average time to fixation. This ensures that the fixation process is fast compared to the time-scale of mutation and migration, and, as a result, when looking at a randomly chosen locus, subpopulations are homogeneous except for short periods of time right after a migration event or a mutation event. This suggests that if we accelerate time properly, we can neglect intra-subpopulation diversity and approximate our model by a population based model.

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

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

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

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

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

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

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

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

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

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

In particular,

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

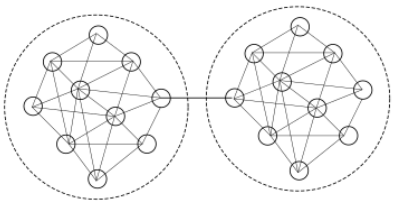
This result can be seen as a law of large numbers over the chromosome. Although the loci are linked (through recombination) and they do not fix independently, when considering a large number of them, they become decorrelated, regardless of the value of λ. (Note that the limiting process does not depend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

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

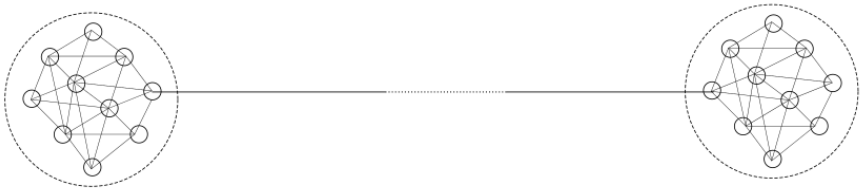
1.5 Consequences of our result

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

One striking illustration of the previous discussion is presented in Section 6, where we consider an example where a geographic bottleneck is dramatically amplified in our new



(a) Geographic distances



(b) Genetic distances

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

metric. See Figure 1.5 and Theorem 6.1 for a more precise statement. If we consider, as in [YI13], that two populations are different species if their genetic distance reaches a certain threshold, that will mean that this metapopulation structure promotes the emergence of two different species, each one corresponding to the population in one subgraph. Very often, parapatric speciation is believed to occur only in the presence of reduced gene flow. Our example shows that in the presence of a geographic bottleneck, genetic differentiation is manly driven by the geographical structure of the population, i.e., even if the gene flow between two neighbouring subpopulations is approximatively identical in the graph, the genetic distance is dramatically amplified at the bottleneck (see Figure 1.5).

We note that using the hitting time of random walks as a metric on graphs is not new, and has been a popular tool in graph analysis (see [DS84] and [KR93]). For example, the commute distance, which is the time it takes a random walk to travel from vertex i to j and back, is commonly used in many fields such as machine learning [VLRH14], clustering [YVW$^+$05], social network analysis [LNK03], image processing [QH05] or drug design [Iva00, Roy04]. In our case the genetic distance is given by the Laplace transform of the hitting time between two random walks, which was already suggested as a metric on graphs by [HSJ15]. In that paper the authors claimed that this metrics preserves the cluster structure of the graph. In the example alluded above (Section 6), we found that

our metric reinforces the cluster structure of the metapopulation graph. In other words, a clustered geographic structure tends to increase genetic differentiation.

1.6 Discussion and open problems

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

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

1.7 Outline

In Section 2, we show that in the rare mutation-rare migration regime (i.e. when $\gamma \to 0$ whereas ϵ remains constant), the individual based model (IBM) described above converges to a population based model (PBM) (see Theorem 2.2). This PBM is a generalization of the model proposed by Yamaguchi and Iwasa [YI13, YI15] in three ways. First, it is an extension of their model from two to an arbitrary number of subpopulations, which is not trivial from a mathematical point of view. Second, in [YI13, YI15], the authors only assumed that the migrant alleles are fixed independently at every locus. Too make the model more realistic, we took into account genetic linkage, which introduces a non-trivial spatial correlation between loci (along the chromosome). Finally, we suppose that the loci are distributed randomly along the chromosome (and not in a regular fashion). Section 2 is interesting on its own since it provides a theoretical justification of the model proposed in [YI13, YI15].

In Section 3 and 4, we study the PBM in the large population - long chromosome limit

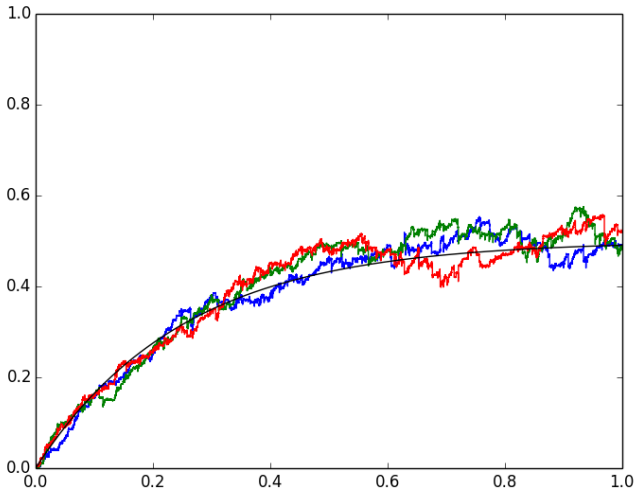


Figure III.2 – Simulation of the individual based model, for $d = 3$, $N_1 = N_2 = N_3 = 1$, $\epsilon = 0.005$, $\gamma = 2e^{-6}$, $l^\epsilon = 100$, $\lambda = 10$. The black curve corresponds to $D_t(i,j)$ (see Theorem 1.1). The blue, green and red curves correspond to the three genetic distances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

(i.e. when $\epsilon \to 0$). We properly introduce the main tool used to study the population based model – the genetic partition probability measure – and show an ergodic theorem related to this process (see Theorem 3.1).

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

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

2 Approximation by a population based model

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

Before going into the description of our model, we start with a definition. It is well known that the Moran model reaches fixation in finite time, i.e., after a (random) finite time, every individual in the population carries the same genetic material, and from that time on, the system remains trapped in this configuration (see [KO69, Kim68]).

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

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

— For every $i \in E$: fix a new mutation in population i at rate $b_\infty l^\epsilon$, the locus being chosen uniformly at random along the chromosome.

— For every $i, j \in E$ and every $S \subseteq \{1, \cdots, l^\epsilon\}$: at every locus in S, fix simultaneously the alleles from population i in population j at rate $\frac{1}{\epsilon} M_{ij} \mathbb{P}\left(\mathcal{F}_j^{L^\epsilon,\lambda} = S\right)$.

