Stochastic Models in Population Genetics: The Impact of Selection and Recombination

Dissertation zur Erlangung
des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsprogramm
"PhD School of Mathematical Sciences (SMS)"
der Georg-August University School of Science (GAUSS)

vorgelegt von
Rebekka Brink-Spalink
aus Buxtehude

Göttingen, 2014

Betreuungsausschuss:

Prof. Dr. Anja Sturm,
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Henner Simianer,
Abteilung Tierzucht und Haustiergenetik, Universität Göttingen

Prof. Dr. Tatyana Krivobokova,
Institut für Mathematische Stochastik, Universität Göttingen

Mitglieder der Prüfungskommission:

Referentin:
Prof. Dr. Anja Sturm,
Institut für Mathematische Stochastik, Universität Göttingen

Korreferent:
Prof. Dr. Dominic Schuhmacher
Institut für Mathematische Stochastik, Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Dorothea Bahns,
Mathematisches Institut, Universität Göttingen

PD Dr. Ulf-Rainer Fiebig,
Institut für Mathematische Stochastik, Universität Göttingen

Jun.-Prof. Dr. Felix Krahmer,
Institut für Numerische und Angewandte Mathematik, Universität Göttingen

Jun.-Prof. Dr. Andrea Krajina,
Institut für Mathematische Stochastik, Universität Göttingen

Tag der mündlichen Prüfung: 23. Januar 2015

Acknowledgements

Ich möchte mich gern an dieser Stelle bei einigen der vielen Leute bedanken, die mich auf dem Weg zu dieser Arbeit begleitet haben.
Zuallererst herzlichen Dank an meine Betreuerin Prof. Dr. Anja Sturm für die hervorragende Unterstützung und Förderung sowie für das Ermöglichen vieler Konferenzbesuche und der Forschungsaufenthalte in Frankreich. Ohne ihre motivierende Art, die stete Diskussionsbereitschaft und Inspiration und die nötige Motivation wäre diese Arbeit wohl so nicht zustande gekommen.
Herzlicher Dank auch an Prof. Dr. Dominic Schuhmacher für die bereitwillige Übernahme des Korreferats.
Ich möchte außerdem der DFG für die finanzielle Förderung sowie allen Mitgliedern des Graduiertenkollegs 1644 danken, die zu meiner Promotion frischen Wind nach Göttingen gebracht haben und die mir viele interessante, entfernte und doch verwandte Forschungsgebiete näher gebracht haben.
Vielen Dank an dieser Stelle auch an Charline Smadi, c'était un grand plaisir de travailler avec toi!
I also wish to thank all members of the IMS of the past four years who added to the excellent working environment and the cheerful atmosphere. I will treasure not only the creative climate but also various birthday cakes and "kicker"-battles. Besonders hervorheben möchte ich meine Langzeit-Bürokollegen Till und Hannes, ohne die die Arbeits- und Pausenzeit nur halb so schön gewesen wäre! Außerdem geht spezieller Dank an Thomas für viele ideengebende Diskussionen und das Korrekturlesen von großen Teilen meiner Aufzeichnungen.
Da der Abschluss meiner Promotion auch ein Abschied aus Göttingen bedeutet, möchte ich an dieser Stelle allen meinen Freunden danken, die mich in den letzten Jahren im Alltag und Urlaub, beim Arbeiten und beim Kartenspielen, beim Feiern und Wohnen begleitet haben. Es war eine unfassbar schöne Zeit!
Zuletzt möchte ich mich ganz besonders bei meinem Freund Christoph, meinen Eltern, meinen Schwestern, Schwagern, Neffen und Nichten für ihre Unterstützung und Liebe auf meinem bisherigen Weg bedanken. Es ist doch alles einfacher, wenn man weiß, dass man einen solchen Rückhalt hat!

DANKE!

Contents

Summary

The focus of this thesis is on the extension of mathematical models in population genetics in order to capture the effects of highly skewed offspring distributions or of strong selection on the genetic variation within a population. The population models which are considered in this work in particular include the biological phenomenon called recombination. Roughly speaking, recombination corresponds to a split and subsequent reassembly of the genome during reproduction due to which a new configuration of genes appears in the offspring. The specific interest was to study neutral gene genealogies where the considered genes are in the vicinity of a genetic locus experiencing high selective pressure. The main results of this thesis therefore add to the understanding of the dependence structure of many partially linked loci on the same chromosome when one of them carries a highly advantageous gene.

In the main part of this thesis, Chapter 2, we consider the partition of a sample taken from a population at the end of a so-called selective sweep. More precisely, we approximate the distribution of the ancestral relationships of neutral genetic loci which are partially linked to one locus under selection. The evolution of the population is here described by a Moran model with selection and recombination. After explaining our model in detail, we proceed with presenting the main result, an approximate distribution for the partition of a sample after such a sweep. The order of the error term in the derived approximation is the reciprocal of the logarithm of the population size. This result is an extension of the findings of Schweinsberg and Durrett [36] towards the multi-locus case and gives insight into the precise dependencies of two neutral loci on the same chromosome as the locus under selection. In a related work, Pfaffelhuber and Studeny [31] studied the same three-locus geometry, however, their analysis was based on a different model which requires to first take the limit for infinitely large population sizes meaning that the resulting approximation only allows for a vague interpretation of its actual accuracy. Further, in contrast to them, we here present the sampling formula which can be used in order to construct a typical sample.

In the subsequent Chapter 3 we present joint work with Charline Smadi from CMAP, École Polytechnique, Paris-Saclay. The research question was basically the same as in Chapter 2, namely to approximate the distribution of the neutral genealogies of a sample taken from a population at the end of a selective sweep. Here however, we modeled the evolution of the population by a birth and death process with varying population size and transition rates depending on the genetic types of the individuals and the current state of the population. This model mirrors the influences on a population more realistically and allows to incorporate several biological parameters, such as the influence by competition. This work relies strongly

on the results presented in Chapter 2 and further benefits from the findings of Charline Smadi in [38].

Chapter 4 focuses on two aspects of so-called Λ-coalescent processes which arise as the limit genealogies for large populations with more variable offspring distributions. By means of a genetic data set from a population of Pacific oysters, Eldon and Wakeley showed in [16] that the classical coalescent approach does not capture the reproduction behavior of these marine organism. Their work revealed that there are reproduction events where a significant proportion of a new generation of the population consists of offspring of one single individual and hence a description by a binary coalescent is not suitable. The motivation for the study of Λ-coalescent processes in this present work results from questions in the field of animal breeding as such skewed offspring distributions can also arise when only a few breeding bulls account for most offspring in the total cattle population. We therefore seek to better understand processes with a large variance in the number of offspring per individual and which at the same time incorporate more complex biological processes such as recombination. It is essential to fully understand the dynamics of such processes in order to develop statistical methods which help to determine which processes and distributions best describe the evolution of the population given the genetic data.

In this last chapter, we first give a formula of recursive nature for the expected height of a coalescent tree for a general measure Λ. For particular cases such as the Kingman or the star-shaped coalescent, this quantity is derived easily. For a general measure however, there is no closed formula so that investigations in this area have so far focused on the study of dual processes where asymptotics for the height and length of the process are known. In addition to the formal statement, we provide a MuPAD algorithm which calculates the expectation for measures $\Lambda = \delta_a$ for some $a \in (0, 1]$. Second, we describe a spatial algorithm on how to generate an ancestral recombination graph for these more general coalescent processes. This is an extension of the spatial algorithm of Wiuf and Hein [41] who only considered populations whose evolution can be well described by the Kingman coalescent.

In the following Chapter 1 we will give a broad introduction into stochastic population models which helps to understand the concepts that are developed in the later chapters. We will focus mostly on models with constant population size and only in the end describe a birth and death process modeling a population of varying size. We start with defining a general model for generating new generations which was first introduced by Cannings in [12]. Here, we also state specific examples for different offspring distributions leading to well-known models such as the Wright-Fisher or the Moran model. Further, we investigate the limit behavior for large population sizes which leads to the introduction of coalescent processes. A core part of the introductory chapter will be the description of how biological features such as selection and recombination can be included into the population models since the main results of this thesis are based on the study of exactly these more complex models.

CHAPTER 1

An introduction into models in population genetics

Before we start with the introduction of some mathematical models which are used to describe the evolution of populations through time we will briefly list some definitions of biological terms and processes which are needed along the way.
A population, that is, an aggregation of individuals, is said to be *panmictic*, if all individuals have the same probability of mating with each other (so-called random mating). This means, that neither genetic properties nor physical appearance, nor any sort of environmental or social aspects influence the mating and reproduction ability of an individual.
The *genotype*, that is, the genetic information of an individual is stored in the DNA which consists of coding parts, called *genes*, and non-coding parts. A gene can have different variants, the *alleles*, and the different alleles of one gene can lead to different *phenotypic* traits, that is, a different appearance such as eye color. We call a part in the DNA which encodes a gene a *locus*, with plural *loci*.
In organisms whose cells contain a nucleus, so-called eukaryotes, the DNA is organized in *chromosomes*. Each chromosome consists of two identical *chromatids* which are connected with each other through the *centromere* (see Figure 1.0.1). The number of chromosomes varies amongst the different species.

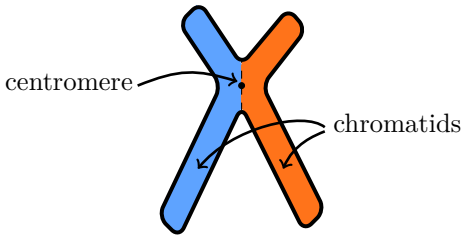


Figure 1.0.1: Schematic drawing of a chromosome.

An organism is called *haploid*, if its cells have a single set of chromosomes, whereas it is *diploid*, if the cells have a double set of chromosome. In this case, generally a *maternal* and a *paternal* chromosome build a pair of so-called *homologous chromosomes*. These have the same genes at the same loci but not necessarily the same alleles as they are inherited from

different individuals.
In case of sexual reproduction, special cells, so-called *gametes*, are formed which fuse with the gametes of the partner during fertilization. In diploid individuals the gametes are formed during a cell-division process called *meiosis*, where the double set of chromosomes is reduced to a haploid one.

1.1 Cannings Model

Throughout this section we will study panmictic haploid populations with fixed population size N over time. We will first define a stochastic population model with well-defined non-overlapping generations in a general form.

Definition 1.1 (Cannings Model)**.** *Label the individuals from 1 to N and define for $i \in [N] := \{1, 2, \dots, N\}$ and $t \in \mathbb{N}_0$ the random variables*

$$\nu_i^t := \#\ \textit{offspring produced by individual } i \textit{ from generation } t.$$

A population model with an offspring mechanism $\nu^t := (\nu_1^t, \dots, \nu_N^t)$ fulfilling the following properties is called a Cannings Model:

1. *(Equilibrium, Markov property) For different t the variables $\nu_1^t, \dots, \nu_N^t$ are independent.*
2. *(Exchangeability) For each t the vector $(\nu_1^t, \dots, \nu_N^t)$ is exchangeable, that is,*

$$(\nu_1^t, \nu_2^t, \dots, \nu_N^t) \sim (\nu_{\pi(1)}^t, \nu_{\pi(2)}^t, \dots, \nu_{\pi(N)}^t)$$

   for any permutation π on $[N]$.

3. *(Constant population size) For each t we have*

$$\sum_{i=1}^{N} \nu_i^t = N.$$

Many of the classical population models can be defined within the above framework.

Example 1.2 (Wright-Fisher Model)**.** In a Wright-Fisher model, a new generation is created by sampling from the previous generation, independently and with replacement. The number of times an individual is drawn corresponds to its number of offspring. In the formulation of a Cannings model, we have $\nu = (\nu_1, \dots, \nu_N) \sim \text{Mult}\left(N, \frac{1}{N}, \dots, \frac{1}{N}\right)$. An alternative interpretation is obtained by considering the process backwards in time: each offspring chooses its parent uniformly at random out of the previous generation, independently of the others. See Figure 1.1.1 for a graphical representation of this model.

