Population genetic models with selection, fluctuating environments and population structure

Dissertation

zur Erlangung des Doktorgrades
an der Fakultät für Mathematik und Physik der

Albert-Ludwigs-Universität

Freiburg im Breisgau

vorgelegt von

Elisabeth Sester-Huss

Januar 2020

Dekan:	Prof. Dr. Wolfgang Soergel
Referent:	Prof. Dr. Peter Pfaffelhuber
Koreferent:	Prof. Dr. Martin Hutzenthaler

Tag der mündlichen Prüfung: 2. März 2020

Abstract

This thesis consists of three parts each dealing with different questions related to population genetics. We start with the study of the effect of natural selection on genealogies. We make use of the theory on tree-valued Fleming-Viot processes that describe the evolution of genealogical trees to compute the Laplace-transform of the tree length both in the neutral and in the selective setting. We show that trees are shorter in the selective case (under the so-called *Laplace-transform-order*) than trees under neutrality – an assumption already widely believed to be true in the field of biology. In the second part we work with a mutation-selection model in a fluctuating environment by introducing a modifier locus determining the mutation rate at a second locus. Fitness acts on the second locus and changes as the environment fluctuates. For a fast flucttutating environment, we obtain limit results for the evolution of allele frequencies and apply them to a two-type setting in which we compute the fixation probability for the higher mutation rate. The last part focuses on analysing human DNA samples and estimating their heritage. The aim is to extend the already existing models for inferring individual admixture proportions – a vector of which each entry corresponds to the fraction of one's genome originating from a certain population. We develop a method that delivers individual admixture proportions of an individual's parents. This enables us to test whether the admixture of two populations has occured only recently or several generations ago. We apply both the already existing method and our new method to the 1000 genomes dataset and test the accuracy of their outputs by computing the distance to their true admixture proportions.

Contents

Introduction

In population genetics, one aims to understand phenomena that are observed in the genetic structure of populations. Factors affecting genetic variation between either genes, several individuals or even whole populations are among others natural selection, mutation and migration. In Chapter 1 and 2 we will be working with a special type of Markov processes, called *Fleming-Viot processes*. We will describe our desired dynamics in terms of generators and obtain the corresponding processes as solutions of a martingale problem.

In Chapter 1 we will rely on the *tree-valued Fleming-Viot process with mutation and selection (TFVMS)* introduced in Depperschmidt et al. (2012). As the name suggests the TFVMS describes the evolution of genealogical trees in a population affected by mutational and selective events. As an application, Depperschmidt et al. (2012) manage to quantitatively analyse a widely held assumption in the field of biology, namely that genealogical distances are shorter in the presence of selection. The heuristics are the following: In a population with different types that are unequally fit, there will be at least one type that will be favored by nature. The consequence is that this fitter type reproduces faster and therefore genealogical distances get shorter because randomly picked individuals now have a more recent common ancestor due to the quick reproduction of the fitter type. Depperschmidt et al. (2012) assume a setting in which an individual carries an allele of either type $\bullet$ or type ${\color{gray}\bullet}$ where $\bullet$ is chosen to be the fitter one. Their machinery then allows them to compute the Laplace-transform of pairwise genealogical distances in the selective case for small selection coefficients α and compare them to distances in the neutral case with no selection. In this two-type setting pairwise distances are in fact shorter under selection in the so-called *Laplace-transform-order*. We will make use of the findings of Depperschmidt et al. (2012) and compare genealogical trees spanned by $n \geq 2$ individuals under selection to those under neutrality. We find that, again, tree lengths tend to be shorter in the selective case.

In Chapter 2 we continue with another biologically motivated subject. Phenomena witnessed in nature are complex and very much intertwined such that when modeling certain processes, we always need to make assumptions that might not completely reflect the observed but that allow us to focus on these very processes that we wish to analyse in more detail. For instance, in Chapter 1 we assume that we have a population in which mutation rates and selection coefficients are fixed. The assumption is that the environment does not change and therefore we have one type that has a fitness advantage – a condition which does not change in the course of time. In reality, this is obviously not the case. The environment is a major factor that influences the fitness of different types. Type $\bullet$ might be fitter in one environment but it might be less fit in another. Therefore, depending on which environment we are in, it might be more beneficial to be the other type. In a world where the environment undergoes changes, fitness of an individual might not be mirrored only by the type of one specific allele but also by the ability to respond to these changes and to produce offspring that are better adapted to the new environment. In other words, during times of adaptation individuals with higher mutation rates are needed to produce mutations of which some have the required fitness to survive and establish themselves in the new surroundings. These individuals that have higher mutation rates are called *mutators* and have been gaining more and more attention especially in microbial evolution. Mutators are observed in many experiments where a bacterial population is exposed to some environment-changing mutagen. After the exposure the frequency of the mutators visibly increases before decreasing again once they have produced a type that is well-adapted to the new environment and that has taken over the population. Chapter 2 focuses on this exact process of changing evironments and the resulting changes in the fitness of individuals. We will obtain results on limiting processes in a fast fluctuating environment and fixation probabilities in the special case where we only consider two types.

In Chapter 1 and Chapter 2 we have studied population models by looking at results that can be observed after a long time period: The evolving genealogical trees and their lengths in equilibrium as well as the role of mutators in a fast fluctuating environment and their fate in form of fixation probabilities. The last chapter deals with more recent events. For an enormous amount of positions on the DNA, their functions have been identified today due to advances in DNA sequencing. Some parts of the DNA indicate population-dependent frequencies which are crucial when analyzing DNA samples with the objective of inferring the owner's genetic heritage. In Chapter 3 we are interested in the detecting of recent admixtures by investigating DNA samples of people with parents originating from two (genetically isolated) populations. We will introduce a model which we will call *recent-admixture model* and is specifically designed for detecting recent admixture between two populations. Given a DNA sample of an individual, we will obtain the individual admixture (IA) of both the parents. This is an extention of the already existing models, which we will refer to as *admixture models*, that give IA of the individual itself rather than those of the parents. With the help of the parents' IA, we obtain additional information on the history of admixture leading eventually to the IA of the child.

Notation

We give a short list with the notation that will be used primarily in Chapter 1 and 2.

1. For a complete and separable metric space, i.e. a Polish space, (E, r), we denote by

$\mathcal{M}(E)$	the space of measurable,
$\mathcal{B}(E)$	the space of bounded, measurable,
$\mathcal{C}_b(E)$	the space of bounded, continuous

   real-valued functions on E (equipped with convergence of uniform on compacta) and for $L > 0$ by

$\mathcal{C}_L(E)$	the space of bounded, real-valued functions with Lipschitz constant L.

   We denote by $\mathcal{P}(E)$ the space of probability measures on (the Borel sets of) E, equipped with the topology of weak convergence denoted by $\Rightarrow$.

2. For $A \subseteq \mathbb{R}$ (equipped with the Euclidean topology) we denote by $\mathcal{D}_E(A)$ the set of continuous càdlàg functions $A \to E$ (equipped with the Skorohod topology).

3. For product spaces $X \times Y \times ...$, we denote by $\pi_X, \pi_Y, ...$ the projection operators.

4. For a random variable Z with distribution ν we write $Z \sim \nu$.

5. For another Polish space (E', r'), a measure $\mu \in \mathcal{P}(E)$ and some $\varphi : E \to E'$, the image measure of μ under φ is denoted by $\varphi_* \mu$. Moreover for $f \in \mathcal{M}(E)$, we write

   $$\langle \mu, f \rangle := \int f(u)\mu(\mathrm{d}u), \tag{0.0.1}$$

   if the right-hand side exists.

6. We will be using the *Landau symbol* $\mathcal{O}(\cdot)$.
   For functions g and h that depend on α, we write

   $$\begin{aligned} & g(\alpha) = h(\alpha) + \mathcal{O}(\alpha^n) \qquad \text{as } \alpha \to 0 \\ \text{if} \qquad & \limsup_{\alpha \to 0} |(g(\alpha) - h(\alpha))/\alpha^n| < \infty. \end{aligned}$$

Chapter 1

Genealogical distances under low levels of selection

As mentioned in the Introduction, the goal of this chapter is to compute the tree length of genealogies under weak selection. In a population where individuals carry different types of alleles it seems reasonable to assume that beneficial alleles spread quicker due to their fitness advantage and therefore genealogical distances between randomly chosen individuals get shorter as the fitter alleles spread. In spite of its rather simple reasoning this widely held assumption is in fact quite difficult to prove.

1.1 Introduction

Classical works on population genetic models focus on describing allelic frequencies by diffusions in the large population limit. In the early 1980s, the study of genealogical trees started to gain more and more attention. Coalescent theory can for instance be used to produce theoretical genealogies in order to compare them to observed data. The earliest contributions on this topic can be ascribed to Kingman (1982). The analysis of genealogical trees is carried out by looking at coalescents backward in time. Starting with the sample of interest, the tree is traced back until enough coalescence events have occurred and we find ourselves at the most recent common ancestor. The coalescent describing the genealogy of a sample in a population under neutrality, i.e. no selection, is called the *Kingman's coalescent.* The study of coalescent theory involves the incorporation of other dynamics that shape the population structure such as mutation or recombination and so on. The object of our interest in the following chapter is the analysis of selection and its effect on genealogical trees, in particular the effect selection has on the tree length. While the mutation process is independent of the genealogical tree and hence, mutational events can be added quite simply by superposing a Poisson process onto the tree, selection affects the coalescent in a more complex way. We speak of selection whenever we observe specific traits that come in different types, that is, we have multiple alleles of which some are fitter than others. A lot of research has been done on this topic (see e.g. Wakeley, 2010 for a review). Krone and Neuhauser (1997) and Neuhauser and Krone (1997) introduce the *Ancestral Selection Graph (ASG)* describing genealogical trees under selection. When looking backward in time on the ASG we encounter splitting events indicating possible ancestry. Beneficial alleles tend to have more offspring meaning that they are more likely to be the true ancestor. Therefore, in order to find the true ancestor we need to additionally go forward in time to fix the types so one can decide which of the possible ancestors is true.

Both the Kingman's coalescent and the ASG focus on fixing one specific time point and analyzing the resulting genealogical tree. Depperschmidt et al. (2012) use a different approach in which instead of looking at a specific genealogical tree that captures allele frequencies at some fixed time they are interested in the evolution of such trees. As time passes the population evolves and with that the relationships of individuals within that population change. They introduce a tree-valued Markov process describing this evolution of genealogies under mutation, selection and random reproduction under stochastically evolving random type distributions and geanealogical distances and call it the *tree-valued Fleming-Viot process with mutation and selection (TFVMS)*. In order to investigate genealogical trees and the distances of individuals, the trees are interpreted as metric spaces where the metric corresponds to the genealogical distances. Furthermore, as we are interested in incorporating the effects of selection we need to implement a set of types – we will call this set I – that *mark* the individuals. These metric spaces are finally equipped with a probability measure allowing us to randomly sample points from our tree turning them to *marked metric measure spaces (mmm-spaces)* which is an extention of the *metric measure spaces (mm-spaces)*, the state space of the tree-valued Fleming-Viot process under neutrality analysed in Greven et al. (2008). In the setting of bi-allelic mutation and low levels of selection, they use generator calculations to obtain the second order Laplace-transform of genealogical distances for small selection coefficients α, generalising the results on the first order Laplace-transform of Krone and Neuhauser. Their results reaffirm the ones from Krone and Neuhauser: The pairwise genealogical distance under selection is in fact shorter compared to the neutral case.

We will apply the same machinery to extend this result to the total tree length spanned by a sample of $n \geq 2$ individuals again for small selection coefficients α.

The chapter is structured as follows: In Section 1.2, we introduce the model we are going to study, i.e. genealogies in the large population limit for a Moran model under additive selection and we give some definitions needed to define the TFVMS given in Section 1.3. Theorem 1.15 and Corollary 1.20 are the main results of Section 1.4 and give the Laplace-transform of the genealogical tree and the expected tree length, respectively. Next in Section 1.5, we extend the results of Section 1.4 to other modes of dominance and again give the Laplace-transform and the expectation of tree lengths in this new case (Theorem 1.26 and Corollary 1.28). Section 1.6 contains all proofs.

This chapter is joint work with Peter Pfaffelhuber and all results have been published in Huss and Pfaffelhuber (2019).

1.2 Graphical description

Before we give formal definitions we will look at a graphical representation of the so-called *tree-valued Moran model with mutation and selection (TMMMS)*, a model capturing the effects of resampling, mutation and selection in a population with finitely many individuals. Detailed descriprions of this model can be found in Depperschmidt et al. (2012). Figure 1.1 illustrates the evolution of genealogical trees in the setting of a TMMMS with population size $N = 5$.

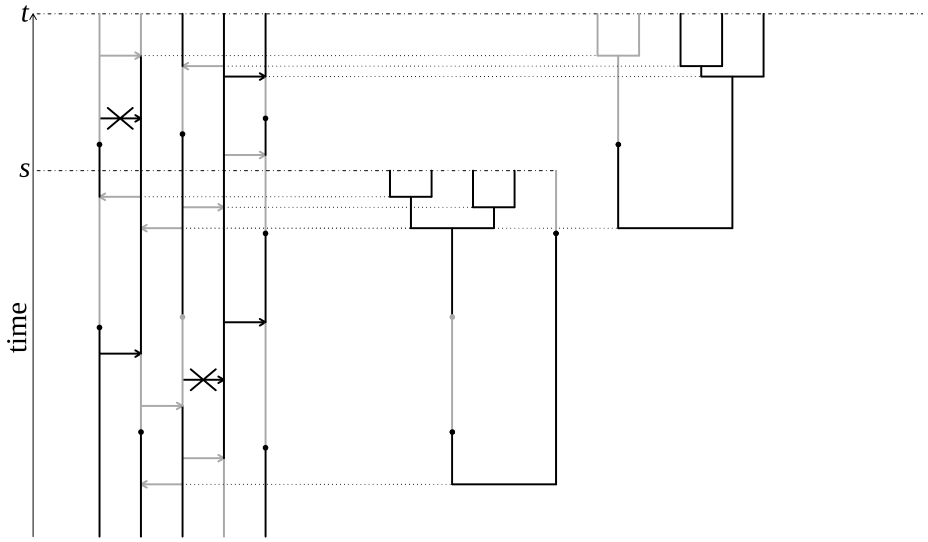


Figure 1.1: Graphical construction of a TMMMS with population size $N = 5$. In the left part of the figure one can see the dynamic interaction happening between the individuals. Bullet points indicate mutation events from the beneficial type • to • and and grey points indicate mutations in the other direction. Grey arrows denote neutral resampling events. Selective events are depicted as black arrows and only take effect if the arrows starts at the beneficial (black) type. Selective arrows starting with the deleterious type cannot be used and are indicated by a cross. On the right, genealogical trees for the population ($N = 5$) at times s and t are drawn.

Remark 1.1 (Dynamics of the Moran model)**.** Each line in the graphical respresentation stands for one of the N (haploid) individuals in the Moran model. The color of the line depicts the type of the individual. In our case, we have the type set $I = \{\bullet, \bullet\}$ where • is advantageous with selection coefficient α. Mutation events happen in both directions. The dynamics are the following:

1. Every pair of individuals resamples at rate γ; upon such a resampling event, one of the two individuals involved dies, the other one reproduces.
2. Every line is hit by a mutation event from • to • at a rate $\vartheta_{\bullet}/2 > 0$, and by a mutation event from • to • at a rate $\vartheta_{\bullet}/2 > 0$.
3. Every line of type • places an offspring on a randomly chosen line at a rate α.

The evolution of trees mentioned in the introduction is captured comprehensively in Figure 1.1. On the left, the relationships within the population change and these changes are visible in the genealogical trees evolving in time illustrated on the right.

The limit $N \to \infty$ gives us the *tree-valued Fleming-Viot process with mutation and selection* (TFVMS) (see Depperschmidt et al., 2012 for a more detailed description).

1.3 TFVMS

In Theorem 1.8 we will obtain the TFVMS as a solution of a so-called *martingale problem.*

Definition 1.2 (Martingale Problem)**.** For some complete and separable metric space (E, r), some linear $G : \mathcal{D}(G) \subseteq \mathcal{C}_b(E) \to \mathcal{C}_b(E)$ and $\mu \in \mathcal{P}(E)$, we say that an E-valued process X solves the $(G, \mathcal{D}(G), \mu)$-*martingale problem* if $X_0 \sim \mu$ and

$$\left(f(X_t) - \int_0^t Gf(X_s)\mathrm{d}s\right)_{t\geq 0}$$

is a martingale for every $f \in \mathcal{D}(G)$. We say that the $(G, \mathcal{D}(G), \mu)$-martingale problem is *well-posed* if there is a unique (in law) process X which solves this martingale problem. We call $\mathcal{D}(G)$ the *domain of* G.

First of all, we will look at the state space of the TFVMS. The state space of this process is called the *I-marked metric measure spaces (mmm-spaces)* where I denotes the set of types. As already mentioned, the metric spaces will represent our genealogical trees where the metric relates to the genealogical distances of individuals. For any Polish space I we define the *marked metric measure space* as follows:

Definition 1.3 (mmm-space)**.** (1) An *I-marked metric measure space*, *I-mmm-space* or *mmm-space*, for short, is a triple (X, r, μ) such that (X, r) is a complete and seperable metric space and $\mu \in \mathcal{P}(X \times I)$. Without loss of generality we assume that $X \subseteq \mathbb{R}$.

(2) An mmm-space (X, r, μ) is called *compact* if $(\mathrm{supp}((\pi_X)_*\mu), r)$ is compact. It is called *ultrametric* if $(\mathrm{supp}((\pi_X)_*\mu), r)$ is ultrametric.

(3) Two mmm-spaces (X, r_X, μ_X) and (Y, r_Y, μ_Y) are *measure-preserving isometric* and I-preserving (or *equivalent*), if there exists a measurable map $\varphi : X \to Y$ such that $r_X(x, x') = r_Y(\varphi(x), \varphi(x'))$ for all $x, x' \in \mathrm{supp}((\pi_X)_*\mu_X)$ and $\widetilde{\varphi}_*\mu_X = \mu_Y$ for $\widetilde{\varphi}(x, u) = (\varphi(x), u)$. The equivalence class of an mmm-space (X, r, μ) is denoted by $\overline{(X, r, \mu)}$.

(4) We define

$$\mathbb{M}^I := \{\overline{(X, r, \mu)} : (X, r, \mu) \text{ mmm-space}\}.$$

Moreover

$$\begin{aligned}
\mathbb{M}^I_c &:= \{\overline{(X, r, \mu)} : (X, r, \mu) \text{ compact mmm-space}\},\\
\mathbb{U}^I &:= \{\overline{(X, r, \mu)} : (X, r, \mu) \text{ ultrametric mmm-space}\},\\
\mathbb{U}^I_c &:= \mathbb{M}^I_c \cap \mathbb{U}^I.
\end{aligned}$$

Generic elements of $\mathbb{M}^I$ $(\mathbb{U}^I)$ are denoted by $\mathfrak{x}, \mathfrak{y}, \ldots$ $(\mathfrak{u}, \ldots)$.

Definition 1.4 (Marked distance matrix distribution)**.** Let (X, r, μ) be an mmm-space, $\mathfrak{x} := \overline{(X, r, \mu)} \in \mathbb{M}^I$ and

$$R^{(X,r)} : \begin{cases} (X \times I)^{\mathbb{N}} \to \mathbb{R}_+^{\binom{\mathbb{N}}{2}} \times I^{\mathbb{N}}, \\ ((x_i, u_i)_{i\geq 1}) \mapsto ((r(x_i, x_j))_{1\leq i<j}, (u_k)_{k\geq 1}). \end{cases} \tag{1.3.1}$$

The *marked distance matrix distribution of* χ is given by

$$\nu^{\chi} := (R^{(X,r)})_{*}\mu^{\mathbb{N}} \in \mathcal{P}(\mathbb{R}_+^{\binom{\mathbb{N}}{2}} \times I^{\mathbb{N}}).$$

A very common way of describing a Markov process is by defining what happens in infinitesimal time. The changes we observe are described by a *generator*:

Definition 1.5 (Generator)**.** Let $X = (X_t)_{t\geq 0}$ be a Markov process. The *generator* G^X of X is defined as

$$\left(G^X f\right)(x) := \lim_{t\to 0} \frac{\mathbb{E}_x[f(X_t) - f(x)]}{t} \tag{1.3.2}$$

for all f for which the right-hand side exists. The set of all functions f for which $(G^X f)(x)$ exists for all $x \in E$, is called the *domain* of G^X and is denoted by $\mathcal{D}(G^X)$.

In the case of TFVMS, the domain consists of so-called *polynomials.*

Definition 1.6 (Polynomials)**.**

(1) For any $n, k \in \mathbb{N}$ we denote by

$$\mathcal{B}_n := \mathcal{B}_n(\mathbb{R}_+^{\binom{\mathbb{N}}{2}} \times I^{\mathbb{N}}), \quad \bar{\mathcal{C}}_n := \mathcal{C}_n(\mathbb{R}_+^{\binom{\mathbb{N}}{2}} \times I^{\mathbb{N}}), \quad \bar{\mathcal{C}}_n^k := \bar{\mathcal{C}}_n^k(\mathbb{R}_+^{\binom{\mathbb{N}}{2}} \times I^{\mathbb{N}})$$

the sets of bounded measurable, continuous as well as continuous and k times continuously differentiable (with respect to all variables in $\mathbb{R}_+^{\binom{\mathbb{N}}{2}}$) functions ϕ on $\mathbb{R}_+^{\binom{\mathbb{N}}{2}} \times I^{\mathbb{N}}$, such that $(\underline{\underline{r}}, \underline{u}) \mapsto \phi(\underline{\underline{r}}, \underline{u})$ depends on the first $\binom{n}{2}$ variables in $\underline{\underline{r}}$ and the first n in $\underline{u}$ only. If $n = 0$, the spaces consist of constant functions.

(2) A function $\Phi : \mathbb{M}^I \to \mathbb{R}$ is a *polynomial* if for some $n \in \mathbb{N}$, there exists $\phi \in \mathcal{B}_n$ such that for all $\chi \in \mathbb{M}^I$,

$$\Phi(\chi) := \Phi^{n,\phi} = \langle \nu^{\chi}, \phi \rangle = \int \phi(\underline{\underline{r}}, \underline{u})\nu^{\chi}(\mathrm{d}\underline{\underline{r}}, \mathrm{d}\underline{u}). \tag{1.3.3}$$

(Recalling the notation given in (0.0.1).)

(3) The *degree* of a polynomial Φ is the smallest number n for which there exists $\phi \in \mathcal{B}_n$ such that (1.3.3) holds.

(4) Writing $\bar{\mathcal{C}}_n^0 := \bar{\mathcal{C}}_n$, we set

$$\begin{aligned} \Pi &:= \bigcup_{n=0}^{\infty} \Pi_n, & \Pi_n &:= \{\Phi^{n,\phi} : \phi \in \mathcal{B}_n\}, & \\ \Pi^k &:= \bigcup_{n=0}^{\infty} \Pi_n^k, & \Pi_n^k &:= \{\Phi^{n,\phi} : \phi \in \bar{\mathcal{C}}_n^k\}, & k = 0, 1. \end{aligned} \tag{1.3.4}$$

In the following, we will define the generator G for the TFVMS. With Theorem 1.8 we will establish that the $(\mathbb{P}_0, G, \Pi_1)$-martingale is well-posed for $\mathrm{P}_0 \in \mathcal{M}_1(\mathbb{U}^I)$.

There are four types of dynamics that influence the tree-valued process: the growth of the tree as time passes, resampling, mutation and selection events. Hence, the generator of the TFVMS is of the form

$$G := G^{\text{grow}} + G^{\text{res}} + G^{\text{mut}} + G^{\text{sel}} \tag{1.3.5}$$

where the operators G^{grow}, G^{res}, G^{mut} and G^{sel} describe the above dynamics. For $\Phi := \Phi^{n,\phi} \in \Pi_n^1$ the different terms are given as follows:

(A) *Growth operator:* During times when neither resampling nor mutation nor selection events happen, the tree grows deterministically. Looking at 2 individuals their tree grows with speed 2 (as the tree length corresponds to the 2 lines in the graphical representation). The change we observe is thus given by

$$G^{\mathrm{grow}}\Phi(u) := \langle \nu^u, \langle \nabla_{\underline{\underline{r}}}\phi, \underline{\underline{2}}\rangle\rangle$$

with

$$\langle \nabla_{\underline{\underline{r}}}\phi, \underline{\underline{2}}\rangle = 2\sum_{1\le i<j}\frac{\partial\phi}{\partial r_{ij}}(\underline{\underline{r}}, \underline{u}).$$

(B) *Resampling operator:* At rate γ, a resampling event occurs and replaces every unordered pair $k \neq l$ meaning that l is replaced by an offspring of k, or k is replaced by an offspring of l. The probability of either one is $\frac{1}{2}$. Therefore one can say that for each ordered pair $k \neq l$, l is replaced by an offspring of k at rate $\frac{\gamma}{2}$. The *resampling operator* is given by

$$G^{\mathrm{res}}\Phi(u) := \frac{\gamma}{2}\sum_{k,l=1}^{n}\langle \nu^u, \phi\circ\theta_{k,l} - \phi\rangle$$

with $\theta_{k,l}(\underline{\underline{r}}, \underline{u}) = (\underline{\underline{\widetilde{r}}}, \hat{\theta}_{k,l}(\underline{u}))$ and

$$\widetilde{r}_{ij} := \begin{cases} r_{ij}, & \text{if } i,j \neq l,\\ r_{i\wedge k, i\vee k}, & \text{if } j = l,\\ r_{j\wedge k, j\vee k}, & \text{if } i = l \end{cases}$$

and the *replacement operator* $\hat{\theta}_{k,l}$ is the map which replaces the l-th component of an infinite sequence by the k-th; that is, for $\underline{u} = (u_1, u_2, \ldots)$,

$$\begin{aligned} \hat{\theta}_{k,l}(\underline{u}) &:= \underline{u}_l^{u_k},\\ \underline{u}_l^{v} &:= (u_1, \ldots, u_{l-1}, v, u_{l+1}, \ldots). \end{aligned}$$

(C) *Mutation operator:* In our setting an individual of type $\bullet$ mutates to ${\color{gray}\bullet}$ at a rate $\vartheta_\bullet/2$ and from ${\color{gray}\bullet}$ to $\bullet$ at a rate $\vartheta_{\color{gray}\bullet}/2$. The mutation stochastic kernel $\beta(\cdot,\cdot)$ on I is then given by

$$\frac{\bar{\vartheta}}{2}\cdot\beta(u, \mathrm{d}v) = \frac{\vartheta_\bullet}{2}\mathbb{1}_{\{v={\color{gray}\bullet}\}} + \frac{\vartheta_{\color{gray}\bullet}}{2}\mathbb{1}_{\{v=\bullet\}} \tag{1.3.6}$$

with mutation rate $\bar{\vartheta} := \vartheta_\bullet + \vartheta_{\color{gray}\bullet}$ and we get

$$G^{\mathrm{mut}}\Phi(u) := \frac{\bar{\vartheta}}{2}\sum_{k=1}^{n}\langle \nu^u, \beta_k\phi - \phi\rangle,$$

such that

$$(\beta_k\phi)(\underline{\underline{r}}, \underline{u}) := \int \phi(\underline{\underline{r}}, \underline{u}_k^v)\beta(u_k, \mathrm{d}v).$$

(D) *Selection operator*: The impact of selection on the genealogical tree depends on the type carried by an individual. For the definition of the selection operator we therefore need to introduce a so-called *fitness function* defined on the set of types, i.e. a function of the form $\chi : I \to [0,1]$. We only have two types, $\bullet$ and ${\color{gray}\bullet}$, where $\bullet$ is the fitter type meaning that selection only occurs when the individual is of type $\bullet$. The fitness function is thus given by $\chi(u) = \mathbb{1}_{\{u=\bullet\}}$. We recall here the graphical representation given in Figure 1.1 where the selective arrows (black arrows) are used only when they start from a black line, hence from an individual of type $\bullet$. We say that the k-th individual has fitness $\alpha\chi_k := \alpha\chi(u_k)$, and χ is the fitness function. As for resampling, selective events occur in a pair of one individual k giving birth and the other, individual l, dying, as given through the function $\theta_{k,l}$ from (B).

$$G^{\text{sel}}\Phi(u) := \alpha \sum_{l<k}^{\infty} \langle \nu^u, \chi_k \left(\phi \circ \theta_{k,l} - \phi\right)\rangle$$

where we can ignore the summands with $l > n$ as ϕ only depends on the first n individuals leading to

$$= \alpha \sum_{\substack{l<k \\ l\leq n}} \langle \nu^u, \chi_k \left(\phi \circ \theta_{k,l} - \phi\right)\rangle$$

where the summands with $k \leq n$ only give a negligible effect and hence only summands with $k > n$ are of interest. Without loss of generality we choose $k = n+1$ and obtain

$$\approx \alpha \sum_{l=1}^{n} \langle \nu^u, \chi_{n+1} \left(\phi \circ \theta_{n+1,l} - \phi\right)\rangle$$

which gives by permuting sampling order of l and $n+1$ in the first term

$$= \alpha \sum_{l=1}^{n} \langle \nu^u, \chi_l \cdot \phi - \chi_{n+1} \cdot \phi\rangle. \tag{1.3.7}$$

Definition 1.7 (Generator of TFVMS)**.** Let G^{grow}, G^{res}, G^{mut} and G^{sel} given as in (A)-(D). The generator of TFVMS is the linear operator on Π with domain Π^1, given by

$$G := G^{\text{grow}} + G^{\text{res}} + G^{\text{mut}} + G^{\text{sel}}. \tag{1.3.8}$$

The next two theorems are results from Depperschmidt et al. (2012), Theorem 1, Theorem 4 and Lemma 8.1.

Theorem 1.8 (Martingale problem is well-posed)**.** *Let* $\mathrm{P}_0 \in \mathcal{M}_1(\mathbb{U}^I)$*,* Π^1 *be as in (1.3.4) and* G *as in (1.3.8).*

(1) The (P_0, G, Π^1)*-martingale problem is well-posed. The unique solution* $\mathcal{U} := (\mathcal{U}_t)_{t\geq 0}$ *is called the tree-valued Fleming-Viot process with mutation and selection (TFVMS).*

(2) $\mathcal{U}$ *has the following properties:*

(a) $\mathrm{P}(t \mapsto \mathcal{U}_t$ *is continuous*$) = 1$,

(b) $\mathrm{P}(\mathcal{U}_t \in \mathbb{U}_c^I$ *for all* $t > 0) = 1$,

(c) $\mathcal{U}$ *is strong Markov.*

Theorem 1.9 (The long-time behaviour and continuity of TFVMS)**.** *Let* $\mathcal{U} = (\mathcal{U}_t)_{t\geq 0}$ *be the TFVMS from Theorem 1.8 with* $\mathcal{U}_0 = u$. *Then*

(a) there exists an $\mathbb{U}_c^I$*-valued random variable* $\mathcal{U}_\infty^\alpha$ *with*

$$\mathcal{U}_t \xRightarrow{t\to\infty} \mathcal{U}_\infty^\alpha. \tag{1.3.9}$$

(b) The law of $\mathcal{U}_\infty^\alpha$ *is the unique invariant distribution of* $\mathcal{U}$. *It depends on all the model parameters but is independent of the initial state.*

(c) For $\Phi \in \Pi^1$,

$$\mathbb{E}\left[\Phi\left(\mathcal{U}_\infty^\alpha\right)\right] - \mathbb{E}[\Phi(\mathcal{U}_\infty^0)] = \mathcal{O}(\alpha) \quad \textit{as } \alpha \to 0. \tag{1.3.10}$$

Remark 1.10 (Other modes of dominance)**.** In Section 1.5 we will be treating a more general case leading to a selection operator denoted by $G^{\mathrm{sel},h}$ instead of G^{sel} where h is called the *dominance coefficient.* The results from Depperschmidt et al. (2012) are in fact proved for this general case and, hence, Theorem 1.8 and Theorem 1.9 also hold in Section 1.5. We will denote the limiting process in the case of other modes of dominiance by $\mathcal{U}_\infty^{\alpha,h}$.

Remark 1.11 (Notation)**.** In order to simplify the notation and still be able to distinguish between the computation of expectations unter selection and neutrality, we write $\mathbb{P}^\alpha(\cdot)$ for the distribution of TFVMS under the selection coefficient α and $\mathbb{P}^0(\cdot)$ for the neutral case, respectively. $\mathbb{E}^\alpha[\cdot]$ and $\mathbb{E}^0[\cdot]$ will denote the corresponding expectations. More precisely,

$$\mathbb{E}^\alpha\left[\Phi\right] := \mathbb{E}[\Phi(\mathcal{U}_\infty^\alpha)] \quad \text{and} \quad \mathbb{E}^0\left[\Phi\right] := \mathbb{E}[\Phi(\mathcal{U}_\infty^0)]. \tag{1.3.11}$$

1.4 Main results

In this section, we finally get to study the object of our interest: the length of genealogical trees. For this we recall some parameters and define an additional Θ.

Remark 1.12 (Relevant parameters)**.** Let

$$\alpha, \gamma, h, \vartheta_\bullet, \vartheta_{\color{gray}\bullet} \geq 0, \quad \text{and} \quad \bar{\vartheta} = \vartheta_\bullet + \vartheta_{\color{gray}\bullet}, \quad \Theta := \frac{\vartheta_{\color{gray}\bullet}}{\bar{\vartheta}}$$

where h is the dominance coefficient, a parameter appearing in results stated in Section 1.5 when handling other modes of dominance.