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

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

as opposed to $d^{\gamma,\epsilon}$ which will refer to the genetic distances in the IBM as described in Section 1.3 (parametrised by γ and ϵ). We note that the definition of the genetic distance in the PBM is consistent with the one in the IBM (see (III.1)) in the sense that if the subpopulations are homogeneous in the IBM, (III.1) is equal to the RHS of the previous equation. We are now ready to state the main result of this section.

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

$$\lim_{\gamma \to 0} \left(d^{\gamma,\epsilon}_{t_1/(\gamma\epsilon)}, \cdots, d^{\gamma,\epsilon}_{t_n/(\gamma\epsilon)}\right) \;=\; \left(d^\epsilon_{t_1}, \cdots, d^\epsilon_{t_n}\right) \quad \textit{in distribution.} \tag{III.3}$$

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

We define a coupling between the IBM and a new PBM that is close (in distribution) to the PBM defined at the beginning of this section. The idea behind the coupling is that, when time is accelerated by $\gamma\epsilon$, and γ is small, in the IBM, the time to fixation after a mutation or migration event is short enough so that the population has reached fixation before the next mutation or migration event takes place. Then, we can decompose the trajectories of the IBM into periods where the population is homogeneous (and waits

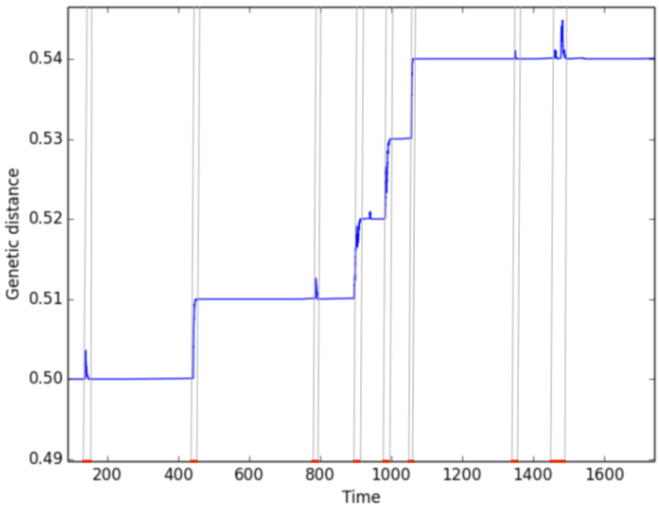


Figure III.3 – The curve represents the genetic distance between two subpopulations for the IBM. The vertical lines represent the decomposition of the trajectories. The fixation phases are represented in red. Simulation with 3 subpopulations, with 100 individuals on each, carrying 100 loci, with $\gamma = 0.0001$, $\forall i, j, M_{i,j} = 1$, $b_\infty = 1$, $\lambda = 5$.

for the next mutation or migration event to take place) and fixation phases (where the dynamics of the population is described by a Moran model). See Figure III.3 for an illustration of this concept.

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

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

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

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

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

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

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

Let us now consider

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

By construction, this process defines a PBM in the sense that at every time t, any subpopulation is composed by genetically homogeneous individuals. Further, since $(T_t^{\gamma,\epsilon}; t \geq 0)$ converges to the identity and mutation and migration events occur at Poisson times, the finite dimensional distributions of $Z^{\gamma,\epsilon}$ are a good approximations of the ones for the IBM.

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

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

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

Since fixation occurs in finite time almost surely, and the distribution of the fixation time only depends on ϵ, we have

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

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

Putting all the previous observations together, one can easily show that the genetic distance in $Z^{\gamma,\epsilon}$ converges (in the finite dimensional distributions sense) to the ones of the

PBM. In particular, we recover the mutation rate on subpopulation i in the PBM

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

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

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

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

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

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

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

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

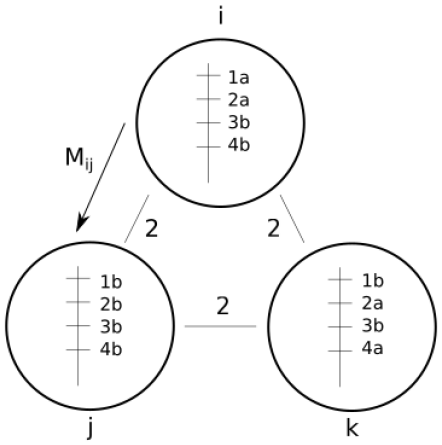
Remark 2.4. *Our model can be seen as a multi-locus Moran model with inhomogeneous reproduction rates. The main difficulty in analysing this model stems from the fact that there exists a non trivial correlation between loci. This correlation is induced by the fact that fixation of migrant alleles can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

3 Large population - long chromosome limit

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

3.1 The genetic partition measure

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



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

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

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

Let $\mathcal{P}_d$ the set of partitions of $\{1, \cdots, d\}$. Let $\pi \in \mathcal{P}_d$ and $i, j \in E$. Define $\mathcal{S}_i(\pi)$ as the element of $\mathcal{P}_d$ obtained from π by making i a singleton (e.g., $\mathcal{S}_2(\{1,2,3\}) = \{1,3\}\{2\}$). Define $\mathcal{I}_{ij}(\pi)$ as the element of $\mathcal{P}_d$ obtained from π by displacing j into the block containing i (e.g., $\mathcal{I}_{2,3}(\{1,3\}\{2\}) = \{1\}\{2,3\}$).

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

In the following, fix $L^\epsilon \in [0,1]^{l^\epsilon}$, the vector containing the positions of the loci. In the PBM parametrised by ϵ, we condition on the loci being located at L^ϵ and for every $k \in \{1, \cdots, l^\epsilon\}$ we let $\Pi_k^{\epsilon, L^\epsilon}(t)$, the partition induced at locus k (see Figure III.4). The vector

$\Pi^{\epsilon,L^\epsilon}(t) = \left(\Pi_k^{\epsilon,L^\epsilon}(t);\ k \in \{1,\cdots,l^\epsilon\}\right)$ describes the genetic composition of the population at time t. According to the description of our dynamics, $\Pi^{\epsilon,L^\epsilon}(t)$ is a Markov chain with the following transition rates:

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

$$\mathcal{S}_i^k(\Pi) = \begin{cases} \Pi_j & \forall j \neq k \\ \mathcal{S}_i(\Pi_j) & j = k \end{cases}.$$

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

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

$$\mathcal{I}_{ij}^S(\Pi) = \begin{cases} \Pi_k & \forall k \notin S \\ \mathcal{I}_{ij}(\Pi_k) & \forall k \in S \end{cases}.$$

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

To summarise, for every test function h, the generator of $\Pi^{\epsilon,L^\epsilon}$ can be written as

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

3.2 Some notation

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

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

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

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

Define

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

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

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

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

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

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

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

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

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

3.3 Convergence of the genetic partition probability measure

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

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

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

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

Recall that the expression of the generator associated to the allelic partition at a given locus k is independent on the position of locus k and on λ. Also recall that $\epsilon n_i^\epsilon \to N_i$, and

thus

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

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

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

In the light of (III.7), the following theorem can be interpreted as an ergodic theorem. We show that the (dynamical) empirical measure constructed from the allelic partitions along the chromosome converges to the probability measure of a single locus. Although in the IBM the different loci are linked and do not fix independently (as already mentioned in Remark 2.4), as the number of loci tends to infinity, they become decorrelated. In the large population - long chromosome limit, the following result indicates that the model behaves as if infinitely many loci evolved independently according to the (one-locus) Moran model with generator G provided in (III.7).