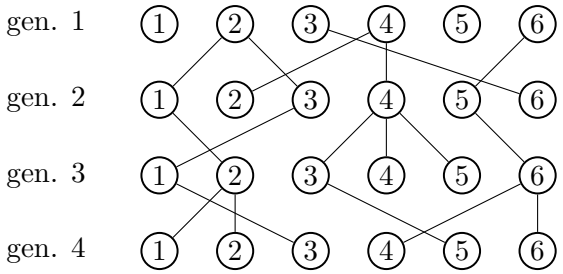


Figure 1.1.1: Four generations under a Wright-Fisher Model.

♢

Example 1.3 (Moran Model - discrete generation version)**.** In one step of the Moran model, a pair of individuals is chosen uniformly at random, the first one reproduces and has one offspring, the second one dies, all other individuals just persist. A generation in the sense of the Cannings Model thus corresponds to the reproduction of only one individual. We therefore speak of steps in the Moran model rather than generations. If we define

$$P_{210} := \{P_N \cdot (2, 1, \ldots, 1, 0)^T \mid P_N \text{ a } N \times N \text{ permutation matrix}\}.$$

then the offspring distribution $\nu = (\nu_1, \ldots, \nu_N) \sim \mathcal{U}(P_{210})$ describes the evolution of the population under a Moran Model.

A version of the Moran Model is obtained when drawing twice from the population *with*

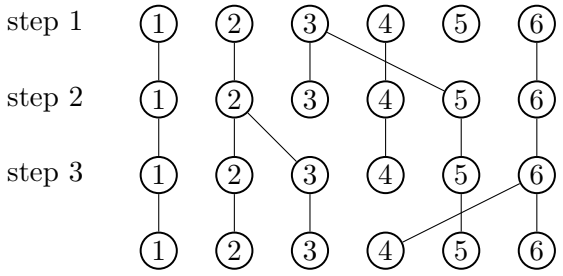


Figure 1.1.2: Three steps of a Moran Model.

replacement. Then however, the offspring distribution is different as the vector $(1, 1, \ldots, 1)$ has positive probability.

The classical Moran model is defined by the same reproduction mechanism but with exponentially distributed random times between two events. More precisely, each individual i has a clock which rings with rate 1, that is, after a time $t_i \sim \text{Exp}(1)$, independently of all other clocks. When this clocks rings, individual i is replaced by an offspring of a uniformly chosen individual from the population (including i itself). This implies that in a population of size

N, an event happens after a time $t \sim \text{Exp}(N)$. As this rate does not change the reproduction mechanism and is constant over time, we can impose the specific event times afterwards and stick with the definition of the discrete time Moran model. ♢

Example 1.4 (Modified Moran model with skewed offspring distribution (cf. [16]))**.** We can generalize the above defined Moran Model towards a skewed offspring distribution with higher variance. This can be realized by choosing a random number of individuals which are replaced by the offspring of only one individual: we here first choose randomly the number $U-1$ of killed individuals and then uniformly at random the labels of the corresponding individuals. The variable U can have any distribution on the set $\{0, 1, 2, \ldots, N\}$. As an example, we restate the distribution from [16]:

$$\mathbb{P}_U(u) = \begin{cases} 1 - N^{-\gamma} & \text{if } u = 2 \\ N^{-\gamma} & \text{if } u = N\psi \\ 0 & \text{otherwise,} \end{cases}$$

with constants $\gamma \geqslant 0$ and $\psi \in (0, 1)$ which control the frequency and the extent of larger merging events, respectively. Conditional on $U = u$, the offspring vector then is in the set

$$P_{u10} := \{P_N \cdot (u, \underbrace{1, \ldots, 1}_{N-u}, \underbrace{0, \ldots, 0}_{u-1})^T \mid P_N \text{ a } N \times N \text{ permutation matrix}\}.$$

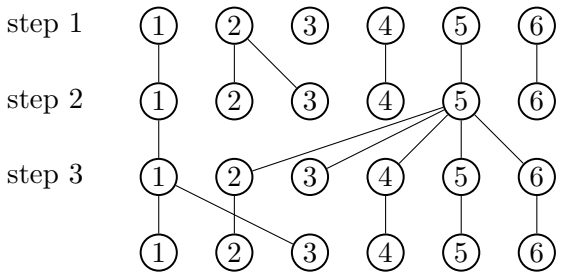


Figure 1.1.3: Three steps under the modified Moran Model.

♢

In these population models with exchangeable offspring distribution, it is easy to impose different, neutral genotypes on the individuals of the first generation. We can then study the evolution of the number of a given type over time, assuming that an offspring always copies the type of its parent. By *neutral* we mean, that no type has any advantage over any other type (otherwise, the exchangeability would not hold anymore). By definition of the Cannings model we can see that the process $X = (X_t)_{t \in \mathbb{N}}$ which indicates the proportion of individuals with the distinguished type at each time step t is a Markov chain with transition rates

determined by the offspring distribution. If we continued in this direction, an appropriate time scaling of the process allows us to set up the generator of the corresponding continuous-time Markov process. This combined with taking the limit of large population sizes shows that the frequency of alleles of a given type can be described by a diffusion process given through the following stochastic differential equation:

$$dX_t = \sqrt{X_t(1 - X_t)}dB_t, \tag{1.1}$$

with B_t a standard Brownian motion. As the focus of this thesis lies on different aspects, namely genealogies rather than frequencies, we do not go into details regarding the derivation of the above equation or results which have been obtained by this approach. We will only later, in Chapter 2 compare our findings with the results from [31] obtained by the study of this so-called diffusion approximation.
At this point, we pursue another approach motivated by the following observation: when looking at the above realizations of the population models, we notice that not all individuals from the starting generation will have passed on their genetic material to an individual from the last/present generation. Consider as an example the ancestral lineages of generation 4 in the Wright-Fisher Model from Figure 1.1.1:

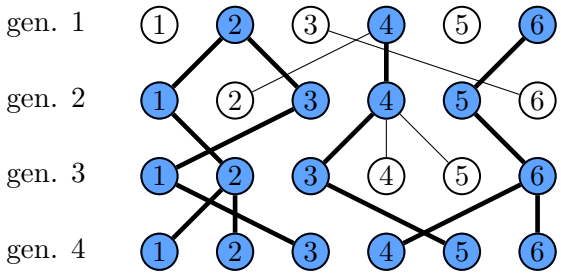


Figure 1.1.4: Ancestral lineages in a Wright-Fisher Model.

Already those four generations indicate that by simulating the evolution of a population forwards in time, we create lots of genetic material which is not relevant for any individual living at the present time. Therefore, depending on the aspects that we want to study, it may be more efficient to consider a population backwards in time, and generate its genealogy starting with the individuals from the present generation. Whenever two or more individuals are offspring from the same parent in the forward-in-time population model, they find a common ancestor in the backward-in-time model and their ancestral lineages merge into one. This is one of the reasons for the analysis of so-called coalescent processes in the field of population genetics.

1.2 The limit of large populations - coalescent processes

The great strength of the above formulation of the Cannings Model is the related theory concerning the limit behavior for large populations. We will shortly introduce coalescent processes as a self-contained concept and then draw the connection to population models in general and in particular to the discrete models introduced above. For further details on exchangeable random partitions and general coagulation processes see [33, 5].

Definition 1.5. *Denote with $\mathcal{P}_{\mathbb{N}}$ the space of all partitions of $\mathbb{N}$, and analogously $\mathcal{P}_{[n]} =: \mathcal{P}_n$ the space of all partitions of the set $[n]$. The number of blocks of a partition $\xi \in \mathcal{P}_{\mathcal{N}}$, for $\mathcal{N} \in \{\mathbb{N}, [n]\}$, is denoted by $|\xi|$. We further define the following relations for $\xi, \eta \in \mathcal{P}_n$:*

$$\begin{aligned} \xi \subset \eta \quad &\text{if} \quad \eta \text{ can be created from } \xi \text{ by merging blocks of } \xi, \\ \xi \overset{k}{\prec} \eta \quad &\text{if} \quad \xi \subset \eta \text{ and } \eta \text{ is created from } \xi \text{ by merging exactly} \\ &\qquad k \text{ blocks of } \xi \text{ into 1 block, } |\xi| = |\eta| + (k-1), \end{aligned} \tag{1.2}$$

where we will write $\prec$ instead of $\overset{2}{\prec}$.

Definition 1.6. *We call a continuous-time process $\Pi^{\mathcal{N}} := (\Pi_t^{\mathcal{N}})_{t \geqslant 0}$, with values $\Pi_t^{\mathcal{N}}$ in $\mathcal{P}_{\mathcal{N}}$, $\mathcal{N} \in \{\mathbb{N}, [n]\}$ a coalescent process if*

1. *$\Pi^{\mathcal{N}}$ is a Markov process,*
2. *the transition rates $q_{\xi\eta}$ are positive if and only if $\xi \subset \eta$,*
3. *$\Pi^{\mathcal{N}}$ is exchangeable: $\Pi^{[n]}$ is called exchangeable if any partition of $[n]$ into $k \leqslant n$ blocks of sizes $b_1, \ldots, b_k$ has the same probability, independent of the contents of the blocks but only dependent on the number of blocks and their sizes. $\Pi^{\mathbb{N}}$ is called exchangeable if the restriction to $[n]$ given by $\Pi^{[n]}$ is exchangeable for all $n \in \mathbb{N}$.*

We further call $\Pi^{\mathcal{N}}$ consistent if for any k with $[k] \subset \mathcal{N}$ its projection onto the smaller set $[k]$ is again a coalescent process where the probabilities of a certain partition can be derived as marginal probabilities from the original process.

The name *coalescent process* is motivated by the dynamics of such a process: we only observe the merging or coalescence of blocks. An event where such a fusion of blocks occurs is called a coalescent event.

Remark 1.7. *As we can identify every partition in $\mathcal{P}_n$ with an equivalence relation on $[n]$ and vice versa, we also denote the set of all equivalence relations on $[n]$ by $\mathcal{P}_n$.*

Obviously, once the coalescent process reaches the state where there is only one block left, it never leaves it again. In terms of genealogies, this motivates the following definition:

Definition 1.8. *The first time when $\Pi^{[n]}$ is equal to the trivial partition with one block,*

$$T_{mrca}^{\Pi^{[n]}} := \inf\{t \geqslant 0 : \ |\Pi_t^{[n]}| = 1\} \tag{1.3}$$

is called the time to the most recent common ancestor. It is the first time that all sampled lineages can be traced back to the same ancestor and thus, it is the height of the coalescent tree (see also Definition 1.13).

We can analogously define this height for coalescent processes on $\mathcal{P}_{\mathbb{N}}$ but then $T_{mrca}^{\Pi^{\mathbb{N}}}$ is not necessarily finite, depending on the rates $q_{\xi\eta}$ of that process (see [35] for a necessary and sufficient condition for this property, which is called *coming down from infinity*). In this work, we only consider genealogies of finite samples out of a population.
We will now define two specific coalescent processes which are widely known in the setting of population genetics. The first one is the well-studied Kingman coalescent where coalescent events always involve exactly two blocks, the second one the more recently studied and more flexible Λ-coalescent process, introduced by Pitman in [32] and Sagitov in [34].