Up to this point, the form of the functions $\Phi \in \Pi^1$ is kept quite general. The only premise we make is the fact that the functions solely depend on a finite sample. To compute the Laplace-transforms of the length of genealogical trees we need to choose our functions appropriately. Let us sample $n+j$ points from the ultra-metric tree u with $n, j \geq 0$. For some $0 \leq i \leq n$ and $\lambda \geq 0$ we define the function ϕ_{ij}^n

$$\phi_{ij}^n(\underline{\underline{r}}, \underline{u}) := e^{-\lambda L_n(\underline{\underline{r}})} \cdot \mathbb{1}_{\{u_1=\bullet\}} \cdots \mathbb{1}_{\{u_i=\bullet\}} \cdot \mathbb{1}_{\{u_{n+1}=\bullet\}} \cdots \mathbb{1}_{\{u_{n+j}=\bullet\}} \tag{1.4.1}$$

where L_n denotes the length of the genealogy spanned by n points. By definition, ϕ^n_{ij} is a function only dependent of $n+j$ points so we have $\Phi^n_{ij} := \Phi^{n+j,\phi^n_{ij}} \in \bar{\mathcal{C}}^1_{n+j}$. Taking the expectation of this function, i.e. $\mathbb{E}^\alpha[\Phi^n_{ij}]$ gives us the Laplace-transform of the subtree length spanned by n of the total $n+j$ sampled points where i points within and j points outside the subsample are of type $\bullet$. We recall that the subject of our interest is the Laplace-transform of the length of a tree spanned by n points, hence we wish to compute

$$\mathbb{E}^\alpha\left[\Phi^n_{00}\right] = \mathbb{E}\left[e^{-\lambda L_n(\underline{\underline{r}})}\right]. \tag{1.4.2}$$

Applying the generator G from (1.5.1) on Φ^n_{ij} gives the following:

(A) *Tree growth:* By simply deriving ϕ^n_{ij} with respect to $\underline{\underline{r}}$ gives

$$G^{\text{grow}}\Phi^n_{ij} = -n\lambda\Phi^n_{ij}\mathbb{1}_{\{n\geq 2\}}.$$

(B) *Resampling:* For resampling events we first define the sets $I = \{1,...,i\}$, $H = \{i+1,...,n\}$ and $J = \{n+1,...,n+j\}$. We need to distinguish between coalescence events

1.) among I at rate $\binom{i}{2}$;

2.) among $I \cup H$ with at most one partner within I at rate $i(n-i) + \binom{n-i}{2}$;

3.) with one partner within I and the second among J at rate ij;

4.) with one partner in H and the second among J at rate $(n-i)j$;

5.) and among J at rate $\binom{j}{2}$.

Only if two individuals among $I \cup H$ coalesce, n decreases. This gives

$$\begin{aligned} G^{\text{res}}\Phi^n_{ij} = \gamma\Bigg(&\binom{i}{2}\left(\Phi^{n-1}_{i-1,j} - \Phi^n_{ij}\right) + i(n-i)\left(\Phi^{n-1}_{ij} - \Phi^n_{ij}\right) + \binom{n-i}{2}\left(\Phi^{n-1}_{ij} - \Phi^n_{ij}\right) \\ &+ ij\left(\Phi^n_{i,j-1} - \Phi^n_{ij}\right) + (n-i)j\left(\Phi^n_{i+1,j-1} - \Phi^n_{ij}\right) + \binom{j}{2}\left(\Phi^n_{i,j-1} - \Phi^n_{ij}\right)\Bigg). \end{aligned}$$

(C) *Mutation:* Mutation events do not have any effects on the length of the tree but on the number of individuals of different types. Mutations from $\bullet$ to ${\color{gray}\bullet}$ happen at rate $\vartheta_\bullet/2$ and from ${\color{gray}\bullet}$ to $\bullet$ at rate $\vartheta_{\color{gray}\bullet}/2$, hence the effects we observe are

$$G^{\text{mut}}\Phi^n_{ij} = i\frac{1}{2}\left(\vartheta_{\color{gray}\bullet}\Phi^n_{i-1,j} - \bar{\vartheta}\Phi^n_{ij}\right) + j\frac{1}{2}\left(\vartheta_{\color{gray}\bullet}\Phi^n_{i,j-1} - \bar{\vartheta}\Phi^n_{ij}\right).$$

(D) *Selection:* For selection we have

$$\begin{aligned} G^{\text{sel}}\Phi^n_{i,j}(u) &= \alpha\sum_{k=1}^{n+j}\langle\nu^u, \chi_k\cdot\phi^n_{i,j} - \chi_{n+j+1}\cdot\phi^n_{i,j}\rangle \\ &= \alpha\left(\sum_{k=1}^{n+j}\langle\nu^u, \chi_k\cdot\phi^n_{i,j}\rangle - (n+j)\langle\nu^u, \chi_{n+j+1}\cdot\phi^n_{i,j}\rangle\right). \end{aligned}$$

For $1 \leq k \leq i$ we have

$$\chi_k \cdot \phi^n_{i,j} = \mathbb{1}_{\{u_k=\bullet\}} \cdot e^{-\lambda L_n} \cdot \mathbb{1}_{\{u_1=\cdots=u_k=\cdots=u_i=u_{n+1}=\cdots=u_{n+j}=\bullet\}} = \phi^n_{i,j}.$$

Similarly we get

$$\begin{aligned} \chi_k \cdot \phi^n_{i,j} &= \phi^n_{i+1,j} \quad \text{for } i+1 \leq k \leq n, \\ \chi_k \cdot \phi^n_{i,j} &= \phi^n_{i,j} \quad \text{for } n+1 \leq k \leq n+j, \\ \chi_{n+j+1} \cdot \phi^n_{i,j} &= \phi^n_{i,j+1}. \end{aligned}$$

Overall it holds

$$\begin{aligned} G^{\text{sel}}\Phi^n_{i,j} &= \alpha \cdot \left((i+j)\Phi^n_{ij} + (n-i)\Phi^n_{i+1,j} - (n+j)\Phi^n_{i,j+1}\right) \\ &= \alpha \cdot \left((i+j)\left(\Phi^n_{ij} - \Phi^n_{i,j+1}\right) + (n-i)\left(\Phi^n_{i+1,j} - \Phi^n_{i,j+1}\right)\right). \end{aligned}$$

Remark 1.13 (Effect of G on Φ^n_{ij})**.** Overall we have the following total effect of

$$\begin{aligned} G\Phi^n_{ij} = &- n\lambda\Phi^n_{ij}\mathbb{1}_{n\geq 2} + i\frac{1}{2}\left(\vartheta_{\bullet}\Phi^n_{i-1,j} - \bar{\vartheta}\Phi^n_{ij}\right) + j\frac{1}{2}\left(\vartheta_{\bullet}\Phi^n_{i,j-1} - \bar{\vartheta}\Phi^n_{ij}\right) \\ &+ \gamma\Bigg(\binom{i}{2}\left(\Phi^{n-1}_{i-1,j} - \Phi^n_{ij}\right) + i(n-i)\left(\Phi^{n-1}_{ij} - \Phi^n_{ij}\right) + \binom{n-i}{2}\left(\Phi^{n-1}_{ij} - \Phi^n_{ij}\right) \\ &+ ij\left(\Phi^n_{i,j-1} - \Phi^n_{ij}\right) + (n-i)j\left(\Phi^n_{i+1,j-1} - \Phi^n_{ij}\right) + \binom{j}{2}\left(\Phi^n_{i,j-1} - \Phi^n_{ij}\right)\Bigg) \\ &+ \alpha\left((i+j)\left(\Phi^n_{ij} - \Phi^n_{i,j+1}\right) + (n-i)\left(\Phi^n_{i+1,j} - \Phi^n_{i,j+1}\right)\right). \end{aligned}$$

Remark 1.14 (Tree length under neutrality)**.** We note that in the absence of selection, L_n does not depend on the mutational mechanism and $L_n \sim \text{Law}\left(\sum_{k=2}^n kT_k\right)$ in equilibrium, where $T_k \sim \text{Exp}(\mu_k)$ with $\mu_k = \binom{k}{2}\gamma$, $k = 2, ..., n$ are the inter-coalescence times in the tree (see e.g. (3.25) in Wakeley, 2008). It holds

$$\mathbb{E}\left[e^{tT_k}\right] = \frac{\mu_k}{\mu_k - t} = \frac{\binom{k}{2}\gamma}{\binom{k}{2}\gamma - t} = \frac{k(k-1)\gamma}{k(k-1)\gamma - 2t} \tag{1.4.3}$$

for $t < \mu_k$. In particular, for $\lambda \geq 0$,

$$\begin{aligned} f_n := \mathbb{E}^0\left[e^{-\lambda L_n}\right] &= \mathbb{E}^0\left[e^{-\lambda\sum_{k=2}^n kT_k}\right] = \prod_{k=2}^n \mathbb{E}^0\left[e^{-\lambda k T_k}\right] \\ &= \prod_{k=2}^n \frac{k(k-1)\gamma}{k(k-1)\gamma + 2\lambda} = \prod_{k=2}^n \frac{(k-1)\gamma}{(k-1)\gamma + 2\lambda} \end{aligned} \tag{1.4.4}$$

with $f_1 = 1$ since the empty product is defined to be 1.

We are now ready to give our first main result, which gives a recursion for an approximation of the Laplace-transform of the tree length under selection for small α.

Theorem 1.15 (Genealogical distances under additive selection). *Let $\lambda \geq 0$ and $\alpha, \gamma, \bar{\vartheta}$ and Θ be given as in Remark 1.12 and L_n denote the total tree length of a sample of size n. Further we define*

$$x_n := \mathbb{E}^\alpha[e^{-\lambda L_n}] - \mathbb{E}^0[e^{-\lambda L_n}].$$

Then as $\alpha \to 0$, $x_1, x_2, \dots$ satisfy the recursion $x_1 = 0$ and

$$\left(\gamma\binom{n}{2} + n\lambda\right) \cdot x_n = \gamma\binom{n}{2} \cdot x_{n-1} + \alpha^2 n \cdot a_n, \tag{1.4.5}$$

where $a_1, a_2, \dots$ satisfy the recursion $a_1 = 0$ and

$$\left(\gamma\binom{n+1}{2} + \frac{\bar{\vartheta}}{2} + n\lambda\right) \cdot a_n = \gamma\binom{n}{2} \cdot a_{n-1} + \Theta(1-\Theta)b_n + \mathcal{O}(\alpha), \tag{1.4.6}$$

where $b_1, b_2, \dots$ satisfy the recursion $b_1 = 0$ and

$$\left(\gamma\binom{n+2}{2} + \bar{\vartheta} + n\lambda\right) \cdot b_n = \gamma\binom{n}{2} \cdot b_{n-1} + \gamma\binom{n}{2} \cdot c_{n-1} + \gamma(n-1) \cdot d_n \tag{1.4.7}$$

where $c_1, c_2, \dots$ satisfy the recursion $c_1 = 0$ and

$$\left(\gamma\binom{n+2}{2} + \bar{\vartheta} + n\lambda\right) \cdot c_n = \gamma\binom{n}{2} \cdot c_{n-1} + 2\gamma \cdot e_n + \gamma d_n, \tag{1.4.8}$$

where $e_1, e_2, \dots$ satisfy a recursion $e_1 = 0$ and

$$\left(\gamma\binom{n+1}{2} + \bar{\vartheta} + n\lambda\right) \cdot e_n = \gamma\binom{n}{2} \cdot e_{n-1} + \gamma d_n \tag{1.4.9}$$

and finally (recall (1.4.4)*)*

$$d_n = f_{n-1} - f_n - g_{n-1} + g_n \tag{1.4.10}$$

with $g_1 = 1/(1 + 2\bar{\vartheta})$ and

$$g_n = \frac{2(n+1)}{(n-1)} \sum_{i=2}^{n} \frac{1}{(i+1)i} \cdot \prod_{k=2}^{i-1} \frac{(k-1)\gamma}{(k-1)\gamma + 2\lambda} \cdot \prod_{k=i}^{n} \frac{k(k-1)\gamma}{k(k-1)\gamma + 2(k\lambda + \bar{\vartheta})}. \tag{1.4.11}$$

Remark 1.16 (Solving the recursions). All recursions for $x_n, a_n, b_n, c_n, e_n, h_n$ are of the form

$$\mu_n = \gamma_n \cdot \mu_{n-1} + \nu_n$$

with $\mu_1 = 0$ and can readily be solved by writing

$$\begin{aligned}\mu_n &= \nu_n + \gamma_n \cdot (\nu_{n-1} + \gamma_{n-1} \cdot (\nu_{n-2} + \gamma_{n-2} \cdot (\cdots \nu_2 + \gamma_2 \cdot 0))) \\ &= \sum_{k=2}^{n} \nu_k \prod_{m=k+1}^{n} \gamma_m\end{aligned}$$

with $\prod_\emptyset := 1$.

In the proof of Theorem 1.15 we will be looking at coalescents of size $(n+j)$ and their first-step decompositions. To be able to clarify the size of the coalescent at any point of the computations we will introduce some notation. We will use a superscript $(n+j)$ to indicate the size of the coalescent studied at the moment.

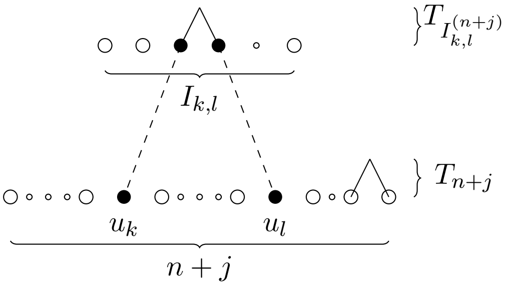
Remark 1.17 (Notation)**.** Starting with an $(n+j)$-coalescent where $k, l \in \{1, ..., n+j\}$ with $k \neq l$ and $u_k = u_l = \bullet$, let $I_{k,l}^{(n+j)}$ denote the number of individuals in the coalescent right before individual k and l coalesce.
Then the *genealogical distance* between individual k and l is given by

$$R_{kl}^{(n+j)} := \sum_{i=I_{kl}^{(n+j)}}^{n+j} T_i.$$

Note that

$$R_{kl}^{(n+j)} = T_{n+j} + R_{kl}^{(n+j-1)}. \tag{1.4.12}$$

A graphic showing the $(n+j)$-coalescent as described above is depicted below:



Furthermore we have the probability

$$\begin{aligned}
&\mathbb{P}^0(\{u_k = u_l = \bullet\}) \\
=&\left(1 - e^{-\frac{\bar\vartheta}{2}R_{kl}^{(n+2)}}\right)^2 \cdot \Theta^2 + 2 \cdot \left(1 - e^{-\frac{\bar\vartheta}{2}R_{kl}^{(n+2)}}\right) \cdot e^{-\frac{\bar\vartheta}{2}R_{kl}^{(n+2)}} \cdot \Theta^2 + e^{-\bar\vartheta R_{kl}^{(n+2)}} \cdot \Theta \\
=&\left(1 - 2e^{-\frac{\bar\vartheta}{2}R_{kl}^{(n+2)}} + e^{-\bar\vartheta R_{kl}^{(n+2)}} + 2e^{-\frac{\bar\vartheta}{2}R_{kl}^{(n+2)}} - 2e^{-\bar\vartheta R_{kl}^{(n+2)}}\right) \cdot \Theta^2 + e^{-\bar\vartheta R_{kl}^{(n+2)}} \cdot \Theta \\
=&\left(1 - e^{-\bar\vartheta R_{kl}^{(n+2)}}\right) \cdot \Theta^2 + e^{-\bar\vartheta R_{kl}^{(n+2)}} \cdot \Theta \\
=&\Theta^2 + \Theta(1-\Theta) \cdot e^{-\bar\vartheta R_{kl}^{(n+2)}}.
\end{aligned}$$

Hence we get

$$\mathbb{E}\left[e^{-\lambda L_n} \cdot \mathbb{1}_{\{u_k = u_l = \bullet\}}\right] = \Theta^2 \cdot \mathbb{E}\left[e^{-\lambda L_n}\right] + \Theta(1-\Theta) \cdot \mathbb{E}\left[e^{-\lambda L_n} \cdot e^{-\bar\vartheta R_{kl}^{(n+2)}}\right]. \tag{1.4.13}$$

Remark 1.18 (Interpretations)**.** During the proof of Theorem 1.15 we will obtain explicit formulas for $a_n, b_n, c_n, ...$ which are given by

$$\begin{aligned}
a_n &:= \frac{1}{\alpha}\mathbb{E}^\alpha\left[\Phi_{10}^n - \Phi_{01}^n\right] = \mathbb{E}^\alpha\left[e^{-\lambda L_n}\left(\mathbb{1}_{\{u_1=\bullet\}} - \mathbb{1}_{\{u_{n+1}=\bullet\}}\right)\right], \\
b_n &:= \mathbb{E}^0\left[(n-1)\Phi_{20}^n - 2n\Phi_{11}^n + (n+1)\Phi_{02}^n\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}^0\left[e^{-\lambda L_n^{(n+2)}}\left((n-1)e^{-\bar\vartheta R_{12}^{(n+2)}} - 2ne^{-\bar\vartheta R_{1,n+1}^{(n+2)}} + (n+1)e^{-\bar\vartheta R_{n+1,n+2}^{(n+2)}}\right)\right], \\
c_n &:= \mathbb{E}^0\left[e^{-\lambda L_{n-1}^{(n+1)}}\left(e^{-\bar\vartheta R_{12}^{(n+1)}} - 2e^{-\bar\vartheta R_{1,n}^{(n+1)}} + e^{-\bar\vartheta R_{n,n+1}^{(n+1)}}\right)\right], \qquad (1.4.14)\\
d_n &:= \mathbb{E}^0\left[\left(e^{-\lambda L_{n-1}^{(n+1)}} - e^{-\lambda L_n^{(n+1)}}\right)\left(1 - e^{-\bar\vartheta R_{12}^{(n+1)}}\right)\right], \\
e_n &:= \mathbb{E}^0\left[e^{-\lambda L_n^{(n+1)}}\left(e^{-\bar\vartheta R_{12}^{(n+1)}} - e^{-\bar\vartheta R_{1,n+1}^{(n+1)}}\right)\right], \\
g_n &:= \mathbb{E}^0\left[e^{-\lambda L_n}e^{-\bar\vartheta R_{12}}\right].
\end{aligned}
$$

Moreover, in Theorem 1.26, another quantity will arise, which is

$$
h_n = e_n - c_n = \mathbb{E}^0\left[e^{-\lambda L_n^{(n+2)}}\left(e^{-\bar\vartheta R_{1,n+1}^{(n+2)}} - e^{-\bar\vartheta R_{n+1,n+2}^{(n+2)}}\right)\right]. \qquad (1.4.15)
$$

The absence of any superscripts indicating the size of the coalescent in the definition of g_n is no mistake but intentional. We will see in the proof that g_n is actually of the form

$$
g_n := \mathbb{E}^0\left[e^{-\lambda L_n^{(n+1)}}e^{-\bar\vartheta R_{12}^{(n+1)}}\right].
$$

However, since all relevant parameters such as the tree length (L_n) and the genealogical distance of indivudual 1 and 2 ($R_{12}^{(n+1)}$) are given within the sample n, we can omit the superscript $(n+1)$ as it makes no difference in which coalescent the above expression is computed in. This follows from an application of the subsampling formula given in Saunders et al. (1984) where the authors examine the behaviour of the number of distinct ancestors in a subsample in a larger sample.

We also note that only a_n is computed in the model with selection while all other expressions are determined within the neutral model (as indicated by the superscripts α and 0).

The initial value d_2 is given through the initial condition $f_1 = 1$, as well as $f_2 = \frac{\gamma}{\gamma+2\lambda}$, g_1 and g_2.

Remark 1.19 (Comparing neutral and selective genealogies).

1. We note that for $\alpha = 0$, (1.4.5) gives precisely (1.4.4). Moreover, there is no linear term in α in the recursion (1.4.5) which is consistent with the results of Theorem 4.26 in Krone and Neuhauser (1997) and Theorem 5 in Depperschmidt et al. (2012). For other modes of dominance, however, a linear term arises (see Theorem 1.26).

2. While x_n and a_n are quantities within the selected genealogies, all other quantities can be computed under neutrality. However, if one would like to obtain finer results, i.e. specify the $\mathcal{O}(\alpha^3)$-term in (1.4.5), more quantities within selected genealogies would have to be computed. In principle, this is straightforward using our approach to the proof of Theorem 1.15.

3. With the recursions given in Theorem 1.15 we can tackle the question of whether genealogies in selective models are shorter compared to ones in neutral models or not. We recall that in order for genealogical distances under selection to be shorter in the *Laplace-transform-order* it needs to hold

$$
\mathbb{E}^\alpha\left[e^{-\lambda L_n}\right] \geq \mathbb{E}^0\left[e^{-\lambda L_n}\right], \qquad (1.4.16)
$$

in other words $x_n := \mathbb{E}^{\alpha}[e^{-\lambda L_n}] - \mathbb{E}^{0}[e^{-\lambda L_n}]$ needs to be positive. The positivity of x_n ultimately depends on the positivity of d_n which can be easily shown: For g_n we have that

$$\begin{aligned}
g_n &= \frac{2(n+1)}{(n-1)} \sum_{i=2}^{n} \frac{1}{(i+1)i} \cdot \prod_{k=2}^{i-1} \frac{(k-1)\gamma}{(k-1)\gamma + 2\lambda} \cdot \prod_{k=i}^{n} \frac{k(k-1)\gamma}{k(k-1)\gamma + 2(k\lambda + \bar{\vartheta})} \\
&= \frac{2(n+1)}{(n-1)} \cdot \frac{n}{n} \cdot \frac{(n-2)}{(n-2)} \cdot \frac{n(n-1)\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})} \\
&\qquad \cdot \sum_{i=2}^{n-1} \frac{1}{(i+1)i} \cdot \prod_{k=2}^{i-1} \frac{(k-1)\gamma}{(k-1)\gamma + 2\lambda} \cdot \prod_{k=i}^{n-1} \frac{k(k-1)\gamma}{k(k-1)\gamma + 2(k\lambda + \bar{\vartheta})} \\
&\quad + \frac{2(n+1)}{(n-1)} \cdot \frac{1}{(n+1)n} \cdot \prod_{k=2}^{n-1} \frac{(k-1)\gamma}{(k-1)\gamma + 2\lambda} \cdot \frac{n(n-1)\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})} \\
&= \frac{(n+1)(n-2)\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})} \cdot g_{n-1} + \frac{2\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})} \cdot f_{n-1}.
\end{aligned}$$

Further for any $k \in \mathbb{N}$

$$\frac{k(k-1)\gamma}{k(k-1)\gamma + 2(k\lambda + \bar{\vartheta})} \leq \frac{(k-1)\gamma}{(k-1)\gamma + 2\lambda}.$$

Therefore it holds

$$g_n \leq \frac{2(n+1)}{(n-1)} \cdot \sum_{i=2}^{n} \frac{1}{(i+1)i} \cdot f_n = f_n.$$

Overall we get

$$\begin{aligned}
d_n &= f_{n-1} - f_n - g_{n-1} + g_n \\
&= f_{n-1} - \frac{(n-1)\gamma}{(n-1)\gamma + 2\lambda} \cdot f_{n-1} - g_{n-1} \\
&\quad + \frac{(n+1)(n-2)\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})} \cdot g_{n-1} + \frac{2\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})} \cdot f_{n-1} \\
&= \left(1 - \frac{(n-1)\gamma}{(n-1)\gamma + 2\lambda} + \frac{2\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})}\right) \cdot f_{n-1} \\
&\quad - \left(1 - \frac{(n+1)(n-2)\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})}\right) \cdot g_{n-1} \\
&= \left(\frac{2\lambda}{(n-1)\gamma + 2\lambda} + \frac{2\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})}\right) \cdot f_{n-1} - \frac{(2\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})} \cdot g_{n-1} \\
&\geq \left(\frac{2\lambda}{(n-1)\gamma + 2\lambda} + \frac{2\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})} - \frac{(2\gamma}{n(n-1)\gamma + 2(n\lambda + \bar{\vartheta})}\right) \cdot f_{n-1} \\
&= \frac{2\lambda}{(n-1)\gamma + 2\lambda} \cdot f_{n-1} \\
&\geq 0.
\end{aligned}$$

Therefore we can show that for small $\alpha > 0$ the genealogical tree is in fact shorter than under neutrality under the Lapace-transform-order.

Since we can directly obtain expected tree lengths from the Laplace-transforms in Theorem 1.15, we obtain also a recursion for expected tree lengths by using that $\mathbb{E}^0[L_n]-\mathbb{E}^\alpha[L_n] = \frac{\partial}{\partial\lambda}x_n|_{\lambda=0}$.

Corollary 1.20 (Expected tree length under additive selection)**.** *With $\alpha, \gamma, \bar{\vartheta}, \Theta$ and L_n as in Theorem 1.15, let*

$$\widetilde{x}_n := \mathbb{E}^0[L_n] - \mathbb{E}^\alpha[L_n].$$

Then, $\widetilde{x}_1, \widetilde{x}_2, ...$ satisfy the recursion $\widetilde{x}_1 = 0$ and

$$\gamma\binom{n}{2} \cdot \widetilde{x}_n = \gamma\binom{n}{2} \cdot \widetilde{x}_{n-1} + \alpha^2 n \cdot \widetilde{a}_n, \qquad n = 2, 3, ...$$

where $\widetilde{a}_1, \widetilde{a}_2, ...$ satisfy the recursion $\widetilde{a}_1 = 0$ and

$$\left(\gamma\binom{n+1}{2} + \frac{\bar{\vartheta}}{2}\right) \cdot \widetilde{a}_n = \gamma\binom{n}{2} \cdot \widetilde{a}_{n-1} + \Theta(1-\Theta) \cdot \widetilde{b}_n + \mathcal{O}(\alpha), \qquad n = 2, 3, ...$$

where $\widetilde{b}_1, \widetilde{b}_2, ...$ satisfy the recursion $\widetilde{b}_1 = 0$ and

$$\left(\gamma\binom{n+2}{2} + \bar{\vartheta}\right) \cdot \widetilde{b}_n = \gamma\binom{n}{2} \cdot \widetilde{b}_{n-1} + \gamma\binom{n}{2} \cdot \widetilde{c}_{n-1} + \gamma(n-1) \cdot \widetilde{d}_n$$

where $\widetilde{c}_1, \widetilde{c}_2, ...$ satisfy the recursion $\widetilde{c}_1 = 0$ and

$$\left(\gamma\binom{n+2}{2} + \bar{\vartheta}\right) \cdot \widetilde{c}_n = \gamma\binom{n}{2} \cdot \widetilde{c}_{n-1} + 2\gamma \cdot \widetilde{e}_n + \gamma\widetilde{d}_n,$$

where $\widetilde{e}_1, \widetilde{e}_2, ...$ satisfy a recursion $\widetilde{e}_1 = 0$ and

$$\left(\gamma\binom{n+1}{2} + \bar{\vartheta}\right) \cdot \widetilde{e}_n = \gamma\binom{n}{2} \cdot \widetilde{e}_{n-1} + \gamma\widetilde{d}_n$$

and finally

$$\widetilde{d}_n = \frac{2}{(n-1)\gamma} - \widetilde{g}_{n-1} + \widetilde{g}_n \tag{1.4.17}$$

with $\widetilde{g}_1 = 0$ and

$$\widetilde{g}_n = \frac{-4(n+1)}{(n-1)} \sum_{i=2}^{n} \frac{1}{(i+1)i} \Bigg[\left(\frac{1}{\gamma} \cdot \sum_{k=2}^{i-1} \frac{1}{(k-1)} \right) \cdot \prod_{k=i}^{n} \frac{k(k-1)\gamma}{k(k-1)\gamma + 2\bar{\vartheta}} \tag{1.4.18}$$
$$+ \sum_{k=i}^{n} \frac{k^2(k-1)\gamma}{(k(k-1)\gamma + 2\bar{\vartheta})^2} \cdot \prod_{\substack{l=i \\ l\neq k}}^{n} \frac{k(k-1)\gamma}{k(k-1)\gamma + 2\bar{\vartheta}} \Bigg].$$

The following result, the special case $n = 2$, was already obtained in Theorem 5 of Depperschmidt et al. (2012).

Corollary 1.21 (Genealogical distance of two individuals under additive selection)**.** *With* $\alpha, \gamma, \bar{\vartheta}, \Theta$ *and* L_n *as in Theorem 1.15,*

$$\begin{aligned}\mathbb{E}^\alpha[e^{-\lambda L_2}] &= \frac{\gamma}{\gamma + 2\lambda}\\ &\quad + 8\alpha^2\lambda \frac{\Theta(1-\Theta)\gamma\bar{\vartheta}(2\gamma + 2\lambda + \bar{\vartheta})}{(6\gamma + 4\lambda + 2\bar{\vartheta})(6\gamma + 2\lambda + 2\bar{\vartheta})(\gamma + 2\lambda)^2(\gamma + \bar{\vartheta})(\gamma + 2\lambda + \bar{\vartheta})} + \mathcal{O}(\alpha^3),\\ \mathbb{E}^\alpha[L_2] &= \frac{1}{\gamma}\left(2 - 8\alpha^2 \frac{\Theta(1-\Theta)\bar{\vartheta}(2\gamma + \bar{\vartheta})}{(6\gamma + \bar{\vartheta})^2(\gamma + \bar{\vartheta})^2}\right) + \mathcal{O}(\alpha^3).\end{aligned}$$

Proof. Applying Theorem 1.15, we get

$$\begin{aligned}g_2 &= \frac{\gamma}{\gamma + 2\lambda + \bar{\vartheta}},\\ d_2 &= 1 - \frac{\gamma}{\gamma + 2\lambda} - \frac{\gamma}{\gamma + \bar{\vartheta}} + \frac{\gamma}{\gamma + 2\lambda + \bar{\vartheta}} = \frac{2\lambda}{\gamma + 2\lambda} - \frac{2\gamma\lambda}{(\gamma + \bar{\vartheta})(\gamma + 2\lambda + \bar{\vartheta})}\\ &= \frac{2\lambda\bar{\vartheta}(2\gamma + 2\lambda + \bar{\vartheta})}{(\gamma + \bar{\vartheta})(\gamma + 2\lambda)(\gamma + 2\lambda + \bar{\vartheta})},\\ a_2 &= \frac{\Theta(1-\Theta)}{3\gamma + 2\lambda + \frac{\bar{\vartheta}}{2}} \cdot b_2 + \mathcal{O}(\alpha) = \frac{2\Theta(1-\Theta)}{6\gamma + 4\lambda + \bar{\vartheta}} \cdot \frac{\gamma}{6\gamma + 2\lambda + \bar{\vartheta}} d_2 + \mathcal{O}(\alpha),\\ x_2 &= \frac{2\alpha^2}{\gamma + 2\lambda} \cdot a_n = \frac{8\alpha^2\Theta(1-\Theta)\gamma\lambda\bar{\vartheta}(2\gamma + 2\lambda + \bar{\vartheta})}{(6\gamma + 4\lambda + \bar{\vartheta})(6\gamma + 2\lambda + \bar{\vartheta})(\gamma + \bar{\vartheta})(\gamma + 2\lambda)^2(\gamma + 2\lambda + \bar{\vartheta})} + \mathcal{O}(\alpha^3).\end{aligned}$$

Overall we get

$$\begin{aligned}\mathbb{E}^\alpha\left[e^{-\lambda L_2}\right] &= \mathbb{E}^0\left[e^{-\lambda L_2}\right] + x_2\\ &= \frac{\gamma}{(\gamma + 2\lambda)} + \frac{8\alpha^2\Theta(1-\Theta)\gamma\lambda\bar{\vartheta}(2\gamma + 2\lambda + \bar{\vartheta})}{(6\gamma + 4\lambda + \bar{\vartheta})(6\gamma + 2\lambda + \bar{\vartheta})(\gamma + \bar{\vartheta})(\gamma + 2\lambda)^2(\gamma + 2\lambda + \bar{\vartheta})} + \mathcal{O}(\alpha^3)\end{aligned}$$

which shows the first equation.
The second equation follows with

$$\begin{aligned}\mathbb{E}^\alpha[L_2] &= -\frac{\partial}{\partial\lambda}\mathbb{E}^\alpha\left[e^{-\lambda L_2}\right]\Bigg|_{\lambda=0}\\ &= \frac{2\gamma}{(\gamma + 2\lambda)^2}\Bigg|_{\lambda=0} - \frac{8\alpha^2\Theta(1-\Theta)\gamma\bar{\vartheta}(2\gamma + 2\lambda + \bar{\vartheta})}{(6\gamma + 4\lambda + \bar{\vartheta})(6\gamma + 2\lambda + \bar{\vartheta})(\gamma + \bar{\vartheta})(\gamma + 2\lambda)^2(\gamma + 2\lambda + \bar{\vartheta})}\Bigg|_{\lambda=0}\\ &\quad - \left(\lambda \cdot \frac{\partial}{\partial\lambda} \frac{8\alpha^2\Theta(1-\Theta)\gamma\bar{\vartheta}(2\gamma + 2\lambda + \bar{\vartheta})}{(6\gamma + 4\lambda + \bar{\vartheta})(6\gamma + 2\lambda + \bar{\vartheta})(\gamma + \bar{\vartheta})(\gamma + 2\lambda)^2(\gamma + 2\lambda + \bar{\vartheta})}\right)\Bigg|_{\lambda=0} + \mathcal{O}(\alpha^3)\\ &= \frac{2}{\gamma} - \frac{8\alpha^2\Theta(1-\Theta)\bar{\vartheta}(2\gamma + \bar{\vartheta})}{(6\gamma + \bar{\vartheta})^2\gamma(\gamma + \bar{\vartheta})^2} + \mathcal{O}(\alpha^3).\end{aligned}$$

□

1.5 Other modes of dominance

In this following section we will try to generalise the results of the previous section by changing the fitness function to capture additional modes of dominance. To do so, we introduce a *dominance coefficient* $h \in (-\infty, \infty)$ and replace 3. from Remark 1.1 by the following interpretation:

3'. Every line picks a random partner and if the pair is a heterozygote, it has fitness advantage αh, and if it is homozygous for $\bullet$, it has fitness advantage α.

By introducing h we are now able to study different degrees of dominance. The cases $h = 1$ and $h = 0$ describe cases of *complete dominance*: If $h = 1$ we have that $\bullet$ is the dominant allele and if $h = 0$, type $\circ$ is the dominant one. The cases $0 < h < 1$, $h > 1$ and $h < 0$ describe cases of *incomplete dominance*, *overdominance* and *underdominance*, respectively. We refer to Chapter 3 from Gillespie (2004) for further details on the topic of natural selection. We assume that $h \geq 0$ in order to obtain positive transition rates, but some modifications also allow for $h < 0$.