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

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

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

Then

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

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

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

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

4 Proof of Theorem 3.1

The idea behind the proof is to condition on $\mathcal{L}^\epsilon = L^\epsilon$, and then decompose the Markov process $(< \xi_t^{\epsilon, L_\epsilon}, v >; t \geq 0)$ into a drift part and a Martingale part. We show that the drift part converges to the solution of the Kolmogorov equation alluded in Theorem 3.1

and that the Martingale part vanishes when $\epsilon \to 0$. The main steps of the computation are outlined in the next subsection. We leave technical details (tightness and second moment computations) until the end of the section.

4.1 Main steps of the proof

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

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

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

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

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

Proof. We define the two following signed measures:

$$\partial_{ij}^S X(\Pi) = X(\mathcal{I}_{ij}^S(\Pi)) - X(\Pi),\quad \partial_i^K X(\Pi) = X(\mathcal{S}_i^K(\Pi)) - X(\Pi). \qquad \text{(III.12)}$$

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

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

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

$$\mathbb{E}_{\lambda,L^\epsilon,j}\left(\left\langle \partial_{ij}^S X(\Pi), v\right\rangle\right) = \frac{1}{n_j^\epsilon}\left\langle \mathcal{I}_{ij} * X(\Pi) - X(\Pi), v\right\rangle,$$

$$\mathbb{E}_{l^\epsilon}\left(\left\langle \partial_i^K X(\Pi)\right\rangle\right) \quad = \quad \frac{1}{l^\epsilon}\left\langle \mathcal{S}_i * X(\Pi) - X(\Pi), v\right\rangle. \tag{III.13}$$

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

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

Secondly, for every $j\in E$,

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

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

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

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

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

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

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

This completes the proof of Lemma 4.1. □

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

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

Since $\left\langle \xi_t^{\epsilon,L^\epsilon}, v\right\rangle$ is bounded, the previous result implies that $M^{\epsilon,L^\epsilon,v}$ is a martingale with respect to $(\mathcal{F}_t^{L^\epsilon})_{t\ge0}$, the filtration generated by $(\Pi^{\epsilon,L^\epsilon}(t);t\ge0)$. Further, the semi-martingale

$\left\langle \xi_t^{\epsilon,L^\epsilon}, v \right\rangle$ admits the following decomposition:

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

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

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

with

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

Proof. Let $h(\Pi) = \langle X(\Pi), v\rangle^2$. Then, by (III.10)

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

Since

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

the previous identities yield

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

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

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

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

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

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

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

Proposition 4.3.

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

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

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

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

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

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

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

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

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

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

Lemma 4.2 and Proposition 4.3 imply that

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

which ends the proof of Theorem 3.1. □

Remark 4.5 (Magnitude of the stochastic fluctuations)**.** *Lemma 4.1 and the proof Proposition 4.3 entail that:*

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

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

In [YI13], the authors proposed a diffusion approximation (only for the case of two subpopulations). Their approximation is based on the simplifying hypothesis that loci are fixed independently on each other – the number of fixed loci (after each migration event) follows a binomial distribution–, and the hypothesis that the number of loci l is s.t. $l >> \frac{1}{\epsilon}$. They found that the magnitude of the stochastic fluctuations was $\sqrt{\epsilon}$.

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

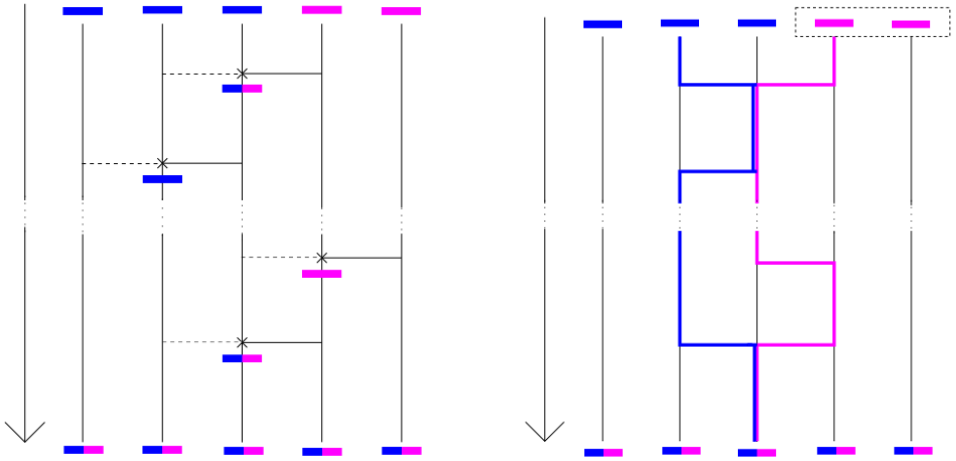
4.2 Proof of Proposition 4.3

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

Lemma 4.6. $\forall\ i, j\ \in E, \forall \lambda > 0,$

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

Before turning to the proof of this result, we recall the definition of the ancestral recombination graph (ARG) (see also [Hud83], [Gri81], [Gri91]) for the case of two loci. Fix $L^\epsilon = \{\ell_1, \cdots \ell_{l^\epsilon}\}$ the positions of the loci in the chromosome, λ the recombination rate, and choose two loci k and k' among the l^ϵ loci. In order to compute the probability that for both loci the allele from the migrant is fixed in the host population – $\mathbb{P}(k, k' \in \mathcal{F}_j^{L^\epsilon,\lambda})$ – we follow backwards in time the genealogy of the corresponding alleles carried by a reference individual in the present population, assuming that a migration event occurred in the past (sufficiently many generations ago, so that we can assume fixation).