Definition 1.9 (Kingman coalescent)**.** *The Kingman coalescent process $\Pi^K := (\Pi_t^K)_{t\geqslant 0}$ is a consistent coalescent process with values in $\mathcal{P}_{\mathbb{N}}$ and initial partition $\Pi_0^K = \Delta_{\mathbb{N}} := \{\{1\},\{2\},\{3\},\dots\}$, the partition of $\mathbb{N}$ into singletons. Furthermore, for $\xi \neq \eta$, the transition rates fulfill*

$$q_{\xi\eta} = \begin{cases} 1 & \text{if } \ \xi \prec \eta \\ 0 & \text{otherwise.} \end{cases} \tag{1.4}$$

In other words, there are only binary mergers and no simultaneous coalescent events. In particular, the rates do not depend on the sizes of the blocks in ξ.
The Kingman n-coalescent process $\Pi^{K,n} := (\Pi_t^{K,n})_{t\geqslant 0}$ for $n \in \mathbb{N}$ is the projection of Π^K on $[n]$.

The concept of partitions may be rather abstract, alternatively, we can keep the following image in mind when it comes to coalescent processes on $\mathcal{P}_n$:

Remark 1.10. *Start drawing a mathematical tree with n leaves which represents the starting configuration Δ_n. The rates of the Kingman n-coalescent process stated in* (1.4) *indicate that any specific pair of lines starting from the n leaves merges with rate one. Therefore, with $k \leqslant n$ lines present, the time until the next merger happens is distributed exponentially with parameter $\binom{k}{2}$. As the process is exchangeable, each pair has the same probability of being chosen at the time of an event, and due to the Markov property of the process, the time until the next merging event is independent of all previous times. As the process stops when there is only one line (or block) left, we obtain a graphical representation as shown in Figure* (1.2.1a).

Definition 1.11 (Λ-coalescent)**.** *Let Λ be a finite measure on $[0,1]$. The Λ-coalescent process $\Pi^\Lambda := (\Pi_t^\Lambda)_{t\geqslant 0}$ is a consistent coalescent process with values in $\mathcal{P}_\mathbb{N}$ and initial partition $\Pi_0^\Lambda = \Delta_\mathbb{N}$, the partition of $\mathbb{N}$ into singletons. Furthermore, for $\xi \neq \eta$, the transition rates fulfill*

$$q_{\xi\eta} = \begin{cases} \int_{[0,1]} x^{k-2}(1-x)^{|\xi|-k}\Lambda(dx) & \text{if } \xi \overset{k}{\prec} \eta \\ 0 & \text{otherwise.} \end{cases} \tag{1.5}$$

Put differently, there are no simultaneous coalescent events and if the process is in state ξ with $|\xi| = b$, any k-tuple of blocks of ξ merges together to one block with rate $\lambda_{b,k} := \int_{[0,1]} x^{k-2}(1-x)^{b-k}\Lambda(dx)$. The consistency of the process holds due to the following property of those rates:

$$\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1}.$$

The Λ-n-coalescent process $\Pi^{\Lambda,n} := (\Pi_t^{\Lambda,n})_{t\geqslant 0}$ for $n \in \mathbb{N}$ is the projection of Π^Λ on $[n]$.

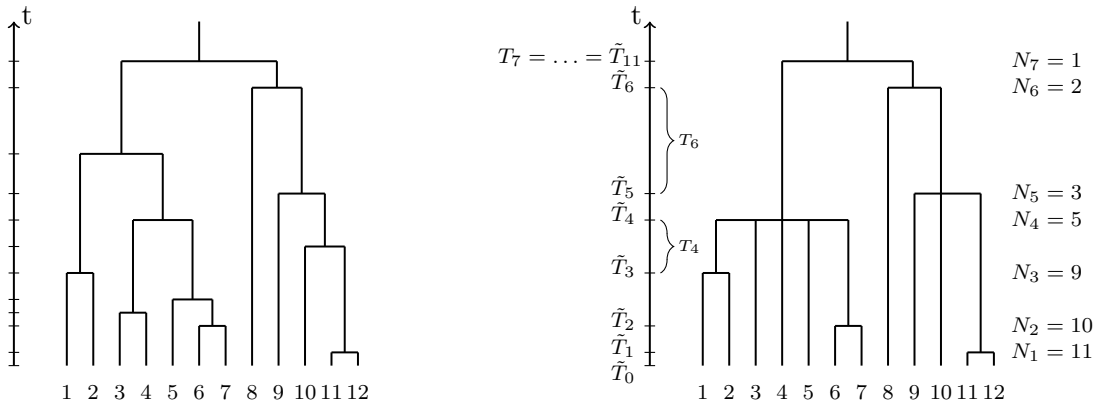
We can think of the Λ-n-coalescent in the same way as described in Remark 1.10 for the Kingman n-coalescent. The only differences are that in one merging event more than two lines may coalesce, and that the times between events are distributed with a different rate, according to the measure Λ. We only list some examples at this point and will discuss the properties of these coalescent processes in detail in Section 4.1 where we in particular give an intuition on the structure of the rates from (1.5).

Example 1.12 (Examples of Λ-coalescent processes)**.** 1. The Kingman coalescent is identical to a Λ-coalescent with $\Lambda = \delta_0$, the Dirac measure at zero.
2. If $\Lambda = \delta_1$, we call the process a star-shaped coalescent as in case of a merging event, all blocks coalesce into one.
3. The coalescent with $\Lambda(dx) = dx$, the uniform measure on the interval $[0,1]$ is called the Bolthausen-Sznitman coalescent, introduced in [10].

If we consider a coalescent process in reversed time, that is, starting with only the root, it is nothing else than a pure birth process in continuous time.

Definition 1.13. *For a coalescent process $\Pi^{[n]}$ define the times of coalescent events recursively for $1 \leqslant k \leqslant n-1$ by*

$$\begin{aligned} \tilde{T}_0^{\Pi^{[n]}} &= 0, \\ \tilde{T}_k^{\Pi^{[n]}} &= \inf\left\{t \geqslant 0 : \ |\Pi_t^{[n]}| < |\Pi_{\tilde{T}_{k-1}}^{[n]}|\right\} \quad \text{if } \ |\Pi_{\tilde{T}_{k-1}}^{[n]}| > 1, \quad \text{and} \\ \tilde{T}_k^{\Pi^{[n]}} &= \tilde{T}_{k-1}^{\Pi^{[n]}} \quad \text{if } \ |\Pi_{\tilde{T}_{k-1}}^{[n]}| = 1. \end{aligned} \tag{1.6}$$



(a) A Kingman coalescent tree. **(b)** A Λ-coalescent tree.

Figure 1.2.1: Coalescent trees.

Further, denote the number of blocks $\Pi^{[n]}$ *has after the* k*-th jumps by*

$$N_k^{\Pi^{[n]}} = |\Pi^{[n]}_{\tilde{T}_k}|, \tag{1.7}$$

and define the duration of time $\Pi_t^{[n]}$ *spends in one state by*

$$T_k^{\Pi^{[n]}} = \tilde{T}_k^{\Pi^{[n]}} - \tilde{T}_{k-1}^{\Pi^{[n]}}. \tag{1.8}$$

We denote by

$$\mathcal{H}^{\Pi^{[n]}} = \tilde{T}_{n-1}^{\Pi^{[n]}} = \sum_{k=1}^{n-1} T_k^{\Pi^{[n]}} \quad \left(= T_{mrca}^{\Pi^{[n]}} \right) \tag{1.9}$$

the height of the coalescent tree corresponding to $\Pi^{[n]}$ *and by*

$$\mathcal{L}^{\Pi^{[n]}} = \sum_{k=1}^{n-1} N_{k-1}^{\Pi^{[n]}} \cdot T_k^{\Pi^{[n]}} \tag{1.10}$$

the (branch) length of the coalescent tree. Whenever there is no ambiguity, we omit the superscript $\Pi^{[n]}$ *or replace it by an* n *in order to ease notation.*

Both quantities, $\mathcal{H}^{\Pi^{[n]}}$ and $\mathcal{L}^{\Pi^{[n]}}$ play an important role in the investigation of biological processes: if for example mutations happen on the same time scale then their impact will be visible in a sample from the population (see Sections 1.3.2 and 4.2).
Note that in case of the Kingman coalescent all states from n to 1 will be attained, that is, we indeed have $n-1$ coalescent events and the $\tilde{T}_k$'s are all distinct. This is not necessarily

the case for a Λ-coalescent where the number of blocks can decrease by more than one at a time. However, in any case we have $\tilde{T}_{n-1}^{\Pi^{[n]}} = T_{mrca}^{\Pi^{[n]}}$. An example of the times $\tilde{T}$ and the corresponding block sizes is given in Figure 1.2.1b.

We will now turn back to the population setting and consider the situation where we want to trace back the ancestry of a sample containing a fixed number n of individuals out of a (infinitely) large population. In this case, there exist some very nice convergence results which in particular show the universality of the Kingman coalescent: many of the standard population models will in fact converge to the same limit object. Although this latter aspect was introduced by Kingman in [27, 26] we here follow the formulation by Möhle, [28]. In his work he considered an even more general population model, a Cannings model with varying population size for the different generations. Here, we will restate his result only in the case where the population size is constant over time and where the distribution of the offspring mechanism ν^t is independent of the generation t. Recall Remark 1.7 where the correspondence between partitions and equivalence relations was addressed.

Theorem 1.14 (Theorem 3.1 (Remark), [28], Kingman)**.** *Consider some Cannings model given through ν with ν^t identically distributed for all t and constant population size N as introduced in Definition 1.1. Fix a sample size $n \in \mathbb{N}$ and denote by $\left(\Pi_m^{\nu,n}\right)_{m\in\mathbb{N}}$ the process with values in $\mathcal{P}_n$ which is created by the following equivalence relation:*

$$i \sim^{\Pi_m^{\nu,n}} j \quad :\Leftrightarrow \quad \begin{array}{l} i \text{ and } j \text{ have a common ancestor } m \text{ generations backward in time} \\ \text{under a realization of } \nu, \end{array} \tag{1.11}$$

and $\Pi_0^{\nu,n} = \Delta_n$. Define

$$c_N := \frac{1}{N-1}\mathbb{E}(\nu_1(\nu_1 - 1)). \tag{1.12}$$

If the below stated conditions (1)-(3) are satisfied, the time-changed process $\left(\Pi_{\lfloor t/c_N \rfloor}^{\nu,n}\right)_{t\geqslant 0}$ converges weakly in $D_{\mathcal{P}_n}([0,\infty))$ to the Kingman n-coalescent process for $N \to \infty$:

$$\begin{aligned} &(1)\ \lim_{N\to\infty} c_N = 0, \\ &(2)\ \lim_{N\to\infty} \frac{1}{N^3 c_N} \sum_{i=1}^{N} \mathbb{E}(\nu_i^{k+1}(\nu_i - 1)) = 0, \text{ for all } k \in \mathbb{N}, \\ &(3)\ \lim_{N\to\infty} \frac{1}{N^4 c_N} \sum_{i,j=1}^{N} \mathbb{E}(\nu_i(\nu_i - 1)\nu_j^2) = 0. \end{aligned}$$

Here, $D_{\mathcal{P}_n}([0,\infty))$ is the Skorokhod space on $[0,\infty)$, that is, the space of all càdlàg (right continuous with left limits) functions on $[0,\infty)$ with values in $\mathcal{P}_n$.

In the following remark we give an intuition on the conditions (1) to (3) from above.

Remark 1.15. *Note that we needed to speed up time by the factor $1/c_N$ in order to see convergence to the Kingman n-coalescent. The value c_N defined in* (1.12) *is the probability of the event A, that two individuals picked at random from the population find a common ancestor in the previous generation:*

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{E}(\mathbf{1}_{\{A\}}) = \mathbb{E}(\mathbb{E}(\mathbf{1}_{\{A\}}|\nu)) \\ &= \mathbb{E}\Big(\sum_{i=1}^{N} \frac{\nu_i}{N}\frac{\nu_i - 1}{N-1}\Big) = \frac{1}{N-1}\frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}(\nu_1^2) - \mathbb{E}(\nu_1)\right) = \mathbb{E}\left(\frac{\nu_1(\nu_1-1)}{N-1}\right) = c_N.\end{aligned}$$

Thus, the new time scale is such that we expect to see one merging event in the sample in one time unit.
The conditions (2) and (3) on the offspring distribution ν stipulate that neither the probability for a parent to have more than two offspring nor the probability of two different parents producing offspring in the same generation is big enough for an occurrence of these events in the limit process.