Before we can formulate results analogous to Theorem 1.15 and Corollary 1.20 we need to adjust the generator to our new situation. The dynamics of tree growth, resampling and mutation events are not affected by the introduction of the dominance coefficient h. Only the selection operator needs to be changed in the following way:

(D') *Selection operator under other modes of dominance*: As we are now dealing with diploid selection we need to consider a fitness function of the form

$$\chi : I \times I \to [0, 1]$$

with $\chi(u, v) = \chi(v, u)$ for all $u, v \in I$.
We recall that n is our sample size. In order to define a fitness function reflecting the action described in 3'. we randomly pick a haploid individual m. Since $n < \infty$, we have that m is outside the sample with high probability, i.e. $m > n$. For any $1 \leq k \leq n$ we have the fitness function

$$\chi_{k,m} = \begin{cases} 1, & u_k = u_m = \bullet, \\ h, & u_k \neq u_m, \\ 0, & u_k = u_m = \circ, \end{cases}$$

which can be rewritten as follows

$$\begin{aligned} \chi_{k,m} &= \mathbb{1}_{\{u_k = u_m = \bullet\}} + h \mathbb{1}_{\{u_k \neq u_m\}} \\ &= \mathbb{1}_{\{u_k = u_m = \bullet\}} + h \left(\mathbb{1}_{\{u_k = \bullet\}} \cdot \mathbb{1}_{\{u_m = \circ\}} + \mathbb{1}_{\{u_k = \circ\}} \cdot \mathbb{1}_{\{u_m = \bullet\}} \right) \\ &= \mathbb{1}_{\{u_k = u_m = \bullet\}} + h \left(\mathbb{1}_{\{u_k = \bullet\}} \cdot (1 - \mathbb{1}_{\{u_m = \bullet\}}) + (1 - \mathbb{1}_{\{u_k = \bullet\}}) \cdot \mathbb{1}_{\{u_m = \bullet\}} \right) \\ &= (1 - 2h) \mathbb{1}_{\{u_k = u_m = \bullet\}} + h \left(\mathbb{1}_{\{u_k = \bullet\}} + \mathbb{1}_{\{u_m = \bullet\}} \right). \end{aligned}$$

Just like in the additive case, selective events occur in a pair of one individual m giving birth and the other, individual k, dying, again, as given through the function $\theta_{m,k}$ from (D) in Section 1.3.

The same reasoning as in (1.3.7) then gives us

$$G^{\text{sel},h} \Phi(u) := \alpha \sum_{k < m, m'}^{\infty} \langle \nu^u, \chi_{m,m'} \left(\phi \circ \theta_{m,k} - \phi \right) \rangle$$

where we can ignore the summands with $k > n$ as ϕ only depends on the first n individuals leading to

$$= \alpha \sum_{\substack{k<m \\ k\leq n}} \langle \nu^u, \chi_{m,m+1} \left(\phi \circ \theta_{m,k} - \phi \right) \rangle$$

where the summands with $m \leq n$ only give a negligible effect and hence only summands with $m > n$ are of interest. Without loss of generality we choose $m = n+1$ and obtain

$$\approx \alpha \sum_{k=1}^{n} \langle \nu^u, \chi_{m+1,m+2} \left(\phi \circ \theta_{n+1,k} - \phi \right) \rangle$$

which gives by permuting sampling order of l and $n+1$ in the first term

$$= \alpha \sum_{k=1}^{n} \langle \nu^u, \chi_{k,n+1} \cdot \phi - \chi_{n+1,n+2} \cdot \phi \rangle.$$

Definition 1.22 (Generator of TFVMS under other modes of dominance)**.** Let G^{grow}, G^{res} and G^{mut} given as in (A), (B) and (C). Let $G^{\text{sel},h}$ be as in (D'). The generator of TFVMS with dominance coefficient h is the linear operator on Π with domain Π^1, given by

$$G^h := G^{\text{grow}} + G^{\text{res}} + G^{\text{mut}} + G^{\text{sel},h}. \tag{1.5.1}$$

Remark 1.23 (Link to additive selection)**.** A more general form of fitness functions is the following:

$$\chi' : I \times I \times \mathbb{R}_+ \to [0,1]$$

with $\chi'(u,v,r) = \chi(v,u)$ for all $u,v \in I, r \in \mathbb{R}_+$ The form of the fitness function arises as we have a process that encodes the type distribution as well as the genealogical tree. Hence, we deal with diploid selection depending also on genealogical distance. $\chi'(u,v,r)$ gives us the fitness of a diploid individual with genotype $\{u,v\}$ if the genealogical distance of the two haploids forming the diploid individual is r. We define

$$\chi'_{k,l}(\underline{\underline{r}}, \underline{u}) := \chi'(u_k, u_l, r_{k\wedge l, k\vee l}).$$

If $\chi'(u,v,r)$ does not depend on r, and if there exists a function $\chi : I \to [0,1]$ such that

$$\chi'(u,v,r) = \chi(u) + \chi(v),$$

we say that selection is *additive* and conclude that with

$$\chi_k(\underline{\underline{r}}, \underline{u}) = \chi(u_k).$$

We obtain

$$G^{\text{sel}}\Phi(u) = \alpha \sum_{k=1}^{n} \langle \nu^u, \phi \cdot \chi'_{k,n+1} - \phi \cdot \chi'_{n+1,n+2} \rangle = \alpha \sum_{k=1}^{n} \langle \nu^u, \phi \cdot \chi_k - \phi \cdot \chi_{n+1} \rangle.$$

Additive selection describes the case that the selective advantage of an individual which is homozygous for • is twice the advantage of a heterozygote. In other words, the fitness of the heterozygote is exactly intermediate between the fitness of the homozygotes. Indeed, for 2α and $h = 1/2$ we have

$$\chi_{k,m} = \frac{1}{2}\left(\mathbb{1}_{\{u_k=\bullet\}} + \mathbb{1}_{\{u_m=\bullet\}}\right) = \frac{1}{2}\left(\chi_k + \chi_m\right)$$

leading to

$$G^{\text{sel},1/2}\Phi(u) = \alpha\sum_{k=1}^{n}\langle \nu^u, \chi_k \cdot \phi - \chi_{n+1}\cdot\phi\rangle = G^{\text{sel}}\Phi(u).$$

Hence the case $h = 1/2$ describes the case of additive selection we studied in Section 1.3.

We recall Remark 1.10 and denote the the limiting process of the TFVMS under other modes of dominance by $\mathcal{U}_\infty^{\alpha,h}$.

Remark 1.24 (Notation)**.** We extend the notation introduced in Remark 1.11 and denote the distribution of TFVMS under the selection coefficient α and dominance coefficient h by $\mathbb{P}^{\alpha,h}(\cdot)$. $\mathbb{E}^{\alpha,h}[\cdot]$ will denote the corresponding expectation. More precisely

$$\mathbb{E}^{\alpha,h}\left[\Phi\right] := \mathbb{E}[\Phi(\mathcal{U}_\infty^{\alpha,h})]. \tag{1.5.2}$$

Recalling that $\mathbb{P}^{\alpha}(\cdot)$ and $\mathbb{E}^{\alpha}[\cdot]$ are the corresponding operators for additive selection, we have, according to Remark 1.23 and using the above notation,

$$\mathbb{P}^{\alpha}(\cdot) = \mathbb{P}^{2\alpha,1/2}(\cdot).$$

To be able to study the Laplace-transform of tree lengths we quickly investigate the effect G^h has on the function $\Phi^n_{ij} := \Phi^{n+j,\phi^n_{ij}} \in \bar{\mathcal{C}}^1_{n+j}$ with

$$\phi^n_{ij}(\underline{r},\underline{u}) := e^{-\lambda L_n(\underline{r})}\cdot \mathbb{1}_{\{u_1=\bullet\}}\cdots\mathbb{1}_{\{u_i=\bullet\}}\cdot\mathbb{1}_{\{u_{n+1}=\bullet\}}\cdots\mathbb{1}_{\{u_{n+j}=\bullet\}}. \tag{1.5.3}$$

Tree growth, resampling and mutation have the same effects as in the in the case of additive selection and we are left to look at the selection operator $G^{\text{sel},h}$:

$$\begin{aligned}G^{\text{sel},h}\Phi^n_{i,j}(u) &= \alpha\sum_{k=1}^{n+j}\langle\nu^u, \chi_{k,n+j+1}\cdot\phi^n_{i,j} - \chi_{n+j+1,n+j+2}\cdot\phi^n_{i,j}\rangle\\ &= \alpha\left(\sum_{k=1}^{n+j}\langle\nu^u,\chi_{k,n+j+1}\cdot\phi^n_{i,j}\rangle - (n+j)\langle\nu^u,\chi_{n+j+1,n+j+2}\cdot\phi^n_{i,j}\rangle\right).\end{aligned}$$

For $1 \le k \le i$ and $U := \{u_1 = \cdots = u_m = \cdots = u_i = u_{n+1} = \cdots = u_{n+j} = \bullet\}$ we have

$$\begin{aligned}&\chi_{k,n+j+1}\cdot\phi^n_{i,j}\\ =&\left((1-2h)\mathbb{1}_{\{u_k=u_{n+j+1}=\bullet\}} + h\left(\mathbb{1}_{\{u_k=\bullet\}} + \mathbb{1}_{\{u_{n+j+1}=\bullet\}}\right)\right)\cdot e^{-\lambda L_n}\cdot\mathbb{1}_U\\ =&(1-2h)\phi^n_{i,j+1} + h\left(\phi^n_{i,j} + \phi^n_{i,j+1}\right).\end{aligned}$$

Similarly we get

$$\chi_{k,n+j+1}\cdot\phi^n_{i,j} = (1-2h)\phi^n_{i+1,j+1} + h\left(\phi^n_{i+1,j} + \phi^n_{i,j+1}\right) \quad \text{for } i+1 \le k \le n,$$

$$\begin{aligned}\chi_{k,n+j+1}\cdot\phi^n_{i,j} &= (1-2h)\phi^n_{i,j+1}+h\left(\phi^n_{i,j}+\phi^n_{i,j+1}\right) \quad \text{for } n+1\le k\le n+j,\\ \chi_{n+j+1,n+j+2}\cdot\phi^n_{i,j} &= (1-2h)\phi^n_{i,j+2}+2h\phi^n_{i,j+1}.\end{aligned}$$

Overall it holds

$$\begin{aligned}G^{\mathrm{sel},h}\Phi^n_{i,j}(u) = \alpha\cdot\Big(&i((1-2h)\Phi^n_{i,j+1}+h(\Phi^n_{ij}+\Phi^n_{i,j+1}))\\ &+(n-i)\cdot((1-2h)\Phi^n_{i+1,j+1}+h(\Phi^n_{i+1,j}+\Phi^n_{i,j+1}))\\ &+j\cdot((1-2h)\Phi^n_{i,j+1}+h(\Phi^n_{ij}+\Phi^n_{i,j+1}))\\ &-(n+j)\cdot((1-2h)\Phi^n_{i,j+2}+2h\Phi^n_{i,j+1})\Big)\\ = \alpha\cdot\mathbb{E}^{\alpha,h}\big[&(i+j)h\Phi^n_{ij}+(i(1-h)+(n-i)h+j(1-h)-2(n+j)h)\Phi^n_{i,j+1}\\ &+(n-i)h\Phi^n_{i+1,j}+(n-i)(1-2h)\Phi^n_{i+1,j+1}-(n+j)(1-2h)\Phi^n_{i,j+2}\big].\end{aligned}$$

Remark 1.25 (Effect of G^h on Φ^n_{ij})**.** Together with $G^{\mathrm{grow}}, G^{\mathrm{res}}$ and G^{mut} we have

$$\begin{aligned}G^h\Phi^n_{ij} =& -n\lambda\Phi^n_{ij}\mathbb{1}_{n\ge2}+i\frac{1}{2}\left(\vartheta_{\bullet}\Phi^n_{i-1,j}-\bar\vartheta\Phi^n_{ij}\right)+j\frac{1}{2}\left(\vartheta_{\bullet}\Phi^n_{i,j-1}-\bar\vartheta\Phi^n_{ij}\right)\\ &+\gamma\Bigg(\binom{i}{2}\left(\Phi^{n-1}_{i-1,j}-\Phi^n_{ij}\right)+i(n-i)\left(\Phi^{n-1}_{ij}-\Phi^n_{ij}\right)+\binom{n-i}{2}\left(\Phi^{n-1}_{ij}-\Phi^n_{ij}\right)\\ &+ij\left(\Phi^n_{i,j-1}-\Phi^n_{ij}\right)+(n-i)j\left(\Phi^n_{i+1,j-1}-\Phi^n_{ij}\right)+\binom{j}{2}\left(\Phi^n_{i,j-1}-\Phi^n_{ij}\right)\Bigg)\\ &+\alpha\cdot\Big[(i+j)h\Phi^n_{i,j}+(i+j)\Phi^n_{i,j+1}-(n+2i+3j)h\Phi^n_{i,j+1}\\ &\quad+(n-i)h\Phi^n_{i+1,j}+(n-i)(1-2h)\Phi^n_{i+1,j+1}-(n+j)(1-2h)\Phi^n_{i,j+2}\Big].\end{aligned}$$

With this we are now ready to give an analogous result on the Laplace-transform of the tree length of a sample of size n under any form of dominance.

Theorem 1.26 (Genealogical distances under any form of dominance)**.** *Let $\lambda\ge0$ and $\alpha,\gamma,h,\bar\vartheta$ and Θ be given as in Remark 1.12 and L_n denote the total tree length of a sample of size n. Further we define*

$$y_n := \mathbb{E}^{\alpha,h}[e^{-\lambda L_n}]-\mathbb{E}^0[e^{-\lambda L_n}].$$

Then as $\alpha\to0$, $y_1,y_2,...$ satisfy the recursion $y_1=0$ and

$$\left(\gamma\binom{n}{2}+n\lambda\right)\cdot y_n=\gamma\binom{n}{2}\cdot y_{n-1}+\alpha n(1-2h)\Theta(1-\Theta)\cdot h_n+\mathcal{O}(\alpha^2),$$

where $h_1,h_2,...$ satisfy the recursion $h_1=0$ and

$$\left(\gamma\binom{n+2}{2}+\bar\vartheta+n\lambda\right)\cdot h_n=\gamma\binom{n}{2}\cdot h_{n-1}+\gamma(n-1)\cdot e_n, \tag{1.5.4}$$

and e_n was given in Theorem 1.15.

Remark 1.27 (Comparing genealogies)**.** 1. Most interestingly, neutral trees differ from trees under additive selection only in order α^2, whereas the difference is in order α for other forms of dominance. While this may be counter-intuitive at first, it can be easily explained. Note that the model actually does not change if we replace α by $-\alpha$ and h by $1-h$ at the same time. By doing so, we just interchange the roles of allele • and •. For $h = 1/2$, this means that our results have to be identical for α and $-\alpha$, leading to a vanishing linear term in (1.4.5). For $h \neq 1/2$, this symmetry does not have to hold, leading to a linear term in α.

2. Similar to our reasoning in Remark 1.19.3, the sign of h_n in the recursion for y_n determines if tree lengths are shorter or longer under selection. We see that the behaviour changes at $h = 1/2$. By construction, h_n is positive because e_n is positive which again follows from the positivity of d_n (from Theorem 1.15) which we have proved in Remark 1.19.3. So if $h < 1/2$, then y_n is positive as well and we see that trees are shorter under selection (in the Laplace-transform-order). If $h > 1/2$, the reverse is true and trees are longer under selection. This result is not surprising for over-dominant selection, $h > 1$, since the advantage of the heterozygote leads to maintenance of heterozygosity or balancing selection, which in turn is known to produce longer genealogical trees.

Corollary 1.28 (Expected tree length under any form of dominance)**.** *With $\alpha, \gamma, h, \bar\vartheta, \Theta$ and L_n as in Theorem 2.9, let*

$$\widetilde{y}_n := \mathbb{E}^0[L_n] - \mathbb{E}^{\alpha,h}[L_n].$$

Then, $\widetilde{y}_1, \widetilde{y}_2, \ldots$ satisfy the recursion $\widetilde{y}_1 = 0$ and

$$\binom{n}{2} \cdot \widetilde{y}_n = \binom{n}{2} \cdot \widetilde{y}_{n-1} + \alpha n(1-2h) \cdot \widetilde{h}_n + \mathcal{O}(\alpha^2),$$

where $\widetilde{h}_1, \widetilde{h}_2, \ldots$ satisfy the recursion $\widetilde{h}_1 = 0$ and

$$\left(\binom{n+2}{2} + 2\bar\vartheta\right) \cdot \widetilde{h}_n = \binom{n}{2} \cdot \widetilde{h}_{n-1} + (n-1) \cdot \widetilde{e}_n, \tag{1.5.5}$$

and $\widetilde{e}_n$ was given in Corollary 1.20.

As an application we give an analogous result on the pairwise distance given in Corollary 1.21.

Corollary 1.29 (Genealogical distance of two individuals under any form of dominance)**.** *With $\alpha, h, \bar\vartheta, \Theta$ and L_n as in Theorem 1.26,*

$$\begin{aligned}\mathbb{E}^{\alpha,h}[e^{-\lambda L_2}] &= \frac{\gamma}{\gamma+2\lambda} + \frac{4\alpha\lambda(1-2h)\Theta(1-\Theta)\bar\vartheta\gamma^2(2\gamma+2\lambda+\bar\vartheta)}{(\gamma+2\lambda)^2(6\gamma+2\lambda+\bar\vartheta)(3\gamma+2\lambda+\bar\vartheta)(\gamma+\bar\vartheta)(\gamma+2\lambda+\bar\vartheta)} + \mathcal{O}(\alpha^2),\\ \mathbb{E}^{\alpha,h}[L_2] &= \frac{2}{\gamma} - \frac{4\alpha(1-2h)\Theta(1-\Theta)\bar\vartheta(2\gamma+\bar\vartheta)}{(1+2\bar\vartheta)^2(6\gamma+\bar\vartheta)(3\gamma+\bar\vartheta)(\gamma+\bar\vartheta)^2} + \mathcal{O}(\alpha^2).\end{aligned}$$

Proof of Corollary 1.29. Applying Theorem 2.9 and with d_2 from the proof of Corollary 1.21, we get

$$e_2 = \frac{\gamma}{3\gamma+2\lambda+\bar\vartheta} \cdot d_2, \qquad h_2 = \frac{\gamma}{6\gamma+2\lambda+\bar\vartheta} \cdot e_2,$$

and it follows

$$\begin{aligned} y_2 &= \frac{2\alpha(1-2h)\Theta(1-\Theta)}{\gamma+2\lambda} \cdot h_2 + \mathcal{O}(\alpha^2) \\ &= \frac{4\alpha(1-2h)\Theta(1-\Theta)\lambda\bar{\vartheta}\gamma^2(2\gamma+2\lambda+\bar{\vartheta})}{(\gamma+2\lambda)^2(6\gamma+2\lambda+\bar{\vartheta})(3\gamma+2\lambda+\bar{\vartheta})(\gamma+\bar{\vartheta})(\gamma+2\lambda+\bar{\vartheta})} + \mathcal{O}(\alpha^2). \end{aligned}$$

Therefore we have

$$\begin{aligned} \mathbb{E}^{\alpha,h}\left[e^{-\lambda L_2}\right] &= \mathbb{E}^0\left[e^{-\lambda L_2}\right] + y_2 \\ &= \frac{\gamma}{(\gamma+2\lambda)} + \frac{4\alpha(1-2h)\Theta(1-\Theta)\lambda\bar{\vartheta}\gamma^2(2\gamma+2\lambda+\bar{\vartheta})}{(\gamma+2\lambda)^2(6\gamma+2\lambda+\bar{\vartheta})(3\gamma+2\lambda+\bar{\vartheta})(\gamma+\bar{\vartheta})(\gamma+2\lambda+\bar{\vartheta})} + \mathcal{O}(\alpha^2). \end{aligned}$$

As in the proof of Corollary 1.20 we compute $\mathbb{E}^{\alpha,h}[L_2] = -\frac{\partial}{\partial\lambda}\mathbb{E}^{\alpha,h}\left[e^{-\lambda L_2}\right]\Big|_{\lambda=0}$ and obtain the desired expression. □

We conclude this section with Figure 1.2 illustrating the results from Corollary 1.20 and Corollary 1.28. In Figure 1.2.(a) we see the effect of additive selection for $n = 50$. We observe a maximum at $\bar{\vartheta} \approx 0.5$ which can be explained as follows: In a population with a very small mutation rate, the beneficial type will almost always be present and hence, there will be no significant difference to the neutral case. Then again, a very high mutation rate causes selection to be inefficient and we are again close to neutrality. We note that $\Theta(1-\Theta)$ has a linear effect on the change in tree length. Figure 1.2(b) depicts the difference in expected tree length for $h = 0$ and variable n. For other dominance coefficients h we obatin similar graphs as $1 - 2h$ only has a linear effect on the recursions.

1.6 Proofs

The proofs of Theorem 1.15 and Theorem 1.26 are based on the following equality:

$$\mathbb{E}[G\Phi(\mathcal{U}_\infty^\alpha)] = 0 \tag{1.6.1}$$

for $\Phi \in \Pi^1$. We recall that $\mathcal{U}$ is the solution of the (P_0, G, Π^1)-martingale problem, i.e.

$$(M_t^\Phi)_{t\geq 0} := \left(\Phi(\mathcal{U}_t) - \Phi(\mathcal{U}_0) - \int_0^t G\Phi(\mathcal{U}_s)\mathrm{d}s\right)_{t\geq 0}$$

is a martingale for $\Phi \in \Pi^1$. Since, according to Theorem 1.9, $\mathcal{U}_t \xrightarrow{t\to\infty} \mathcal{U}_\infty^\alpha$ and the law of $\mathcal{U}_\infty^\alpha$ is an invariant distribution of $\mathcal{U}$, (1.6.1) follows easily from the fact that M_t^Φ has expectation zero for all $t \geq 0$.

Proof of Theorem 1.15. Let $x_n := \mathbb{E}^\alpha[\Phi_{00}^n] - \mathbb{E}^0[\Phi_{00}^n]$.
With equation (1.6.1) we have that

$$\begin{aligned} 0 &= \mathbb{E}\left[G_0\left(\Phi_{00}^n(\mathcal{U}_\infty^\alpha) - \Phi_{00}^n(\mathcal{U}_\infty^0)\right)\right] \\ &= -n\lambda\mathbb{E}^\alpha\left[\Phi_{00}^n\right] + \gamma\binom{n}{2}\mathbb{E}^\alpha\left[\Phi_{00}^{n-1} - \Phi_{00}^n\right] + \alpha n\mathbb{E}^\alpha\left[\Phi_{10}^n - \Phi_{01}^n\right] \end{aligned}$$

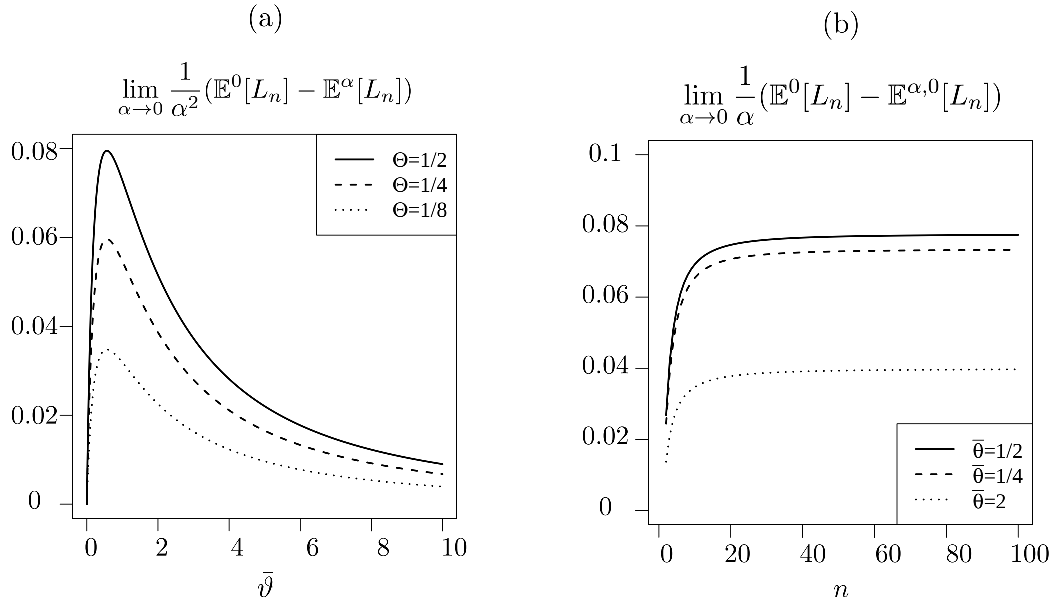


Figure 1.2: Using the recursions from Corollary 1.20 and Corollary 1.28, we see differences in expected tree length. (a) For genic selection and large samples, the effect changes with the total mutation rate $\bar\vartheta$ and is linear in $\Theta(1-\Theta)$. (b) Plot of the change in total tree length for small values of α with $h=0$, dependent on the sample size, and three parameters of $\bar\vartheta$.

$$
\begin{aligned}
&+ n\lambda\mathbb{E}^0\left[\Phi_{00}^n\right] + \gamma\binom{n}{2}\mathbb{E}^0\left[\Phi_{00}^{n-1}-\Phi_{00}^n\right]\\
&= -\left(n\lambda+\gamma\binom{n}{2}\right)\left(\mathbb{E}^\alpha[\Phi_{00}^n]-\mathbb{E}^0[\Phi_{00}^n]\right)+\gamma\binom{n}{2}\left(\mathbb{E}^\alpha[\Phi_{00}^{n-1}]-\mathbb{E}^0[\Phi_{00}^{n-1}]\right)\\
&\quad+\alpha n\mathbb{E}^\alpha\left[\Phi_{10}^n-\Phi_{01}^n\right].
\end{aligned}
$$

Defining $a_n := \frac{1}{\alpha}\mathbb{E}^\alpha\left[\Phi_{10}^n-\Phi_{01}^n\right]$ we get

$$\left(n\lambda+\gamma\binom{n}{2}\right)x_n=\gamma\binom{n}{2}x_{n-1}+\alpha^2 n\cdot a_n.$$

Again with (1.6.1) and additionally (1.3.10) as well as $n+\binom{n}{2}=\binom{n+1}{2}$ we obtain

$$
\begin{aligned}
0 &= \mathbb{E}^\alpha\left[G(\Phi_{10}^n-\Phi_{01}^n\right]\\
&= -n\lambda\mathbb{E}^\alpha\left[\Phi_{10}^n\right]+\frac{1}{2}\left(\vartheta_{\bullet}\mathbb{E}^\alpha\left[\Phi_{00}^n\right]-\bar\vartheta\mathbb{E}^\alpha\left[\Phi_{10}^n\right]\right)\\
&\quad+\gamma\left((n-1)\mathbb{E}^\alpha\left[\Phi_{10}^{n-1}-\Phi_{10}^n\right]+\binom{n-1}{2}\mathbb{E}^\alpha\left[\Phi_{10}^{n-1}-\Phi_{10}^n\right]\right)\\
&\quad+\alpha\Big(\mathbb{E}^\alpha\left[\Phi_{10}^n-\Phi_{11}^n\right]+(n-1)\mathbb{E}^\alpha\left[\Phi_{20}^n-\Phi_{11}^n\right]\Big)\\
&\quad+n\lambda\mathbb{E}^\alpha\left[\Phi_{01}^n\right]-\frac{1}{2}\left(\vartheta_{\bullet}\mathbb{E}^\alpha\left[\Phi_{00}^n\right]-\bar\vartheta\mathbb{E}^\alpha\left[\Phi_{01}^n\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
&-\gamma\left(\binom{n}{2}\mathbb{E}^\alpha\left[\Phi_{01}^{n-1}-\Phi_{01}^{n}\right]+n\mathbb{E}^\alpha\left[\Phi_{10}^{n}-\Phi_{01}^{n}\right]\right)\\
&-\alpha\left(\mathbb{E}^\alpha\left[\Phi_{01}^{n}-\Phi_{02}^{n}\right]+n\mathbb{E}^\alpha\left[\Phi_{11}^{n}-\Phi_{02}^{n}\right]\right)\\
=&-\left(n\lambda+\frac{\bar\vartheta}{2}+\gamma\binom{n+1}{2}\right)\mathbb{E}^\alpha\left[\Phi_{10}^{n}-\Phi_{01}^{n}\right]+\gamma\binom{n}{2}\mathbb{E}^\alpha\left[\Phi_{10}^{n-1}-\Phi_{01}^{n-1}\right]\\
&+\alpha\mathbb{E}^\alpha\left[\Phi_{10}^{n}-\Phi_{01}^{n}\right]+\alpha\underbrace{\mathbb{E}^\alpha\left[(n-1)\Phi_{20}^{n}-2n\Phi_{11}^{n}+(n+1)\Phi_{02}^{n}\right]}_{=\mathbb{E}^0\left[(n-1)\Phi_{20}^{n}-2n\Phi_{11}^{n}+(n+1)\Phi_{02}^{n}\right]+\mathcal{O}(\alpha)}\\
=&-\left(n\lambda+\frac{\bar\vartheta}{2}+\gamma\binom{n+1}{2}\right)\mathbb{E}^\alpha\left[\Phi_{10}^{n}-\Phi_{01}^{n}\right]+\gamma\binom{n}{2}\mathbb{E}^\alpha\left[\Phi_{10}^{n-1}-\Phi_{01}^{n-1}\right]\\
&+\alpha\mathbb{E}^0\left[(n-1)\Phi_{20}^{n}-2n\Phi_{11}^{n}+(n+1)\Phi_{02}^{n}\right]+\mathcal{O}(\alpha^2)\\
=&-\left(n\lambda+\frac{\bar\vartheta}{2}+\gamma\binom{n+1}{2}\right)\mathbb{E}^\alpha\left[\Phi_{10}^{n}-\Phi_{01}^{n}\right]+\gamma\binom{n}{2}\mathbb{E}^\alpha\left[\Phi_{10}^{n-1}-\Phi_{01}^{n-1}\right]\\
&+\alpha\Theta(1-\Theta)\mathbb{E}^0\left[e^{-\lambda L_n^{(n+2)}}\left((n-1)e^{-\bar\vartheta R_{12}^{(n+2)}}-2ne^{-\bar\vartheta R_{1,n+1}^{(n+2)}}+(n+1)e^{-\bar\vartheta R_{n+1,n+2}^{(n+2)}}\right)\right]+\mathcal{O}(\alpha^2)
\end{aligned}
$$

where we use (1.4.13) and the fact that $(n-1)-2n+(n+1)=0$, in the last step.
With $b_n := \mathbb{E}^0\left[e^{-\lambda L_n^{(n+2)}}\left((n-1)e^{-\bar\vartheta R_{12}^{(n+2)}}-2ne^{-\bar\vartheta R_{1,n+1}^{(n+2)}}+(n+1)e^{-\bar\vartheta R_{n+1,n+2}^{(n+2)}}\right)\right]$ we have

$$
\left(n\lambda+\frac{\bar\vartheta}{2}+\gamma\binom{n+1}{2}\right)a_n=\gamma\binom{n}{2}a_{n-1}+\Theta(1-\Theta)b_n+\mathcal{O}(\alpha).
$$

In order to obtain a recursion for b_n, consider a coalescent with $n+2$ lines and distinguish the following cases for the first step:

1. Coalescence of lines among the first n lines, except for lines 1,2 (rate $\binom{n}{2}-1$);
2. Coalescence of lines 1,2 (rate 1);
3. Coalescence of lines $n+1$ and 1 (rate 1);
4. Coalescence of lines $n+1$ and one of $2,...,n$ (rate $n-1$);
5. Coalescence of lines $n+1$ and $n+2$ (rate 1);
6. Coalescence of lines $n+2$ and one of $1,...,n$ (rate n).