(a) Moran model with recombination (b) ARG

Figure III.5 – Realisation of Moran model with recombination and the Ancestral Recombination Graph. In the figures the population size is equal to 5 and $l^\epsilon = 2$. Time goes from top to bottom as indicated by the arrow on the left. In Figure (a) the origins of the arrows indicate the parents, and the tips of the arrows point to their offspring. The dashed arrow corresponds to the father and the solid line to the mother. In (b) the blue and pink lines correspond to the ancestral lineages of two distinct loci belonging to the same chromosome in the extant population.

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

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

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

$\mathbf{A}^{L^\epsilon,\lambda,j} = (A_k^{L^\epsilon,\lambda,j},\ A_{k'}^{L^\epsilon,\lambda,j})$ defines a 2-dimensional stochastic process on $\{1, \cdots, n_j^\epsilon\}$. At time 0, the two particles have the same position (they coincide at a randomly chosen individual as in Figure III.5) and then evolve according to the following dynamics:

— When both particles are occupying the same location z, the group splits into two at rate $r_{k,k'}^{L^\epsilon,\lambda}$ (see (III.17)). Forward in time this corresponds to a reproduction event

where z is replaced by the offspring of x and y. Each individual x reproduces at rate 1 (chooses a random partner y), and with probability $1/n_j^\epsilon$ his offspring replaces individual z. There are n_j^ϵ possible choices for x. Following (III.17), the probability that both loci are inherited from different parents is $r_{k,k'}^{L^\epsilon,\lambda}$, so the rate of fragmentation for loci k, k' is given by $n_j^\epsilon \,.\, \frac{1}{n_j^\epsilon} \,.\, r_{k,k'}^{L^\epsilon,\lambda}$.

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

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

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

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

Proof of Lemma 4.6. Define $(Y^{L^\epsilon,\lambda,j}(t) := 1_{A_k^{L^\epsilon,\lambda,j}(s)=A_{k'}^{L^\epsilon,\lambda,j}(s)}; s \geq 0)$. It is easy to see from the previous description of the dynamics that Y is a Markov chain on $\{0,1\}$ with the following transition rates:

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

and further

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

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

We have:

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

Furthermore, it is straightforward to show that

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

From (III.18), we get that,

$$\mathbb{E}_{\lambda,L^\epsilon,j}(\sum_{k,k'\in\{1,\cdots,l^\epsilon\}} 1_{k\in S}1_{k'\in S}) = \lim_{t\to\infty}\sum_{k,k'\in\{1,\cdots,l^\epsilon\}}\mathbb{P}(A_k^{L^\epsilon,\lambda,j}(t)=A_{k'}^{L^\epsilon,\lambda,j}(t)=1)$$
$$= \frac{2}{(n_j^\epsilon)^2}\sum_{k,k'\in\{1,\cdots,l^\epsilon\}}\frac{1}{\frac{1}{2}(1-e^{-\lambda|\ell_k-\ell_{k'}|})+\frac{2}{n_j^\epsilon}}.$$

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

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

Thus,

$$\mathbb{E}\left(\mathbb{E}_{\lambda,\mathcal{L}^\epsilon,j}(\sum_{k=1}^{l^\epsilon}\sum_{k'=1}^{l^\epsilon}1_{k\in S}1_{k'\in S}\mid\mathcal{L}^\epsilon)\right) \le \frac{2}{(n_j^\epsilon)^2}\int_{[0,1]^{l^\epsilon}}dx_1,\cdots dx_{l^\epsilon}\sum_{k,k'\in\{1,\cdots,l^\epsilon\}}\frac{1}{\alpha|x_k-x_{k'}|+\frac{2}{n_j^\epsilon}}$$
$$\le \frac{2}{(n_j^\epsilon)^2}\sum_{k,k'\in\{1,\cdots,l^\epsilon\}}\int_{[0,1]^2}\frac{dx_kdx_{k'}}{\alpha|x_k-x_{k'}|+\frac{2}{n_j^\epsilon}}.$$

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

$$\frac{1}{(l^\epsilon)^2\epsilon}\mathbb{E}\left(\mathbb{E}_{\lambda,\mathcal{L}^\epsilon,j}(\sum_{k=1}^{l^\epsilon}\sum_{k'=1}^{l^\epsilon}1_{k\in S}1_{k'\in S}\mid\mathcal{L}^\epsilon)\right)$$
$$\le \frac{2\epsilon}{(N_j-\epsilon)^2}\int_{[0,1]^2}\frac{dtds}{\alpha|t-s|+2\epsilon/N_j}$$
$$= \frac{4\epsilon}{(N_j-\epsilon)^2}\int_0^1 ds\int_o^s\frac{dt}{\alpha|t-s|+2\epsilon/N_j}$$
$$= \frac{4\epsilon}{\alpha(N_j-\epsilon)^2}\int_0^1\log\left(\frac{\alpha N_j}{2\epsilon}s+1\right)ds$$
$$= \frac{4\epsilon}{\alpha(N_j-\epsilon)^2}\left((1+\frac{2\epsilon}{\alpha N_j})\log\left(\frac{\alpha N_j}{2\epsilon}+1\right)-1\right)$$
$$\xrightarrow[\epsilon\to0]{} 0.$$

□

We are now ready to prove Proposition 4.3.

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

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

To bound the second term in the RHS, we note that, by definition, $\mathcal{S}_i^k(\Pi)$ and Π only

differ in one component, so from the definition of $\partial_i^K X(\Pi)$ (see (III.12)), it is not hard to see that

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

It follows that,

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

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

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

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

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

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

randomising the positions of the loci and using Lemma 4.6 the term on the RHS also converges and can also be bounded from above, which completes the proof. □

4.3 Tightness: Proof of Proposition 4.4

We follow closely [FM04]. It is sufficient to prove that for every $v \in \mathcal{M}_d$, the projected process $(\langle \xi^\epsilon, v \rangle\ ;\ \epsilon > 0)$ is tight. To this end, we use Aldous criterium (see [Ald89]). In the following, we define

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

We first note that

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

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

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

are bounded from above by two functions in δ (uniformly in the choice of $\tau^\epsilon, \sigma^\epsilon$ and ϵ) going to 0 as δ go to 0. The rest of the proof is dedicated to proving those two inequalities. We start with the martingale part. First,

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

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

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

where $m^{\epsilon,L^\epsilon,v}(\Pi)$ was defined in Lemma 4.2 and where the second line follows from the fact that τ^ϵ and σ^ϵ are stopping times for the filtration $(\mathcal{G}_t^\epsilon)_{t\ge 0}$. If there exists C_1 such that

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

then,

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

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

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

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

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

so (III.22) is proved.