Lemma 1.16. *The Wright-Fisher model as introduced in Example 1.2 and the Moran model from Example 1.3 both fulfill the conditions (1)-(3) and thus converge weakly to the Kingman coalescent as N tends to infinity.*

Proof. The proof can be done by straightforward calculations and we will therefore not state it here but only derive the value of c_N for both offspring distributions as it also gives insight in the needed time-change.
For the first introduced formulation of the Moran model we have $\mathbb{E}(\nu_1) = 2/N + (N-2)/N = 1$ and $\mathbb{E}(\nu_1^2) = 4/N + (N-2)/N = 1 + 2/N$ and thus $c_N = 2/(N(N-1))$.
In the Wright-Fisher model, each ν_i is distributed binomially with parameters N and $1/N$. Hence, $\mathbb{E}(\nu_1) = 1$, $\mathbb{E}(\nu_1^2) = 2 - 1/N$ and thus $c_N = 1/N$. □

As we will later focus on Λ-coalescents, we here state a more general convergence result as it was formulated by Schweinsberg in [37]:

Proposition 1.17 (Proposition 3, [37])**.** *Suppose*

$$\lim_{N\to\infty} c_N = 0 \quad \textit{and} \quad \lim_{N\to\infty} \frac{\mathbb{E}\big[(\nu_1(\nu_1-1)\cdot\nu_2(\nu_2-1)\big]}{N^2 c_N} = 0. \tag{1.13}$$

Also, assume that for some probability measure Λ on $[0,1]$, we have

$$\lim_{N\to\infty} \frac{N}{c_N}\mathbb{P}(\nu_1 > Nx) = \int_x^1 y^{-2}\Lambda(dy) \tag{1.14}$$

for all $x \in (0,1)$ at which the limit function is continuous. Fix a sample size $n \in \mathbb{N}$ and denote by $\left(\Pi_m^{\nu,n}\right)_{m\in\mathbb{N}}$ the process with values in $\mathcal{P}_n$ which is created by the equivalence relation

given in (1.11) *with* $\Pi_0^{\nu,n} = \Delta_n$. *Then, as* $N \to \infty$, *the process* $(\Pi^{\nu,n}_{\lfloor t/c_N \rfloor})_{t\geqslant 0}$ *converges to* $(\Pi_t^{\Lambda,n})_{t\geqslant 0}$ *which has the same law as the restriction to* $[n]$ *of a* Λ*-coalescent.*

The moment condition stated in (1.13) can be interpreted as follows: in the limit of large populations and with respect to the new time scale, the probability that two individuals reproduce at the same time should be negligible. This implies that we will see no simultaneous mergers in the limit process. Again, the time change is given through c_N.

Example 1.18. Let us resume the example of the modified Moran model introduced in Example 1.4 and consider different values for the parameter γ which regulated the occurrence of a multiple merger. For $\gamma > 2$, the authors from [16] showed that the limit process is indeed a Kingman coalescent. However, as observed in [16], if $\gamma < 2$, the probability for a merging event with more than two lineages is high enough such that those events can not be ignored in the limit process. We will here apply Proposition 1.17 in order to reach the same conclusion. Let $0 < \gamma < 2$ and $\psi \in (0, 1)$.

$$c_N = \frac{\mathbb{E}\big[\mathbb{E}[\nu_1^2 - \nu_1 \mid U]\big]}{N-1} = \frac{\mathbb{E}\big[\mathbb{E}[\frac{U^2}{N} + 1 - \frac{U}{N} - 1]\big]}{N-1} = \tfrac{1}{N(N-1)}\big[2(1-N^{-\gamma}) + \psi N^{1-\gamma}(\psi N - 1)\big],$$

and thus $c_N \to 0$ as $N \to \infty$ (holds true for all $\gamma > 0$). Further, the discrete model does not allow for simultaneous mergers and thus $\mathbb{E}[\nu_1(\nu_1 - 1) \cdot \nu_2(\nu_2 - 1)] = 0$. Now we want to look for a measure Λ which might fulfill (1.14). By definition of the distribution of U we find $\mathbb{P}(\nu_1 > Nx) \equiv 0$ for all $x \in (0, \frac{2}{N}) \cup [\psi, 1)$. Now, for $x \in [2/N, \psi)$, we have

$$\begin{aligned}\frac{N}{c_N}\mathbb{P}(\nu_1 > Nx) = \frac{N}{c_N}\mathbb{E}\Big[\frac{1}{N}\mathbf{1}_{\{U=N\psi\}}\Big] = \frac{1}{c_N}N^{-\gamma} &= \frac{1}{\frac{1}{N(N-1)}\big(2(N^\gamma - 1) + \psi N(\psi N - 1)\big)} \\ &\to \frac{1}{\psi^2},\ N \to \infty, \quad \text{as } \gamma < 2.\end{aligned}$$

Hence, for $\Lambda = \delta_\psi$, the Dirac measure at ψ, the equality in (1.14) holds true for all $x \in (0, 1)$. With time scaled by $1/c_N$, the ancestral process of an n-sample following the offspring distribution ν therefore converges to $(\Pi_t^{\delta_\psi,n})_{t\geqslant 0}$ as $N \to \infty$. In particular, for $0 < \gamma < 2$, the two-mergers will not occur in the limit process.
As mentioned by Eldon and Wakeley, the case $\gamma = 2$ is special in the sense that both kind of mergers will be seen in the limit process. This can be deduced from the following calculation:

$$\frac{N}{c_N}\mathbb{P}(\nu_1 > Nx) = \frac{1}{c_N}N^{-2} = \frac{1}{\frac{1}{N(N-1)}\big(2(N^2 - 1) + \psi N(\psi N - 1)\big)} \to \frac{1}{2 + \psi^2},\ N \to \infty,$$

and hence, $\Lambda = \frac{2}{2+\psi^2}\delta_0 + \frac{\psi^2}{2+\psi^2}\delta_\psi$ fulfills (1.14) for $x \in (0, 1)$. ♢

With this theory of convergence at hand, there are now two possible ways to reflect the evolution of a population for large, constant population sizes. On the one hand, we can

study the discrete-time model and specify in every detail how the reproduction mechanism is supposed to work. On the other, we can first take the limit for infinite populations and then work with the limit process. Often, the latter is chosen since the analysis of coalescent processes is usually simpler and by now there is a vast variety of theoretical work to rely on. In addition, many discrete models actually converge to the same limit and hence, the study of the one limit object allows to draw conclusions for several schemes.

1.3 Biological complications: recombination, mutation and selection

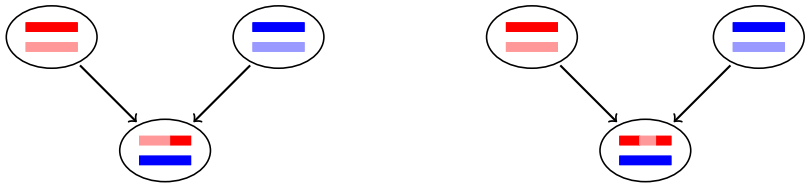
In this section, we want to introduce some biological processes which influence the evolution of a population and which we have ignored up to this point. We will consider the discrete population models introduced before and also show how these biological complications can be incorporated into processes in continuous time. We will in particular introduce and study the impact of recombination as this mechanism is one of the key aspects of this work.

1.3.1 Recombination

In contrast to Section 1.1, we will first explain recombination for diploid populations, that is, populations where the individuals have a double chromosome set. In the models which are introduced later we however consider the framework of sexually reproducing haploid populations, that is, we consider a gene pool of single chromosomes and disrespect the diploid pairing. For a rigorous mathematical model for diploid individuals see for example [8].

Definition 1.19. *Recombination denotes the chromosomal crossover (or crossing-over) which may happen during the meiosis when the parental gametes are formed. Before the cell division starts, each chromosome is duplicated and during this process the two chromosomes of one parent can entangle and thus interchange genetic material. The resulting chromatids after meiosis then are not a simple copy of one parental chromatid but are a unique combination of genetic material from both of the parental chromosomes.*

Remark 1.20. *1. If we go back to the level of one chromosome, we see that in contrast to the previously described models it can have up to two parent chromosomes, while each locus is still copied from exactly one individual. Note that in particular in the case of a double (or multiple) crossover as shown in Figure 1.3.1b, the offspring's chromosome still only copied material from the two maternal (or paternal) chromosomes.*
2. When we later, in Chapter 2, indeed restrict ourselves to the chromosomal level, we will consider population models where the two parents of an offspring are drawn randomly from



(a) Single crossover. **(b)** Double crossover.

Figure 1.3.1: A pair of homologous chromosomes of the offspring.

all the individuals in the population. We have to keep in mind though that this is only an approximation. In the diploid scenario, the second parent of the newborn's chromosome is known to be the second homologous chromosome of the first parent and not some randomly chosen partner.

Let us now consider different ways of modeling the evolution of a population when taking care of the phenomenon of recombination. We will give an example of a discrete time model in a similar spirit as in the Examples 1.31 and 1.32 and then introduce two different approaches of modeling the process in continuous time.

Definition 1.21. *Consider a neutral, discrete-in-time model with constant population size $2N$. In each individual, the genealogical relationships of $l \in \mathbb{N}$ distinguished loci are considered. At each locus L_j, an individual can have one of n_j possible alleles. We study the alignment $L_1 - L_2 - \cdots - L_l$ and assume that during a reproduction event, a recombination happens (independently of everything else) between each adjacent pair of loci L_j and L_{j+1} with probability r_j, $j = 1, 2, \ldots, l-1$. A new generation at time $t+1$ is then created as follows:*

- *each individual i from generation t has ν_i offspring, where $\sum_{j=1}^{2N} \nu_j = 2N$ and ν follows some distribution (as in Definition 1.1)*
- *independently for each offspring, the number and places of recombinations are given by the outcome of $l-1$ independent Bernoulli random variables $B_1, \ldots, B_{l-1}$ with $B_j \sim$ Bernoulli(r_j),*
- *define the recombination points $m_1, m_2, \ldots$ by $m_1 := \min\{k \geqslant 1 : B_k = 1\}$ and for $j > 1$ $m_j := \min\{k > m_{j-1} : B_k = 1\}$*
- *if there is no recombination (all $B_j \equiv 0$), the offspring copies all alleles from the parent,*
- *if there is at least one recombination, a second parent is chosen uniformly at random from all individuals of the t-th generation and the offspring copies the alleles from both its parents as follows:*

– *the alleles at the loci* $1, 2, \ldots, m_1, m_2+1, \ldots, m_3, m_4+1, \ldots$ *are copied from the first parent*

– *the alleles at the remaining loci* $(m_1+1, \ldots, m_2, m_3+1, \ldots, m_4, m_5+1, \ldots)$ *are copied from the second parent*

Figure 1.3.2: Example of the offspring's alleles for $l = 7$.

- parent's alleles: ●,
- second parent's alleles: ■
- $(B_1, \ldots, B_6) = (1, 0, 0, 1, 0, 1)$

$\Rightarrow$ offspring's alleles: ●■■■●●■

For a better understanding, we will exemplarily describe a Moran model with two loci and two alleles in the next example.

Example 1.22. For a population size of now $2N$ individuals, we use the offspring distribution from Example 1.3 in which the Moran model was introduced. Let us consider $l = 2$ loci, with possible alleles A, a at L_1 and B, b at L_2. Recall that one step of this model consists of uniformly choosing one individual which reproduces and one individual which is replaced by the offspring of the former. Here, with probability r_1, the offspring copies the allele at L_2 not from the parent but from an again uniformly chosen individual of the parental generation. If for example we have drawn an (A, B)-individual as the parent and an (A, b)-individual as the second parent, the offspring is of type (A, B) with probability $1 - r_1$ and of type (A, b) with probability r_1. See Figure 1.3.3 for an illustration.