As all following quantities are computed under neutrality, i.e. $\alpha=0$, we will write $\mathbb{E}[\cdot]$ instead of $\mathbb{E}^0[\cdot]$.

$$
\begin{aligned}
&\mathbb{E}\left[e^{-\lambda L_n^{(n+2)}}\left((n-1)e^{-\bar\vartheta R_{12}^{(n+2)}}-2ne^{-\bar\vartheta R_{1,n+1}^{(n+2)}}+(n+1)e^{-\bar\vartheta R_{n+1,n+2}^{(n+2)}}\right)\right]\\
&=\frac{\left(\binom{n}{2}-1\right)}{\binom{n+2}{2}}\mathbb{E}\Bigg[e^{-n\lambda T_{n+2}}e^{-\lambda L_{n-1}^{(n+1)}}\Bigg((n-1)e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{12}^{(n+1)}}\\
&\qquad\qquad -2ne^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{1,n}^{(n+1)}}+(n+1)e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{n,n+1}^{(n+1)}}\Bigg)\Bigg]\\
&+\frac{1}{\binom{n+2}{2}}\cdot\mathbb{E}\Bigg[e^{-n\lambda T_{n+2}}e^{-\lambda L_{n-1}^{(n+1)}}\Bigg((n-1)e^{-\bar\vartheta T_{n+2}}\\
&\qquad\qquad -2ne^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{1,n}^{(n+1)}}+(n+1)e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{n,n+1}^{(n+1)}}\Bigg)\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{\binom{n+2}{2}} \cdot \mathbb{E}\Big[e^{-n\lambda T_{n+2}} e^{-\lambda L_n^{(n+1)}} \Big((n-1)e^{-\bar{\vartheta} T_{n+2}} e^{-\bar{\vartheta} R_{12}^{(n+1)}} \\
&\qquad\qquad\qquad -2ne^{-\bar{\vartheta} T_{n+2}} + (n+1)e^{-\bar{\vartheta} T_{n+2}} e^{-\bar{\vartheta} R_{1,n+1}^{(n+1)}}\Big)\Big] \\
&+ \frac{(n-1)}{\binom{n+2}{2}} \cdot \mathbb{E}\Big[e^{-n\lambda T_{n+2}} e^{-\lambda L_n^{(n+1)}} \Big((n-1)e^{-\bar{\vartheta} T_{n+2}} e^{-\bar{\vartheta} R_{12}^{(n+1)}} \\
&\qquad\qquad\qquad -2ne^{-\bar{\vartheta} T_{n+2}} e^{-\bar{\vartheta} R_{12}^{(n+1)}} + (n+1)e^{-\bar{\vartheta} T_{n+2}} e^{-\bar{\vartheta} R_{1,n+1}^{(n+1)}}\Big)\Big] \\
&+ \frac{1}{\binom{n+2}{2}} \cdot \mathbb{E}\Big[e^{-n\lambda T_{n+2}} e^{-\lambda L_n^{(n+1)}} \Big((n-1)e^{-\bar{\vartheta} T_{n+2}} e^{-\bar{\vartheta} R_{12}^{(n+1)}} \\
&\qquad\qquad\qquad -2ne^{-\bar{\vartheta} T_{n+2}} e^{-\bar{\vartheta} R_{1,n+1}^{(n+1)}} + (n+1)e^{-\bar{\vartheta} T_{n+2}}\Big)\Big] \\
&+ \frac{n}{\binom{n+2}{2}} \cdot \mathbb{E}\Big[e^{-n\lambda T_{n+2}} e^{-\lambda L_n^{(n+1)}} \Big((n-1)e^{-\bar{\vartheta} T_{n+2}} e^{-\bar{\vartheta} R_{12}^{(n+1)}} \\
&\qquad\qquad\qquad -2ne^{-\bar{\vartheta} T_{n+2}} e^{-\bar{\vartheta} R_{1,n+1}^{(n+1)}} + (n+1)e^{-\bar{\vartheta} T_{n+2}} e^{-\bar{\vartheta} R_{1,n+1}^{(n+1)}}\Big)\Big] \\
&= \frac{\mathbb{E}\Big[e^{-n\lambda T_{n+2}} e^{-\bar{\vartheta} T_{n+2}}\Big]}{\binom{n+2}{2}} \\
&\quad \cdot \Bigg\{ \binom{n}{2} \cdot \mathbb{E}\Big[e^{-\lambda L_{n-1}^{(n+1)}} \Big((n-1)e^{-\bar{\vartheta} R_{12}^{(n+1)}} - 2ne^{-\bar{\vartheta} R_{1,n}^{(n+1)}} + (n+1)e^{-\bar{\vartheta} R_{n,n+1}^{(n+1)}}\Big)\Big] \\
&\qquad + (n-1)\mathbb{E}\Big[e^{-\lambda L_{n-1}^{(n+1)}} \Big(1 - e^{-\bar{\vartheta} R_{12}^{(n+1)}}\Big)\Big] \\
&\qquad + \mathbb{E}\Bigg[e^{-\lambda L_n^{(n+1)}} \cdot \Bigg[\Big((n-1) + (n-1)(n-1-2n) + (n-1) + n(n-1)\Big) \cdot e^{-\bar{\vartheta} R_{12}^{(n+1)}} \\
&\qquad\qquad + \Big((n+1) + (n-1)(n+1) - 2n + n(-2n+n+1)\Big) \cdot e^{-\bar{\vartheta} R_{1,n+1}^{(n+1)}} \\
&\qquad\qquad + \Big(-2n + (n+1)\Big)\Bigg]\Bigg]\Bigg\} \\
&= \frac{1}{\binom{n+2}{2}} \cdot \frac{\gamma\binom{n+2}{2}}{\gamma\binom{n+2}{2} + (n\lambda + \bar{\vartheta})} \\
&\quad \cdot \Bigg\{ \binom{n}{2} \cdot \mathbb{E}\Big[e^{-\lambda L_{n-1}^{(n+1)}} \Big((n-1)e^{-\bar{\vartheta} R_{12}^{(n+1)}} - 2ne^{-\bar{\vartheta} R_{1,n}^{(n+1)}} + (n+1)e^{-\bar{\vartheta} R_{n,n+1}^{(n+1)}}\Big)\Big] \\
&\qquad + (n-1)\mathbb{E}\Big[e^{-\lambda L_{n-1}^{(n+1)}} \Big(1 - e^{-\bar{\vartheta} R_{12}^{(n+1)}}\Big)\Big] \\
&\qquad + \mathbb{E}\Big[e^{-\lambda L_n^{(n+1)}} \Big((n-1)e^{-\bar{\vartheta} R_{12}^{(n+1)}} - (n-1)\Big)\Big] \Bigg\} \\
&= \frac{\gamma}{\gamma\binom{n+2}{2} + (n\lambda + \bar{\vartheta})} \\
&\quad \cdot \Bigg\{ \binom{n}{2} \cdot \mathbb{E}\Big[e^{-\lambda L_{n-1}^{(n+1)}} \Big((n-2)e^{-\bar{\vartheta} R_{12}^{(n+1)}} - 2(n-1)e^{-\bar{\vartheta} R_{1,n}^{(n+1)}} + ne^{-\bar{\vartheta} R_{n,n+1}^{(n+1)}}\Big)\Big]
\end{aligned}
$$

$$\begin{aligned}
&+\binom{n}{2}\cdot\mathbb{E}\left[e^{-\lambda L_{n-1}^{(n+1)}}\left(e^{-\bar\vartheta R_{12}^{(n+1)}}-2e^{-\bar\vartheta R_{1,n}^{(n+1)}}+e^{-\bar\vartheta R_{n,n+1}^{(n+1)}}\right)\right]\\
&+(n-1)\mathbb{E}\left[\left(e^{-\lambda L_{n-1}^{(n+1)}}-e^{-\lambda L_{n}^{(n+1)}}\right)\left(1-e^{-\bar\vartheta R_{12}^{(n+1)}}\right)\right]\Bigg\}.
\end{aligned}$$

With

$$c_n := \mathbb{E}\left[e^{-\lambda L_{n-1}^{(n+1)}}\left(e^{-\bar\vartheta R_{12}^{(n+1)}}-2e^{-\bar\vartheta R_{1,n}^{(n+1)}}+e^{-\bar\vartheta R_{n,n+1}^{(n+1)}}\right)\right]$$

and

$$d_n := \mathbb{E}\left[\left(e^{-\lambda L_{n-1}^{(n+1)}}-e^{-\lambda L_{n}^{(n+1)}}\right)\left(1-e^{-\bar\vartheta R_{12}^{(n+1)}}\right)\right]$$

we get

$$\left(\gamma\binom{n+2}{2}+n\lambda+\bar\vartheta\right)\cdot b_n=\gamma\binom{n}{2}\cdot b_{n-1}+\gamma\binom{n}{2}\cdot c_{n-1}+\gamma(n-1)\cdot d_n.$$

For c_n, we use the same coalescent, and by distinguishing the six cases, we write

$$\begin{aligned}
&\mathbb{E}\left[e^{-\lambda L_{n}^{(n+2)}}\left(e^{-\bar\vartheta R_{12}^{(n+2)}}-2e^{-\bar\vartheta R_{1,n+1}^{(n+2)}}+e^{-\bar\vartheta R_{n+1,n+2}^{(n+2)}}\right)\right]\\
=&\frac{\binom{n}{2}-1}{\binom{n+2}{2}}\cdot\mathbb{E}\Bigg[e^{-n\lambda T_{n+2}}e^{-\lambda L_{n-1}^{(n+1)}}\Bigg(e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{12}^{(n+1)}}\\
&\qquad\qquad -2e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{1,n}^{(n+1)}}+e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{n,n+1}^{(n+1)}}\Bigg)\Bigg]\\
&+\frac{1}{\binom{n+2}{2}}\cdot\mathbb{E}\left[e^{-n\lambda T_{n+2}}e^{-\lambda L_{n-1}^{(n+1)}}\left(e^{-\bar\vartheta T_{n+2}}-2e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{1,n}^{(n+1)}}+e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{n,n+1}^{(n+1)}}\right)\right]\\
&+\frac{1}{\binom{n+2}{2}}\cdot\mathbb{E}\left[e^{-n\lambda T_{n+2}}e^{-\lambda L_{n}^{(n+1)}}\left(e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{12}^{(n+1)}}-2e^{-\bar\vartheta T_{n+2}}+e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{1,n+1}^{(n+1)}}\right)\right]\\
&+\frac{(n-1)}{\binom{n+2}{2}}\cdot\mathbb{E}\Bigg[e^{-n\lambda T_{n+2}}e^{-\lambda L_{n}^{(n+1)}}\Bigg(e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{12}^{(n+1)}}\\
&\qquad\qquad -2e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{12}^{(n+1)}}+e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{1,n+1}^{(n+1)}}\Bigg)\Bigg]\\
&+\frac{1}{\binom{n+2}{2}}\cdot\mathbb{E}\left[e^{-n\lambda T_{n+2}}e^{-\lambda L_{n}^{(n+1)}}\left(e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{12}^{(n+1)}}-2e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{1,n+1}^{(n+1)}}+e^{-\bar\vartheta T_{n+2}}\right)\right]\\
&+\frac{n}{\binom{n+2}{2}}\cdot\mathbb{E}\Bigg[e^{-n\lambda T_{n+2}}e^{-\lambda L_{n}^{(n+1)}}\Bigg(e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{12}^{(n+1)}}\\
&\qquad\qquad -2e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{1,n+1}^{(n+1)}}+e^{-\bar\vartheta T_{n+2}}e^{-\bar\vartheta R_{1,n+1}^{(n+1)}}\Bigg)\Bigg]\\
=&\frac{\mathbb{E}\left[e^{-(n\lambda+\bar\vartheta)T_{n+2}}\right]}{\binom{n+2}{2}}\\
&\cdot\Bigg\{\binom{n}{2}\cdot\mathbb{E}\left[e^{-\lambda L_{n-1}^{(n+1)}}\left(e^{-\bar\vartheta R_{12}^{(n+1)}}-2e^{-\bar\vartheta R_{1,n}^{(n+1)}}+e^{-\bar\vartheta R_{n,n+1}^{(n+1)}}\right)\right]\\
&\quad+\mathbb{E}\left[e^{-\lambda L_{n-1}^{(n+1)}}\left(1-e^{-\bar\vartheta R_{12}^{(n+1)}}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}\left[e^{-\lambda L_n^{(n+1)}}\left(\big(1-(n-1)+1+n\big)\cdot e^{-\bar{\vartheta}R_{12}^{(n+1)}}\right.\right.\\
&\qquad\qquad\qquad\qquad\left.\left.+\big(1+(n-1)-2-n\big)\cdot e^{-\bar{\vartheta}R_{1,n+1}^{(n+1)}}-1\right)\right]\Bigg\}\\
=&\frac{1}{\binom{n+2}{2}}\cdot\frac{\gamma\binom{n+2}{2}}{\gamma\binom{n+2}{2}+(n\lambda+\bar{\vartheta})}\\
&\cdot\Bigg\{\binom{n}{2}\cdot\mathbb{E}\left[e^{-\lambda L_{n-1}^{(n+1)}}\left(e^{-\bar{\vartheta}R_{12}^{(n+1)}}-2e^{-\bar{\vartheta}R_{1,n}^{(n+1)}}+e^{-\bar{\vartheta}R_{n,n+1}^{(n+1)}}\right)\right]\\
&+\mathbb{E}\left[e^{-\lambda L_{n-1}^{(n+1)}}\left(1-e^{-\bar{\vartheta}R_{12}^{(n+1)}}\right)\right]\\
&+2\mathbb{E}\left[e^{-\lambda L_n^{(n+1)}}\left(e^{-\bar{\vartheta}R_{12}^{(n+1)}}-e^{-\bar{\vartheta}R_{1,n+1}^{(n+1)}}\right)\right]+\mathbb{E}\left[e^{-\lambda L_n^{(n+1)}}\left(1-e^{-\bar{\vartheta}R_{12}^{(n+1)}}\right)\right]\Bigg\}\\
=&\frac{\gamma}{\gamma\binom{n+2}{2}+(n\lambda+\bar{\vartheta})}\\
&\cdot\Bigg\{\binom{n}{2}\cdot\mathbb{E}\left[\left(e^{-\lambda L_{n-1}^{(n+1)}}-e^{-\lambda L_n^{(n+1)}}\right)\left(e^{-\bar{\vartheta}R_{12}^{(n+1)}}-2e^{-\bar{\vartheta}R_{1,n}^{(n+1)}}+e^{-\bar{\vartheta}R_{n,n+1}^{(n+1)}}\right)\right]\\
&+\mathbb{E}\left[e^{-\lambda L_{n-1}^{(n+1)}}\left(1-e^{-\bar{\vartheta}R_{12}^{(n+1)}}\right)\right]+2\mathbb{E}\left[e^{-\lambda L_n^{(n+1)}}\cdot\left(e^{-\bar{\vartheta}R_{12}^{(n+1)}}-e^{-\bar{\vartheta}R_{1,n+1}^{(n+1)}}\right)\right]\Bigg\}.
\end{aligned}
$$

We define

$$
e_n:=\mathbb{E}\left[e^{-\lambda L_n^{(n+1)}}\left(e^{-\bar{\vartheta}R_{12}^{(n+1)}}-e^{-\bar{\vartheta}R_{1,n+1}^{(n+1)}}\right)\right]
$$

and get

$$
\left(\gamma\binom{n+2}{2}+n\lambda+\bar{\vartheta}\right)\cdot c_n=\gamma\binom{n}{2}\cdot c_{n-1}+\gamma d_n+2\gamma e_n.
$$

For d_n we compute

$$
\begin{aligned}
\mathbb{P}(I_{12}^{(n)}=i)&=\frac{\binom{n}{2}-1}{\binom{n}{2}}\cdot\frac{\binom{n-1}{2}-1}{\binom{n-1}{2}}\cdots\frac{\binom{i+1}{2}-1}{\binom{i+1}{2}}\cdot\frac{1}{\binom{i}{2}}=\frac{1}{\binom{i}{2}}\cdot\prod_{k=i+1}^{n}\frac{\binom{k}{2}-1}{\binom{k}{2}}\\
&=\frac{2}{i(i-1)}\cdot\prod_{k=i+1}^{n}\frac{(k+1)(k-2)}{k(k-1)}=\frac{2(n+1)}{(n-1)(i+1)i}.
\end{aligned}
$$

Then we have (by omitting the superscript $(n+1)$ as explained in Remark 1.18)

$$
\begin{aligned}
g_n&:=\mathbb{E}\left[e^{-\lambda L_n}e^{-\bar{\vartheta}R_{12}}\right]\\
&=\mathbb{E}\left[e^{-\lambda\sum_{k=2}^{n}kT_k}e^{-\bar{\vartheta}\sum_{k=I_{12}^{(n)}}^{n}T_k}\right]=\mathbb{E}\left[e^{-\left(\lambda\sum_{k=2}^{I_{12}^{(n)}-1}kT_k+\sum_{k=I_{12}^{(n)}}^{n}(k\lambda+\bar{\vartheta})T_k\right)}\right]\\
&=\sum_{i=2}^{n}\mathbb{P}(I_{12}^{(n)}=i)\cdot\prod_{k=2}^{i-1}\frac{k(k-1)\gamma}{k(k-1)\gamma+2k\lambda}\cdot\prod_{k=i}^{n}\frac{k(k-1)\gamma}{k(k-1)\gamma+2(k\lambda+\bar{\vartheta})}
\end{aligned}
$$

$$= \frac{2(n+1)}{(n-1)} \sum_{i=2}^{n} \frac{1}{(i+1)i} \cdot \prod_{k=2}^{i-1} \frac{(k-1)\gamma}{(k-1)\gamma + 2\lambda} \cdot \prod_{k=i}^{n} \frac{k(k-1)\gamma}{k(k-1)\gamma + 2(k\lambda + \bar{\vartheta})}$$

and we obtain that

$$d_n = f_{n-1} - f_n - g_{n-1} + g_n.$$

For g_1 we have

$$g_1 = \mathbb{E}\left[e^{-\bar{\vartheta} T_2}\right] = \frac{\gamma}{\gamma + \bar{\vartheta}}.$$

Finally, for e_n, we again use a recursion. Consider a coalescent with $n+1$ lines and make a first-step-analysis. In this first step, we distinguish four cases:

1. Coalescence of lines 1 or 2 with one of $3, ..., n$; rate $\binom{n}{2} - 1$
2. Coalescence of lines 1 and 2; rate 1
3. Coalescence of lines $n+1$ and 1; rate 1
4. Coalescence of lines $n+1$ and one of $2, ..., n$; rate $n-1$.

$$\begin{aligned}
&\mathbb{E}\left[e^{-\lambda L_n^{(n+1)}} \cdot \left(e^{-\bar{\vartheta} R_{12}^{(n+1)}} - e^{-\bar{\vartheta} R_{1,n+1}^{(n+1)}}\right)\right] \\
=&\frac{\binom{n}{2} - 1}{\binom{n+1}{2}} \cdot \mathbb{E}\left[e^{-n\lambda T_{n+1}} e^{-\lambda L_{n-1}^{(n)}} \cdot \left(e^{-\bar{\vartheta} T_{n+1}} e^{-\bar{\vartheta} R_{12}^{(n)}} - e^{-\bar{\vartheta} T_{n+1}} e^{-\bar{\vartheta} R_{1,n}^{(n)}}\right)\right] \\
&+ \frac{1}{\binom{n+1}{2}} \cdot \mathbb{E}\left[e^{-n\lambda T_{n+1}} e^{-\lambda L_{n-1}^{(n)}} \cdot \left(e^{-\bar{\vartheta} T_{n+1}} - e^{-\bar{\vartheta} T_{n+1}} e^{-\bar{\vartheta} R_{1,n}^{(n)}}\right)\right] \\
&+ \frac{\binom{n}{2} - 1}{\binom{n+1}{2}} \cdot \mathbb{E}\left[e^{-n\lambda T_{n+1}} e^{-\lambda L_n^{(n)}} \cdot \left(e^{-\bar{\vartheta} T_{n+1}} e^{-\bar{\vartheta} R_{12}^{(n)}} - e^{-\bar{\vartheta} T_{n+1}}\right)\right] \\
&+ \frac{\binom{n}{2} - 1}{\binom{n+1}{2}} \cdot \mathbb{E}\left[e^{-n\lambda T_{n+1}} e^{-\lambda L_n^{(n)}} \cdot \left(e^{-\bar{\vartheta} T_{n+1}} e^{-\bar{\vartheta} R_{12}^{(n)}} - e^{-\bar{\vartheta} T_{n+1}} e^{-\bar{\vartheta} R_{12}^{(n)}}\right)\right] \\
=&\frac{\mathbb{E}\left[e^{-(n\lambda + \bar{\vartheta}) T_{n+1}}\right]}{\binom{n+1}{2}} \cdot \left\{ \binom{n}{2} \cdot \mathbb{E}\left[e^{-\lambda L_{n-1}^{(n)}} \cdot \left(e^{-\bar{\vartheta} R_{12}^{(n)}} - e^{-\bar{\vartheta} R_{1,n}^{(n)}}\right)\right] \right. \\
&\qquad \left. + \mathbb{E}\left[e^{-\lambda L_{n-1}^{(n)}} \cdot \left(1 - e^{-\bar{\vartheta} R_{12}^{(n)}}\right)\right] - \mathbb{E}\left[e^{-\lambda L_n^{(n)}} \cdot \left(1 - e^{-\bar{\vartheta} R_{12}^{(n)}}\right)\right] \right\} \\
=&\frac{\gamma}{\gamma \binom{n+1}{2} + n\lambda + \bar{\vartheta}} \\
&\cdot \left\{ \binom{n}{2} \cdot \mathbb{E}\left[e^{-\lambda L_{n-1}^{(n)}} \cdot \left(e^{-\bar{\vartheta} R_{12}^{(n)}} - e^{-\bar{\vartheta} R_{1,n}^{(n)}}\right)\right] + \mathbb{E}\left[\left(e^{-\lambda L_{n-1}^{(n)}} - e^{-\lambda L_n^{(n)}}\right) \cdot \left(1 - e^{-\bar{\vartheta} R_{12}^{(n)}}\right)\right] \right\}.
\end{aligned}$$

Hence we have

$$\left(\gamma \binom{n+1}{2} + n\lambda + \bar{\vartheta}\right) \cdot e_n = \gamma \binom{n}{2} \cdot e_{n-1} + \gamma d_n.$$

□

Proof of Corollary 1.20. We first note that

$$x_n|_{\lambda=0} = a_n|_{\lambda=0} = b_n|_{\lambda=0} = c_n|_{\lambda=0} = d_n|_{\lambda=0} = e_n|_{\lambda=0} = 0.$$

We define $\widetilde{x}_n := \mathbb{E}^0[L_n] - \mathbb{E}^\alpha[L_n]$ and observe

$$\begin{aligned}
\widetilde{x}_n &= \frac{\partial}{\partial\lambda} x_n \bigg|_{\lambda=0} \\
&= \frac{\partial}{\partial\lambda}\left(\frac{\gamma\binom{n}{2}}{\gamma\binom{n}{2}+n\lambda}x_{n-1} + \alpha^2\cdot\frac{n}{\gamma\binom{n}{2}+n\lambda}a_n\right)\Bigg|_{\lambda=0} \\
&= \left(\frac{\partial}{\partial\lambda}\frac{\gamma\binom{n}{2}}{\gamma\binom{n}{2}+n\lambda}\right)\Bigg|_{\lambda=0}\cdot x_{n-1}|_{\lambda=0} + \frac{\gamma\binom{n}{2}}{\gamma\binom{n}{2}+n\lambda}\Bigg|_{\lambda=0}\cdot\left(\frac{\partial}{\partial\lambda}x_{n-1}\right)\Bigg|_{\lambda=0} \\
&\quad + \alpha^2\cdot\left[\left(\frac{\partial}{\partial\lambda}\frac{n}{\gamma\binom{n}{2}+n\lambda}\right)\Bigg|_{\lambda=0}\cdot a_n|_{\lambda=0} + \frac{n}{\gamma\binom{n}{2}+n\lambda}\Bigg|_{\lambda=0}\cdot\left(\frac{\partial}{\partial\lambda}a_n\right)\Bigg|_{\lambda=0}\right] \\
&= \left(\frac{\partial}{\partial\lambda}x_{n-1}\right)\Bigg|_{\lambda=0} + \alpha^2\frac{n}{\gamma\binom{n}{2}}\cdot\left(\frac{\partial}{\partial\lambda}a_n\right)\Bigg|_{\lambda=0}.
\end{aligned}$$

With

$$\widetilde{a}_n := \left(\frac{\partial}{\partial\lambda}a_n\right)\Bigg|_{\lambda=0}$$

we get

$$\gamma\binom{n}{2}\widetilde{x}_n = \gamma\binom{n}{2}\widetilde{x}_{n-1} + \alpha^2 n\widetilde{a}_n.$$

With the same steps as before for the computation of $\widetilde{x}_n$ we obtain

$$\begin{aligned}
\left(\gamma\binom{n+1}{2}+\frac{\bar\vartheta}{2}\right)\widetilde{a}_n &= \gamma\binom{n}{2}\widetilde{a}_{n-1} + \widetilde{b}_n + \mathcal{O}(\alpha), \\
\left(\gamma\binom{n+2}{2}+\bar\vartheta\right)\widetilde{b}_n &= \gamma\binom{n}{2}\widetilde{b}_{n-1} + \gamma\binom{n}{2}\widetilde{c}_{n-1} + \gamma(n-1)\widetilde{d}_n, \\
\left(\gamma\binom{n+2}{2}+\bar\vartheta\right)\widetilde{c}_n &= \gamma\binom{n}{2}\widetilde{c}_{n-1} + 2\gamma\widetilde{e}_n + \gamma\widetilde{d}_n, \\
\left(\gamma\binom{n+1}{2}+\bar\vartheta\right)\widetilde{e}_n &= \gamma\binom{n}{2}\widetilde{e}_{n-1} + \gamma\widetilde{d}_n,
\end{aligned}$$

with

$$\widetilde{b}_n := \left(\frac{\partial}{\partial\lambda}b_n\right)\Bigg|_{\lambda=0}, \quad \widetilde{c}_n := \left(\frac{\partial}{\partial\lambda}c_n\right)\Bigg|_{\lambda=0}, \quad \widetilde{d}_n := \left(\frac{\partial}{\partial\lambda}d_n\right)\Bigg|_{\lambda=0}, \quad \widetilde{e}_n := \left(\frac{\partial}{\partial\lambda}e_n\right)\Bigg|_{\lambda=0}.$$

For the computation of $\widetilde{d}_n$ we first define $\widetilde{f}_n := \left(\frac{\partial}{\partial\lambda} f_n\right)|_{\lambda=0}$ as well as $\widetilde{g}_n := \left(\frac{\partial}{\partial\lambda} g_n\right)|_{\lambda=0}$. We recall that

$$f_n = \prod_{k=2}^{n} \frac{(k-1)\gamma}{(k-1)\gamma + 2\lambda} = \frac{(n-1)\gamma}{(n-1)\gamma + 2\lambda} \cdot f_{n-1}$$

so with $f_n|_{\lambda=0} = 1$ we obtain

$$\begin{aligned}
\widetilde{f}_n &:= \left(\frac{\partial}{\partial\lambda} f_n\right)\Bigg|_{\lambda=0} \\
&= \left(\frac{\partial}{\partial\lambda}\frac{(n-1)\gamma}{(n-1)\gamma+2\lambda}\right)\Bigg|_{\lambda=0} \cdot f_{n-1}|_{\lambda=0} + \frac{(n-1)\gamma}{(n-1)\gamma+2\lambda}\Bigg|_{\lambda=0} \cdot \left(\frac{\partial}{\partial\lambda} f_{n-1}\right)\Bigg|_{\lambda=0} \\
&= -\frac{2}{(n-1)\gamma} + \widetilde{f}_{n-1} \\
&= -\frac{2}{\gamma}\sum_{k=2}^{n} \frac{1}{(k-1)}.
\end{aligned}$$

With equation (1.4.10) we obtain

$$\widetilde{d}_n = \widetilde{f}_{n-1} - \widetilde{f}_n - \widetilde{g}_{n-1} + \widetilde{g}_n = \frac{2}{(n-1)\gamma} - \widetilde{g}_{n-1} + \widetilde{g}_n.$$

Further with

$$\begin{aligned}
&\left(\frac{\partial}{\partial\lambda}\prod_{k=i}^{n}\frac{k(k-1)\gamma}{k(k-1)\gamma+2(k\lambda+\bar{\vartheta})}\right)\Bigg|_{\lambda=0} \\
=&\left(\frac{\partial}{\partial\lambda}\frac{n(n-1)\gamma}{n(n-1)\gamma+2(n\lambda+\bar{\vartheta})}\right)\Bigg|_{\lambda=0} \cdot \prod_{k=i}^{n-1}\frac{k(k-1)\gamma}{k(k-1)\gamma+2(k\lambda+\bar{\vartheta})}\Bigg|_{\lambda=0} \\
&+ \frac{n(n-1)\gamma}{n(n-1)\gamma+2(n\lambda+\bar{\vartheta})}\Bigg|_{\lambda=0} \cdot \left(\frac{\partial}{\partial\lambda}\prod_{k=i}^{n-1}\frac{k(k-1)\gamma}{k(k-1)\gamma+2(k\lambda+\bar{\vartheta})}\right)\Bigg|_{\lambda=0} \\
=&\frac{-2n\cdot n(n-1)\gamma}{(n(n-1)\gamma+2\bar{\vartheta})^2} \cdot \prod_{k=i}^{n-1}\frac{k(k-1)\gamma}{k(k-1)\gamma+2\bar{\vartheta}} \\
&+ \frac{n(n-1)\gamma}{n(n-1)\gamma+2\bar{\vartheta}} \cdot \left(\frac{\partial}{\partial\lambda}\prod_{k=i}^{n-1}\frac{k(k-1)\gamma}{k(k-1)\gamma+2(k\lambda+\bar{\vartheta})}\right)\Bigg|_{\lambda=0} \\
=&-2\sum_{k=i}^{n}\frac{k^2(k-1)\gamma}{(k(k-1)\gamma+2\bar{\vartheta})^2} \cdot \prod_{\substack{l=i \\ l\neq k}}^{n}\frac{k(k-1)\gamma}{k(k-1)\gamma+2\bar{\vartheta}}
\end{aligned}$$

we get

$$\widetilde{g}_n = \frac{2(n+1)}{(n-1)}\sum_{i=2}^{n}\frac{1}{(i+1)i}\left[\left(\frac{\partial}{\partial\lambda}\prod_{k=2}^{i-1}\frac{(k-1)\gamma}{(k-1)\gamma+2\lambda}\right)\Bigg|_{\lambda=0} \cdot \prod_{k=i}^{n}\frac{k(k-1)\gamma}{k(k-1)\gamma+2(k\lambda+\bar{\vartheta})}\Bigg|_{\lambda=0}\right.$$

$$
\begin{aligned}
&\qquad\qquad + \prod_{k=2}^{i-1} \frac{(k-1)\gamma}{(k-1)\gamma+2\lambda}\bigg|_{\lambda=0} \left(\frac{\partial}{\partial\lambda} \prod_{k=i}^{n} \frac{k(k-1)\gamma}{k(k-1)\gamma+2(k\lambda+\bar{\vartheta})} \right)\bigg|_{\lambda=0} \Bigg] \\
&= \frac{-4(n+1)}{(n-1)} \sum_{i=2}^{n} \frac{1}{(i+1)i} \Bigg[\left(\frac{1}{\gamma} \cdot \sum_{k=2}^{i-1} \frac{1}{(k-1)} \right) \cdot \prod_{k=i}^{n} \frac{k(k-1)\gamma}{k(k-1)\gamma+2\bar{\vartheta}} \\
&\qquad\qquad + \sum_{k=i}^{n} \frac{k^2(k-1)\gamma}{(k(k-1)\gamma+2\bar{\vartheta})^2} \cdot \prod_{\substack{l=i \\ l\neq k}}^{n} \frac{k(k-1)\gamma}{k(k-1)\gamma+2\bar{\vartheta}} \Bigg].
\end{aligned}
$$

□

Proof of Theorem 1.26. Let $y_n := \mathbb{E}^{\alpha,h}[\Phi^n_{00}] - \mathbb{E}^0[\Phi^n_{00}]$.
Again with (1.6.1) and (1.3.10) we have

$$
\begin{aligned}
0 &= \mathbb{E}\left[G^{\alpha,h}\Phi^n_{00}(\mathcal{U}^\alpha_\infty) - G^{\alpha,h}\Phi^n_{00}(\mathcal{U}^0_\infty) \right] \\
&= -n\lambda \mathbb{E}^{\alpha,h}\left[\Phi^n_{00}\right] + \gamma\binom{n}{2}\left(\mathbb{E}^{\alpha,h}\left[\Phi^{n-1}_{00}\right] - \mathbb{E}^{\alpha,h}\left[\Phi^n_{00}\right] \right) \\
&\quad + \alpha\cdot\mathbb{E}^{\alpha,h}\Big[-nh\Phi^n_{0,1} + nh\Phi^n_{1,0} + n(1-2h)\Phi^n_{1,1} - n(1-2h)\Phi^n_{0,2} \Big] \\
&\quad + n\lambda\mathbb{E}^0\left[\Phi^n_{00}\right] + \gamma\binom{n}{2}\left(\mathbb{E}^0\left[\Phi^{n-1}_{00}\right] - \mathbb{E}^0\left[\Phi^n_{00}\right] \right) \\
&= -\left(n\lambda + \gamma\binom{n}{2} \right)\left(\mathbb{E}^{\alpha,h}[\Phi^n_{00}] - \mathbb{E}^0[\Phi^n_{00}] \right) + \gamma\binom{n}{2}\left(\mathbb{E}^{\alpha,h}[\Phi^{n-1}_{00}] - \mathbb{E}^0[\Phi^{n-1}_{00}] \right) \\
&\quad + \alpha n\left(h\mathbb{E}^{\alpha,h}\left[\Phi^n_{10} - \Phi^n_{01}\right] + (1-2h)\mathbb{E}^{\alpha,h}\left[\Phi^n_{11} - \Phi^n_{02}\right] \right) \\
&= -\left(n\lambda + \gamma\binom{n}{2} \right)\left(\mathbb{E}^{\alpha,h}[\Phi^n_{00}] - \mathbb{E}^0[\Phi^n_{00}] \right) + \gamma\binom{n}{2}\left(\mathbb{E}^{\alpha,h}[\Phi^{n-1}_{00}] - \mathbb{E}^0[\Phi^{n-1}_{00}] \right) \\
&\quad + \alpha n\left(h\mathbb{E}^0\left[\Phi^n_{10} - \Phi^n_{01}\right] + (1-2h)\mathbb{E}^0\left[\Phi^n_{11} - \Phi^n_{02}\right] + \mathcal{O}(\alpha) \right) \\
&= -\left(n\lambda + \gamma\binom{n}{2} \right)\left(\mathbb{E}^{\alpha,h}[\Phi^n_{00}] - \mathbb{E}^0[\Phi^n_{00}] \right) + \gamma\binom{n}{2}\left(\mathbb{E}^{\alpha,h}[\Phi^{n-1}_{00}] - \mathbb{E}^0[\Phi^{n-1}_{00}] \right) \\
&\quad + \alpha n\Big(h\mathbb{E}^0\left[\Phi^n_{10} - \Phi^n_{01}\right] \\
&\qquad\quad + (1-2h)\Theta(1-\Theta)\mathbb{E}^0\left[e^{-\lambda L^{(n+2)}_n}\cdot e^{-\bar{\vartheta}R^{(n+2)}_{1,n+1}} - e^{-\lambda L^{(n+2)}_n}\cdot e^{-\bar{\vartheta}R^{(n+2)}_{n+1,n+2}} \right] + \mathcal{O}(\alpha) \Big)
\end{aligned}
$$

where we use (1.4.13) in the very last step.
Defining $h_n := \mathbb{E}^0\left[e^{-\lambda L^{(n+2)}_n}\cdot e^{-\bar{\vartheta}R^{(n+2)}_{1,n+1}} - e^{-\lambda L^{(n+2)}_n}\cdot e^{-\bar{\vartheta}R^{(n+2)}_{n+1,n+2}} \right]$ we get

$$
\left(n\lambda + \gamma\binom{n}{2} \right) y_n = \gamma\binom{n}{2} y_{n-1} + \alpha n(1-2h)\Theta(1-\Theta)\cdot h_n + \mathcal{O}(\alpha^2).
$$

We use the same notation as in the proof of Theorem 1.15. Consider a coalescent with $n+2$ lines . We distinguish the following cases:

1. Coalescence of lines among the first n lines (rate $\binom{n}{2}$);

2. Coalescence of lines $n+1$ and 1 (rate 1);

3. Coalescence of lines $n+1$ and one of $2, ..., n$ (rate $n-1$);
4. Coalescence of lines $n+1$ and $n+2$ (rate 1);
5. Coalescence of lines $n+2$ and one of $1, ..., n$ (rate n).