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

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

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

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

So,

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

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

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

5 Proof of Theorem 1.1 and more

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

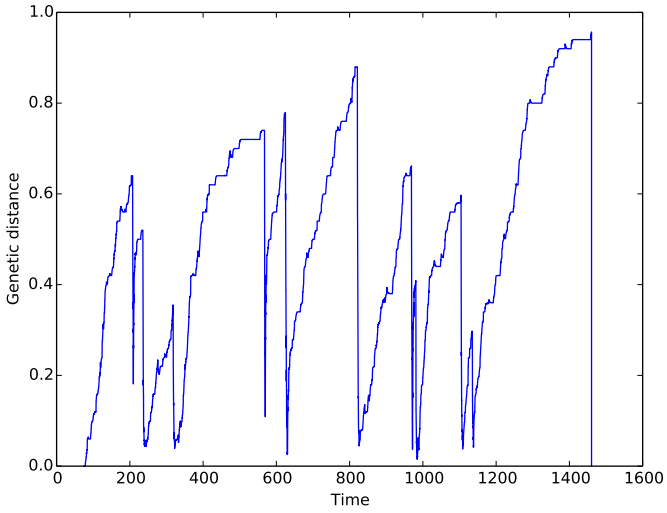


Figure III.6 – Simulation of the individual based model, for $d = 2$, $N_1 = N_2 = 1$, $\epsilon = 0.01$, $\gamma = 0.005$, $l^\epsilon = 100$, $\lambda = 0.5$. With this set of parameters, Theorem 1.1, predicts that the genetic distance at equilibrium should be 0.5. In this simulation, the mean genetic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

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

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

For every $t \geq 0$, define

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

where $\tau_{ij} = \inf\{t \geq 0 \;:\; S^i(t) = S^j(t)\}$. We have the following generalization of of Theorem 1.1.

Theorem 5.1.

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

In particular,

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

Proof. We start by proving that,

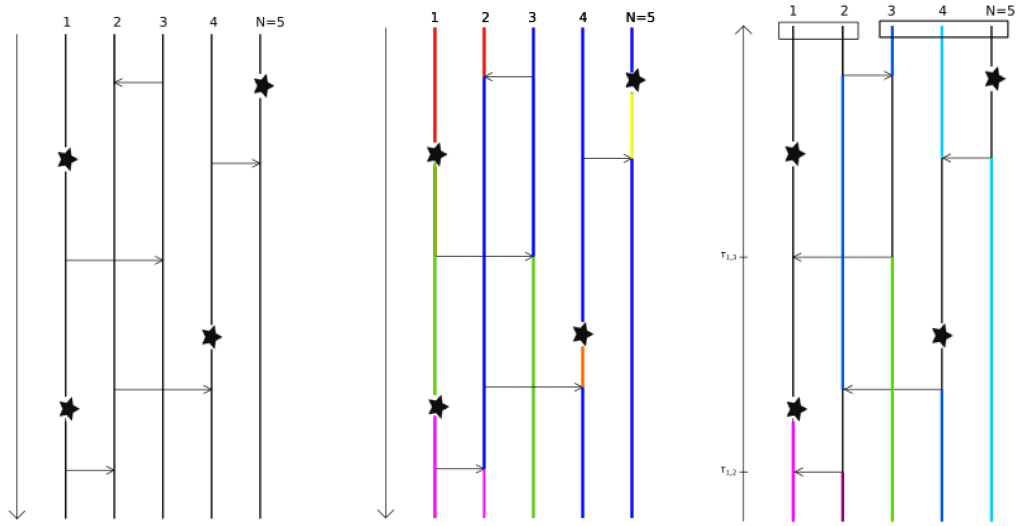
$$(d_t^\epsilon(i,j); t \geq 0) \underset{\epsilon \to 0}{\Longrightarrow} (D_t(i,j); t \geq 0) \text{ in the weak topology,} \tag{III.24}$$

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

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

— $(B_n^i, n \geq 1)$, with intensity measure $b_\infty dt$, that corresponds to mutation events at site i. At each point (i, B_n^i) we draw a $\star$ in the graphical representation (Figure III.7(a)).

— $(T_n^{i,j}, n \geq 1)$, with intensity measure $\tilde{M}_{ji} dt$, that corresponds to reproduction events, where j is replaced by i. We draw an arrow from $(i, T_n^{i,j})$ to $(j, T_n^{i,j})$ in the graphical representation to indicate that lineage j inherits the type of lineage i.



(a) Graphical representation of the Moran model

(b) Genetic partitioning process

(c) Ancestral lineages

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

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

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

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

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

(ii) $\tau_{ij}^t \geq t$ and $S^i(0) \sim_{\pi_0} S^j(0)$ and S_t^i and S_t^j their is no $\star$ in their paths.

From here, it is easy to check that:

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

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

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

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

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

This completes the proof of Theorem 1.1. □

6 An example: a population with a geographic bottleneck

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

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

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

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

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

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

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

— (II) if the vertices belong to the interior of the same subgraph (either $\mathcal{G}_1$ or $\mathcal{G}_2$).

— $(I\bar{I})$ if the vertices belong to the interior of distinct subgraphs.

— (IB) if one of the vertex is in the interior of a subgraph, and the other vertex belongs to the boundary point of the same subgraph.

— $(I\bar{B})$, $(B\bar{B})$ are defined analogously.

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

□

Bibliography

[AAA+13] R. Abbott, D. Albach, S. Ansell, J. W. Arntzen, S. J E Baird, N. Bierne, J. Boughman, A. Brelsford, C. A. Buerkle, R. Buggs, R. K. Butlin, U. Dieckmann, F. Eroukhmanoff, A. Grill, S. H. Cahan, J. S. Hermansen, G. Hewitt, A. G. Hudson, C. Jiggins, J. Jones, B. Keller, T. Marczewski, J. Mallet, P. Martinez-Rodriguez, M. Möst, S. Mullen, R. Nichols, A. W. Nolte, C. Parisod, K. Pfennig, A. M. Rice, M. G. Ritchie, B. Seifert, C. M. Smadja, R. Stelkens, J. M. Szymura, R. Väinölä, J. B. W. Wolf, and D. Zinner. Hybridization and speciation. *Journal of Evolutionary Biology*, 26:229–246, 2013.

[Ald89] D. Aldous. Stopping times and tightness. ii. *Ann. Probab.*, 17(2):586–595, 04 1989.

[Arr98] R. Arratia. On the central role of the scale invariant poisson processes on (0,infty). In *Microsurveys in discrete probability (Princeton, NJ, 1997)*, volume 41 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 21–41. Amer. Math. Soc., Providence, RI, 1998.