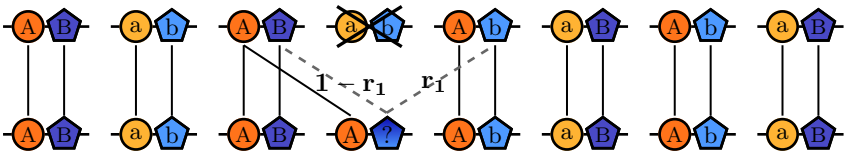


Figure 1.3.3: A step in a Moran model with recombination, $2N = 8$, type of offspring at second locus depends on whether there is a recombination.

♢

In continuous time, one would like to generate an analog to a coalescent tree in order to describe the genealogy of a sample (or the whole population). Trying to interpret the genealogy

of some group of individuals by means of a coalescent brings problems as soon as a recombination causes one chromosome with two parents. Such ancestry can no longer be described by a tree but we need a graph in order to capture the different ancestral paths of the different parts of one chromosome. The resulting construction is then called an *ancestral recombination graph*, or short, ARG. As already mentioned in Remark 1.20, each single locus in the genome can still be traced back to exactly one individual from each previous generation, although the whole chromosome may have two ancestors. Thus, we can indeed build a coalescent tree for each locus of the present population. However, all those coalescent trees will be highly correlated as neighboring loci are inherited together if there is no recombination. As pointed out in Durrett [14], it is rather hard to obtain analytical results considering the properties of genealogical processes which account for recombination. While there are explicit formulas for the covariance of the coalescent trees for partially linked loci in case of a sample size of two individuals, most other results for a higher number of individuals are of a recursive nature. Instead of reproducing known analytic results here, we concentrate on the algorithmic construction of processes with recombination.
However, before we do so, we need to think about the correct time scaling of the recombination probability, similar to the line of thought from the previous Section 1.2.
In Theorem 1.14 and Proposition 1.17 we saw that, for constant population size $2N$, we need to scale time in the discrete process by a factor of $1/c_{2N}$ in order to obtain convergence to a coalescence process. In the discrete model with recombination, each individual recombines (independent of all others and previous events) with probability r within the time span of one generation. Thus, during $1/c_{2N}$ generations the expected number of recombinations for an individual is r/c_{2N}. Inspired by this, define the *scaled recombination rate* (also called the population recombination rate)

$$\frac{\rho}{2} := \frac{r}{c_{2N}}, \tag{1.15}$$

and denote by T^c and T^d the time until a recombination happens in one ancestral lineage with respect to the continuous and the discrete process, respectively. Then, if T^c is greater than t time units, T^d needs to be greater than t/c_{2N} time units. This implies:

$$\begin{aligned}\mathbb{P}(T^c > t) &= \mathbb{P}(T^d > t/c_{2N}) = (1-r)^{t/c_{2N}} \\ &= \Big(1 - \frac{r/c_{2N}}{1/c_{2N}}\Big)^{\frac{t}{c_{2N}}} = \Big[\Big(1 - \frac{\rho/2}{\frac{1}{c_{2N}}}\Big)^{\frac{1}{c_{2N}}}\Big]^t \to \exp\big(-\rho t/2\big),\end{aligned} \tag{1.16}$$

which is the probability that an $\mathrm{Exp}(\rho/2)$-distributed random variable exceeds t. Hence, for the rescaled process in continuous time, the time until one lineage experiences a recombination is distributed exponentially with parameter $\rho/2$, which is in general assumed to be finite.

Remark 1.23. *In biology and as well in mathematical models describing biological processes, often we do not measure the length of some DNA sequence with respect to the physical distance between the endpoints or the number of base pairs included in the sequence. Instead, the length is expressed in the expected number of recombination points along the sequence. A unit of that length measure is called a centimorgan, or a map unit. More precisely, one centimorgan (cM) is equal to that distance between positions on the chromosome in which the expected number of recombinations per generation is equal to* 0.01*.*
The existence of so-called recombination hotspots on the genome implies that those two distance measures, number of base pairs and centimorgan, are in general not directly proportional to each other.

In the following we will measure distances between genetic loci in the expected number of recombinations occurring between them.

Remark 1.24. *Recall Definition 1.21 and assume that the probability to recombine is identical between all loci, that is,* $r_1 = r_2 = \ldots = r_{l-1} = r$*. Then the number of recombinations on one sequence in one step is binomially distributed:*

$$\mathbb{P}(\#\ \textit{recombinations}\ = n) = \binom{l-1}{n} r^n (1-r)^{l-1-n}.$$

As each generation is created through the same mechanism, this can be extended to the following thought: suppose we are interested in the number of recombinations in a genealogy of time length $1/c_{2N}$ *(keep in mind that we will scale time by* $1/c_{2N}$*). Then,*

$$\begin{aligned}&\mathbb{P}(\#\ \textit{recombinations during time}\ 1/c_{2N}\ = n)\\ &= \binom{(l-1)/c_{2N}}{n} r^n (1-r)^{(l-1)/c_{2N}-n}.\end{aligned}$$

If we assume the following limit behavior,

$$1/c_{2N} \to \infty,\ l \to \infty,\ r \to 0\ \ \textit{such that}\ \ r \cdot (l-1)/c_{2N} \to \rho/2,$$

then, by the Poisson limit theorem, we get

$$\mathbb{P}(\#\ \textit{recombinations during time}\ 1/c_{2N}\ = n) \to \exp(-\rho/2)\frac{(\rho/2)^n}{n!}. \tag{1.17}$$

Thus, the number of recombinations in 1 unit of the new time is Poisson distributed with parameter $\rho/2$*.*

Note particularly that in the rescaled process with time sped up, there are almost surely no simultaneous recombination and coalescence as both events can be modeled by Poisson Point

processes (with intensity proportional to the Lebesgue measure).

The construction of an ancestral recombination graph (ARG)

We will now describe two possible algorithms for the construction of an ARG and focus here on Kingman coalescents with binary mergers only, as defined in Definition 1.9. An algorithm for the ARG in case of a more general Λ-coalescent is then given in Section 4.3.

The ARG as a birth and death process. The first approach, as described by Griffiths [20], Griffiths and Marjoram [21], is to directly generate a graph starting at time 0 with the current sample of individuals and going back in time until all loci in the sample have found a common ancestor. The genealogical line of an individual then can split whenever a recombination produces two different ancestors of the genetic material of that individual, or the line merges with another lineage in a coalescent event. We will here only consider a two-locus version with a one-step recombination probability of r between those two loci. This model however can be extended straightforwardly towards multiple loci with different strength of linkage between them.

Algorithm 1.25 (cf.[20],[21])**.** The ARG can be constructed as a birth and death process where the rates depend on the number of ancestral lineages present at a given time. If we use the scaled recombination rate from (1.15), the rates of the process are as follows:

$$\text{birth rate } = k\rho/2, \qquad \text{death rate } = \binom{k}{2},$$

in case there are k lines present in the graph. A birth corresponds to a recombination of the affected lineage, a death to the coalescence of two lineages. ♢

Note that the death rate is quadratic in k compared to the linear birth rate, which implies that the process will almost surely reach the state where there is only one line left in the graph. Although the process could be continued from that time on (as the birth rate is positive) we stop the process at the first time it reaches one line as no further information on the genealogy is gained from that point on.
The death rate is known from Remark 1.10, the birth rate follows the intuition from Equation (1.16). For a better understanding of the shape of such a process, see Figure 1.3.4 in which we consider a two-locus model and recombinations are illustrated by stars. We further indicated the embedded coalescent tree which describes the history of the rightmost locus by slightly shifted orange lines. The leftmost locus only finds a common ancestor when the whole graph reaches the state of one line.
This straightforward construction however can lead to the creation of edges in the graph which contain no genetic material ancestral to the sample, such as for example the dashed

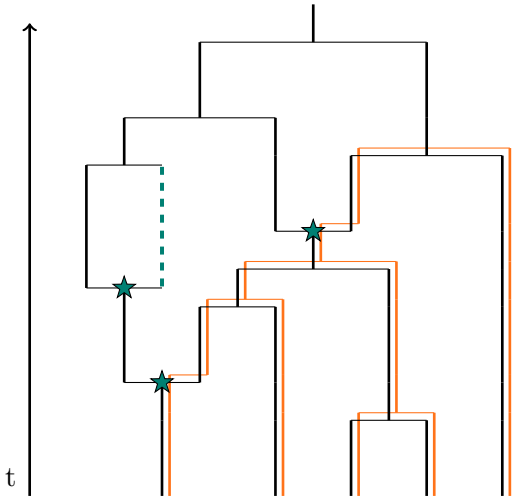


Figure 1.3.4: Ancestral recombination graph as a birth- and death process.

line in Figure 1.3.4, a result of two successive recombinations. The true time until the most recent common ancestor of all sampled genetic material is found can therefore be smaller than the generated height of the graph. This is an important difference between ancestral recombination graphs and the coalescent trees from Section 1.2.
As we will later focus only on the second possibility to construct the ancestral recombination graph, we will give no further detail to the properties or extensions of the above birth- and death process but continue with the definition of a spatial algorithm.

A spatial algorithm for the construction of the ARG. The second idea was introduced by Wiuf and Hein in [41], and can be thought of as "walking along the genome" instead of going back in time. In this algorithm, each sampled sequence is taken from the same part of the genome for all individuals and the sequences are thought of as a continuous interval rather than a discrete alignment of loci. The basic idea is to first generate a coalescent tree which models the ancestry of the leftmost point in the observed part of the gene sequences. Then, in order to check whether this tree also describes the history of the neighboring parts, we need to define a process which defines the number, point and time of the next recombination event. Here, "next" is meant with respect to the spatial position in the genome, independent of time. If there is a recombination, we adapt the existing coalescent in a specific way and subsequently continue this process for all successive loci, each time based on the most recent coalescent tree.
We borrow the notation from [41], where the authors start from a constant population size model for N diploid individuals, each with two sequences which consist of L nucleotides. The next generation is then formed by sampling with replacement N pairs, that is $2N$ sequences, of the previous generation and allowing the pairs to recombine between any two nucleotides

with probability r. Similar as described in Remark 1.24, the limit of large populations is then taken as follows:

$$\begin{aligned} &\text{time is measured in } 1/c_{2N} = 2N \text{ generations and} \\ &N \to \infty,\ L \to \infty \ \text{ and } \ r \to 0 \ \text{ such that } \ 2NLr \to \rho/2. \end{aligned} \tag{1.18}$$

This way, the authors measure the sequence length in expected number of recombinations per $2N$ generations (cf. Remark 1.23). The findings from Remark 1.24 and in particular (1.17) then justify the following, which is stated in Section 1 in [41]: let X be the sequence length until a recombination occurs, and b the total branch length of the complete genealogy, then

$$\begin{aligned} &\mathbb{P}(\text{no recombination in the genealogy} \mid b) = \exp(-b\rho/2), \\ &\text{and for } x < \rho/2 \text{ we have} \qquad \mathbb{P}(X > x \mid b) = \exp(-bx), \end{aligned}$$

that is, X given b is exponentially distributed with parameter b, truncated at $\rho/2$. As all generations are constructed in the same way and the individuals are exchangeable, the recombination event takes place uniformly at random over all ancestral lineages: denote by T the location of the event on the graph, then $T \sim \mathcal{U}(0, b)$.
It is helpful to define the concept of a local tree before we continue.

Definition 1.26. *Let L be the length of each of the sampled sequences, and $p \in [0, L]$ a position in the genome (which can be either a point or be identified with a gene). Then the local tree at p, denoted by $\mathcal{T}(p)$, is the coalescent tree which describes the genealogy of the sample with respect to the point p.*

We can now state a pseudo-code of the spatial algorithm.