Again writing $\mathbb{E}[.]$ instead of $\mathbb{E}^0[.]$ we obtain

$$\begin{aligned}
&\mathbb{E}\left[e^{-\lambda L_n^{(n+2)}}\left(e^{-\bar\vartheta R_{1,n+1}^{(n+2)}} - e^{-\bar\vartheta R_{n+1,n+2}^{(n+2)}}\right)\right]\\
=&\frac{\binom{n}{2}}{\binom{n+2}{2}} \cdot \mathbb{E}\left[e^{-n\lambda T_{n+2}} e^{-\lambda L_{n-1}^{(n+1)}}\left(e^{-\bar\vartheta T_{n+2}} e^{-\bar\vartheta R_{1,n}^{(n+1)}} - e^{-\bar\vartheta T_{n+2}} e^{-\bar\vartheta R_{n,n+1}^{(n+1)}}\right)\right]\\
&+\frac{1}{\binom{n+2}{2}} \cdot \mathbb{E}\left[e^{-n\lambda T_{n+2}} e^{-\lambda L_n^{(n+1)}}\left(e^{-\bar\vartheta T_{n+2}} - e^{-\bar\vartheta T_{n+2}} e^{-\bar\vartheta R_{1,n+1}^{(n+1)}}\right)\right]\\
&+\frac{(n-1)}{\binom{n+2}{2}} \cdot \mathbb{E}\left[e^{-n\lambda T_{n+2}} e^{-\lambda L_n^{(n+1)}}\left(e^{-\bar\vartheta T_{n+2}} e^{-\bar\vartheta R_{12}^{(n+1)}} - e^{-\bar\vartheta T_{n+2}} e^{-\bar\vartheta R_{1,n+1}^{(n+1)}}\right)\right]\\
&+\frac{1}{\binom{n+2}{2}} \cdot \mathbb{E}\left[e^{-n\lambda T_{n+2}} e^{-\lambda L_n^{(n+1)}}\left(e^{-\bar\vartheta T_{n+2}} e^{-\bar\vartheta R_{1,n+1}^{(n+1)}} - e^{-\bar\vartheta T_{n+2}}\right)\right]\\
&+\frac{n}{\binom{n+2}{2}} \cdot \mathbb{E}\left[e^{-n\lambda T_{n+2}} e^{-\lambda L_n^{(n+1)}}\left(e^{-\bar\vartheta T_{n+2}} e^{-\bar\vartheta R_{1,n+1}^{(n+1)}} - e^{-\bar\vartheta T_{n+2}} e^{-\bar\vartheta R_{1,n+1}^{(n+1)}}\right)\right]\\
=&\frac{\mathbb{E}\left[e^{-(n\lambda+\bar\vartheta)T_{n+2}}\right]}{\binom{n+2}{2}}\\
&\cdot\Bigg\{\binom{n}{2}\mathbb{E}\left[e^{-\lambda L_{n-1}^{(n+1)}}\left(e^{-\bar\vartheta R_{1,n}^{(n+1)}} - e^{-\bar\vartheta R_{n,n+1}^{(n+1)}}\right)\right]\\
&\quad+\mathbb{E}\left[e^{-\lambda L_n^{(n+1)}} \cdot \Big((n-1)\cdot e^{-\bar\vartheta R_{12}^{(n+1)}} + \big(-1-(n-1)+1\big)\cdot e^{-\bar\vartheta R_{1,n+1}^{(n+1)}}\Big)\right]\Bigg\}\\
=&\frac{1}{\binom{n+2}{2}} \cdot \frac{\gamma\binom{n+2}{2}}{\gamma\binom{n+2}{2} + n\lambda + \bar\vartheta}\\
&\cdot\Bigg\{\binom{n}{2}\mathbb{E}\left[e^{-\lambda L_{n-1}^{(n+1)}}\left(e^{-\bar\vartheta R_{1,n}^{(n+1)}} - e^{-\bar\vartheta R_{n,n+1}^{(n+1)}}\right)\right]\\
&\quad+(n-1)\mathbb{E}\left[e^{-\lambda L_n^{(n+1)}}\left(e^{-\bar\vartheta R_{12}^{(n+1)}} - e^{-\bar\vartheta R_{1,n+1}^{(n+1)}}\right)\right]\Bigg\}.
\end{aligned}$$

Hence we have

$$\left(\gamma\binom{n+2}{2} + n\lambda + \bar\vartheta\right) \cdot h_n = \gamma\binom{n}{2} \cdot h_{n-1} + \gamma(n-1) \cdot e_n.$$

□

Proof of Corollary 1.28. As in Corollary 2.13 we first note that

$$y_n|_{\lambda=0} = 0 \quad \text{and} \quad h_n|_{\lambda=0} = 0$$

and define

$$\widetilde{y}_n := \frac{\partial}{\partial\lambda} y_n \bigg|_{\lambda=0} \quad \text{and} \quad \widetilde{h}_n := \frac{\partial}{\partial\lambda} h_n \bigg|_{\lambda=0}.$$

The same calculations as in the proof of Corollary 2.13 give

$$\begin{aligned}
\widetilde{y}_n &= \widetilde{y}_{n-1} + \alpha(1-2h)\frac{n}{\gamma\binom{n}{2}}\widetilde{h}_n \\
\widetilde{h}_n &= \frac{\gamma\binom{n}{2}}{\gamma\binom{n+2}{2}+\vartheta}\widetilde{h}_{n-1} + \frac{(n-1)}{\gamma\binom{n+2}{2}+\vartheta}\widetilde{e}_n.
\end{aligned}$$

□

Chapter 2

Modifiers of mutation rate in selectively fluctuating environments

2.1 Introduction

In order for evolution to take place it is necessary that new alleles are created to drive genetic diversity. The main force behind this process are mutations which are DNA copying errors that can result in the creation of such new types. The understanding of mutation and the rates at which they occur is therefore of great importance when trying to understand evolutionary processes. In population genetic models such as the Moran model, the Wright-Fisher diffusion or the Fleming-Viot process, a common assumption is a constant mutation rate which is furthermore presumed to be quite low. The reasoning behind this is the following: Most mutations have been detected as being deleterious (Fisher, 1930). An individual with a high mutation rate would therefore be at a disadvantage relative to ones that rarely mutate as they would produce too many types that are not able to compete with the others. However, this argument is only adequate when we assume that these individuals live in an environment which does not undergo any changes. Having a low mutation rate causes individuals that are well-adapted to maintain their selective advantage while individuals with high mutation rates fail to establish themselves within that population as they produce too many deleterious mutations. The assumption of a constant low mutation rates therefore comes together with the assumption of an environment which continuously prefers one type over the other and has long been the subject of discussion. One of the earliest works dealing with this subject goes back to Sturtevant (1937) discussing the fact that mutation rates can differ even within taxa and that genes affecting the mutation rate succumb to selection. Many experiments show that populations which are exposed to environmental changes exhibit an increasing frequency in individuals with high mutation rates, also often called *mutators*. In an environment that challenges the population by exposing it to new surroundings that call for different and new types, mutators are necessary in order to increase the chance of creating individuals with traits needed to survive under the new circumstances (Denamur and Matic, 2006). Though once the environment has found a constant state, higher mutation rates are no longer needed and lower mutation rates are again favored (Wielgoss et al., 2013). The study of the rise and fall in frequency of these mutators and their role in adaptive evolution has been gaining more and more attention over the years especially for microbial evolution, see e.g. Tenaillon et al. (2001) for a review. In an early experiment, Mao et al. (1997) observe a complete takeover

of an *Escherichia coli* population by mutators after successive exposure to a mutagen. After only four rounds of treatments, the population consists to 100% of mutators.
Wielgoss et al. (2013) investigated the long-term effect on mutators. In Lenski's Long-Term Evolution Experiment, 12 populations of *E. coli* have been evolving in a glucose-limited medium for, by now, more than 60000 generations. In one of the populations Wielgoss et al. (2013) notice a decrease in the mutation rate after mutators had established themselves for about 10000 generations. These are only few of many other experiments where the assumption of a constant mutation rate is not appropriate. A fluctuating environment requires individuals that are able to produce new types in order to create a population capable of survival. Motivated by this perception and the above insights from biology, our goal is to complement the view of constant mutation rates by assuming that the mutation rate itself is driven by evolution by using modifier theory, where an additional neutral modifier locus determines the mutation rate at a second locus.

This chapter is structured as follows: In section 2.2 we will first derive a bivariate process that describes the mutation rate and type space in the first variable and the fitness of the type in the second variable which will act according to a fluctuating environment. The process is defined as a solution to a well-posed martingale problem and is called a *Fleming-Viot process with mutation modifier and fluctuating selection.* Our first result is the convergence of this process to a unique limit in the case of a fast fluctuating environment (Theorem 2.6).
To show how the results can be applied we continue in section 2.3 with a special 2-type case where only two mutation rates and two types at the second locus exist. In Theorem 2.12, we compute the fixation probability of the high mutating type depending on the two mutation rates.

This chapter relies on joint work with Peter Pfaffelhuber and Franz Baumdicker. All results have been submitted and are under review. A preprint is available in Baumdicker et al. (2019).

As this chapter deals with so-called *Feller processes*, we recall some definitions. For a more detailed elaboration of this topic we refer to Kallenberg (2002) and Ethier and Kurtz (1986). For a Markov process X we can define a family of operators $(S_t)_{t\geq 0}$ by

$$S_t f(x) := \mathbb{E}_x[f(X_t)] \tag{2.1.1}$$

for a bounded measuralbe function f on E. The *Chapman-Kolmogorov equation* states that it holds

$$S_t S_s = S_{t+s}. \tag{2.1.2}$$

Families of operators satisfying (2.1.2) are called *semigroups.* Further $(S_t)_{t\geq 0}$

1. is a *contraction* if $||S_t f|| = \sup \mathbb{E}_x[f(X_t)] \leq ||f||$,
2. and *positive* if $S_t f(x) = \mathbb{E}_x[f(X_t)] \geq 0$ for $f \geq 0$,
3. and is said to have a *conservative generator* if $S_t 1 = 1$.

If it further holds $S_t f(x) = \mathbb{E}_x[f(X_t)] \xrightarrow{t\to 0} f(x)$, then the semigroup is called *strongly continuous.*

Remark 2.1 (Feller process)**.** A positive, strongly continuous contraction semigroup with conservative generator and $S_t f$ a continuous function on E for all $t \geq 0$ and continuous functions f on E is called a *Feller semigroup.* Reversely, if E is locally compact and separable, a Feller semigroup corresponds to a strong Markov process with sample paths in $\mathcal{D}_E([0,\infty))$; see Ethier and Kurtz (1986), Section 4.3. Such processes are therefore also called *Feller processes.*

2.2 A Fleming-Viot system with mutation modifier and fast fluctuating selection

The model we will be dealing with includes mutation and selection and additionally a fluctuating environment. More precisely, individuals in a large population are assumed to have a modifier locus (A-locus) determining the mutation rate $u \in [0, \vartheta]$ at a second locus (B-locus) with types $v \in [0, 1]$. In addition, the environment fluctuates, meaning that individual types change their fitness at some high rate. Fitness only depends on the type of the B-locus. First of all will derive a bivariate process describing this very scenario. The first variable will correspond to the two loci and the second variable to the fitness of the type at the B-locus. We will be working with a Markov process (X, Z), or rather a sequence of such processes and its limiting process with state space $E := \mathcal{P}([0, \vartheta] \times I) \times \mathcal{C}_L(I)$ for some $L > 0$ and $I := [0, 1]$. Let (u, v) be some sample from the first variable $X_t \in \mathcal{P}([0, \vartheta] \times I)$ at time t where u denotes the allele at the first locus (which we call A-locus), while v is the allele at the second locus (the B-locus). The variable u takes values in $[0, \vartheta]$ and equals the mutation rate of the sampled individual at the B-locus. Upon a mutation (which happens at rate u) the allele at the B-locus is drawn according to a transition kernel $\beta(v, .)$ on I. The second variable $Z_t \in \mathcal{C}_L(I)$ corresponds to the fitness function according to which selection acts on the B-locus. Fluctuations in the environment have impact on the fitness function Z_t which changes along a Poisson process to independent draws from $\nu \in \mathcal{P}(\mathcal{C}_L(I))$. We assume that $\mathbb{E}_\nu[Z(v)] = 0$ for all $v \in I$, i.e. on average, no allele at the B-locus has a fitness advantage.

We collect all assumptions and some notation in the following remark.

Remark 2.2 (Assumption, state space and notation)**.**

1. Let

$$\begin{aligned} \vartheta &\geq 0, && \text{(maximal mutation rate at B-locus)},\\ \alpha &\geq 0, && \text{(selection intensity)},\\ L &\geq 0, && \text{(Lipschitz constant for fitness function)},\\ \sigma &> 0, && \text{(rate of environmental change)},\\ \nu &\in \mathcal{P}(\mathcal{C}_L(I)), && \text{(distribution of random fitness)}, \end{aligned}$$

   and β a transition kernel from I to I (mutation kernel at the B-locus). Throughout, we assume that

$$\mathbb{E}_\nu[Z(v)] = 0, \qquad v \in I. \tag{2.2.1}$$

2. The state space of the Markov process in the next definition will be $E := \mathcal{P}([0,\vartheta] \times I) \times \mathcal{C}_L(I)$. This space is equipped with the product topology, where $\mathcal{C}_L(I)$ is equipped with the topology of uniform convergence, and $\mathcal{P}([0,\vartheta] \times I)$ is equipped with the topology of weak convergence. Note that E is locally compact.

3. For $(u,v) \in [0,\vartheta] \times I$, we say that u is the allele at the A-locus and v is the allele at the B-locus. Denote by $\pi_A : [0,\vartheta] \times I \to [0,\vartheta]$ and $\pi_B : [0,\vartheta] \times I \to I$ the projections on the first and second coordinate, i.e. the A- and B-locus, respectively. More generally, for $k = 1, ..., n$, $\pi_{k,A}$ ($\pi_{k,B}$) is the projection of $([0,\vartheta] \times I)^n$ to the k-th entry at the A-locus (B-locus).

4. For a transition kernel β from I to I and $\phi \in \mathcal{C}(([0,\vartheta] \times I)^n)$, we set, for $u \in [0,\vartheta]^n$,

$$\beta_{k,B}\phi(u, v_1, ..., v_n) := \int \beta(v_k, \mathrm{d}v')\phi(u, v_1, ..., v_{k-1}, v', v_{k+1}, ..., v_n)$$

5. For $z \in \mathcal{C}_L(I)$ and $v \in I^n$, we set $z_k(v) := z(v_k)$.

We give the martingale problem for the process (X^N, Z^N) for some $N = 1, 2, ...$ We refer to Remark 1.2 for a definition of a martingale problem.

Definition 2.3 (Martingale problem for the Fleming-Viot process with mutation modifier and fluctuating selection)**.** For $(u,v) \in ([0,\vartheta] \times I)^n$ with $u = (u_1, ..., u_n), v = (v_1, ..., v_n)$ and $1 \leq k, l \leq n$, we set

$$\theta_{k,l}(u) := (u_1, ..., u_{l-1}, u_k, u_l, ..., u_{n-1}), \qquad \theta_{k,l}(u,v) := (\theta_{k,l}(u), \theta_{k,l}(v)).$$

For the domain of the generator of (X^N, Z^N), we define the set of functions

$$\begin{aligned}\Pi := \{(x,z) \mapsto \Phi(x)\Psi(z) : \Phi(x) = \Phi^{n,\phi}(x) = \langle x^n, \phi\rangle, \Psi(z) = \Psi^{m,u}(z) = z(u_1)\cdots z(u_m),\\ m, n = 1, 2, ..., \phi \in \mathcal{C}(([0,\vartheta] \times I)^n), u = (u_1, ..., u_m) \in I^m\}.\end{aligned}$$

The generator then reads

$$G_N = G^{\mathrm{res}} + G^{\mathrm{mut}} + N \cdot G^{\mathrm{sel}} + N^2 \cdot G^{\mathrm{env}}$$

with

$$\begin{aligned}G^{\mathrm{res}}\Phi(x)\Psi(z) &= \Psi(z) \cdot \sum_{k,l=1}^{n} \langle x^n, \phi \circ \theta_{kl} - \phi\rangle,\\ G^{\mathrm{mut}}\Phi(x)\Psi(z) &= \Psi(z) \cdot \sum_{k=1}^{n} \langle x^n, \pi_{k,A} \cdot (\beta_{k,B}\phi - \phi)\rangle,\\ G^{\mathrm{sel}}\Phi(x)\Psi(z) &= \Psi(z) \cdot \alpha \sum_{k=1}^{n} \langle x^{n+1}, \phi \cdot (z_k - z_{n+1})\rangle,\\ G^{\mathrm{env}}\Phi(x)\Psi(z) &= \Phi(x) \cdot \sigma \cdot (\mathbb{E}_\nu[\Psi(Z)] - \Psi(z)).\end{aligned}$$

Then, for $E := \mathcal{P}([0,\vartheta] \times I) \times \mathcal{C}(I)$ and $\mu \in \mathcal{P}(E)$, we call every E-valued process (X^N, Z^N) such that $(X^N(0), Z^N(0)) \sim \mu$ and

$$\Big(\Phi(X_t^N)\Psi(Z_t^N) - \int_0^t G_N\Phi(X_s^N)\Psi(Z_s^N)\Big)_{t\geq 0}$$

is a martingale, the *Fleming-Viot process with mutation modifier and fluctuating selection.* Its martingale problem is called the (G_N, Π, μ)-martingale problem.

Remark 2.4 (Some notes on the generator terms).

1. We have already seen G^{res} and G^{sel} in Chapter 1.

2. For the mutation operator, we note that

$$(\pi_{k,A} \cdot \beta_{k,B}\phi)(u, v) = u_k \cdot \int \beta(v_k, \mathrm{d}v')\phi(u, v_1, ..., v_{k-1}, v', v_{k+1}, ..., v_n).$$

   Hence, the state at the A-locus, u_k, equals the mutation rate at the B-locus.

3. The generator describing the change in environment is the common generator for a Poisson process with rate σ; see Ethier and Kurtz (1986), Section 4.2.

Before moving on to our results for fast fluctuating environments we briefly check the well-posedness of the martingale problem given in Definition 2.3.

Lemma 2.5. *For $N = 1, 2, ...$ and $\mu \in \mathcal{P}(E)$, the (G_N, Π, μ)-martingale problem is well-posed. This solution (X^N, Z^N) is strongly continuous, i.e. $(X_t^N, Z_t^N) \overset{t\to 0}{\Longrightarrow} (X_0^N, Z_0^N)$ and has the Feller property, i.e. $x \mapsto \mathbb{E}_x[f(X_t^N, Z_t^N)]$ is continuous for every $f \in \mathcal{C}(E)$.*

Proof. First, fix N and let (X^N, Z^N) be some solution of the (G_N, Π, μ)-martingale problem. By setting $\Phi = 1$ we obtain that

$$\Big(\Psi(Z_t^N) - N^2\sigma \int_0^t (\mathbb{E}_\nu[\Psi(Z)] - \Psi(Z_s^N))ds\Big)_{t\geq 0}$$

is a martingale problem. As mentioned in Remark 2.4, this is the usual generator of a Poisson process. More precisely, Z^N is a Markov jump process, which jumps from z to $Z \sim \nu$ at rate $N^2\sigma$. Next, we construct X^N conditional on Z^N. The process Z^N is a piece-wise constant process with constant jump rate $N^2\sigma$, hence jump points do not accumulate, i.e. Z^N is non-explosive. Therefore we can solve the resulting martingale problem for X^N (conditional on Z^N) uniquely between jumps of Z^N. This means that we only require the well-posedness of the martingale problem for $\sigma = 0$ which is again a classical result in mathematical population genetics; see e.g. Ethier and Kurtz (1993). In summary, by this two-step procedure, we obtain existence and uniqueness of the (G_N, Π, μ)-martingale problem. □

In the next theorem we obtain general limit results for the evolution of the allele frequency distribution for rapidly fluctuating environments.

Theorem 2.6 (Convergence for fast fluctuating environment). *Given that $X_0^N \overset{N\to\infty}{\Longrightarrow} X_0 \sim \mu_1$ and $6\alpha^2/\sigma < 1$, we find that $X^N \overset{N\to\infty}{\Longrightarrow} X$, the unique solution of the (G, Π_1, μ_1) martingale problem, where*

$$\Pi_1 := \{x \mapsto \Phi(x) : \Phi(x) = \Phi^{n,\phi}(x) = \langle x^n, \phi\rangle, n = 1, 2, ..., \phi \in \mathcal{C}(([0, \vartheta] \times [0, 1])^n)\}$$

and, setting

$$\chi_{k,l}(v) := \chi(v_k, v_l) := \mathbb{E}_\nu[Z(v_k)Z(v_l)],$$

with

$$G = G^{res} + G^{mut} + \overline{G}^{sel}$$

where G^{res} and G^{mut} are as in Definition 2.3 and, for $\Phi = \Phi^{n,\phi}$,

$$\overline{G}^{sel}\Phi(x) = \frac{\alpha^2}{\sigma} \sum_{\substack{k,l=1 \\ k\neq l}}^{n} \big\langle x^{n+2}, \phi \cdot (\chi_{kl} - \chi_{n+1,n+2})\big\rangle + 2n\frac{\alpha^2}{\sigma} \sum_{k=1}^{n} \big\langle x^{n+2}, \phi \cdot (\chi_{n+1,n+2} - \chi_{k,n+1}\big\rangle \\ + \frac{\alpha^2}{\sigma} \sum_{k=1}^{n} \big\langle x^{n+2}, \phi \cdot (\chi_{kk} - \chi_{n+1,n+1})\big\rangle. \tag{2.2.2}$$

The proof of Theorem 2.6 requires a corollary from Ethier and Kurtz (1986) that deals with the approximation of strongly continuous contraction semigroups.

Corollary 2.7 (Corollary 1.7.8, Ethier and Kurtz (1986))**.** *Let us briefly recall this result. For some locally compact and separable (E, r), let $L := \mathcal{C}_b(E)$, equipped with the topology of uniform convergence on compacts. For operators G_i with domain $\mathcal{D}(G_i)$, $i = 0, 1, 2$, assume the following:*

1. *G_2 generates a strongly continuous contraction semigroup $(S_t)_{t\geq 0}$ on L, such that*

$$\lim_{\lambda\to 0+} \lambda \int_0^\infty e^{-\lambda t} S_t f dt =: Pf \text{ exists for all } f \in L;$$

2. *$\mathcal{D} := \mathcal{D}(G_0) \cap \mathcal{D}(G_1) \cap \mathcal{D}(G_2)$ is a core for G_2;*
3. *For N sufficiently large, $G_0 + N\cdot G_1 + N^2\cdot G_2$ generates a strongly continuous contraction semigroup $(T^N(t))_{t\geq 0}$ on L.*

For $f \in D \subseteq \{f : \mathcal{D}(G_0) \cap \mathcal{D}(G_1) : G_2 f = 0\}$, set

$$D_f := \{h \in \mathcal{D} : G_2 h = -G_1 f\}$$

and define for any $f \in D$ and $h \in D_f$

$$\bar{G}f = PG_0 f + PG_1 h. \tag{2.2.3}$$

Then, $\bar{G}$ is dissipative and if its closure generates a strongly continuous contraction semigroup $(T(t))_{t\geq 0}$ on $\bar{D}$, then $T^N(t)f \xrightarrow{N\to\infty} T(t)f$ for all $t \geq 0$, uniformly on bounded intervals.

Remark 2.8 (Outline of the proof)**.** Before actually moving on to the proof of Theorem 2.6, we will give a short sketch of what needs to be done.

1. For the desired convergence we need to apply Corollary 2.7 the following way: We are dealing with the special situation that $E = E_1 \times E_2$,

   (A1) G_2 has the form
   $$G_2 f(x, z) = \sigma\big(\mathbb{E}_\nu[f(x, Z)] - f(x, z)\big)$$
   for some $\nu \in \mathcal{P}(E_2)$ and $\sigma > 0$,

(A2) G_1 satisfies $\mathbb{E}_\nu[G_1 f(x,Z)] = 0$ if f only depends on x.

In this situation, G_2 generates a strongly continuous contraction semigroup $(S_t)_{t\geq 0}$ on $\mathcal{C}_b(E)$, which has the form

$$S_t f(x,z) = e^{-\sigma t} f(x,z) + (1-e^{-\sigma t})\mathbb{E}_\nu[f(x,Z)].$$

Clearly, since $S_t f(x,z) = e^{-\sigma t} f(x,z) + (1-e^{-\sigma t})\mathbb{E}_\nu[f(x,Z)]$,

$$\begin{aligned}\lambda\int_0^\infty e^{-\lambda t} S_t f(x,z)dt &= \frac{\lambda}{\lambda+\sigma}(f(x,z) - \mathbb{E}_\nu[f(x,Z)]) + \lambda\int_0^\infty e^{-\lambda t}\mathbb{E}_\nu[f(x,Z)]dt \\ &\xrightarrow{\lambda\to 0} \mathbb{E}_\nu[f(x,Z)] =: Pf(x,z).\end{aligned}$$

For any $(x,z)\mapsto f(x,z)$ only dependent on x with $G_2 f = 0$, we choose $h = \frac{1}{\sigma}G_1 f$. By (A1) and (A2) it holds

$$G_2 h(x,z) = \frac{1}{\sigma}\cdot G_2 G_1 f(x,z) = \frac{1}{\sigma}\cdot\sigma\big(\mathbb{E}_\nu[G_1 f(x,Z)] - G_1 f(x,z)\big) = -G_1 f(x,z),$$

i.e. $h = \frac{1}{\sigma}G_1 f$ is a solution of $G_2 h = -G_1 f$. In total, we find that (abusing notation by writing $x\mapsto f(x)$ if f only depends on x), (2.2.3) transforms to

$$\bar{G}f(x) = \mathbb{E}_\nu\big[G_0 f(x,Z) + \tfrac{1}{\sigma}G_1 G_1 f(x,Z)\big]. \tag{2.2.4}$$

To show convergence we need to prove that $\bar{G}$ generates a strongly continuous contraction semigroup which is implied by well-posedness of the $(\bar{G},D)$-martingale problem.

2. It remains to show well-posedness of the $\bar{G}$-martingale problem as well as the Feller property. At least, existence of a solution of the martingale problem follows by general theory; see Chapter 4.5 of Ethier and Kurtz (1986), provided that the Markov processes X^N with semigroups T^N satisfy the compact containment condition. Indeed, since $||\frac{1}{N}h|| \xrightarrow{N\to\infty} 0$ and

$$\begin{aligned}(G_0 + N\cdot G_1 + N^2\cdot G_2)(f+\tfrac{1}{N}h) &= G_0 f(x) + G_1 h(x,z) + N\cdot(G_1 f + G_2 h) + o(1) \\ &= G_0 f(x) + G_1 h + o(1),\end{aligned}$$

we find generator convergence.
For uniqueness and the Feller property, we will be using a duality argument (see Chapter 4.4 in Ethier and Kurtz (1986)). Duality has been proven to be a useful tool when having to characterise certain processes. One very important application is the case when proving uniqueness of a solution of a martingale problem. Recall that X (i.e. a solution of the $(\bar{G},D)$-martingale problem) is *dual* to some stochastic process Y with (separable) state space Υ with respect to $H: E\times\Upsilon\to\mathbb{R}$ bounded and measurable, if

$$\mathbb{E}_x[H(X_t,y)] = \mathbb{E}_y[H(x,Y_t)]$$

for all t,x,y. Proposition 4.4.7 of Ethier and Kurtz (1986) states that the existence of a dual process Y is sufficient to guarantee the uniqueness of X: If $\Pi := \{H(.,y) : y\in\Upsilon\}\subseteq D$ and Y is a Markov process with generator G_Y, and if $H(x,.)$ is in the domain of G_Y for all x, the latter equality is implied by

$$\bar{G}H(.,y)(x) = G_Y H(x,.)(y), \tag{2.2.5}$$

since

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}[H(X_s, Y_{t-s})] = \mathbb{E}[\bar{G}H(., Y_{t-s})(X_s) - G_Y H(X_s, .)(Y_{t-s})] = 0$$

on a probability space where X and Y are independent. If Π is separating, existence of Y implies uniqueness of the $(\bar{G}, D)$-martingale problem. Moreover, if H is bounded and continuous, we find that $x \mapsto \mathbb{E}_x[H(X_t, y)] = \mathbb{E}_y[H(x, Y_t)]$ is continuous by dominated convergence. If Π is convergence determining and Y is Feller, this implies that X is Feller as well.

Proof of Theorem 2.6. We use Remark 2.8.1 with

$$\begin{aligned} E_1 &= \mathcal{P}([0,\vartheta] \times I), \quad E_2 = \mathcal{C}_L(I), \\ G_0 &= G^{\text{res}} + G^{\text{mut}}, \quad G_1 = G^{\text{sel}} G_2 = G^{\text{env}}. \end{aligned}$$

(A1) obviously holds due to the form of G^{env} in Definition 2.3. If Φ only depends on x, (A2) is satisfied since $G^{\text{sel}}\Phi$ depends on z only linearly and because of (2.2.1). If $\Phi\Psi \in \Pi$ with $\Phi = \Phi^{n,\phi}, \Psi = \Psi^{m,u}$ only depends on x, we have that $\Psi = \text{const}$ and as in Remark 2.8 $h = -\frac{1}{\sigma}G^{\text{sel}}\Phi$ solves $G^{\text{env}}h = -G^{\text{sel}}\Phi$. Therefore, (2.2.4) gives

$$\bar{G}\Phi(x) = G^{\text{res}}\Phi(x) + G^{\text{mut}}\Phi(x) + \frac{1}{\sigma}\mathbb{E}_\nu[G^{\text{sel}}G^{\text{sel}}\Phi(x, Z)].$$

In order to compute that last term, we define for $v \in I^n$

$$\chi_{k,l}(v) := \chi(v_k, v_l) := \mathbb{E}_\nu[Z(v_k)Z(v_l)].$$

Using the symmetry relationship $\langle x^{n+2}, \phi \cdot Z_{n+1}\rangle = \langle x^{n+2}, \phi \cdot Z_{n+2}\rangle$ we obtain for ϕ depending only on the first n coordinates at both loci

$$\begin{aligned} \overline{G}^{\text{sel}}\Phi(x) &:= \frac{1}{\sigma}\mathbb{E}_\nu[G^{\text{sel}}G^{\text{sel}}\Phi(x, Z)] \\ &= \frac{\alpha}{\sigma}\sum_{l=1}^{n}\mathbb{E}_\nu\big[G^{\text{sel}}\big\langle x^{n+1}, \phi \cdot (Z_l - Z_{n+1})\big\rangle\big] \\ &= \frac{\alpha^2}{\sigma}\sum_{l=1}^{n}\sum_{k=1}^{n+1}\mathbb{E}_\nu\Big[\big\langle x^{n+2}, \phi \cdot (Z_l - Z_{n+1}) \cdot (Z_k - Z_{n+2})\big\rangle\Big] \\ &= \frac{\alpha^2}{\sigma}\sum_{\substack{k,l=1\\k\neq l}}^{n}\big\langle x^{n+2}, \phi \cdot (\chi_{kl} - 2\chi_{k,n+1} + \chi_{n+1,n+2})\big\rangle \\ &\qquad + \frac{\alpha^2}{\sigma}\sum_{l=1}^{n}\big\langle x^{n+2}, \phi \cdot (\chi_{l,l} - 2\chi_{l,n+1} + 2\chi_{n+1,n+2} - \chi_{n+1,n+1})\big\rangle \\ &= \frac{\alpha^2}{\sigma}\sum_{\substack{k,l=1\\k\neq l}}^{n}\big\langle x^{n+2}, \phi \cdot (\chi_{kl} - \chi_{n+1,n+2} - 2\chi_{k,n+1} + 2\chi_{n+1,n+2})\big\rangle \\ &\qquad + \frac{\alpha^2}{\sigma}\sum_{l=1}^{n}\big\langle x^{n+2}, \phi \cdot (\chi_{l,l} - \chi_{n+1,n+1} - 2\chi_{l,n+1} + 2\chi_{n+1,n+2})\big\rangle \end{aligned}$$

$$= \frac{\alpha^2}{\sigma} \sum_{\substack{k,l=1 \\ k\neq l}}^{n} \big\langle x^{n+2}, \phi \cdot (\chi_{kl} - \chi_{n+1,n+2})\big\rangle + 2n\frac{\alpha^2}{\sigma} \sum_{k=1}^{n} \big\langle x^{n+2}, \phi \cdot (\chi_{n+1,n+2} - \chi_{k,n+1}\big\rangle$$
$$+ \frac{\alpha^2}{\sigma} \sum_{k=1}^{n} \big\langle x^{n+2}, \phi \cdot (\chi_{kk} - \chi_{n+1,n+1})\big\rangle.$$

This corresponds to the term given in (2.2.2). Existence of the (G, Π_1)-martingale problem follows as in Remark 2.8.2.