[Bar09] N.H. Barton. Why sex and recombination? *Cold Spring Harb Symp Quant Biol.*, 74:187–95, 2009.

[Bar10] N. H. Barton. Genetic linkage and natural selection. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1552):2559–2569, 2010.

[BB11] B.L. Browning and S.R. Browning. A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics*, 88(2):173–182, 2011.

[BBI11] A. Brelsford, Mila B., and D.E. Irwin. Hybrid origin of audubon's warbler. *Molecular Ecology*, 20:2380–2389, 2011.

[BC98] N.H. Barton and B. Charlesworth. Why sex and recombination? *Science*, 281(5385):1986–90, 1998.

[Ben53] J.H. Bennett. Junctions in inbreeding. *Genetica*, 26(1):392–406, 1953.

[Ber09] N. Beresticky. Recent progress in coalescent theory. In *Ensaios Matemáticos [Mathematical Surveys] 16*. Sociedade Brasileira de Matemática, 2009.

[BGZT12] M.D. Brown, C.G. Glazner, C. Zheng, and E.A. Thompson. Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics*, 190(4):1447–1460, 2012.

[Bil68] P. Billingsley. *Convergence of probability measures*. Wiley Series in probability and Mathematical Statistics: Tracts on probability and statistics. Wiley, 1968.

[Bla41] A. P. Blair. Variation, isolating mechanisms, and hybridization in certain toads. *Genetics*, 26(4):398–417, 1941.

[Bla48] D. Blackwell. A renewal theorem. *Duke Math. J.*, 15(1):145–150, 1948.

[BR08] C. A. Buerkle and L. H. Rieseberg. The rate of genome stabilization in homoploid hybrid species. *Evolution*, 62(2):266–275, 2008.

[BWK10] A. Bobrowski, T. Wojdyła, and M. Kimmel. Asymptotic behavior of a Moran model with mutations, drift and recombination among multiple loci. *Journal of Mathematical Biology*, 61(3):455–473, Sep 2010.

[CBT00] J.A. Coyne, N.H. Barton, and M. Turelli. Is Wright's shifting balance process important in evolution? *Evolution*, 54:307–317, 2000.

[CDRM15] T. Capblancq, L. Després, D. Rioux, and J. Mavárez. Hybridization promotes speciation in coenonympha butterflies. *Molecular Ecology*, 24(24):6209–6222, 2015.

[Cla04] A.G. Clark. The role of haplotypes in candidate gene studies. *Genetic Epidemilogy*, 27(4):321–333, 2004.

[CO04] J.A. Coyne and H.A. Orr. *Speciation*. Sinauer Associates, 2004.

[CT02] N. H. Chapman and E. A. Thompson. The effect of population history on the lengths of ancestral chromosome segments. *Genetics*, 162(1):449–458, 2002.

[CT03] N.H. Chapman and E.A. Thompson. A model for the length of tracts of identity by descent in finite random mating populations. *Theoretical Population Biology*, 64(2):141–150, 2003.

[DJ94] M.-P. Dubuisson and A. K. Jain. A modified hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 566–568. IEEE, 1994.

[Dob37] T.G. Dobhansky. *Genetics and the origin of species*. Columbia University Press, 1937.

[dQ98] K. de Queiroz. *Endless Forms: Species and Speciation*. Oxford University Press, 1998.

[dQ05] K. de Queiroz. Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences*, 102(suppl 1):6600–6607, 2005.

[DS84] P.G. Doyle and J.L. Snell. *Random Walks and Electric Networks*, volume 22. Mathematical Association of America, 1 edition, 1984.

[Dur08] R. Durrett. *Probability Models for DNA Sequence Evolution.* Springer, 2 edition, 2008.

[EK05] S.N. Ethier and T.G. Kurtz. *Markov processes: characterization and convergence.* Wiley Interscience, 2005.

[EN89] S.N. Ethier and T. Nagylaki. Diffusion approximations of the two-locus Wright-Fisher model. *J. Math. Biol.*, 27:17–28, 1989.

[EPB16] M. Esser, S. Probst, and E. Baake. Partitioning, duality, and linkage disequilibria in the moran model with recombination. *Journal of mathematical biology*, 73(1):161—197, July 2016.

[Eth11] A. Etheridge. *Some Mathematical Models from Population Genetics: École D'Été de Probabilités de Saint-Flour XXXIX-2009.* Lecture Notes in Mathematics. Springer, 2011.

[Fel51] W. Feller. Diffusion processes in genetics. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 227–246, 1951.

[Fel74] J. Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78(2):737–756, 1974.

[Fis30] R.A. Fisher. *The Genetical Theory of Natural Selection.* Oxford Univ. Press, 1930.

[Fis49] R. A. Fisher. *The Theory of Inbreeding.* Oliver and Boyd, 1949.

[Fis54] R. A. Fisher. A fuller theory of "junctions" in inbreeding. *Heredity*, 8:187–197, 1954.

[Fis59] R. A. Fisher. An algebraically exact examination of junction formation and transmission in parent-offspring inbreeding. *Heredity*, 13:179–186, 1959.

[FM04] N. Fournier and S. Méléard. A microscopic probabilistic description of a locally regulated population and macroscopic approximations. *Ann. Appl. Probab.*, 14(4):1880–1919, 11 2004.

[GAG00] S. Gavrilets, R. Acton, and J. Gravner. Dynamics of speciation and diversification in a metapopulation. *Evolution*, 54:1493–1501, 2000.

[Gal64] J. Gale. Some applications of the theory of junctions. *Biometrics*, pages 85–117, 1964.

[Gav97] S. Gavrilets. Evolution and speciation on holey adaptive landscapes. *Trends Ecol Evol.*, 12(8):307–312, 1997.

[Gav04] S. Gavrilets. *Fitness Landscapes and the Origin of Species*, volume 41. Princeton University Press, monographs in population biology edition, 2004.

[Gav05] S. Gavrilets. Adaptive speciation: it is not that simple. *Evolution*, 59:696–699, 2005.

[GG97] S. Gavrilets and J. Gravner. Percolation on the fitness hypercube and the evolution of reproductive isolation. *J Theor Biol.*, 184(1):51–64, 1997.

[GHA97] P.H. Gouyon, J.P. Henry, and J. Arnould. *Les avatars du gène : la théorie néodarwinienne de l'évolution.* Collection Regards sur la science, Belin, 1997.

[Gil01] N.W. Gillham. Evolution by jumps: Francis Galton and William Bateson and the mechanism of evolutionary change. *Genetics*, 159(4):1383–1392, 2001.