Algorithm 1.27 (cf. [41])**.** Assume the model from [41] as described above with limit behavior of the parameters as stated in (1.18). Each of the (finitely many) sampled sequences is thus identified with the interval $[0, \rho/2]$.

1. Generate a local (Kingman) coalescent tree $\mathcal{T}(0)$ for position 0.
2. Determine the total branch length $b(0)$ of $\mathcal{T}(0)$ (as defined in Definition 1.13).
3. Put $B_0 = b(0)$, $P_0 = 0$ and $G_0 = \mathcal{T}(0)$ and repeat the next steps for $i = 1, 2, \ldots$ while $P_{i-1} \leqslant \rho/2$:
4. Choose the next recombination point P_i on the sequence through $p_i \sim \mathrm{Exp}(B_{i-1})$ and then $P_i = P_{i-1} + p_i$. Break if $P_i > \rho/2$.
5. Draw the location (time and lineage) via $t_i \sim \mathcal{U}(0, B_{i-1})$.

6. Split the concerned line a position t_i in two and thus create a new edge e_i.
7. Coalesce the new recombined edge e_i to the graph G_{i-1} according to the rates of a Kingman coalescent: if there are k lines present in G_{i-1}, the rate for coalescence of e_i to the graph is equal to k (as we have a pair coalescence rate of 1 and there are k pairs of e_i and one other line of the graph). The edge e_i starts at t_i and ends at the point where it coalesces with the previous graph.
8. Set $G_i = G_{i-1} \cup e_i$ and B_i the total branch length of G_i.

♢

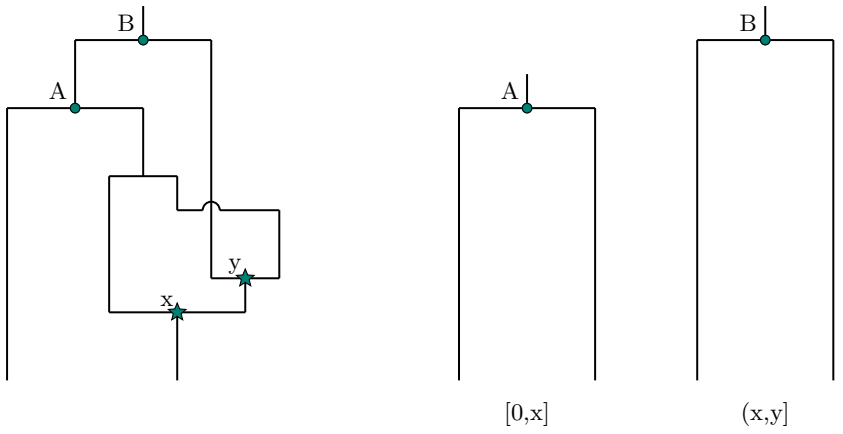
We will look at an example in a more complex context in Chapter 4 and finish this section on recombination with a few remarks on the spatial algorithm.

Remark 1.28. *1. B_i is strictly greater than B_{i-1} but not necessarily equal to $B_{i-1} + |e_i|$. It may happen that the new edge coalesces only with the root, thus the height of the ancestral coalescence graph can increase.*
2. As the branch length is strictly increasing, the recombination points will be closer on the sequence the further the algorithm advances. However, many of them will fall at a position t on some edge e of the graph where there is no ancestral material affected by a recombination at a breaking point p, as the edge e might only describe the ancestry of positions $\tilde{p} < p$. We still have to continue the algorithm in those cases as it might happen that later, a new edge coalesces with that edge e and endows it with ancestral material which is indeed affected by a recombination at p. Only in the very end we can delete edges (and recombinations on them) which have no influence on the ancestry of the sample.
3. The authors of [41] show that their graph is embedded in the graph resulting from Algorithm 1.25 and thus the algorithm as described above will indeed stop after a finite time almost surely.
4. We need to keep track of the whole graphs G_i and can not formulate the algorithm as a Markov process on the set of local trees $\{\mathcal{T}(p),\ p \in [0, \rho/2]\}$. This can be seen by the example given in Figure 1.3.5 where in the graph construction, the positions $(y, \rho/2]$ on the sequences find a common ancestor at the same point A as positions $[0, x]$. Now, if we only kept track of the last local tree, $\mathcal{T}((x, y])$, the event that the tree for $(y, \rho/2]$ indeed has the same shape as the tree for $[0, x]$ would have probability zero, in contrast to the positive probability obtained in the graph construction.

In Chapter 4 we will see which changes are necessary in order to adapt this algorithm for the use in the context of Λ-coalescents.

1.3.2 Mutations

As this thesis focuses mainly on the impact of recombination and selection on the genealogy of a sample, we will only briefly discuss the process of mutation. So far, the offspring's genome



(a) ARG for two individuals, $x < y$. **(b)** Local trees.

Figure 1.3.5: A counterexample.

was an exact copy of the DNA of its parent(s). In real life, this copying process holds the risks of errors which can lead to a mutation.

Definition 1.29. *The word mutation is a generic term for any event that leads to changes in the DNA sequence of an individual. This could be for example a transversion or transition of one nucleotide to another, such as the substitution of a G by an A, or the insertion or deletion of parts of the DNA. We speak of synonymous or silent mutations if the change in the DNA does not lead to a change of the thus coded protein.*

For an extensive list of the different types of mutations, see for example Chapter 5 of [22]. We will here consider only neutral mutations which lead to no advantage or disadvantage of the affected individual. Only in the next section, where we introduce the concept of selection, we discuss some models which take care of the influence of advantageous gene mutations. As neutral mutations do not change the character of the underlying population model, it is very easy to include them into our theory for Cannings models and the coalescent.
Regardless of the exact effect of the mutation, we can incorporate this biological process into the discrete time models by allowing each offspring to mutate with a certain probability u, independently of all others. As in Section 1.3.1, we can define the *scaled mutation rate*

$$\frac{\theta}{2} := \frac{u}{c_N}, \tag{1.19}$$

and can determine the probability of witnessing a certain number of mutations in the corresponding continuous time process in the same way as in Remark 1.24. By the Poisson limit

theorem we get

$$\begin{aligned}&\mathbb{P}(\#\text{ mutations in } 1/c_N \text{ generations } = m)\\ &= \binom{1/c_N}{m} u^m (1-u)^{1/c_N - m} \to \exp\left(-\theta/2\right) \frac{\left(\frac{\theta}{2}\right)^m}{m!},\end{aligned}$$

which implies that for a coalescent tree with total branch length b we can uniformly distribute a Poisson$(b\theta/2)$-distributed number of mutations on the tree. This is one of the reasons why it is interesting to consider the height and in particular the length of coalescent processes.

Depending on the point of interest, we can choose between two different types of models, both introduced by Kimura in [25] and [24]: the first is the so-called *infinite alleles model* (IAM), the second is called the *infinite sites model* (ISM).
For the IAM we assume that every new mutation leads to a different characteristic of the concerned gene, that is, each mutation leads to a new allele. This makes it rather simple to model the allele types of individuals sampled from a present population as only the first encountered mutation, when following the ancestry back in time, is relevant and fully determines the type of the individual.

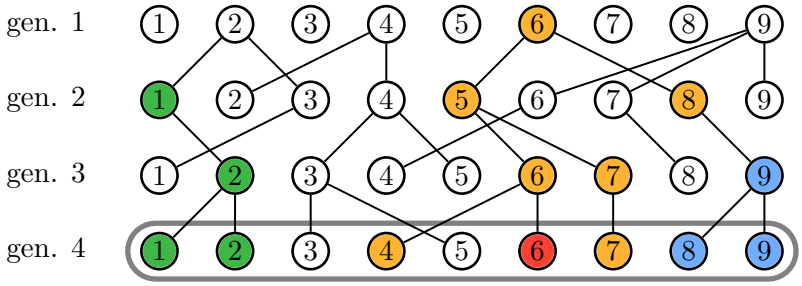


Figure 1.3.6: Wright-Fisher Model with mutation, IAM; different mutations in different colors; sample from present time in gray box.

The importance of the second model, the ISM, increased substantially when the newer sequencing methods led to a finer decoding of the DNA, enhancing the demand for a more accurate model which can capture sequences of mutations. As the name suggests, we here assume an infinite number of sites in the considered sequences which implies that every new mutation hits a new site with probability 1. The present generation from the example in Figure 1.3.6 then could be represented as shown in Figure 1.3.7

There exists a vast amount of literature concerning these two different models, which also gives theoretical answers about the distribution of types in a sample. As this topic is not within the scope of this work, we continue with the concept of selection.

Figure 1.3.7: Sample from present time under the IAM; different mutations in different colors, all mutations recorded.

1.3.3 Selection

When recalling the Definition 1.1 of a Cannings model, we see that we might run into problems when adding the concept of selection to the reproduction process. Even if the assumption of constant population size can be met approximately, the request for exchangeability clearly contradicts the idea of selection, which is, that some individuals have a higher chance of reproducing than others. We will therefore not try to fit models with selection into the Cannings framework.

Within this thesis we will only consider non-spatial models and assume that selection is expressed only through the genotype of an individual and not by its phenotype or environmental properties. In order to study the impact of different selection mechanisms we therefore label all individuals in the population with respect to their genotype. We will in particular focus on a so-called one-locus-two-allele model where each individual has either the allele A or the allele B at some distinguished locus in the genome. Based on this simple model one can then add more loci, more possible alleles or other structures, such as interactions between the different allele combinations at a number of loci.

Let us first consider only one locus. If the gene expression at the locus is not neutral, it has some influence on the fertility or lethality of an individual and hence, it exerts some kind of selective pressure. This pressure can either be negative or positive and leads to so-called directional selection. In this directional selection, one genotype is favored over the other which leads to fixation of the advantageous genotype and thus at the same time to extinction of the other type. Here, the process of *favoring* one allele can be expressed by different mechanisms. Either, an individual having the favorable allele is less likely to die than an individual with the other type, or the probability to produce offspring is higher, or else, the number of offspring produced by an individual with the advantageous type is higher. Before we give an example of a population model with directional selection we introduce the notion of relative fitness and the selection coefficient for a population with two possible genotypes A and B.

Definition 1.30 (see Chapter 6 in [22])**.** *Let $N_G(t)$ be the number of individuals in the population with type $G \in \{A, B\}$ at time t. We suppose that the growth rate of the population is genotype specific and that those rates λ_G for $G \in \{A, B\}$ do not change over time. Then λ_G is said to be the absolute fitness of the genotype G and*

$$N_G(t+1) = \lambda_G \cdot N_G(t), \quad G \in \{A, B\}.$$

λ_G can therefore be understood as the average number of offspring produced by an individual of type G. The relative fitness w_A^B of the genotype B with respect to the genotype A is given through the fraction of the absolute fitnesses, $w_A^B = \lambda_B/\lambda_A$, and the selection coefficient is then defined by

$$s^B := 1 - w_A^B, \; s^A := 1 - w_B^A.$$

Note that s^G can be positive or negative, depending on which allele is favored over the other. We will later drop the type-dependent sub- and superscripts whenever the reference is clear.

Example 1.31 (Wright-Fisher model with selection)**.** Recall the Example 1.2 of the Wright-Fisher model and let $s > 0$. Selection in a two-allele model where individuals with allele A have relative fitness $(1+s) : 1$ with respect to type-B-individuals, can be expressed through a change in the probability of sampling an individual with a specific genotype from the generation at time t to be a parent of an individual from generation $t+1$:

$$\mathbb{P}(A\text{-individual is sampled}) = \frac{N_A(t)(1+s)}{(1+s)N_A(t) + N_B(t)},$$
$$\mathbb{P}(B\text{-individual is sampled}) = \frac{N_B(t)}{(1+s)N_A(t) + N_B(t)}.$$

Note that we could have replaced $N_B(t)$ by $N - N_A(t)$ here as the Wright-Fisher model assumes constant population size N. ♢

Example 1.32 (Moran model with selection)**.** We here present two different ways of constructing a Moran model with selection where the relative fitness of the A-allele is $1+s$. The first version is taken from [40] and very much like the above Wright-Fisher model with selection whereas the second example is introduced (with different parameter and additional complication) in [36]. Recall that a step in the Moran model is equivalent to choosing one individual which reproduces and one individual which dies.