For uniqueness, we use duality. The dual process will be similar to the one of the TFVMS given in Depperschmidt et al. (2012). The goal is to use (2.2.5), and therefore, we have to rewrite the generator terms. We define for $u = (u_1, u_2, ...)$

$$\begin{aligned} \bar\alpha_l(u) &= (u_{i-1_{\{i>l\}}}) = (u_1, ..., u_l, u_l, u_{l+1}, ...), \\ \alpha_l(u) &= (u_{i+1_{\{i\geq l\}}}) = (u_1, ..., u_{l-1}, u_{l+1}, u_{l+2}, ...). \end{aligned}$$

We note that, for ϕ depending only on the first n coordinates, and $1 \leq k \neq l \leq n$

$$\begin{aligned} \langle x^n, \phi \circ \theta_{kl}\rangle &= \langle x^{n-1}, \phi \circ \theta_{kl} \circ \bar\alpha_l\rangle, \\ \mathbb{E}_\nu[\langle x^{n+1}, \phi \cdot Z_{n+1}\rangle] &= \mathbb{E}_\nu[\langle x^{n+1}, (\phi \circ \alpha_k) \cdot Z_k\rangle], \\ \langle x^{n+2}, \phi \cdot \chi_{n+1,n+2}\rangle &= \langle x^{n+2}, (\phi \circ \alpha_k) \cdot \chi_{k,n+2}\rangle = \langle x^{n+2}, (\phi \circ \alpha_k \circ \alpha_l) \cdot \chi_{k,l}\rangle, \end{aligned}$$

holds, since integrating with respect to the product measure x^n does not depend on the order of coordinates.

Therefore, we can write for $\Phi = \Phi^{n,\phi}$

$$\begin{aligned} G^{\mathrm{res}}\langle x^n, \phi\rangle &= \sum_{\substack{k,l=1 \\ k\neq l}}^{n} \langle x^{n-1}, \phi \circ \theta_{k,l} \circ \bar\alpha_l\rangle - \langle x^n, \phi\rangle, \\ G^{\mathrm{mut}}\langle x^n, \phi\rangle &= \vartheta \cdot \sum_{k=1}^{n} \Big\langle x^n, \frac{\pi_{k,A}}{\vartheta} \cdot \beta_{k,B}\phi + \Big(1 - \frac{\pi_{k,A}}{\vartheta}\Big) \cdot \phi\Big\rangle - \langle x^n, \phi\rangle, \\ \overline{G}^{\mathrm{sel}}\langle x^n, \phi\rangle &= \frac{\alpha^2}{\sigma} \sum_{\substack{k,l=1 \\ k\neq l}}^{n} (\langle x^{n+2}, \phi \cdot \chi_{kl} + (\phi \circ \alpha_k \circ \alpha_l) \cdot (1 - \chi_{kl})\rangle - \langle x^n, \phi\rangle) \qquad (2.2.6) \\ &\quad + 2n\frac{\alpha^2}{\sigma} \sum_{k=1}^{n} (\langle x^{n+2}, (\phi \circ \alpha_k) \cdot \chi_{k,n+2} + \phi \cdot (1 - \chi_{k,n+2})\rangle - \langle x^n, \phi\rangle) \\ &\quad + \frac{\alpha^2}{\sigma} \sum_{k=1}^{n} \big(\langle x^{n+2}, (\phi \cdot \chi_{k,k} + (\phi \circ \alpha_k) \cdot (1 - \chi_{k,k})\rangle - \langle x^n, \phi\rangle\big). \end{aligned}$$

With this reformulation, we can construct a dual process $\Xi = (\xi_t)_{t\geq 0}$ which will be function-valued. The state space is

$$\Upsilon = \bigcup_{n=0}^{\infty} \Upsilon_n \qquad \text{with} \qquad \Upsilon_n = \mathcal{C}(([0, \vartheta] \times [0, 1])^n).$$

The process Ξ is a pure jump process with transitions from $\xi \in \Upsilon_n$ to

$$\begin{cases} \xi \circ \theta_{k,l} \circ \bar{\alpha}_l \in \Upsilon_{n-1} & \text{at rate 1 for each unordered pair } 1 \leq k \neq l \leq n, \\ \frac{\pi_{k,A}}{\vartheta} \cdot \beta_{k,B} \cdot \xi + \Big(1 - \frac{\pi_{k,A}}{\vartheta}\Big) \cdot \xi \in \Upsilon_n & \text{at rate } \vartheta \text{ for each } 1 \leq k \leq n, \\ \xi \cdot \chi_{kl} + (\xi \circ \alpha_k \circ \alpha_l) \cdot (1 - \chi_{kl}) \in \Upsilon_{n+2} & \text{at rate } \frac{\alpha^2}{\sigma} \text{ for each unordered pair } 1 \leq k \neq l \leq n, \\ (\xi \circ \alpha_k) \cdot \chi_{k,n+2} + \xi \cdot (1 - \chi_{k,n+2}) \in \Upsilon_{n+2} & \text{at rate } 2n\frac{\alpha^2}{\sigma} \text{ for each } 1 \leq k \leq n, \\ \xi \cdot \chi_{k,k} + (\xi \circ \alpha_k) \cdot (1 - \chi_{k,k}) \in \Upsilon_{n+1} & \text{at rate } \frac{\alpha^2}{\sigma} \text{ for each } 1 \leq k \leq n. \end{cases}$$

Then, for the duality function $H(\cdot, \cdot)$ with

$$\begin{aligned} H : E \times \Upsilon &\to \mathbb{R} \\ (x, \xi) &\mapsto \langle x^n, \xi \rangle, \quad \text{for } \xi \in \Upsilon_n, \end{aligned}$$

we have established (2.2.5), i.e. the generator of Ξ for $\xi \in \Upsilon_n$ is $(G^{\text{res}} + G^{\text{mut}} + \overline{G}^{\text{sel}})\langle x^n, \xi \rangle$ with $G^{\text{res}}, G^{\text{mut}}$ and $\overline{G}^{\text{sel}}$ as the right-hand sides in (2.2.6). In other words, Ξ and X, a solution of the G-martingale problem are dual, provided that existence for Ξ can be guaranteed. Here, we have to take into account that the number of dependent variables, n, can explode. This number decreases at rate $n(n-1)$ and increases by two at rate $(n(n-1) + 2n^2)\alpha^2/\sigma$ and by one at rate $\alpha^2/\sigma n$. Therefore, explosion cannot occur for $6\alpha^2/\sigma < 1$ and from Proposition 4.4.7 of Ethier and Kurtz (1986), uniqueness for the G-martingale problem follows in this case. Since $\{H(., \xi) : \xi \in \Upsilon\}$ is separating and convergence determining (see e.g. Example 5 in Depperschmidt et al., 2019), we have shown that $\bar{G}$ generates a strongly continuous contraction semigroup and the proof of Theorem 2.6 is complete; see Remark 2.8.2. □

2.3 Specialization to a finite dimensional system

We will now specialise Theorem 2.6 to a finite-dimensional system. Precisely, since we have two loci, the minimal number of dimensions is 2 × 2. So, only four types will be present, which will be denoted $\ell 0, \ell 1, h0, h1$. For $0 \leq \vartheta_\ell \leq \vartheta_h \leq \vartheta$, their frequencies are given through $x \in \mathcal{P}([0, \vartheta] \times I)$ by

$$x_{ai} := \Phi_{ai}(x) := x(\{\vartheta_a\} \times \{i\}) = \langle x, 1_{\{\vartheta_a\} \times \{i\}} \rangle, \qquad (a, i) \in \{\ell, h\} \times \{0, 1\}.$$

For mutation, we consider the case that each mutation event (either at rate ϑ_ℓ or ϑ_h) results in type 0 at the B-locus with probability $r \in [0, 1]$. For selection, let $z : \{0, 1\} \to \{-\frac{1}{2}, \frac{1}{2}\}$ be given by $z(0) = \frac{1}{2}, z(1) = -\frac{1}{2}$ and

$$\nu = \tfrac{1}{2}(\delta_z + \delta_{-z}).$$

Consider the solution X^N of the martingale problem from Definition 2.3 in this case, which exists uniquely by Lemma 2.5. Letting $X^N_{ai}, (a, i) \in \{\ell, h\} \times \{0, 1\}$ be as above, using the martingale representation theorem (see e.g. Theorem 16.12. of Kallenberg, 2002), it is straightforward to see that $X^N = (X^N_{\ell0}, X^N_{\ell1}, X^N_{h0}, X^N_{h1})$ is a weak solution of the system of SDEs

$$\begin{aligned}
\mathrm{d}X^N_{\ell0} &= \alpha N Z^N X^N_{\ell0} X^N_1 \mathrm{d}t + \vartheta_\ell(rX^N_{\ell1} - (1-r)X^N_{\ell0})\mathrm{d}t \\
&\qquad + \sqrt{X^N_{\ell0}X^N_{\ell1}}\mathrm{d}W_1 + \sqrt{X^N_{\ell0}X^N_{h0}}\mathrm{d}W_2 + \sqrt{X^N_{\ell0}X^N_{h1}}\mathrm{d}W_3, \\
\mathrm{d}X^N_{\ell1} &= -\alpha N Z^N X^N_{\ell1} X^N_0 \mathrm{d}t + \vartheta_\ell((1-r)X^N_{\ell0} - rX^N_{\ell1})\mathrm{d}t \\
&\qquad - \sqrt{X^N_{\ell1}X^N_{\ell0}}\mathrm{d}W_1 + \sqrt{X^N_{\ell1}X^N_{h0}}\mathrm{d}W_4 + \sqrt{X^N_{\ell1}X^N_{h1}}\mathrm{d}W_5, \\
\mathrm{d}X^N_{h0} &= \alpha N Z^N X^N_{h0} X^N_1 \mathrm{d}t + \vartheta_h(rX^N_{h1} - (1-r)X^N_{h0})\mathrm{d}t \\
&\qquad - \sqrt{X^N_{h0}X^N_{\ell0}}\mathrm{d}W_2 - \sqrt{X^N_{h0}X^N_{\ell1}}\mathrm{d}W_4 + \sqrt{X^N_{h0}X^N_{h1}}\mathrm{d}W_6, \\
\mathrm{d}X^N_{h1} &= -\alpha N Z^N X^N_{h1} X^N_0 \mathrm{d}t + \vartheta_h((1-r)X^N_{h0} - rX^N_{h1})\mathrm{d}t \\
&\qquad - \sqrt{X^N_{h1}X^N_{\ell0}}\mathrm{d}W_3 - \sqrt{X^N_{h1}X^N_{\ell1}}\mathrm{d}W_5 - \sqrt{X^N_{h1}X^N_{h0}}\mathrm{d}W_6,
\end{aligned} \tag{2.3.1}$$

with $X^N_i = X^N_{hi} + X^N_{\ell i}$, $i = 0, 1$, independent Brownian motions $W_1, ..., W_6$, and Z^N (the fitness difference between types 0 and 1) changes from -1 to $+1$ and back at rate $N^2\frac{\sigma}{2}$.

Theorem 2.9 (Convergence for fast fluctuating environment)**.** *For weak solutions* $(X^N)_{N=1,2,...}$ *of* (2.3.1)*, assume that* $X^N(0) \overset{n\to\infty}{\Longrightarrow} X_0$ *and* $2\alpha^2/\sigma < 1$*. Then,* $(X^N_{\ell0}, X^N_{\ell1}, X^N_{h0}, X^N_{h1}) \overset{N\to\infty}{\Longrightarrow} X = (X_{\ell0}, X_{\ell1}, X_{h0}, X_{h1})$ *which is the unique weak solution of*

$$\begin{aligned}
\mathrm{d}X_{\ell0} &= \tfrac{\alpha^2}{\sigma} X_{\ell0}X_1(X_1 - X_0)\mathrm{d}t + \theta_\ell(rX_{\ell1} - (1-r)X_{\ell0})\mathrm{d}t \\
&\quad + \sqrt{X_{\ell0}X_{\ell1}}\mathrm{d}W_1 + \sqrt{X_{\ell0}X_{h0}}\mathrm{d}W_2 + \sqrt{X_{\ell0}X_{h1}}\mathrm{d}W_3 + \alpha\sqrt{2/\sigma}X_{\ell0}X_1\mathrm{d}W, \\
\mathrm{d}X_{\ell1} &= \tfrac{\alpha^2}{\sigma} X_{\ell1}X_0(X_0 - X_1)\mathrm{d}t + \theta_\ell((1-r)X_{\ell0} - rX_{\ell1})\mathrm{d}t \\
&\quad - \sqrt{X_{\ell1}X_{\ell0}}\mathrm{d}W_1 + \sqrt{X_{\ell1}X_{h0}}\mathrm{d}W_4 + \sqrt{X_{\ell1}X_{h1}}\mathrm{d}W_5 - \alpha\sqrt{2/\sigma}X_{\ell1}X_0\mathrm{d}W \\
\mathrm{d}X_{h0} &= \tfrac{\alpha^2}{\sigma} X_{h0}X_1(X_1 - X_0)\mathrm{d}t + \theta_h(rX_{h1} - (1-r)X_{h0})\mathrm{d}t \\
&\quad - \sqrt{X_{h0}X_{\ell0}}\mathrm{d}W_2 - \sqrt{X_{h0}X_{\ell1}}\mathrm{d}W_4 + \sqrt{X_{h0}X_{h1}}\mathrm{d}W_6 + \alpha\sqrt{2/\sigma}X_{h0}X_1\mathrm{d}W, \\
\mathrm{d}X_{h1} &= \tfrac{\alpha^2}{\sigma} X_{h1}X_0(X_0 - X_1)\mathrm{d}t + \theta_h((1-r)X_{h0} - rX_{h1})\mathrm{d}t \\
&\quad - \sqrt{X_{h1}X_{\ell0}}\mathrm{d}W_3 - \sqrt{X_{h1}X_{\ell1}}\mathrm{d}W_5 - \sqrt{X_{h1}X_{h0}}\mathrm{d}W_6 - \alpha\sqrt{2/\sigma}X_{h1}X_0\mathrm{d}W,
\end{aligned} \tag{2.3.2}$$

with independent Brownian motions $W, W_1, ..., W_6$ *and some initial condition* X_0*.*

Remark 2.10 (Evolution of X_h and X_0)**.** Writing $X_h = X_{h0} + X_{h1}$ and $X_\ell = 1 - X_h$, we also have

$$\begin{aligned}
\mathrm{d}X_h = \frac{\alpha^2}{\sigma}&(X_{h0}X_{\ell1} - X_{h1}X_{\ell0})(X_1 - X_0)\mathrm{d}t \\
&+ \sqrt{X_hX_\ell}\mathrm{d}W' + \alpha\sqrt{2/\sigma}(X_{h0}X_{\ell1} - X_{h1}X_{\ell0})\mathrm{d}W,
\end{aligned} \tag{2.3.3}$$

with independent Brownian motions W, W'. In the same way we can set $X_0 = X_{h0} + X_{\ell0}$ and $X_1 = 1 - X_0$, and get

$$\begin{aligned}
\mathrm{d}X_0 = \frac{\alpha^2}{\sigma}X_0X_1(X_1 - X_0)\mathrm{d}t + \vartheta_\ell(r - X_0) + (\vartheta_h - \vartheta_\ell)(rX_{h1} - (1-r)X_{h0})\mathrm{d}t \\
+ \sqrt{X_0X_1}\mathrm{d}W'' + \alpha\sqrt{2/\sigma}X_0X_1\mathrm{d}W
\end{aligned}$$

with independent Brownian motions W, W''.

Remark 2.11 (Comparison with Gillespie (1981))**.** Gillespie has considered a similar diffusion for a mutation modifier locus in diploids Gillespie (1981). While the mutation rates differ in Gillespie's model compared to the as we do not consider heterozygotes in our haploid model, the remaining diffusion terms of a symmetric semi-dominant model from Gillespie are similar to our setting.

To see this consider equation (5) in Gillespie (1981). The variable $p_1 = 1 - q_1$ corresponds to our X_0, and $p_2 = 1 - q_2$ to X_h. In the symmetric semi-dominant model Gillespie set $A = 0$ and $B = 2$. Thus, ignoring all terms with mutation rates, we get

$$\begin{aligned} \mathrm{d}p_1 &= p_1 q_1 \left(A + B \left(\frac{1}{2} - p_1 \right) \right) + p_1 q_1 \mathrm{d}W \\ &= X_0 X_1 (X_1 - X_0)\mathrm{d}t + X_0 X_1 \mathrm{d}W, \end{aligned}$$

and

$$\mathrm{d}p_2 = D \left(A + B \left(\frac{1}{2} - p_1 \right) \right) \mathrm{d}t + D\mathrm{d}W = D(X_1 - X_0)\mathrm{d}t + D\mathrm{d}W$$

for linkage disequilibrium $D := (X_{h0}X_{\ell 1} - X_{h1}X_{\ell 0})$. Furthermore, we can use Itô's lemma to get

$$\mathrm{d}(X_{h0}X_{\ell 1}) = X_{\ell 1}X_{h0}(X_1 - X_0)^2\mathrm{d}t - X_{h0}X_1X_{\ell 1}X_0\mathrm{d}t + X_{h0}X_{\ell 1}(X_1 - X_0)\mathrm{d}W$$

and

$$\begin{aligned} \mathrm{d}D &= \mathrm{d}(X_{h0}X_{\ell 1} - X_{h1}X_{\ell 0}) \\ &= (X_{h0}X_{\ell 1} - X_{h1}X_{\ell 0})(X_1 - X_0)^2\mathrm{d}t - (X_{h0}X_{\ell 1} - X_{h1}X_{\ell 0})X_1X_0\mathrm{d}t \\ &\qquad + (X_{h0}X_{\ell 1} - X_{h1}X_{\ell 0})(X_1 - X_0)\mathrm{d}W \\ &= D(q_1 - p_1)^2\mathrm{d}t - Dp_1q_1\mathrm{d}t + D(p_1 - q_1)\mathrm{d}W. \end{aligned}$$

The special case presented here is thus a haploid version of the symmetric semi-dominant model in Gillespie's work.

Proof of Theorem 2.9. Since X^N weakly solves (2.3.1) if and only if it solves the martingale problem from Definition 2.3, we need to show that a solution of the limiting martingale problem from Theorem 2.6 solves (2.3.2). By the martingale representation Theorem (see e.g. Theorem 16.12. of Kallenberg, 2002), it is enough to show that (with $X = (X_{\ell 0}, X_{\ell 1}, X_{h0}, X_{h1})$ a solution of the limiting martingale problem) X is a semimartingale with $X = X_0 + M + A$, where $A = (A_{\ell 0}, A_{\ell 1}, A_{h0}, A_{h1})$ is a process of finite variation with

$$\begin{aligned} A_{a0}(t) &= \int_0^t \vartheta_a(rX_{a,1}(s) - (1-r)X_{a0}(s)) + \frac{\alpha^2}{\sigma}X_{a0}(s)X_1(s)(X_1(s) - X_0(s))\mathrm{d}s, \\ A_{a1}(t) &= \int_0^t \vartheta_a((1-r)X_{a0}(s) - rX_{a1}(s)) + \frac{\alpha^2}{\sigma}X_{a1}(s)X_0(s)(X_0(s) - X_1(s))\mathrm{d}s, \end{aligned} \tag{2.3.4}$$

and $M = (M_{\ell 0}, M_{\ell 1}, M_{h0}, M_{h1})$ is a martingale with covariation

$$\begin{aligned}&[M_{ai}, M_{bj}](t)\\&= \int_0^t \Big((\delta_{ai,bj} - X_{ai}(s))X_{bj}(s) + (-1)^{i+j}\frac{2\alpha^2}{\sigma}X_{ai}(s)X_{bj}(s)X_{1-i}(s)X_{1-j}(s)\Big)\mathrm{d}s. \qquad (2.3.5)\end{aligned}$$

As a general fact (see e.g. Corollary 4.6 in Depperschmidt et al., 2012),

$$\begin{aligned}A_{ai}(t) &= \int_0^t G\Phi_{ai}(X(s))\mathrm{d}s, && (2.3.6)\\ [M_{ai}, M_{bj}](t) &= \int_0^t G\Phi_{ai}\Phi_{bj}(X(s))\\ &\qquad - \Phi_{ai}(X(s))G\Phi_{bj}(X(s)) - \Phi_{bj}(X(s))G\Phi_{ai}(X(s))\mathrm{d}s. && (2.3.7)\end{aligned}$$

While the first term in (2.3.4) is due to G^{mut}, the first term in (2.3.5) is due to G^{res}. For the remaining terms, we need to evaluate the operator $\bar{G}^{\mathrm{sel}}$. First, for $v \in \{0,1\}^n$ and $Z \sim \nu$,

$$\chi_{kl}(v) = \mathbb{E}_\nu[Z(v_k)Z(v_\ell)] = \tfrac{1}{4}\big(\mathbb{1}_{\{v_k=v_l\}} - \mathbb{1}_{\{v_k\neq v_l\}}\big) = \tfrac{1}{2}\mathbb{1}_{\{v_k=v_l\}} - \tfrac{1}{4}.$$

Plugging this into (2.2.2), we obtain

$$\begin{aligned}\overline{G}^{\mathrm{sel}}\Phi_{ai}(x) &= \frac{\alpha^2}{\sigma}\langle x^3, \mathbb{1}_{\{\vartheta_a\}\times\{i\}}(u_1,v_1)(\mathbb{1}_{\{v_2=v_3\}} - \mathbb{1}_{\{v_1=v_2\}})\rangle,\\ &= \frac{\alpha^2}{\sigma}\big(x_{ai}(1-2x_0x_1) - x_{ai}x_i\big) = \frac{\alpha^2}{\sigma}x_{ai}(x_{1-i} - 2x_ix_{1-i})) = \frac{\alpha^2}{\sigma}x_{ai}x_{1-i}(x_{1-i} - x_i),\end{aligned}$$

which shows (2.3.4) due to (2.3.6) and

$$\begin{aligned}&\overline{G}^{\mathrm{sel}}\Phi_{ai}\Phi_{bj}(x) - \Phi_{ai}(x)\overline{G}^{\mathrm{sel}}\Phi_{bj}(x) - \Phi_{bj}(x)\overline{G}^{\mathrm{sel}}\Phi_{ai}(x)\\ &\qquad= \frac{\alpha^2}{\sigma}\Big(\langle x^2, 1_{\{\vartheta_a\}\times\{i\}}(u_1,v_1)1_{\{\vartheta_b\}\times\{j\}}(u_2,v_2)1_{v_1=v_2}\rangle - x_{ai}x_{bj}(1-2x_0x_1)\\ &\qquad\qquad + x_{ai}x_{bj}(x_{1-i}(x_{1-i}-x_i) + x_{1-j}(x_{1-j}-x_j))\Big)\\ &\qquad= \frac{\alpha^2}{\sigma}x_{ai}x_{bj}(\delta_{ij} - 1 + x_{1-i}^2 + x_{1-j}^2).\end{aligned}$$

Now, for $i = j$, this gives

$$= \frac{2\alpha^2}{\sigma}x_{ai}x_{bi}x_{1-i}^2,$$

whereas for $i \neq j$, we have

$$= \frac{\alpha^2}{\sigma}x_{ai}x_{bj}(-1 + x_0^2 + x_1^2) = -\frac{2\alpha^2}{\sigma}x_{ai}x_{bj}x_0x_1,$$

which gives in total for arbitrary $i, j \in \{0,1\}$

$$= (-1)^{i+j}\frac{2\alpha^2}{\sigma}x_{ai}x_{bj}x_{1-i}x_{1-j},$$

which finally gives (2.3.5) due to (2.3.7) and the proof is complete. □

Recall $X_h = X_{h0} + X_{h1}$ and $X_\ell = 1 - X_h$. We now give a result on the fixation probability of X_h.

Theorem 2.12 (Fixation probability)**.** *Let X be the solution of* (2.3.2) *with initial condition*

$$\begin{aligned} &X_h(0) = x, && X_{h0}(0) = px, && X_{\ell 0}(0) = q(1-x), \\ &X_\ell(0) = (1-x), && X_{h1}(0) = (1-p)x, && X_{\ell 1}(0) = (1-q)(1-x). \end{aligned}$$

Let $r \in [0,1]$ be the probability that a mutation event results in type 0. Then,

$$\begin{aligned} &\lim_{\frac{\alpha^2}{\sigma}\to 0} \frac{\sigma}{\alpha^2}(\mathbb{P}(X_h(\infty)=1) - x) \\ =&x(1-x) \\ &\cdot \Bigg[\frac{(2r-1)(q-r)(1+2\vartheta_\ell)}{(1+\vartheta_\ell)(3+2\vartheta_\ell)} - \frac{(2r-1)(p-r)(1+2\vartheta_h)}{(1+\vartheta_h)(3+2\vartheta_h)} \\ &\quad + 2\Bigg((1-x)\frac{(q-r)^2}{3+2\vartheta_\ell} - x\frac{(p-r)^2}{3+2\vartheta_h} + (2x-1)\frac{(p-r)(q-r)}{3+\vartheta_\ell+\vartheta_h} \\ &\qquad\qquad + r(1-r)\left(\frac{1}{3+2\vartheta_\ell} - \frac{1}{3+2\vartheta_h}\right)\Bigg)\Bigg]. \end{aligned} \tag{2.3.8}$$

Actually, a straightforward calculation leads to a different form of the last formula.

Corollary 2.13 (Different form of the fixation probability)**.** *For the same situation as in Theorem 2.12,* (2.3.8) *can also be written as*

$$\begin{aligned} &\lim_{\frac{\alpha^2}{\sigma}\to 0} \frac{\sigma}{\alpha^2}(\mathbb{P}(X_h(\infty)=1) - x) \\ =&x(1-x) \\ &\cdot \Bigg[(p-q)\cdot\left(\frac{(1-2r)(1+2\vartheta_l)}{(3+2\vartheta_l)(1+\vartheta_l)} + \frac{2(1-x)(r-q)}{(3+2\vartheta_l)} + \frac{2x(r-p)}{(3+2\vartheta_h)}\right) \\ &\quad -(1-2r)(r-p)(\vartheta_h - \vartheta_l) \\ &\quad \cdot\left(\frac{(2-\vartheta_h\vartheta_l)}{(2+\vartheta_l)(1+\vartheta_l)(2+\vartheta_h)(1+\vartheta_h)} - \frac{2(7+2\vartheta_l+2\vartheta_h)}{(2+\vartheta_h)(3+2\vartheta_h)(2+\vartheta_l)(3+2\vartheta_l)}\right) \\ &\quad +\frac{2(r-q)(r-p)(\vartheta_h-\vartheta_l)}{(3+\vartheta_h+\vartheta_l)}\cdot\left(\frac{(1-x)}{(3+2\vartheta_l)} + \frac{x}{(3+2\vartheta_h)}\right) + \frac{4r(1-r)(\vartheta_h-\vartheta_l)}{(3+2\vartheta_l)(3+2\vartheta_h)}\Bigg]. \end{aligned} \tag{2.3.9}$$

Remark 2.14 (Checking the fixation probability)**.** Some symmetries in (2.3.8) (or equivalently in (2.3.9)) can directly be seen:

- The right-hand side changes sign if we exchange $\vartheta_h \leftrightarrow \vartheta_\ell$, $p \leftrightarrow q$ and $x \leftrightarrow 1-x$, since the roles of X_h and X_ℓ are simply exchanged.
- If $p = q = r = 0$ or $p = q = r = 1$, the right-hand side is 0.
- If $\vartheta_h = \vartheta_\ell = 0$, the result does not depend on r since there are no mutations.
- If $\vartheta_h = \vartheta_\ell$ and $p = q$, the right-hand side is 0 since X_h and X_ℓ are the same (in distribution).

Another interesting case is $p = q = r$, which means that both X_h and X_ℓ are in their mutational balance already at time 0. In this case, we find that

$$\mathbb{P}(X_h(\infty) = 1) \approx x + 4x(1-x)\frac{\alpha^2}{\sigma}\frac{r(1-r)(\vartheta_h - \vartheta_l)}{(3+2\vartheta_l)(3+2\vartheta_h)}$$

for small α^2/σ. This means that the fixation probability of X_h is greater than under neutrality (i.e. for $\alpha^2 = 0$) if and only if $\vartheta_h > \vartheta_\ell$.

Remark 2.15 (Computing moments under neutrality)**.** In the proof of Theorem 2.12, we will have to compute moments of X under neutral evolution, i.e. $\alpha^2/\sigma = 0$ in (2.3.2). Since the evolution of X is only driven by mutation and resampling then, such moments can be computed using the coalescent (Durrett, 2008), which is dual to the solution of (2.3.2). Assume we aim to compute an n-th moment of $X(t)$, i.e. $\mathbb{E}[X_{a_1 i_1}(t) \cdots X_{a_n i_n}(t)]$ for some $a_1, ..., a_n \in \{\ell, h\}$ and $i_1, ..., i_n \in \{0, 1\}$. Then, the coalescent starts with n lineages, any (unordered) pair of lineages coalesces independently at rate 1, and the resulting lineages, stopped after having evolved for time t, are assigned some type, randomly chosen from $X(0)$. Mutations are modeled on top of this tree structure, and we have to deal with all cases such that lineage k is assigned type $a_k i_k$, for $k = 1, ..., n$. Since there is no mutation transforming ℓ to h and back, lineages assigned with ℓ must not coalesce with lineages with h, and ancestors of ℓ (h) must be of type ℓ (h). On all such events, mutation from 0 to 1 and back (at rates ϑ_h and ϑ_ℓ, depending on the type at the first locus) determines types at the second locus. These arguments will be used below starting in (2.3.11).

Proof of Theorem 2.12. We will use the equality (recall (2.3.3))

$$\begin{aligned}\mathbb{P}_x(X_h(\infty) = 1) &= \mathbb{E}_x[X_h(\infty)] = x + \int_0^\infty \mathbb{E}[GX_h(t)]\mathrm{d}t \qquad (2.3.10)\\ &= x + \frac{\alpha^2}{\sigma}\int_0^\infty \mathbb{E}[(X_{h0}(t)X_{\ell 1}(t) - X_{h1}(t)X_{\ell 0}(t))(X_1(t) - X_0(t))]\mathrm{d}t,\end{aligned}$$

together with

$$\begin{aligned}&(X_{h0}X_{\ell 1} - X_{h1}X_{\ell 0})(X_1 - X_0)\\ =&(X_{h0}X_{\ell 1} + X_{h0}X_{\ell 0} - X_{h1}X_{\ell 0} - X_{h0}X_{\ell 0})(1 - 2X_0)\\ =&(X_\ell X_{h0} - X_h X_{\ell 0}) + 2((X_h X_{h0} X_{\ell 0} - X_\ell X_{h0} X_{\ell 0}) + (X_h X_{\ell 0}^2 - X_\ell X_{h0}^2)).\end{aligned}$$

Since we are studying the case of low α^2/σ, and the integral in (2.3.10) is continuous in α^2/σ, we only need to evaluate the integral at $\alpha^2/\sigma = 0$. From (2.3.2), we see that we need to study neutral evolution with the same mutation mechanism. We will write $\mathbb{P}(.)$ for the corresponding probability measure and $\mathbb{E}[.]$ for the expectation under neutral evolution. Following Remark 2.15, we start with

$$\begin{aligned}&\mathbb{E}[X_h(t)] = X_h(0), \qquad \mathbb{E}[X_\ell(t)] = X_\ell(0)\\ &\mathbb{E}[X_{h0}(t)] = e^{-\vartheta_h t}X_{h0}(0) + (1 - e^{-\vartheta_h t})rX_h(0) = x(r + e^{-\vartheta_h t}(p - r)), \qquad (2.3.11)\\ &\mathbb{E}[X_{\ell 0}(t)] = (1-x)(r + e^{-\vartheta_\ell t}(q - r)),\end{aligned}$$

since either no mutation at the B-locus happened by time t and the ancestor at time 0 had type 0, or a mutation occurred which resulted in a type 0 at the B-locus. Then, for

$\mathbb{E}[X_\ell(t)X_{h0}(t)]$, note that coalescence of the two corresponding lines must not have occurred by time t since mutation cannot transform ℓ to h or back. The same argument applies to $\mathbb{E}[X_h(t)X_{\ell 0}(t)]$, hence

$$\begin{aligned}
&\int_0^\infty \mathbb{E}[X_\ell(t)X_{h0}(t) - X_h(t)X_{\ell 0}(t)]\mathrm{d}t \\
&= \int_0^\infty e^{-t}((1-x)x(r+e^{-\vartheta_h t}(p-r)) - x(1-x)(r+e^{-\vartheta_\ell t}(q-r))\mathrm{d}t \\
&= x(1-x)\Big(\frac{p-r}{1+\vartheta_h} - \frac{q-r}{1+\vartheta_\ell}\Big). \qquad (2.3.12)
\end{aligned}$$

For $\mathbb{E}[X_hX_{h0}X_{\ell 0} - X_\ell X_{h0}X_{\ell 0}]$, coalescence may occur between the two h-lines in the first and the two ℓ-lines in the second term. However, on the event that such a coalescence occurs, $\mathbb{E}[X_hX_{h0}X_{\ell 0}, \text{coal}] = \mathbb{E}[X_{h0}X_{\ell 0}, \text{coal}] = \mathbb{E}[X_\ell X_{h0}X_{\ell 0}, \text{coal}]$, i.e. this case cancels. Hence,

$$\begin{aligned}
&\int_0^\infty \mathbb{E}[X_h(t)X_{h0}(t)X_{\ell 0}(t) - X_\ell(t)X_{h0}(t)X_{\ell 0}(t)]\mathrm{d}t \\
&= \int_0^\infty e^{-3t}x(1-x)(2x-1)(r+e^{-\vartheta_h t}(p-r))(r+e^{-\vartheta_\ell t}(q-r))\mathrm{d}t \\
&= x(1-x)(2x-1)\Big(\frac{r^2}{3} + \frac{r(p-r)}{3+\vartheta_h} + \frac{r(q-r)}{3+\vartheta_\ell} + \frac{(p-r)(q-r)}{3+\vartheta_h+\vartheta_\ell}\Big). \qquad (2.3.13)
\end{aligned}$$