[GJL16] R. C. Griffiths, P. A. Jenkins, and S. Lessard. A coalescent dual process for a Wright-Fisher diffusion with recombination and its applications to haplotype partitioning. *Theor. Popul. Biol.*, 112:126–138, 2016.

[GLV98] S. Gavrilets, H. Li, and M.D. Vose. Rapid parapatric speciation on holey adaptive landscapes. *Proc. R. Soc. Lond. B*, 265:1483–1489, 1998.

[GLV00] S. Gavrilets, H. Li, and M.D. Vose. Patterns of parapatric speciation. *Evolution*, 54(4):1126–1134, 2000.

[GM97] R.C. Griffiths and P. Marjoram. An ancestral recombination graph. In P. Donnelly and S. Tavaré, editors, *Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications*, volume 87, pages 257–270. 1997.

[Gor31] M. Gordon. Hereditary basis of melanosis in hybrid fishes. *Amer. J. Cancer.*, 15:1495–1523, 1931.

[Gra81] V. Grant. *Plant speciation.* Columbia University Press, 1981.

[Gri81] R. C. Griffiths. Neutral two-locus multiple allele models with recombination. *Theor. Popul. Biol.*, 19(2):169–186, 1981.

[Gri91] R. C. Griffiths. The two-locus ancestral graph. In I.V. Basawa and R. L. Taylor, editors, *Selected Proceeedings of the Symposium on Applied Probability*, pages 100–117. Institute of Mathematical Statistics, 1991.

[HFQ+09] X. Huang, Q. Feng, Q. Qian, Q. Zhao, L. Wang, A. Wang, J. Guan, D. Fan, Q. Weng, T. Huang, G. Dong, T. Sang, and B.B. Han. High-throughput genotyping by whole-genome resequencing. *Genome Research*, 19(6):1068–1076, 2009.

[HSJ15] T. B. Hashimoto, Y. Sun, and T. S. Jaakkola. From random walks to distances on unweighted graphs. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 3429–3437, Cambridge, MA, USA, 2015. MIT Press.

[Hud83] R.R. Hudson. Properties of the neutral model with intragenic recombination. *Theor.Pop.Biol.*, 23(2):213–201, 1983.

[Iva00] O. Ivanciuc. Qsar and qspr molecular descriptors computed from the resistance distance and electrical conductance matrices. *ACH Models in Chemistry*, 5/6(137):607–632, 2000.

[JNT18] T. Janzen, W. Nolte, A, and A. Traulsen. The breakdown of genomic ancestry blocks in hybrid lineages given a finite number of recombination sites. *Evolution*, 72(4), 2018.

[Kal02] O. Kallenberg. *Foundations of modern probability.* Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[Kar82] S. Karlin. Classification of selection-migration structures and conditions for a protected polymorphism. *Evol. Biol.*, pages 61–204, 1982.

[KDS+07] S. Koblmüller, N. Duftner, K. M. Sefc, M. Aibara, M. Stipacek, M. Blanc, B. Egger, and C. Sturmbauer. Reticulate phylogeny of gastropod-shell-breeding cichlids from lake tanganyika–the result of repeated introgressive hybridization. *BMC Evolutionary Biology*, 7(1):7, 2007.

[Kim53] M. Kimura. Stepping-stone model of population. *Annual Report of the National Institute of Genetics*, 3:62–63, 1953.

[Kim68] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217, 1968.

[Kin82] J.F.C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

[KO69] M. Kimura and T. Ohta. The average number of generations until extinction of an individual mutant gene in finite population. *Genetics*, 3(63), 1969.

[KR93] D. Klein and M. Randic. Resistance distance. *Journal of Mathematical Chemistry*, 12:81 – 95, 1993.

[KS13] C. Kleibler and J. Stoyanov. Multivariate distributions and the moment problem. *Journal of Multivariate Analysis*, 113:7–18, 2013.

[KWG+13] I. Keller, C.E. Wagner, L. Greuter, S. Mwaiko, O.M. Selz, A. Sivasundar, S. Wittwer, and O. Seehausen. Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of lake victoria cichlid fishes. *Molecular Ecology*, 22(11):2848–2863, 2013.

[Lam05] A. Lambert. The branching process with logistic growth. *Ann. Appl. Prob.*, 15:1506–1535, 2005.

[LNK03] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. *International Conference on Information and Knowledge Management (CIKM)*, pages 556–559, 2003.

[LS08] I. Lobo and K. Shaw. Thomas Hunt Morgan, genetic recombination, and gene mapping. *Nature Education*, 1(1):205, 2008.

[M99] M. Mohle. The concept of duality and applications to markov processes arising in neutral population sgenetics models. *Bernoulli*, 5(5):761–777, 1999.

[M01] M. Mohle. Forward and backward diffusion approximations for haploid exchangeable population models. *Stochastic Processes and their Applications*, 95(1):133–149, 2001.

[Mal48] G. Malécot. *Les mathématiques de l'hérédité*. Barnéoud frères, 1948.

[Mas12] J. P. Masly. 170 years of "lock-and-key": Genital morphology and reproductive isolation. *Int. J. Evol. Biol.*, 2012(247352), 2012.

[May42] E. Mayr. *Systematics and the Origin of Species*. Columbia University Press, 1942.

[MH11] O. C. Martin and F. Hospital. Distribution of parental genome blocks in recombinant inbred lines. *Genetics*, 189(2):645–654, 2011.

[MHWS05] A.K. MacLeod, C.S. Haley, J.A. Woolliams, and P. Stam. Marker densities and the mapping of ancestral junctions. *Genetical research*, 85(01):69–79, 2005.

[Mik08] I. Miko. Gregor Mendel and the principles of inheritance. *Nature Education*, 1(1):134, 2008.

[MLAR+08] R. McQuillan, A.-L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic, and L. et al. Barac-Lauc. Runs of homozygosity in european populations. *The American Journal of Human Genetics*, 83(3):359–372, 2008.

[MLL+16] P. Muir, S. Li, S. Lou, D. Wang, D. J. Spakowicz, L. Salichos, J. Zhang, G. M. Weinstock, F. Isaacs, J. Rozowsky, and M. Gerstein. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology*, 17(1):53, 2016.

[MPS17] V. Miró Pina and E. Schertzer. How does geographical distance translate into genetic distance? *arXiv:1703.00357*, 2017.

[MSB+06] J. Mavárez, C. A Salazar, E. Bermingham, C. Salcedo, C. D. Jiggins, and M. Linares. Speciation by hybridization in heliconius butterflies. *Nature*, 441(7095):868–871, 2006.