Model 1. We assume that the selective pressure acts through the choice of the parent, whereas the killed individual is chosen independently of the genotype. Suppose, $N_A(t) = k$ and hence $N_B(t) = N - k$. Then the following transition probabilities follow for the process $N_A(t)$:

$$\mathbb{P}(k \to k+1) = \frac{k(1+s)}{k(1+s) + (N-k)} \frac{N-k}{N}$$
$$\mathbb{P}(k \to k-1) = \frac{N-k}{k(1+s) + (N-k)} \frac{k}{N}$$
$$\mathbb{P}(k \to k) = \frac{k(1+s)}{k(1+s) + (N-k)} \frac{k}{N} + \frac{N-k}{k(1+s) + (N-k)} \frac{N-k}{N} = 1 - \frac{(2+s)k(N-k)}{N - ks},$$

thus the ratio of the absolute fitnesses is indeed

$$\frac{\mathbb{P}(k \to k+1)}{\mathbb{P}(k \to k-1)} = \frac{k(1+s)(N-k)}{(N-k)k} = \frac{1+s}{1}.$$

Model 2. In this version, we suppose that both parent and killed individual are chosen uniformly at random regardless of their types. However, whenever the parent is of type B whereas the killed individual has the advantageous type A, the whole event is declined with probability $s/(1+s)$ (in [36] this probability is set to be s, resulting in a relative fitness of $1:(1-s)$). We get the following transition probabilities:

$$\begin{aligned}
\mathbb{P}(k \to k+1) &= \frac{k}{N} \cdot \frac{N-k}{N} \\
\mathbb{P}(k \to k-1) &= \frac{N-k}{N} \cdot \frac{k}{N} \cdot \Big(1 - \frac{s}{1+s}\Big) \\
\mathbb{P}(k \to k) &= \frac{N-k}{N} \cdot \frac{k}{N} \cdot \frac{s}{1+s} + \frac{k^2}{N^2} + \frac{(N-k)^2}{N^2} = 1 - \frac{\frac{2+s}{1+s}k(N-k)}{N^2}.
\end{aligned}$$

The calculation of the relative fitness results again in $1+s$, as in the first version. However, despite the fact that in both cases we considered a Moran model with two types and relative fitness $1+s$, we get two different models as we have different transition probabilities. Note that in the first model, the selection clearly acted through a higher birth rate of A-individuals. In the second model, the interpretation is not that easy as both parent and dying individual are chosen uniformly at random. The decline of a proposed replacement can either mean that we reject the death of an A-individual, and thus selection acts through survival probabilities, or that we reject the birth of a B-individual, which would imply a higher fertility of the A-individuals compared to the B-individuals. A further discussion of this aspect is postponed to Section 3. ♢

There are different possibilities to describe a process with selection in the limit of large population sizes. On the one hand, there is the so-called *ancestral selection graph*, which was rigorously introduced by Krone and Neuhauser in [29] and [30]. It is similarly constructed as the ancestral recombination graph, only that in this case a branching backwards in time is interpreted as a selection event rather than a recombination.
On the other hand, one can also include selection into the diffusion approximation. Coming from a discrete process such as the Wright-Fisher model with selection, we obtain a different stochastic differential equation than (1.1) for the frequency process of the advantageous type. Assuming a scaling $s = \alpha/N$ of the selection coefficient and rescaling time appropriately, we obtain in the limit of $N \to \infty$

$$dX_t = \sqrt{X_t(1-X_t)}dB_t + \alpha X_t(1-X_t)dt. \tag{1.20}$$

As both approaches are not pursued within this work and only used as a reference during a comparison of our results with related work in Section 2.3.2, we do not go into further details here.

1.4 The evolution of a population of varying size as a birth and death process

So far, we have focused on populations with constant size over time. In this last introductory section we want to briefly mention another possible way to describe the evolution of a population to which we will go back in Chapter 3. For once, our focus does not lie on the determination of the genealogies of individuals of the present time but we consider a forward in time model capturing the number of individuals of a certain type which are present in the population at a time. The evolution through time is assumed to be a Markovian birth and death process: each individual of the population gives birth or dies at a certain rate which may depend on the genotype of the individual and can further include dependencies on the current state of the population expressed through a competition kernel which acts within the total death rate of an individual. Such an approach is called a model for Darwinian evolution, as it assumes type-dependent reproduction- and survival success including competition. Its mathematical properties were in particular rigorously studied in Fournier and Méléard [19] and Champagnat, [13]. We will here define a simple model for an asexual haploid population without recombination or mutation and later, in Chapter 3, consider a more advanced adaptation of it. Further, we will state, without proof, some results from [13] which help to understand the underlying dynamics of the model which will be studied in Chapter 3.

Notation and Population dynamics

We suppose that each individual has some genotype α from a type space $\mathcal{A}$ which is supposed to be finite. In this simple case, we consider $\mathcal{A} = \{A, a\}$. The following microscopic biological parameters define the dynamics of the population process:

Definition 1.33 (cf. page 3 (1129) in [13]). *For $\alpha, \alpha' \in \mathcal{A}$ let*

$$\begin{aligned} &b(\alpha) \in \mathbb{R}_+ \text{ be the rate of birth from an individual of type } \alpha, \\ &D(\alpha) \in \mathbb{R}_+ \text{ be the rate of natural death of an individual of type } \alpha, \\ &C(\alpha, \alpha') \in \mathbb{R}_+ \text{ be the competition kernel evaluated at } (\alpha, \alpha'). \text{ It} \\ &\qquad \text{defines the pressure felt by an individual of type } \alpha \\ &\qquad \text{from an individual of type } \alpha'. \end{aligned} \tag{1.21}$$

Further, $K \in \mathbb{N}$ is a parameter which rescales the competition kernel C. The scaling parameter K reflects the ability of the habitat to nourish and accommodate a population and is often referred to as the *carrying capacity*.
At each time, the state of the (finite) population scaled by the parameter K can be described by a finite point measure ν_t^K on $\mathcal{A}$. Let $N_\alpha(t)$ denote the number of individuals of type α present in the population at time t (note that in the varying population size model we actually need to keep track of both values N_A and N_a over time). If there are in total $N(t) = N_A(t) + N_a(t)$ individuals present, we have

$$\nu_t^K = \frac{1}{K} \sum_{i=1}^{N(t)} \delta_{\alpha_i} = \frac{1}{K} [N_A(t)\delta_A + N_a(t)\delta_a], \tag{1.22}$$

where α_i denotes the type of the i-th individual. The death rate of an individual of type α is composed of the rate for natural death and competition and it can be calculated by using (1.22):

$$d(\alpha) = D(\alpha) + \int_{\mathcal{A}} C(\alpha, \alpha')\nu_t^K(d\alpha') = D(\alpha) + C(\alpha, A)\frac{N_A(t)}{K} + C(\alpha, a)\frac{N_a(t)}{K}. \tag{1.23}$$

We can now define a birth and death Markov process as follows:

Definition 1.34 (Definition 1(b), [13])**.** *Let $K \geqslant 1$. Using the notation from Definition 1.33 we denote by*

$$\mathbf{Q}^{(K)} = \mathbf{Q}^{(K)}(b(A), b(a), D(A), D(a), C(A,A), C(A,a), C(a,A), C(a,a), N_A^{(K)}(0), N_a^{(K)}(0)) \tag{1.24}$$

the law of the birth and death Markov process

$$\begin{aligned} &N^{(K)} = (N_A^{(K)}, N_a^{(K)}) = \big(N_A^{(K)}(t), N_a^{(K)}(t)\big)_{t \geqslant 0}, \quad \textit{with} \\ &\big(N_A^{(K)}(t), N_a^{(K)}(t)\big) = \Big(\frac{N_A(t)}{K}, \frac{N_a(t)}{K}\Big). \end{aligned} \tag{1.25}$$

$N^{(K)}$ has values in $\left[\frac{\mathbb{N}}{K}\right]^2$, initial value $(N_A^{(K)}(0), N_a^{(K)}(0))$ and transition rates as follows:

$$\begin{aligned} &\textit{with rate } n_A b(A) \textit{ from } (n_A/K, n_a/K) \textit{ to } ((n_A+1)/K, n_a/K), \\ &\textit{with rate } n_a b(a) \textit{ from } (n_A/K, n_a/K) \textit{ to } (n_A/K, (n_a+1)/K), \\ &\textit{with rate } n_A(D(A) + C(A,A)n_A/K + C(A,a)n_a/K) \textit{ from } (n_A/K, n_a/K) \\ &\qquad \textit{to } ((n_A-1)/K, n_a/K), \\ &\textit{with rate } n_a(D(a) + C(a,a)n_A/K + C(a,A)n_a/K) \textit{ from } (n_A/K, n_a/K) \\ &\qquad \textit{to } (n_A/K, (n_a-1)/K). \end{aligned} \tag{1.26}$$

Results for the limit of large carrying capacities

Define

$$\bar{n}_\alpha = \frac{b(\alpha) - D(\alpha)}{C(\alpha, \alpha)}. \tag{1.27}$$

We now state a Proposition from [13] which describes the behavior of the process defined in (1.25) in the limit of $K \to \infty$. It can be proven by applying results from Chapter 11.1 and 11.2 in [18].

Proposition 1.35 (Proposition 2(b), [13]). *Recall* (1.22) *and* (1.25) *and assume* $\nu_0^{(K)} = N_A^{(K)}(0)\delta_A + N_a^{(K)}(0)\delta_a$. *Then, for any* $t \geqslant 0$, $\nu_t^{(K)} = N_A^{(K)}(t)\delta_A + N_a^{(K)}(t)\delta_a$ *and* $N_A^{(K)}$ *from* (1.25) *has the law* $\mathbf{Q}^{(K)}$ *as defined in* (1.24). *Assume that*

$$N_A^{(K)}(0) \to n_A(0), \quad \textit{and} \quad N_a^{(K)}(0) \to n_a(0) \quad \textit{in probability for} \quad K \to \infty. \tag{1.28}$$

Then, for any $T < \infty$, $(N_A^{(K)}, N_a^{(K)})$ *converges for* $K \to \infty$ *in probability on* $[0, T]$ *in the uniform norm to the deterministic solution* (n_A, n_a) *of the system*

$$\begin{aligned} \dot{n}_A &= \big[b(A) - D(A) - C(A, A)n_A - C(A, a)n_a\big]n_A, \\ \dot{n}_a &= \big[b(a) - D(a) - C(a, a)n_a - C(a, A)n_A\big]n_a \\ &\textit{and initial condition } (n_A(0), n_a(0)). \end{aligned} \tag{1.29}$$

The system (1.29) *has at least three steady states* $(0, 0)$ *(unstable), and* $(\bar{n}_A, 0)$, $(0, \bar{n}_a)$ *with* $\bar{n}_\alpha$ *from* (1.27).