For $\mathbb{E}[X_h(t)X_{\ell 0}(t)^2 - X_\ell(t)X_{h0}(t)^2]$, either no coalescence occurs, or colescence occurs between the two ℓ-lins (h-lines) in the first (second) term. In this case, either no mutation occurs on both branches to the most recent common ancestor, and this has type $\ell 0$ ($h0$), or mutation occurs on exactly on one branch, or on both branches. So,

$$\begin{aligned}
&\int_0^\infty \mathbb{E}\left[X_h(t)X_{\ell 0}(t)^2 - X_\ell(t)X_{h0}(t)^2\right]\mathrm{d}t \\
&= \int_0^\infty e^{-3t}x(1-x)\left((1-x)(r+e^{-\vartheta_\ell t}(q-r))^2 - x(r+e^{-\vartheta_h t}(p-r))^2\right)\mathrm{d}t \\
&\quad + \int_0^\infty \int_0^t e^{-3s}e^{-(t-s)} \\
&\quad \cdot \Big[x\Big(e^{-2\vartheta_\ell s}\mathbb{E}[X_{\ell 0}(t-s)] + \underbrace{2e^{-\vartheta_\ell s}(1-e^{-\vartheta_\ell s})r\mathbb{E}[X_{\ell 0}(t-s)] + (1-e^{-\vartheta_\ell s})^2r^2(1-x)}_{=(1-e^{-\vartheta_\ell s})r(2X_{\ell 0}(t)-(1-e^{-\vartheta_\ell s})r(1-x))}\Big) \\
&\qquad - (1-x)\Big(e^{-2\vartheta_h s}\mathbb{E}[X_{h0}(t-s)] + 2e^{-\vartheta_h s}(1-e^{-\vartheta_h s})r\mathbb{E}[X_{h0}(t-s)] \\
&\qquad\qquad + (1-e^{-\vartheta_h s})^2r^2x\Big)\Big]\mathrm{d}s\mathrm{d}t \\
&= x(1-x) \qquad (2.3.14) \\
&\quad \cdot \left((1-2x)\frac{r^2}{3} + (1-x)\left(\frac{2r(q-r)}{3+\vartheta_\ell} + \frac{(q-r)^2}{3+2\vartheta_\ell}\right) - x\left(\frac{2r(p-r)}{3+\vartheta_h} + \frac{(p-r)^2}{3+2\vartheta_h}\right)\right) \qquad (2.3.15) \\
&\quad + x(1-x)\left[\int_0^\infty \int_s^\infty e^{-3s}e^{-(t-s)}\left(e^{-2\vartheta_\ell s}(r+e^{-\vartheta_\ell(t-s)}(q-r))\right.\right. \qquad (2.3.16) \\
&\qquad\qquad \left. -e^{-2\vartheta_h s}(r+e^{-\vartheta_h(t-s)}(p-r))\right)\mathrm{d}t\mathrm{d}s \\
&\quad +r\int_0^\infty \int_s^\infty e^{-3s}e^{-(t-s)}\Big(2(1-e^{-\vartheta_\ell s})(r+e^{-\vartheta_\ell t}(q-r)) - (1-e^{-\vartheta_\ell s})^2r \qquad (2.3.17)
\end{aligned}$$

$$-2(1-e^{-\vartheta_h s})(r+e^{-\vartheta_h t}(p-r))+(1-e^{-\vartheta_h s})^2 r\Big)\mathrm{d}t\mathrm{d}s\Big].$$

Now, for (2.3.16)

$$\begin{aligned}
&\int_0^\infty\int_s^\infty e^{-3s}e^{-(t-s)}\Big(e^{-2\vartheta_\ell s}(r+e^{-\vartheta_\ell(t-s)}(q-r))-e^{-2\vartheta_h s}(r+e^{-\vartheta_h(t-s)}(p-r))\mathrm{d}t\mathrm{d}s\\
&\quad=\int_0^\infty\int_0^\infty e^{-3s}e^{-t}\Big(e^{-2\vartheta_\ell s}(r+e^{-\vartheta_\ell t}(q-r))-e^{-2\vartheta_h s}(r+e^{-\vartheta_h t}(p-r))\mathrm{d}t\mathrm{d}s\\
&\quad=\frac{1}{3+2\vartheta_\ell}\Big(r+\frac{q-r}{1+\vartheta_\ell}\Big)-\frac{1}{3+2\vartheta_h}\Big(r+\frac{p-r}{1+\vartheta_h}\Big)\\
&\quad=\frac{r\vartheta_\ell+q}{(3+2\vartheta_\ell)(1+\vartheta_\ell)}-\frac{r\vartheta_h+p}{(3+2\vartheta_h)(1+\vartheta_h)}
\end{aligned}\tag{2.3.18}$$

and for (2.3.17),

$$\begin{aligned}
&\int_0^\infty\int_s^\infty e^{-3s}e^{-(t-s)}\Big(2(1-e^{-\vartheta_\ell s})(r+e^{-\vartheta_\ell t}(q-r))-(1-e^{-\vartheta_\ell s})^2 r\\
&\qquad\qquad-2(1-e^{-\vartheta_h s})(r+e^{-\vartheta_h t}(p-r))+(1-e^{-\vartheta_h s})^2 r)\Big)\mathrm{d}t\mathrm{d}s\\
&=\int_0^\infty\int_s^\infty e^{-2s}e^{-t}\Big(2(1-e^{-\vartheta_\ell s})e^{-\vartheta_\ell t}(q-r)-e^{-2\vartheta_\ell s}r\\
&\qquad\qquad-2(1-e^{-\vartheta_h s})e^{-\vartheta_h t}(p-r)+e^{-2\vartheta_h s}r)\Big)\mathrm{d}t\mathrm{d}s\\
&=\int_0^\infty e^{-3s}\Big(2(1-e^{-\vartheta_\ell s})\frac{1}{1+\vartheta_\ell}e^{-\vartheta_\ell s}(q-r)-e^{-2\vartheta_\ell s}r\\
&\qquad\qquad-2(1-e^{-\vartheta_h s})e^{-\vartheta_h s}\frac{1}{1+\vartheta_h}(p-r)+e^{-2\vartheta_h s}r)\Big)\mathrm{d}t\mathrm{d}s\\
&=\frac{2(q-r)}{1+\vartheta_\ell}\Big(\frac{1}{3+\vartheta_\ell}-\frac{1}{3+2\vartheta_\ell}\Big)-\frac{2(p-r)}{1+\vartheta_h}\Big(\frac{1}{3+\vartheta_h}-\frac{1}{3+2\vartheta_h}\Big)+\frac{r}{3+2\vartheta_h}-\frac{r}{3+2\vartheta_\ell}\\
&=\frac{2\vartheta_\ell(q-r)}{(1+\vartheta_\ell)(3+\vartheta_\ell)(3+2\vartheta_\ell)}-\frac{2\vartheta_h(p-r)}{(1+\vartheta_h)(3+\vartheta_h)(3+2\vartheta_h)}+\frac{r}{3+2\vartheta_h}-\frac{r}{3+2\vartheta_\ell}.
\end{aligned}\tag{2.3.19}$$

Summing $(2.3.12)+2\cdot(2.3.13)+2\cdot(2.3.15)+2x(1-x)\cdot(2.3.18)+2x(1-x)r\cdot(2.3.19)$ gives

$$\begin{aligned}
&\int_0^\infty\mathbb{E}[(X_{h0}(t)X_{\ell 1}(t)-X_{h1}(t)X_{\ell 0}(t))(X_1(t)-X_0(t))]\mathrm{d}t\\
&=x(1-x)\\
&\quad\cdot\Bigg[\frac{p-r}{1+\vartheta_h}-\frac{q-r}{1+\vartheta_\ell}+2\Big(\frac{r(q-r)}{3+\vartheta_\ell}+(1-x)\frac{(q-r)^2}{3+2\vartheta_\ell}\Big)\\
&\qquad-2\Big(\frac{r(p-r)}{3+\vartheta_h}+x\frac{(p-r)^2}{3+2\vartheta_h}\Big)+(2x-1)\frac{2(p-r)(q-r)}{3+\vartheta_\ell+\vartheta_h}\\
&\qquad+2\Bigg(\Big(\frac{r\vartheta_\ell+q}{(3+2\vartheta_\ell)(1+\vartheta_\ell)}+\frac{r(1-r)-r}{3+2\vartheta_\ell}\Big)-\Big(\frac{r\vartheta_h+p}{(3+2\vartheta_h)(1+\vartheta_h)}+\frac{r(1-r)-r}{3+2\vartheta_h}\Big)\\
&\qquad+\frac{4\vartheta_\ell r(q-r)}{(1+\vartheta_\ell)(3+\vartheta_\ell)(3+2\vartheta_\ell)}-\frac{4\vartheta_h r(p-r)}{(1+\vartheta_h)(3+\vartheta_h)(3+2\vartheta_h)}\Bigg)\Bigg]\\
&=x(1-x)
\end{aligned}$$

$$
\begin{aligned}
&\cdot\Bigg[\frac{p-r}{1+\vartheta_h}-\frac{q-r}{1+\vartheta_\ell}+2\bigg((1-x)\frac{(q-r)^2}{3+2\vartheta_\ell}-x\frac{(p-r)^2}{3+2\vartheta_h}+(2x-1)\frac{(p-r)(q-r)}{3+\vartheta_\ell+\vartheta_h}\bigg)\\
&\quad+\frac{2r(q-r)}{3+\vartheta_\ell}\underbrace{\bigg(1+\frac{2\vartheta_\ell}{(1+\vartheta_\ell)(3+2\vartheta_\ell)}\bigg)}_{=\frac{(1+2\vartheta_\ell)(3+\vartheta_\ell)}{(1+\vartheta_\ell)(3+2\vartheta_\ell)}}-\frac{2r(p-r)}{3+\vartheta_h}\bigg(1+\frac{2\vartheta_h}{(1+\vartheta_h)(3+2\vartheta_h)}\bigg)\\
&\quad+2\bigg(\frac{q-r}{(3+2\vartheta_\ell)(1+\vartheta_\ell)}-\frac{p-r}{(3+2\vartheta_h)(1+\vartheta_h)}+r(1-r)\bigg(\frac{1}{3+2\vartheta_\ell}-\frac{1}{3+2\vartheta_h}\bigg)\bigg)\Bigg]\\
&=x(1-x)\\
&\cdot\Bigg[\frac{(p-r)(1+2\vartheta_h)}{(1+\vartheta_h)(3+2\vartheta_h)}-\frac{(q-r)(1+2\vartheta_\ell)}{(1+\vartheta_\ell)(3+2\vartheta_\ell)}\\
&\quad+2\bigg((1-x)\frac{(q-r)^2}{3+2\vartheta_\ell}-x\frac{(p-r)^2}{3+2\vartheta_h}+(2x-1)\frac{(p-r)(q-r)}{3+\vartheta_\ell+\vartheta_h}\bigg)\\
&\quad+\frac{2r(q-r)(1+2\vartheta_\ell)}{(1+\vartheta_\ell)(3+2\vartheta_\ell)}-\frac{2r(p-r)(1+2\vartheta_h)}{(1+\vartheta_h)(3+2\vartheta_h)}+2r(1-r)\bigg(\frac{1}{3+2\vartheta_\ell}-\frac{1}{3+2\vartheta_h}\bigg)\Bigg]\\
&=x(1-x)\\
&\cdot\Bigg[\frac{(2r-1)(q-r)(1+2\vartheta_\ell)}{(1+\vartheta_\ell)(3+2\vartheta_\ell)}-\frac{(2r-1)(p-r)(1+2\vartheta_h)}{(1+\vartheta_h)(3+2\vartheta_h)}\\
&\quad+2\bigg((1-x)\frac{(q-r)^2}{3+2\vartheta_\ell}-x\frac{(p-r)^2}{3+2\vartheta_h}+(2x-1)\frac{(p-r)(q-r)}{3+\vartheta_\ell+\vartheta_h}\\
&\qquad\qquad\qquad+r(1-r)\bigg(\frac{1}{3+2\vartheta_\ell}-\frac{1}{3+2\vartheta_h}\bigg)\bigg)\Bigg]
\end{aligned}
$$

which together with (2.3.10) shows (2.3.8). □

Chapter 3

Recent-admixture model

3.1 Introduction

The human genome consists of a little over three billion base pairs that are composed of two nucleobases and form the building blocks of the DNA. Sequences of these base pairs are called genes that code for either proteins or RNA. By sequencing DNA samples from multiple people one can detect single-nucleotide changes that are responsible for the variation we see in humans. Single-nucleotide changes of which two or more variations are present in at least 1% of the population are classified as so-called *single-nucleotide polymorphisms (SNPs)*. Either on its own or in combination with other SNPs, they determine susceptibility to various diseases, the way of responding to specific drugs and many other things. Therefore the detecting and the understanding of SNPs and their impact are of great importance. As a result of major technological advances in the past decades leading to faster and more inexpensive ways of *reading off* the DNA we have been able to identify more and more SNPs.

The sheer amount of data nowadays enables us to analyse the human genome in various aspects. The aspect we want to focus on in this chapter is the study of *biogeographical ancestry*. Among those SNPs, one can find some whose frequencies are highly dependent on the population where the individual carrying the SNP comes from. These are called *ancestry-informative markers* (AIMs) and these are the very markers at the DNA which are used when analysing DNA samples with special focus on genetic ancestry. When talking about genetic ancestry we make a clear distinction between biogeographical ancestry and ethnicity, the latter being more of a social construct independent from genetics. Genetic ancestry is a very broad term that tackles various problems found in population genetics such as identifying population structure, assigning individuals to specific populations and so on; see Liu et al. (2013) for a more detailed review. In this chapter we will focus on the issue of inferring one's individual admixture (IA) proportions, i.e. we would like to assign to each ancestral population the fraction of our genome originating from that very population. One can observe an increasing interest of the general public caused by, amongst others, companies such as *23andMe* that deliver information on the genetic ancestry of their customers. However, inference of IA also has application in, for instance, forensics where DNA traces left at the crime scene are investigated. Due to the increasing amount of available DNA data, this approach has become a well-established research field in forensic genetics (see e.g. Phillips et al. (2016); Eduardoff et al. (2016); Kidd et al. (2017)).

STRUCTURE, which is a Bayesian approach using MCMC developed by Pritchard et al.

(2000), is probably the most widely used program when estimating IA proportions, later followed by software such as ADMIXTURE by Alexander et al. (2009) and FRAPPE by Tang et al. (2005), which use the faster likelihood-based approach. Their models are based on the idea that allelic states at each marker are indepentent from each other - also known as the Hardy-Weinberg equilibrium. However, when we have an individual whose parents come from two different populations, this individual's genome will exhibit a high frequency of heterozygotes also known as the Wahlund effect (Wahlund (1928)). Hence, the Hardy-Weinberg assumption does not apply in cases of recently admixed individuals. In this chapter, we extend the likelihood-model behind STRUCTURE or ADMIXTURE and FRAPPE in order to account for recent admixture. We do so by introducing two IA vectors, one for the mother and the other for the father. In doing so, we deliberately use the information that each marker of a child's genome consists of one allele passed on by the mother and the other passed on by the father.

This chapter is structured as follows: We will first introduce the admixture model on which softwares such as STRUCTURE, ADMIXTURE and FRAPPE rely on. We then extend this model as described above and call it the recent-admixture model. The rest of this chapter deals with the application of our method on the 1000 genomes dataset and the comparison of the results obtained by the admixture model and the recent-admixture model.

This chapter is based on a collaboration with Peter Pfaffelhuber, Franz Baumdicker, Fabian Staubach and Denise Syndercombe-Court which is still work in progress. The implementation of our methods can be downloaded from `https://github.com/pfaffelh/recent-admixture`.

3.2 Theory

In this section we will first recall the admixture model, which is the basis for the sofware STRUCTURE, ADMIXTURE and FRAPPE. First we write down the admixture model and derive a method to estimate IA in the case when allele frequencies within populations are not updated. Secondly, we introduce the recent-admixture model, where an individual is allowed to have parents with different admixture proportions. We take the following notation for the reference database:

$$\begin{aligned} K &: \text{number of ancestral populations,}\\ M &: \text{number of markers,}\\ p_{mk} &: \text{frequency of allele 1 at (bi-allelic) marker } m \text{ in population } k. \end{aligned}$$

In addition, we consider one additional diploid genome $(G_{m1}, G_{m2})_{m=1,\dots,M}$, or $(G_m)_{m=1,\dots,M}$ with $G_m = G_{m1} + G_{m2}$ if phase is not known. The goal is to estimate admixture proportions $(q_k)_{k=1,\dots,K}$ (or $(q_k^M)_{k=1,\dots,K}, (q_k^P)_{k=1,\dots,K}$) of this additional genome, where q_k (and q_k^M and q_k^P) is the fraction of the genome originating from population k.

Remark 3.1 (Estimation of allele frequencies)**.** Beside the estimation of IA, programs such as STRUCTURE, ADMIXTURE and FRAPPE also aim to simultaneously update the allele frequencies in ancestral populations meaning that a new trace - which we wish to obtain information on ancestry from - changes allelic frequencies in the acestral populations during the runtime. As a result, these programs require long computation times. In our recent-admixture model, we will skip this step and we will only update our IAs. Our reasoning is the following: In forensics we mostly have a large reference database and only one or a few new traces, therefore, the impact of these new traces on the allelic frequencies is negligible. In other words, we take the computationally easier way and take allele frequencies for ancestral populations as fixed, and only change the IAs of the new traces during the runtime.

3.2.1 The admixture model

Suppose we observe $(G_{ma})_{m=1,...,M;a=1,2}$. The main goal of the admixture model is to find an IA vector $q = (q_k)_{k=1,...,K}$ that maximises the log-likelihood (see also (1) and (2) of Tang et al. (2005))

$$\ell(q|G) = \sum_{m=1}^{M} \sum_{a=1,2} \log\left(\sum_{k=1}^{K} \alpha_{mak} q_k\right),$$

where

$$\alpha_{mak} := \begin{cases} p_{mk}, & \text{if } G_{ma} = 1, \\ 1 - p_{mk}, & \text{if } G_{ma} = 0 \end{cases} \tag{3.2.1}$$

is the frequency of the observed allele in copy $a = 1, 2$ of marker m in population k. (Note that $e^{\ell(q|G)}$ is the probabiltiy of observing $(G_{ma})_{m=1,...,M;a=1,2}$, if every allele is picked independently from population k with probability q_k.) Assuming that phase is not known, and with

$$\alpha_{mkl} := \alpha_{m1k}\alpha_{m2l} = \begin{cases} p_{mk}p_{ml}, & \text{if } G_{m1} + G_{m2} = 2, \\ p_{mk}(1 - p_{ml}) + (1 - p_{mk})p_{ml}, & \text{if } G_{m1} + G_{m2} = 1, \\ (1 - p_{mk})(1 - p_{ml}), & \text{if } G_{m1} + G_{m2} = 0, \end{cases} \tag{3.2.2}$$

note that the log-likelihood can as well be written as

$$\ell(q|G) = \sum_{m=1}^{M} \log\left(\sum_{k,l=1}^{K} \alpha_{mkl} q_k q_l\right).$$

Provided that each allele observed in the trace comes from population k with probability q_k, the probability to observe allele 1 at marker m is $\beta_m(q) := \sum_k p_{mk} q_k$. By distinguishing between the three states for G_m we easily obtain

$$\sum_{k,l=1}^{K} \alpha_{mkl} q_k q_l = \begin{cases} \beta_m(q)^2, & \text{for } G_m = 2, \\ 2\beta_m(q)(1 - \beta_m(q)), & \text{for } G_m = 1, \\ (1 - \beta_m(q))^2, & \text{for } G_m = 0, \end{cases} \tag{3.2.3}$$

such that

$$\ell(q|G) = \sum_{m=1}^{M} \log\left(\binom{2}{G_m} \beta_m(q)^{G_m} (1 - \beta_m(q))^{2-G_m}\right). \tag{3.2.4}$$

Lemma 3.2. *The maximum of $q \mapsto \ell(q|G)$ under the constraint $\sum_{k=1}^{K} q_k = 1$ solves*

$$\frac{1}{M} \sum_{m=1}^{M} \sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'=1}^{K} \alpha_{mk'l'} q_{k'} q_{l'}} = 1, \qquad k = 1, ..., K. \tag{3.2.5}$$

Remark 3.3. 1. Let us add a set of *Ancestry Uninformative Markers*, i.e. a set of markers $\tilde{m} = 1, ..., \tilde{M}$, with frequencies not depending on the population, i.e. $\alpha_{\tilde{m}kl} = \alpha_{\tilde{m}k'l'} = \alpha_{\tilde{m}}$ for $k, l, k', l' = 1, ..., K$. Ideally, adding such a set of marker should not influence the maximum of our likelihood function. Indeed this holds since

$$\sum_{\tilde{m}=1}^{\tilde{M}} \sum_{l=1}^{K} \frac{\alpha_{\tilde{m}kl} q_l}{\sum_{k',l'=1}^{K} \alpha_{\tilde{m}k'l'} q_{k'} q_{l'}} = \sum_{\tilde{m}=1}^{\tilde{M}} \frac{\alpha_{\tilde{m}} \sum_{l=1}^{K} q_l}{\alpha_{\tilde{m}} \sum_{k',l'=1}^{K} q_{k'} q_{l'}} = \tilde{M}.$$

2. Fixing $\beta_m := \sum_{k=1}^{K} p_{mk} q_k$, we take a look at the left-hand side of (3.2.5).

For $G_m = 2$, it holds

$$\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'} \alpha_{mk'l'} q_{k'} q_{l'}} = \sum_{l=1}^{K} \frac{p_{mk} p_{ml} q_l}{\sum_{k',l'} p_{mk'} p_{ml'} q_{k'} q_{l'}} = \frac{p_{mk}}{\beta_m},$$

for $G_m = 1$, we have

$$\begin{aligned} \sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'} \alpha_{mk'l'} q_{k'} q_{l'}} &= \sum_{l=1}^{K} \frac{(p_{mk}(1 - p_{ml}) + (1 - p_{mk}) p_{ml}) q_l}{\sum_{k',l'} (p_{mk'}(1 - p_{ml'}) + (1 - p_{mk'}) p_{ml'}) q_{k'} q_{l'}} \\ &= \frac{p_{mk}(1 - \beta_m) + (1 - p_{mk}) \beta_m}{2\beta_m (1 - \beta_m)} = \frac{1}{2} \Big(\frac{p_{mk}}{\beta_m} + \frac{1 - p_{mk}}{1 - \beta_m} \Big) \end{aligned}$$

and finally for $G_m = 0$

$$\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'} \alpha_{mk'l'} q_{k'} q_{l'}} = \sum_{l=1}^{K} \frac{(1 - p_{mk})(1 - p_{ml}) q_l}{\sum_{k',l'} (1 - p_{mk'})(1 - p_{ml'}) q_{k'} q_{l'}} = \frac{1 - p_{mk}}{1 - \beta_m}.$$

Therefore, the maximum q we are looking for in Lemma 3.2 needs to solve

$$\frac{1}{2M} \sum_{m=1}^{M} \left(G_m \frac{p_{mk}}{\beta_m} + (2 - G_m) \frac{1 - p_{mk}}{1 - \beta_m} \right) = 1. \tag{3.2.6}$$

3. With the help of (3.2.6), we can turn the maximization problem from (3.2.5) into the fixation problem where we search for $\hat{q} = (\hat{q}_k)_{k=1,...,K}$ such that $\hat{q} = f_k(\hat{q})$ for

$$f_k(q) = \frac{1}{2M} \sum_{m=1}^{M} \Big(G_m \frac{p_{mk}}{\beta_m} + (2 - G_m) \frac{1 - p_{mk}}{1 - \beta_m} \Big) q_k, \qquad k = 1, ..., K. \tag{3.2.7}$$

A solution can be computed by iteratively computing $q_{n+1} = (f_k(q_n))_{k=1,...,K}$ until convergence. (In our implementation, we continue the iteration until $|q_{n+1} - q_n| < \varepsilon$ for $\varepsilon = 10^{-6}$.) We note that this approach is essentially the same as in the EM-algorithm from Tang et al. (2005), but without carrying out the maximization step.

Proof of Lemma 3.2. We use the theory of *Lagrange multipliers* which is a method for finding local extrema under certain constraints. In our case we need to maximise ℓ over q under the constraint $\sum_k q_k = 1$. We recall that

$$\ell(q|G) = \sum_{m=1}^{M} \log \left(\sum_{k,l=1}^{K} \alpha_{mkl} q_k q_l \right)$$

and we obtain

$$\frac{\partial \ell(q|G)}{\partial q_k} = \sum_{m=1}^{M} \sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'=1}^{K} \alpha_{mk'l'} q_{k'} q_{l'}}.$$

We therefore have to solve the system of equations

$$\begin{aligned} \lambda &= \sum_{m=1}^{M} \sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'=1}^{K} \alpha_{mk'l'} q_{k'} q_{l'}}, \qquad k = 1, ..., K, \\ 1 &= \sum_{k=1}^{K} q_k. \end{aligned} \tag{3.2.8}$$

It is easy to eliminate λ, since using both equations from (3.2.8) gives

$$\lambda = \lambda \sum_{k=1}^{K} q_k = \sum_{m=1}^{M} \sum_{k,l=1}^{K} \frac{\alpha_{mkl} q_k q_l}{\sum_{k',l'=1}^{K} \alpha_{mk'l'} q_{k'} q_{l'}} = M.$$

Dividing both sides by M, we are left with finding $q = (q_k)_{k=1,...,K}$ such that

$$\frac{1}{M} \sum_{m=1}^{M} \sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'=1}^{K} \alpha_{mk'l'} q_{k'} q_{l'}} = 1, \qquad k = 1, ..., K. \tag{3.2.9}$$

□

3.2.2 The recent-admixture model

In the previous admixture model we obtain an IA vector $q = (q_k)_{k=1,...,K}$. We extend this model by distinguishing between maternally and paternally inherited alleles meaning that for each new trace, we will not only find an IA vector q but two IA vectors, one belonging to the mother and the other belonging to the father. The maternally inherited alleles come with IA q^M, and the paternally inherited alleles with q^P. In Section 3.4 we will see that taking the average of q^M and q^P, we obtain IA which is highly similar to q in the admixture model in non-admixed and admixed individuals. Furthermore the IA estimated from the recent-admixture model in recently admixed individuals, in particular if the two parents come from different populations, is more accurate than for the admixture model. As we estimate not only the IA of the individual but the IAs of the parents we obtain even more information on the heritage of the individual and we also gain information on the form of recent admixture.

As we said, our next goal is to find IAs, $q^M = (q_k^M)_{k=1,...,K}, q^P = (q_k^P)_{k=1,...,K}$, of the sampled individual's parents. The new log-likelihood is then

$$\ell(q^M, q^P|G) = \sum_{m=1}^{M} \log \Big(\sum_{k,l=1}^{K} \alpha_{mkl} q_k^M q_l^P \Big),$$

where α_{mkl} is given as in (3.2.2).

Choosing $q^M = q^P = q$ we have that $\ell(q^M, q^P|G) = \ell(q|G)$. Thus the admixture model is a special case in the recent-admixture model. $q^M = q^P = q$ in the recent-admixture model gives the admixture model.

Just like in the admixture model (see (3.2.3)), we can rewrite the log-likelihood and we get

$$\begin{aligned}\ell(q^M, q^P|G) = \sum_{m=1}^{M} \log \Big(& 1_{G_m=2}\beta_m(q^M)\beta_m(q^P) \\ & + 1_{G_m=1}(\beta_m(q^M)(1-\beta_m(q^P)) + (1-\beta_m(q^M))\beta_m(q^P)) \\ & + 1_{G_m=0}(1-\beta_m(q^M))(1-\beta_m(q^P))\Big).\end{aligned} \tag{3.2.10}$$

Lemma 3.4. *The maximum of* $q \mapsto \ell(q^P, q^M, G)$ *under the constraint* $\sum_{k=1}^K q_k^M = \sum_{k=1}^K q_k^P = 1$ *solves for* $k = 1, ..., K$

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{K}\frac{\alpha_{mkl}q_l^M}{\sum_{k',l'=1}^K \alpha_{mk'l'}q_{k'}^M q_{l'}^P} = \frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{K}\frac{\alpha_{mkl}q_l^P}{\sum_{k',l'=1}^K \alpha_{mk'l'}q_{k'}^P q_{l'}^M} = 1. \tag{3.2.11}$$

Remark 3.5.

1. Note that (3.2.11) is symmetric in q^M and q^P, i.e. if (q^M, q^P) solves (3.2.11), another solution is given by (q^P, q^M).

2. As in Remark 3.3 we have a closer look at the left-hand side of (3.2.11). For $\beta_m(q) := \sum_k p_{mk}q_k$, we have for $G_m = 2$

$$\sum_{l=1}^{K}\frac{\alpha_{mkl}q_l^M}{\sum_{k',l'=1}^K \alpha_{mk'l'}q_{k'}^M q_{l'}^P} = \frac{p_{mk}\beta_m(q^M)}{\beta_m(q^M)\beta_m(q^P)} = \frac{p_{mk}}{\beta_m(q^P)},$$

for $G_m = 1$

$$\sum_{l=1}^{K}\frac{\alpha_{mkl}q_l^M}{\sum_{k',l'=1}^K \alpha_{mk'l'}q_{k'}^M q_{l'}^P} = \frac{p_{mk}(1-\beta_m(q^M)) + (1-p_{mk})\beta_m(q^M)}{\beta_m(q^M)(1-\beta_m(q^P)) + (1-\beta_m(q^M))\beta_m(q^P)}$$

and finally for $G_m = 0$

$$\sum_{l=1}^{K}\frac{\alpha_{mkl}q_l^M}{\sum_{k',l'=1}^K \alpha_{mk'l'}q_{k'}^M q_{l'}^P} = \frac{(1-p_{mk})(1-\beta_m(q^M))}{(1-\beta_m(q^M))(1-\beta_m(q^P))} = \frac{1-p_{mk}}{1-\beta_m(q^P)}.$$

3. Just like in the admixture model, we can turn (3.2.11) into fixed point equations to find $\hat{q}^M, \hat{q}^P$ such that $\hat{q}^P = f_k(\hat{q}^M, \hat{q}^P)$ and $\hat{q}^M = f_k(\hat{q}^M, \hat{q}^P)$ for $f(q, q') = (f_k(q, q'))_{k=1,...,K}$ with

$$\begin{aligned}f_k(q, q') := \frac{1}{M}\sum_{m=1}^{M}\Big(& \mathbb{1}_{\{G_m=2\}}\frac{p_{mk}}{\beta_m(q')} \\ & + \mathbb{1}_{\{G_m=1\}}\frac{(p_{mk}(1-\beta_m(q)) + (1-p_{mk})\beta_m(q))}{\beta_m(q)(1-\beta_m(q')) + (1-\beta_m(q))\beta_m(q')} \\ & + \mathbb{1}_{\{G_m=0\}}\frac{(1-p_{mk})}{1-\beta_m(q')}\Big)q'_k.\end{aligned} \tag{3.2.12}$$

In our implementation, we iteratively compute

$$q^P_{n+1} = f(q^M_n, q^P_n) \quad \text{and} \quad q^M_{n+1} = f(q^P_{n+1}, q^M_n)$$

until convergence.

Proof of Lemma 3.4. Again, we use Lagrange multipliers. Since

$$\begin{aligned}
\frac{\partial \ell(q^P, q^M|G)}{\partial q^P_k} &= \sum_{m=1}^{M}\sum_{l=1}^{K} \frac{\alpha_{mak} q^M_l}{\sum_{k',l'} \alpha_{mk'l'} q^P_{k'} q^M_{k'}}, \\
\frac{\partial \ell(q^P, q^M|G)}{\partial q^M_k} &= \sum_{m=1}^{M}\sum_{l=1}^{K} \frac{\alpha_{mak} q^P_l}{\sum_{k',l'} \alpha_{mk'l'} q^P_{k'} q^M_{k'}},
\end{aligned}$$

we have to solve the system of equations

$$\begin{aligned}
\lambda &= \sum_{m=1}^{M}\sum_{l=1}^{K} \frac{\alpha_{mak} q^M_l}{\sum_{k',l'} \alpha_{mk'l'} q^P_{k'} q^M_{k'}}, \qquad & k = 1, ..., K, \\
\rho &= \sum_{m=1}^{M}\sum_{l=1}^{K} \frac{\alpha_{mak} q^P_l}{\sum_{k',l'} \alpha_{mk'l'} q^P_{k'} q^M_{k'}}, \qquad & k = 1, ..., K, \\
1 &= \sum_{k=1}^{K} q^P_k = \sum_{k=1}^{K} q^M_k.
\end{aligned} \tag{3.2.13}$$

Again, it is easy to eliminate λ and ρ, since with (3.2.13) we have

$$\begin{aligned}
\lambda &= \lambda \sum_{k=1}^{K} q^P_k = \sum_{m=1}^{M}\sum_{k,l}^{K} \frac{\alpha_{mkl} q^P_k q^M_l}{\sum^K_{k',l'} \alpha_{mk'l'} q^P_{k'} q^M_{l'}} = M, \\
\rho &= \rho \sum_{k=1}^{K} q^M_k = \sum_{m=1}^{M}\sum_{k,l}^{K} \frac{\alpha_{mkl} q^P_k q^M_l}{\sum^K_{k',l'} \alpha_{mk'l'} q^P_{k'} q^M_{l'}} = M.
\end{aligned}$$

Dividing by M, we are left with finding q^P and q^M such that

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M} \frac{\alpha_{mkl} q^P_l}{\sum^K_{k',l'} \alpha_{mk'l'} q^P_{k'} q^M_{l'}} &= 1, \qquad & k = 1, ..., K, \\
\frac{1}{M}\sum_{m=1}^{M} \frac{\alpha_{mkl} q^M_l}{\sum^K_{k',l'} \alpha_{mk'l'} q^P_{k'} q^M_{l'}} &= 1, \qquad & k = 1, ..., K.
\end{aligned}$$

□

3.3 Application to data

Now that we have introduced both the admixture and the recent-admixture models we want to apply these methods to real data. We used the 1000 Genomes data to produce admixed individuals and test the accuracy of both models.

3.3.1 Human data

We downloaded the 1000 Genomes data (phase 3) from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`, as well as information on the sampling locations from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/integrated_call_samples_v3.20130502.ALL.panel` 1000 Genomes Project Consortium et al. (2015). This data consists of

$$\begin{aligned}&661 \text{ Africans (AFR)},\\ &347 \text{ Admixed Americans (AMR)},\\ &504 \text{ East Asians (EAS)},\\ &503 \text{ Europeans (EUR)},\\ &489 \text{ South-East Asians (SAS)}.\end{aligned} \tag{3.3.1}$$

In the following we will be excluding all Admixed Americans (AMRs) since they are known to have an admixed background (Eduardoff et al. (2016); Pfaffelhuber et al. (2019)). The dataset comes with approximately 80 million SNPs. We will be using two subsets known as the EUROFORGENE AIMset (Phillips et al. (2014)) and the Kidd AIMset (Kidd et al. (2014)), respectively. The EUROFORGENE AIMset contains 128 SNPs that are able to distinguish between continental groups. Since our models are only designed for bi-allelic SNPs, we will ignore all the tri-allelic SNPs (rs17287498, rs2069945, rs2184030, rs433342, rs4540055, rs5030240), as well as rs12402499, which is not available in the 1000 Genomes dataset. The Kidd AIMset consists of 55 bi-allelic SNPs all available in the 1000 Genomes dataset.