[Mul42] H.J. Muller. Isolating mechanisms, evolution and temperature. *Biol. Symp.*, 811:71–125, 1942.

[Mul64] H.J. Muller. The relation of recombination to mutational advance. *Mutat. Res.*, 106:2–9, 1964.

[NFST05] A.W. Nolte, J. Freyhof, K.C. Stemshorn, and D. Tautz. An invasive lineage of sculpins, cottus sp. (pisces, teleostei) in the rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. *Proceedings of the Royal Society B*, 272:2379–2387, Oct 2005.

[NMW83] M. Nei, T. Maruyama, and C.I. Wu. Models of evolution of reproductive isolation. *Genetics*, 103:557–579, 1983.

[Noe97] A.J. Noest. Instability of the sexual continuum. *Proc. R. Soc. Lond. B*, 264:1389–1393, 1997.

[O'C08] C. O'Connor. Isolating hereditary material: Frederick Griffith, Oswald Avery, Alfred Hershey, and Martha Chase. *Nature Education*, 1(1):105, 2008.

[OG06] S.P. Otto and A.C. Gerstein. Why have sex? the population genetics of sex and recombination. *Biochemical Society Transactions*, 34(4), 2006.

[OM08] C. O'Connor and I. Miko. Developing the chromosome theory. *Nature Education*, 1(1):44, 2008.

[Orr95] H.A. Orr. The population genetics of speciation: The evolution of hybrid incompatibilities. *Genetics*, 139:1805–1815, 1995.

[PHN10] J. E. Pool, Jensen J. D. Hellmann, I., and R. Nielsen. ?population genetic inference from genomic sequence variation. *Genome Res.*, 20(3):291–300, 2010.

[PMN10] E. E. Patton, D. L. Mitchell, and R. S. Nairn. Genetic and environmental melanoma models in fish. *Pigment Cell and Melanoma Research*, 23(3):314–337, 2010.

[Pra08] L. Pray. Discovery of DNA structure and function: Watson and Crick. *Nature Education*, 1(1):100, 2008.

[QH05] H. Qiu and E.R. Hancock. Image segmentation using commute times. *Proceedings of the 16th British Machine Vision Conference (BMVC)*, pages 929–938, 2005.

[QHZ+14] Z. Qi, L. Huang, R. Zhu, D. Xin, C. Liu, X. Han, H. Jiang, W. Hong, G. Hu, H. Zheng, and Q. Chen. A high-density genetic map for soybean based on specific length amplified fragment sequencing. *PLoS ONE*, 9(8):e104871, 2014.

[RC13] P. Ralph and G. Coop. The geography of recent genetic ancestry across Europe. *PLoS biology*, 11:e1001555, 2013.

[RCB17] H. Ringbauer, G. Coop, and N.H. Barton. Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205(3):1335–1351, 2017.

[Roy04] K. Roy. Topological descriptors in drug design and modeling studies. *Molecular Diversity*, 4(8):321–323, 2004.

[SFC+06] T. Singer, Y. Fan, H.-S. Chang, T. Zhu, S. P. Hazen, and S. P. Briggs. A high-resolution map of arabidopsis recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genetics*, 2(9):e144, 2006.

[Sla08] M. Slatkin. Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews. Genetics*, 9(6):477–485, 2008.

[SMSBM05] D. Schwarz, B. M. Matta, N. L. Shakir-Botteri, and B. A. McPheron. Host shift to an invasive plant triggers rapid animal hybrid speciation. *Nature*, 436(7050):546–549, 2005.

[SRH+02] P.C. Sabeti, D.E. Reich, J.M. Higgins, H.Z. Levine, D.J. Richter, S.F. Schaffner, S.B. Gabriel, J.V. Platko, N.J. Patterson, G.J. McDonald, H.C. Ackerman, S.J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E.S. Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419:832–837, 2002.

[ST50] J.A. Shohat and J.D. Tamarkin. *The Problem of Moments*. American Mathematical Society, revised edition, 1950.

[Sta80] P. Stam. The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genetical Research*, 35(2):131–155, 1980.

[TAK+13] M.J. Truco, H. Ashrafi, A. Kozik, H. van Leeuwen, J. Bowers, S. R. C. Wo, K. Stoffel, H. Xu, T. Hill, A. Van Deynze, and R. W. Michelmore. An ultra-high-density, transcript-based, genetic map of lettuce. *G3: Genes| Genomes| Genetics*, 3(4):617–631, 2013.

[TEPB17] H. Teotónio, S. Estes, P. C. Phillips, and C. F. Baer. Experimental evolution with caenorhabditis nematodes. *Genetics*, 2(206):691–716, 2017.

[VLRH14] U. Von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large random neighborhood graphs. *Journal of Machine Learning Research*, 15(1):1751–1798, 2014.

[VS11] V. D. Vacquier and W. J. Swanson. Selection in the rapid evolution of gamete recognition proteins in marine invertebrates. *Cold Spring Harbor Perspectives in Biology*, 3(11), 2011.

[Wak16] J. Wakeley. *Coalescent Theory: An Introduction*. Macmillan Learning, 2016.

[WH97] C. Wiuf and H. Hein. On the number of ancestor to a DNA sequence. *Genetics*, 147:1459–1468, 1997.

[Wri32] S. Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress on Genetics*, pages 355–366, 1932.

[Wri43] S. Wright. Isolation by distance. *Genetics*, 28(2):114, 1943.

[WvLK+06] M.A.L. West, H. van Leeuwen, A. Kozik, D. J. Kliebenstein, R.W. Doerge, D. A. St. Clair, and R. W. Michelmore. High-density haplotyping with microarray-based expression and single feature polymorphism markers in arabidopsis. *Genome Research*, 16(6):787–795, 2006.

[XFY+10] W. Xie, Q. Feng, H. Yu, X. Huang, Q. Zhao, Y. Xing, S. Yu, B. Han, and Q. Zhang. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proceedings of the National Academy of Sciences*, 107(23):10578–10583, 2010.

[YI13] R. Yamaguchi and Y. Iwasa. First passage time to allopatric speciation. *Interface Focus*, 3(6), 2013.

[YI15] R. Yamaguchi and Y. Iwasa. Smallness of the number of incompatibility loci can facilitate parapatric speciation. *Journal of Theoretical Biology*, 405:36–45, 2015.

[YVW+05] L. Yen, D. Vanvyve, F. Wouters, F. Fouss, M. Verleysen, and M. Saerens. Clustering using a random walk based distance measure. *In Proceedings of the 13th Symposium on Artificial Neural Networks (ESANN)*, pages 317–324, 2005.