3.3.2 Obtaining admixed individuals in silico

As mentioned above, we created admixed individuals from the 1000 Genomes dataset to be able to test our method. To do so, we first choose genomes $\tilde{G} = (\tilde{G}_m)_{m=1,\dots,M}$ from population k and $\bar{G} = (\bar{G}_m)_{m=1,\dots,M}$ from population k' as the parents. We then obtain the genome $G = (G_m)_{m=1,\dots,M}$ of a first generation admixed individual from populations k and k' by

$$G_m = X_m + Y_m$$

where

$$X_m = \begin{cases} 1 & \text{with probability } \tilde{G}_m/2, \\ 0 & \text{with probability } (2-\tilde{G}_m)/2, \end{cases} \qquad Y_m = \begin{cases} 1 & \text{with probability } \bar{G}_m/2, \\ 0 & \text{with probability } (2-\bar{G}_m)/2. \end{cases} \tag{3.3.2}$$

We note that by iterating this procedure we can also model second generation admixed individuals etc. in silico.

Remark 3.6 (Notation)**.** For a second generation admixed individual with maternal grandparents coming from population K_1 and K_2 and paternal grandparents coming from population K_3 and K_4 we denote the heritage of this second generation admixed individual as

$$K_1, K_2/K_3, K_4.$$

Remark 3.7 (Second generation heritage)**.** Using the population labels AFR, EAS, EUR, SAS as in (3.3.1) and the notation from Remark 3.6, all cases for second generation admixed individuals fall into one of the following seven categories:

(A) 4 non-admixed cases, e.g. AFR, AFR/ AFR, AFR;

(B) 6 admixed cases with admixture ratio 50:50, where both parents are non-admixed, e.g. AFR, AFR/ EAS, EAS;

(C) 6 admixed cases with admixture ratio 50:50, where both parents are admixed, e.g. AFR, EAS/ AFR, EAS;

(D) 12 admixed cases with admixture ratio 75:25, e.g. AFR, AFR/ AFR, EAS;

(E) 12 admixed cases with admixture ratio 50:25:25, where one parent is non-admixed, e.g. AFR, AFR/ EAS, EUR;

(F) 12 admixed with admixture ratio 50:25:25, where both parents are admixed, e.g. AFR, EAS/ AFR, EUR;

(G) 3 admixed with admixture ratio 25:25:25:25, e.g. AFR, EAS/ EUR, SAS.

In total we have 55 cases for which we each simulated 500 individuals in silico by picking four grand-parents at random from the populations, creating mother and father from the grand-parents, and creating a new individual from the parents, as described in (3.3.2).

3.4 Results

3.4.1 Comparing results from admixture and recent-admixture

With the help of the admixture and the recent-admixture model we are able to estimate IA of individual and obtain admixture proportions $q = (q_k)_{k=1,...,K}$ and $q^M = (q_k^M)_{k=1,...,K}$ and $q^P = (q_k^P)_{k=1,...,K}$, respectively. In the following we want to investigate how accurate the inferred admixture proportions actually are. To do so we define for $k = 1, ..., K$

$$q_k^{MP} := \tfrac{1}{2}(q_k^M + q_k^P)$$

which gives the fractions of the genome coming from population k in the recent-admixture model. With q_k^{MP} we have a quantity that can be compared to q_k from the admixture model. The vectors q and $q^{MP} = (q_k^{MP})_{k=1,...,K}$ are then compared to the *true* admixture proportions. Clearly, the true ancestry depends on which of the seven cases (A)-(G) we are in. In case (A), i.e. for a non-admixed individual, we have $q_k^{\text{True}} = 1$ for some k, and in case (B), i.e. an admixed individual with parents from populations k and k', we have $q_k^{\text{True}} = q_{k'}^{\text{True}} = 0.5$, and similarly for all other cases (C)-(G). The true admixture proportions of the remaining cases can be found in Remark 3.7. The distances to the true IA for the admixture model and the recent-admixture model are

$$\sum_k |q_k - q_k^{\text{True}}| \quad \text{and} \quad \sum_k |q_k^{MP} - q_k^{\text{True}}|, \tag{3.4.1}$$

respectively.

Remark 3.8. Above we defined $q_k^{MP} := \frac{1}{2}(q_k^M + q_k^P)$ in order to be able to compare the results obtained from the recent-admixture model to the ones obtained from the admixture model. We need to point out though, that in the recent-admixture model, we in fact receive information on q^M and q^P, the admixture proportions of the parents, separately, such that even more information than q^{MP} is contained in the estimates for this model. Looking at Remark 3.7, we can see that individuals from (B) and (C) all have admixture ratios 50:50 but clearly case (B) covers first generation admixture and case (C) corresponds to second generation admixture. In the best scenario, the admixture model delivers in both cases a vector q where two entries are close to 1/2. The recent-admixture model, however, gives us the vectors q^M and q^P where in case of

(B), both q^M and q^P have each one entry close to 1, and
(C), both q^M and q^P have each two entries close to 1/2.

So we can see that the recent-admixture model is in fact especially designed to detect recent admixture events in data.

Figure 3.1 give boxplots of the distances to the true IA for first-generation and second-generation admixtures using the 1000 genomes dataset.

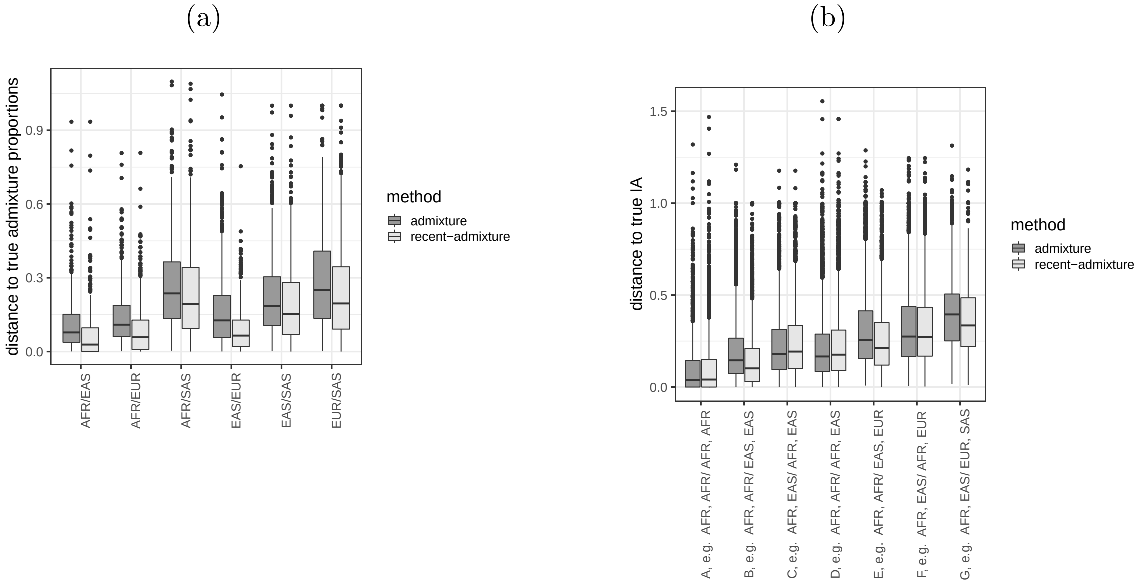


Figure 3.1: For all first generation admixed samples (a) and second generation admixed samples (b), we computed IA from the admixture and recent-admixture model. The distance to the true IA is computed as in (3.4.1). The cases in (B) are as described above.

As can be seen in Figure 3.1.(a), the recent-admixture estimates outperform admixture estimates in all cases for first generation admixed individuals. For second generation admixed individuals, Figure 3.1.(b), the situation is similar but depends on the type of admixture; see cases (B)–(G) above. (A full list of 55 cases is displayed in Figure 3.3 in Section 3.5.) In Figure 3.1.(b), note that column A gives non-admixed samples, and we see that estimates of IA are essentially as accurate in the admixture model and the recent-admixture model. Figure 3.5 depicts the same boxplots for the Kidd AIMset and contains similar results.

3.4.2 A Likelihood-ratio test for recent admixture

Suppose we have a new trace with data $G = (G_m)_{m=1,\dots,M}$. We want to test if G fits significantly better to the recent-admixture model than to the admixture model. Since the admixture model is identical to the recent-admixture model for $q^M = q^P = q$, we are testing the hypothesis

$$H_0 : q^M = q^P \quad \text{against} \quad H_1 : q^M \neq q^P.$$

For this, we take the estimators $\hat{q}$ of q from iteration of (3.2.7), and $\hat{q}^M, \hat{q}^P$ of q^M and q^P from iteration of (3.2.12) and compute the difference of the log-likelihoods

$$\Delta\ell := \ell(\hat{q}^M, \hat{q}^P|G) - \ell(\hat{q}|G)$$

with $\ell(q^M, q^P|G)$ from (3.2.10) and $\ell(q|G)$ from (3.2.4). The larger the $\Delta\ell$, the better the trace fits to the recent-admixture model. Therefore, we need to specify some x to obtain following decision rule

$$\begin{aligned} \text{if } \Delta\ell > x &\Rightarrow \text{ we reject } H_0 \\ \text{and if } \Delta\ell \leq x &\Rightarrow \text{we accept } H_0. \end{aligned}$$

In order to find x, we fix a p-value (1%, say), and set x equal to the p-quantile of the empirical distribution for data in the reference database, meaning, if $p = 1\%$ and the reference dataset contains 1000 samples, we compute all values for $\Delta\ell$ for all samples, and set x to be the 10-th smallest value we obtained.

3.4.3 Power of the Likelihood-ratio test for recent admixture

When fixing the maximal p-value for significance of the likelihood-ratio test for recent-admixture, we obtain the power of the test for all cases of recent admixture. Displaying the false positives (i.e. positively tested non-admixed) against true positives (i.e. positively tested admixed) in cases (B)–(G) for all possible values of p, we obtain the *Receiver-Operation-Characteristic* (ROC) curve. The optimal curve nearly hits 0 false positives with 100% true positives and has an AUC (*Area Under the Curve*) of 1. As we see in Figure 3.2, the power of the test differs with the kind of admixture. For first generation admixed (case (B)), one non-admixed parent (case (E)) and all grandparents from different continents (case (G)), the test is nearly perfect in distinguishing recent-admixture from admixture. If only half of the genome has two different ancestries (cases (D) and (F)), the power is reduced. If the individual is not recently-admixed in first generation, but both parents are (case (C)), power drops even more. In fact, the latter case is not recent-admixture as in our definition, since $q^M = q^P$ should technically hold. For the overall performance of the test, we give some examples in Table 3.1, in particular the power at $p = 1\%$ and AUC in all cases. Results for the Kidd AIMset are similar and given in the SI in Figure 3.6 and Table 3.3.

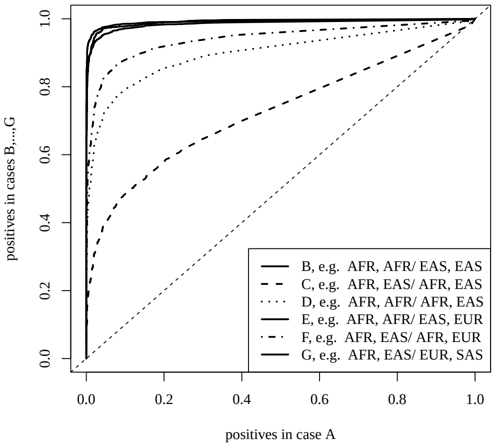


Figure 3.2: Using the EUROFORGEN AIMset, we plot false positives (i.e. positive non-admixed individuals, as in case (A) above, against true positives for all cases of admixture in second generation.

Case	B	C	D	E	F	G
AUC	0.99	0.637	0.872	0.983	0.928	0.99
Power at $p = 0.01$	0.94	0.23	0.51	0.9	0.62	0.9

Table 3.1: Using the same data as in Figure 3.2, we e.g. see that the test for recent admixture turns out to have a p-value below 1% in 94% cases of first generation admixed individuals.

3.4.4 Detecting recent admixture in the 1000 genomes dataset

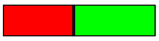
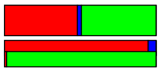
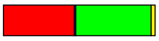
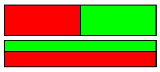
Running both the admixture and the recent-admixture model on the 1000 genomes dataset, we were able to detect samples with large $\Delta\ell$, i.e. individuals that show clear signs for being admixed individuals rather than ancestral.

There are six Africans from population ASW (Americans from Southwest USA), and two from South Asia, one from GIH (Gujarati Indian from Houston, Texas) and one from BEB (Bengali from Bangladesh). We note that it is known that the ASW population is admixed (Eduardoff et al., 2016), but until now, it has not been tested if admixture is recent.

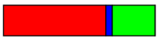
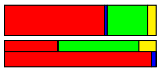
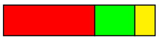
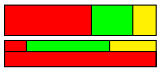
In Table 3.2, we list the following eight most extreme cases which show highly significant results for recent admixture for both AIMsets:

- NA20278: Giving the most significant results for both datasets, this male most probably has parents from African and European ancestry. Note also that q^{MP} and q are very similar for both AIMsets.
- NA20342, NA19625, NA20355: Clearly, one parent has African ancestry. The other parent is most likely partly European.
- NA20274: Our test indicates two parents of different ancestry, one mostly African, the other mostly East-Asian.
- NA20299: Interestingly, the results for both AIMsets differ in this example. One parent has most likely African ancestry, the other is European according to the EUROFORGEN AIMset and South-East Asian according to the Kidd AIMset.
- HG03803: Most likely, one parent of South-East Asian, the other has East-Esian ancestry.
- NA20864: Most likely, one parent of South-East Asian, the other has European ancestry.

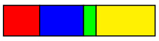
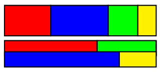
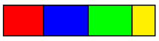
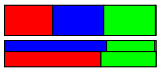
NA20278, AFR, ASW

EURO	ad	q	
$\Delta\ell = 6.69$	r-ad	q^{MP} q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 9.98$	r-ad	q^{MP} q^M, q^P	

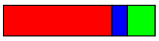
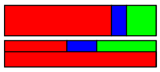
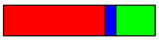
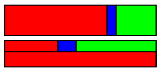
NA20342, AFR, ASW

EURO	ad	q	
$\Delta\ell = 1.82$	r-ad	q^{MP} q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 3.66$	r-ad	q^{MP} q^M, q^P	

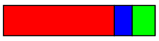
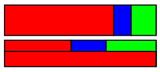
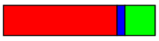
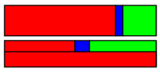
NA20274, AFR, ASW

EURO	ad	q	
$\Delta\ell = 6.56$	r-ad	q^{MP} q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 2.68$	r-ad	q^{MP} q^M, q^P	

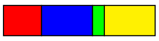
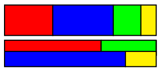
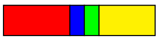
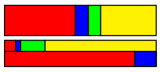
NA19625, AFR, ASW

EURO	ad	q	
$\Delta\ell = 1.52$	r-ad	q^{MP} q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 1.6$	r-ad	q^{MP} q^M, q^P	

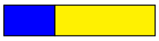
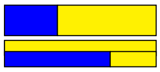
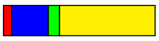
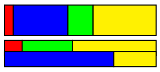
NA20355, AFR, ASW

EURO	ad	q	
$\Delta\ell = 1.11$	r-ad	q^{MP} q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 1.55$	r-ad	q^{MP} q^M, q^P	

NA20299, AFR, ASW

EURO	ad	q	
$\Delta\ell = 6.31$	r-ad	q^{MP} q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 1.52$	r-ad	q^{MP} q^M, q^P	

HG03803, SAS, BEB

EURO	ad	q	
$\Delta\ell = 1.01$	r-ad	q^{MP} q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 1.12$	r-ad	q^{MP} q^M, q^P	

NA20864, SAS, GIH

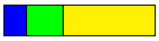
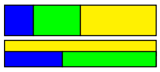
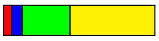
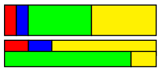
EURO	ad	q	
$\Delta\ell = 0.87$	r-ad	q^{MP} q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 0.95$	r-ad	q^{MP} q^M, q^P	

Table 3.2: The most extreme individuals in the 1000 genomes dataset in terms of a signal for recent admixture. For all individuals, we give IA from the admixture model (ad), given by q, the recent-admixture model (r-ad) $q^M, q^P, q^{MP} = \frac{1}{2}(q^M + q^P)$, for the analysis with the EUROFORGEN and Kidd AIMset. Difference in log-likelihoods for both models is given by $\Delta\ell$. Colors are AFR 🟥, EAS 🟦, EUR 🟩, SAS 🟨.

3.5 Additional results

We complement the results from Section 3.4 by giving the analogous results obtained by using the Kidd AIMset and listing all possible admixed individuals.

3.5.1 Estimation accuracy

For second generation admixed individuals, we have 55 cases, depending on the origin of the grandparents. With the population labels AFR, EAS, EUR, SAS we obtain the following admixtures:

(A) 4 non-admixed cases (with IA 100:0): (AFR, AFR/ AFR, AFR), (EAS, EAS/ EAS, EAS), (EUR, EUR/ EUR, EUR), (SAS, SAS/ SAS, SAS);

(B) 6 admixed cases with admixture ratio 50:50, where both parents are non-admixed: (AFR, AFR/ EAS, EAS), (AFR, AFR/ EUR, EUR), (AFR, AFR/ SAS, SAS), (EAS, EAS/ EUR, EUR), (EAS, EAS/ SAS, SAS), (EUR, EUR/ SAS, SAS);

(C) 6 admixed cases with admixture ratio 50:50, where both parents are admixed: (AFR, EAS/ AFR, EAS), (AFR, EUR/ AFR, EUR), (AFR, SAS/ AFR, SAS), (EAS, EUR/ EAS, EUR), (EAS, SAS/ EAS, SAS), (EUR, SAS/ EUR, SAS);

(D) 12 admixed cases with admixture ratio 75:25: (AFR, AFR/ AFR, EAS), (AFR, AFR/ AFR, EUR), (AFR, AFR/ AFR, SAS), (EAS, EAS/ EAS, AFR), (EAS, EAS/ EAS, EUR), (EAS, EAS/ EAS, SAS), (EUR, EUR/ EUR, AFR), (EUR, EUR/ EUR, EAS), (EUR, EUR/ EUR, SAS), (SAS, SAS/ SAS, AFR), (SAS, SAS/ SAS, EAS), (SAS, SAS/ SAS, EUR);

(E) 12 second generation admixed with admixture ratio 50:25:25, where one parent is non-admixed: (AFR, AFR/ EAS, EUR), (AFR, AFR/ EAS, SAS), (AFR, AFR/ EUR, SAS), (EAS, EAS/ AFR, EUR), (EAS, EAS/ AFR, SAS), (EAS, EAS/ EUR, SAS), (EUR, EUR/ AFR, EAS), (EUR, EUR/ AFR, SAS), (EUR, EUR/ EAS, SAS), (SAS, SAS/ AFR, EAS), (SAS, SAS/ AFR, EUR), (SAS, SAS/ EAS, EUR);

(F) 12 second generation admixed with admixture ratio 50:25:25, where both parents are admixed: (AFR, EAS/ AFR, EUR), (AFR, EAS/ AFR, SAS), (AFR, EUR/ AFR, SAS), (EAS, AFR/ EAS, EUR), (EAS, AFR/ EAS, SAS), (EAS, EUR/ EAS, SAS), (EUR, AFR/ EUR, EAS), (EUR, AFR/ EUR, SAS), (EUR, EAS/ EUR, SAS), (SAS, AFR/ SAS, EAS), (SAS, AFR/ SAS, EUR), (SAS, EAS/ SAS, EUR);

(G) 3 second generation admixed with admixture ratio 25:25:25:25: (AFR, EAS/ EUR, SAS), (AFR, EUR/ EAS, SAS), (AFR, SAS/ EAS, EUR);

Figure 3.3, Figure 3.4 and Figure 3.5 give the distance to the true IA for all of the above cases. See also Figure 3.1 from Section 3.4.

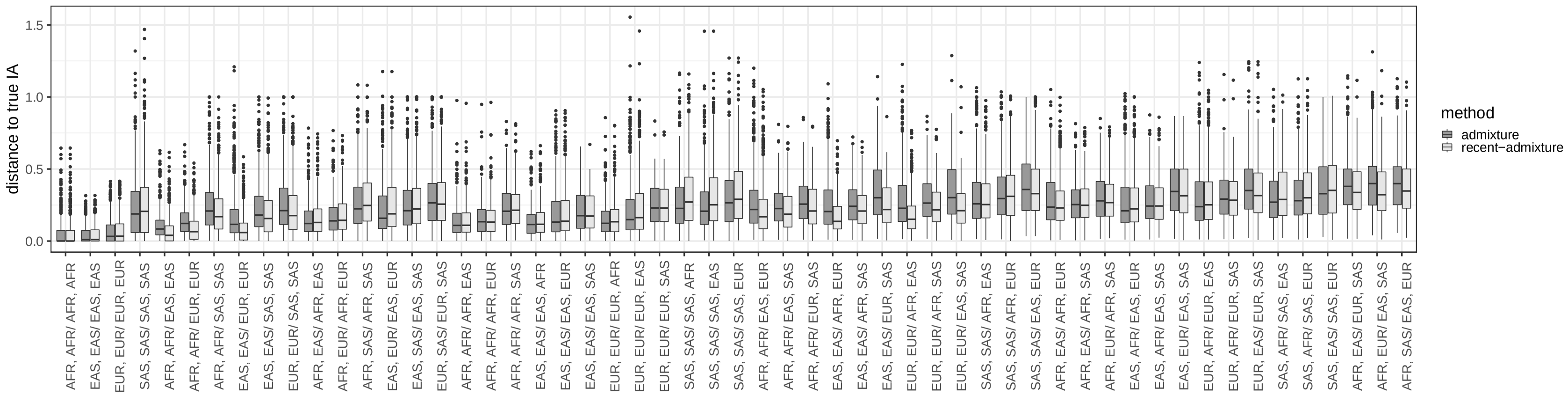


Figure 3.3: For all cases of second generation admixed individuals, we compare estimation accuracy of IA using the EUROFORGEN AIMset.

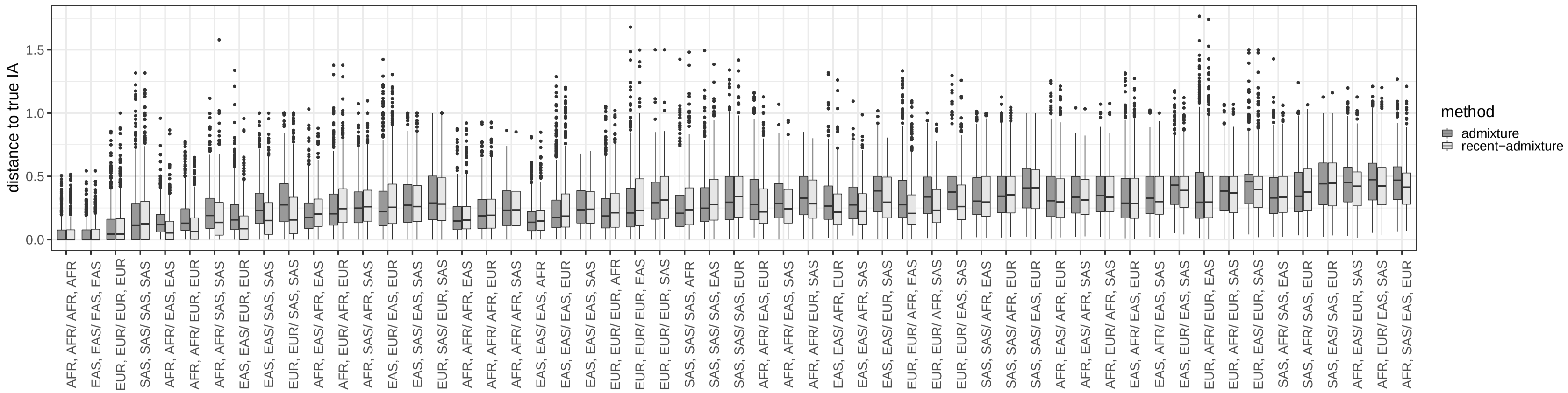


Figure 3.4: Same as in Figure 3.3, but using the Kidd AIMset.

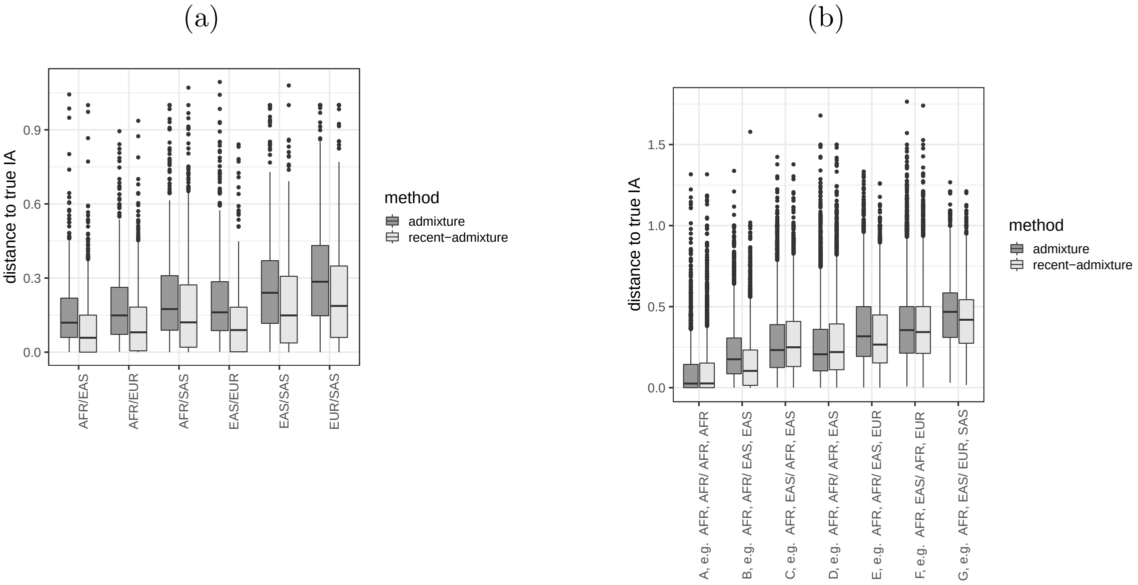


Figure 3.5: Same as in Figure 3.1, but using the Kidd AIMset.

3.5.2 Power of the Likelihood-ratio test using the Kidd AIMset

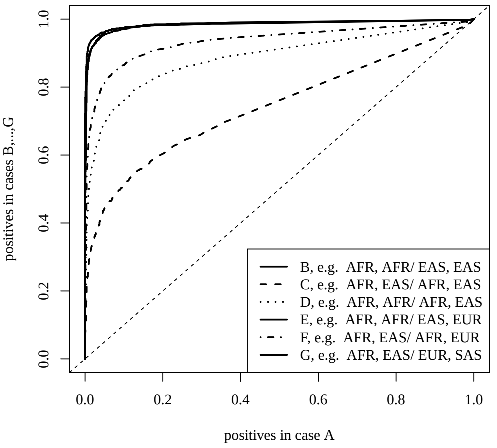


Figure 3.6: Same as in Figure 3.2, but using the Kidd AIMset.

Case	B	C	D	E	F	G
AUC	0.982	0.655	0.855	0.983	0.921	0.982
Power at $p = 0.01$	0.92	0.29	0.5	0.89	0.65	0.89

Table 3.3: Same as Table 3.1, but using the Kidd AIMset.

Acknowledgements

My first an utmost gratitude goes to my supervisor Peter Pfaffelhuber for giving me the opportunity to work on this thesis. I appreciate all the time he has invested in our conversations and for always taking the time to answer my questions. He was always open to new ideas and encouraged me to present my works at various occasions including the numerous interesting workshops and conferences organised by the DFG SPP Priority Programme 1590.

I enjoyed my time at the stochastics department thanks to my wonderful colleagues who have become my dear friends over the last few years. Everyone contributed to a pleasant working environment with lots of fun coffee breaks which I will definitely miss when I'm gone. I need to express my gratitude to the irreplaceable Monika Hattenbach who keeps this place together.

I would like to extend my deep gratitude to my whole family and especially my father for their continuous support and my sister also for proofreading my work. Last but not least, I thank Julian for his love and encouragement especially during times I needed them the most.

Bibliography

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664.

Baumdicker, F., Huss, E., and Pfaffelhuber, P. (2019). Modifiers of mutation rate in selectively fluctuating environments. *arXiv*. 1909.06241.

Denamur, E. and Matic, I. (2006). Evolution of mutation rates in bacteria. *Molecular Microbiology*, 60(4):820–827.

Depperschmidt, A., Greven, A., and Pfaffelhuber, P. (2012). Tree-valued fleming–viot dynamics with mutation and selection. *The Annals of Applied Probability*, 22(6):2560–2615.

Depperschmidt, A., Greven, A., and Pfaffelhuber, P. (2019). Duality and the well-posedness of a martingale problem. *arXiv*, 1904.1564.

Durrett, R. (2008). *Probability models for DNA sequence evolution.* Springer.

Eduardoff, M., Gross, T. E., Santos, C., de la Puente, M., Ballard, D., Strobl, C., Børsting, C., Morling, N., Fusco, L., Hussing, C., Egyed, B., Souto, L., Uacyisrael, J., Court, D.-S., Carracedo, A., Lareu, M. V., Schneider, P. M., Parson, W., Phillips, C., Consortium, E.-N., Parson, W., and Phillips, C. (2016). Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM$^{\text{TM}}$. *Forensic Science International. Genetics*, 23:178–189.

Ethier, S. and Kurtz, T. (1986). *Markov Processes. Characterization and Convergence.* John Wiley, New York.

Ethier, S. and Kurtz, T. (1993). Fleming-Viot processes in population genetics. *SIAM Journal on Control and Optimization*, 31:345–386.

Fisher, R. A. (1930). *The genetical theory of natural selection.* Oxford Clarendon Press.

Gillespie, J. H. (1981). Mutation Modification in a Random Environment. *Evolution*, 35(3):468–476.

Gillespie, J. H. (2004). *Population Genetics: A Concise Guide.* The Johns Hopkins University Press, 2nd edition.

Greven, A., Pfaffelhuber, P., and Winter, A. (2008). Tree-valued resampling dynamics: Martingale problems and applications. *Probability Theory and Related Fields*, 155.

Huss, E. and Pfaffelhuber, P. (2019). Genealogical distances under low levels of selection. *Theoretical Population Biology.*

Kallenberg, O. (2002). *Foundations of Modern Probability. 2nd ed.* Probability and Its Applications. New York, NY: Springer.

Kidd, K. K., Soundararajana, U., Rajeevana, H., Pakstisa, A. J., Moorec, K. N., and Ropero-Millerc, J. D. (2017). The redesigned forensic research/reference on genetics-knowledge base, frog-kb. *Forensic Science International. Genetics*, 33:33–37.

Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F. R., and Kidd, J. R. (2014). Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Science International. Genetics*, 10:23–32.

Kingman, J. (1982). The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248.

Krone, S. and Neuhauser, C. (1997). Ancestral processes with selection. *Theo. Pop. Biol.*, 51:210–237.

Liu, Y., Nyunoya, T., Leng, S., A Belinsky, S., Tesfaigzi, Y., and Bruse, S. (2013). Softwares and methods for estimating genetic ancestry in human populations. *Human genomics*, 7:1.

Mao, E. F., Lane, L., Lee, J., and Miller, J. (1997). Proliferation of mutators in a cell population. *Journal of bacteriology*, 179:417–22.

Neuhauser, C. and Krone, S. (1997). The genealogy of samples in models with selection. *Genetics*, 154:519–534.

Pfaffelhuber, P., Grundner-Culemann, F., Lipphardt, V., and Baumdicker, F. (2019). How to choose sets of ancestry informative markers: A supervised feature selection approach. *Forensic Science International. Genetics*, page minor revision.

Phillips, C., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., Eduardoff, M., Børsting, C., Johansen, P., Fondevila, M., Morling, N., Schneider, P., EUROFORGEN-NoE Consortium, Carracedo, A., and Lareu, M. V. (2014). Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. *Forensic Science International. Genetics*, 11:13–25.

Phillips, C., Santos, C., Fondevila, M., Ángel Carracedo, and Lareu, M. V. (2016). Inference of Ancestry in Forensic Analysis I: Autosomal Ancestry-Informative Marker Sets. In *Forensic DNA Typing Protocols*, volume 1420 of *Methods in Molecular Biology*, pages 233–253. Springer, New York.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–954.

Saunders, I. W., Tavaré, S., and Watterson, G. A. (1984). On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability*, 16(3):471–491.

Sturtevant, A. H. (1937). Essays on evolution. i. on the effects of selection on mutation rate. *The Quarterly Review of Biology*, 12(4):464–467.

Tang, H., Peng, J., Wang, P., and Risch, N. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genet Epidemiol.*, 28:289–301.

Tenaillon, O., Taddei, F., Radman, M., and Matic, I. (2001). Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Research in Microbiology*, 152(1):11–16.

Wahlund, S. (1928). Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas*, 11:65–106.

Wakeley, J. (2008). *Coalescent Theory: An Introduction.* Roberts & Company.

Wakeley, J. (2010). Natural selection and coalescent theory. In *Evolution since Darwin: The First 150 Years*, pages 119–149. Sunderland, MA: Sinauer and Associates.

Wielgoss, S., Barrick, J. E., Tenaillon, O., Wiser, M. J., Dittmar, W. J., Cruveiller, S., Chane-Woon-Ming, B., Médigue, C., Lenski, R. E., and Schneider, D. (2013). Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proceedings of the National Academy of Sciences of the United States of America*, 110(1):222–227.