System (1.29) is composed of two logistic differential equations, known as the competitive Lotka-Volterra equations for two species. The proposition tells us that, in contrast to the limit processes from Section 1.2, the scaled birth and death Markov process converges (under certain conditions) to a deterministic process if we consider the limit $K \to \infty$.
It is however not sufficient to only study the deterministic process in order to gain insight in the structure of a population which experiences the intrusion of a mutant. As Barton already pointed out in [2], random effects have a non-negligible influence as the ancestral lines strongly interfere with each other when the frequency of the mutant population is small. This is reflected in the condition (1.28) which says that the convergence to the deterministic system only holds for sufficiently large starting configurations of the number of a- and A-individuals. Champagnat showed in [13] that, in the case where one sufficiently advantageous mutant a enters a monomorphic A-population, we can divide the invasion process in three phases with different dynamics considering the behavior of the populations of distinct types: an initial phase in which the fraction of a-individuals does not exceed a fixed value $\varepsilon > 0$ and where

the dynamics of the wild-type population is nearly undisturbed by the invading type. A second phase where both types account for a non-negligible percentage of the population and where thanks to the result from Proposition 1.35 the evolution of the population can be well approximated by the deterministic competitive Lotka-Volterra system (1.29). And finally a third phase where the roles of the types are interchanged and the wild-type population is near extinction such that we can no longer apply the above proposition. The durations of the first and third phases of the selective sweep are of order $\log K$ whereas the second phase only lasts an amount of time of order 1. We will return to the ideas developed throughout this section in Chapter 3.

We will now interpret the expression *sufficiently advantageous* in terms of the model parameters. We say that a population is in its *genetic equilibrium*, if the allele frequencies do not change from one generation to the next. Motivated by the statement of the above Proposition 1.35, $\bar{n}_\alpha$ is called the equilibrium density of a monomorphic type α-population. Define the so-called *invasion fitness* of a mutant individual of type α' who enters a monomorphic α-population at equilibrium as

$$S(\alpha', \alpha) = b(\alpha') - D(\alpha') - C(\alpha', \alpha)\bar{n}_\alpha. \tag{1.30}$$

The name of this quantity is explained by the following proposition which states a criterion as to which of the equilibrium states is taken in the limit.

Proposition 1.36 (Proposition 3,[13])**.** *Recall the definition of the invasion fitness* $S(\alpha', \alpha)$ *from* (1.30)*. If the rates from Definition 1.33 fulfill*

$$S(a, A) < 0, \tag{1.31}$$

then $(\bar{n}_A, 0)$ *is a stable steady state of* (1.29)*. If the rates are such that*

$$S(a, A) > 0 \quad \textit{and} \quad S(A, a) < 0, \tag{1.32}$$

then $(\bar{n}_A, 0)$ *is an unstable steady state,* $(0, \bar{n}_a)$ *is a stable steady state, and any solution to* (1.29) *with initial state in* $\mathbb{R}^2_{>0}$ *converges to* $(0, \bar{n}_a)$ *for* $t \to \infty$*.*

The sign of the invasion fitness therefore is crucial for the limit state of a population which experiences the intrusion of a new mutant allele. Further, the proposition tell us that even if we considered a starting configuration for a population with only one mutant individual whose allele a is sufficiently more successful compared to the wild-type A, that is, the rates are such that (1.32) holds true, then Proposition 1.36 says that in the limit $K \to \infty$ the mutant allele will fix in the population and fully replace the wild-type. We will come back to the here presented results in Chapter 3.

CHAPTER 2

The partition of a sample at the end of a selective sweep

In this chapter, we study the influence of an advantageous gene mutation on the genealogy of neighboring loci in a population model with recombination. Precisely, we consider a so-called hard selective sweep where at some time in the past an advantageous mutation B appears in one individual of the otherwise monomorphic diploid population with wild-type b. Conditioning on the fixation of this new allele means conditioning on the completion of a so-called selective sweep after which each individual of the population has type B at the selected locus (SL) and thus, with respect to that locus, all ancestral lineages will go back to the first mutant. In contrast, the genealogy for neighboring loci may differ due to recombination. We here approximate the distribution of the ancestral partition of a sample after a selective sweep from the point of view of two partially linked neutral loci in order to gain insight in the haplotype structure after a selective sweep. By *ancestral partition* we mean that we can distribute the neutral loci of all individuals into equivalence classes defining whether their alleles share a common ancestor at the beginning of the sweep or not. In addition we will mark that equivalence class which contains loci who are descendants from the first, hence we consider marked partitions.

Our results generalize the first approximation of Schweinsberg and Durrett in [36] for one neutral locus partially linked to a selected locus. There the evolution of the population was modeled by a generalized Moran model with selection and recombination, a combination of the model from Example 1.22 and Model 1 from Example 1.32. In the main theorem in [36] a simple approximation for the ancestral partition with respect to the neutral loci was obtain:

Theorem 2.1 (Theorem 1.1, [36]). *Fix a sample size $n \in \mathbb{N}$. Let $\alpha = r\log(2N)/s$, where $r \leqslant C/\log(N)$ is the recombination probability in one reproduction event and the selection coefficient $s \in (0,1)$ is independent of N. Let $p = \mathrm{e}^{-\alpha}$ and let Θ denote the true partition of the sample. Then there exists a positive constant C such that for all marked partitions π of the set $[n]$ we have*

$$\left|\mathbb{P}(\Theta = \pi) - Q_p(\pi)\right| \leqslant C/\log(N).$$

Here, Q_p is the distribution of a so-called p-partition on $[n]$ which is obtained as follows: mark

every element in the set $[n]$ independently with probability p. All marked elements are in the marked block, every non-marked element forms its own equivalence class (where it is the only element).

We will show that we can obtain a similar, but more complex structure in the case that each individual has not one but two neutral loci linked to the locus under selection. In the next section we will give a rigorous description of our three-locus version of the generalized Moran model. Throughout this chapter, we will in detail consider the case where the two neutral loci N1 and N2 are situated to the right of the SL,

$$\mathrm{SL} - \mathrm{N1} - \mathrm{N2} \tag{G1}$$

and describe the model, results and proofs with respect to this alignment. Based on this we then also briefly discuss the partition in case of the other possible geometric alignment (assuming the exchangeability of the two neutral loci):

$$\mathrm{N1} - \mathrm{SL} - \mathrm{N2}. \tag{G2}$$

Whereas many results can be easily extended to the two- or multiple locus case, the relationship between the neutral loci of one individual has to be studied in detail as they are highly dependent in the sense that in the case of (G1), a recombination between SL and N1 also has an impact on the ancestry of N2.
In the case of two neutral loci next to a locus under selection the partition of a sample at the end of a selective sweep was already studied in [31] and later in [39]. However, different assumptions on the model were imposed and the results are based on the diffusion approximation rather than the precise discrete model. We will discuss in detail the impact of the different approaches on the possible ancestral relationships in a sample in Section 2.3.2 and emphasize the benefits of our result. The remainder of this chapter is organized as follows: we give a rigorous definition of our model in Section 2.1 and subsequently present the main results in Section 2.2. In Section 2.3 we state a number of results which address the main issues throughout the proof of the approximation result, followed by a brief discussion of the most interesting aspects of all obtained results in 2.3.2, as well as the actual proof of the result. The remaining part of this chapter is then devoted to the proofs of the auxiliary propositions.

2.1 Model and Notation

Recall the Moran model which was introduced in Example 1.3 and again considered in Examples 1.22 and 1.32. We here consider a multi-locus version of the Moran model in continuous

time with selection and recombination which was introduced in [36]. As a simplification we model the evolution of N diploid individuals by $2N$ haploid individuals (each can be thought of as one chromosome) where in each step each individual gets assigned a label $i \in [2N] := \{1, 2, \ldots, 2N\}$ and an allelic type at the selected locus, B or b. A reproduction event happens with rate $2N$ and consists in drawing three individuals at random with replacement where one of them will be replaced by the offspring of the other two in the next step. In case of no recombination the offspring is an identical copy of the first parent.
The selective advantage of type B individuals is modeled similar as in Example 1.32: if the parent is of type b and the killed individual is of type B we reject the reproduction event with probability s, where s is independent of the population size N. The parameter s is the selection coefficient and models the higher fitness of type B individuals.
In case of a recombination event the offspring copies one or two neutral alleles from the second parent. A recombination happens independently of all other events during reproduction. In the alignment (G1), such a break in the genome happens with probability r_1 between SL and N1 and r_2 between N1 and N2. Whenever a recombination happens, the locus or loci which are separated from the selected locus by this break are copied from the second parent. In case of two recombinations, only the middle part, that is, N1, does not descend from the first parent but is inherited from the second parent.
For the second geometric alignment, (G2), a recombination happens with probability r_j between SL and Nj, $j = 1, 2$. Again, the allele at the neutral locus which separates from the SL due to the recombination is copied from the second parent, whereas the other neutral locus and the selected locus get copies from the first parent's alleles. If two recombinations happen in one reproduction event, both neutral loci are copied from the second parent.

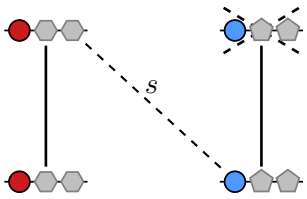


Figure 2.1.1: A declined replacement of a B-individual (red) by a b-individual (blue).

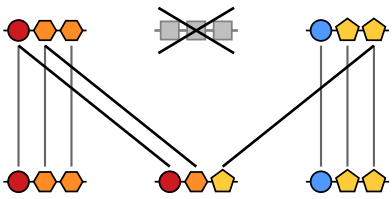


Figure 2.1.2: A recombination which connects N2 to a b-individual, while N1 and SL are copied from the first parent.

If the type at the selected locus to which a neutral locus is linked changes due to a recombination event from B to b when looking one step in the past, we say that the locus *recombined* or *migrated* into the b-population (and analogously for a change from b to B), or that the locus changed its background. If a locus descended from an individual of type b from the beginning of the sweep, it is also said that this locus *escaped* the sweep.
For both geometric alignments we assume that the recombination probabilities are bounded

by a constant over $\log(N)$ which is inspired by the order $\mathcal{O}(\log(N))$ of the expected time duration (with exponentially distributed times between any two events) of the sweep (see for example Theorem 6.3 in [14]). Both assumptions on the parameters are essential for our results and are summarized as follows:

Assumption 2.2. *We assume that $s \in (0,1)$, a constant independent of N, and further*

$$r_j \leqslant \frac{C_j}{\log(2N)}, \quad j = 1,2, \quad \textit{for some constants } C_1, C_2.$$

For simplification, we study the propagation of the B-alleles with time parameter $t \in \mathbb{N}_0$, in contrast to the continuous time Moran model. We can however easily replace the discrete time steps with independent $\mathrm{Exp}(2N)$-distributed random times in order to transfer the results to continuous time. The time-discrete jump chain of the number of B alleles present at a certain time step is denoted by $X := (X_t)_{t \geqslant 0}$, where

$$X_t := \#\{ \text{ individuals with type B in the population at time } t \ \}. \tag{2.1}$$

Define the number of steps it takes the beneficial allele to fix or to be extinct in the population as

$$\tau := \inf\{t : X_t \in \{0, 2N\}\} \tag{2.2}$$

and further let

$$\tau_m := \inf\{t : X_t \geqslant m\} \tag{2.3}$$

be the first time that there are m B-individuals in the population. Similarly, denote the time of the last visit to m of the walk X by

$$\tau_m^* := \sup\{t : X_t \leqslant m\}. \tag{2.4}$$

In the following, we will always condition on the event of a selective sweep:

Definition 2.3. *The measure $\mathbb{P}$ is the probability measure of the above Moran model conditioned on $X_0 = 1$ and $X_\tau = 2N$ (the same holds for $\mathbb{E}$). The unconditioned measure will be denoted by $\mathbb{P}'$.*

We will mostly consider the backwards-in-time process: when stating that some event E happened *before* an event F we mean that for t_E the time E happened and t_F the time F happened, we have $t_E > t_F$.
We use the same evolution describing random variables as in [36] except that we additionally need an indicator for a possible second recombination event:

- $I_{t,1} \sim \mathcal{U}([2N])$ is the label of the individual which will be replaced at time t
- $I_{t,2} \sim \mathcal{U}([2N])$ is the label of the parent of the new individual at time t
- $I_{t,3} \sim \mathcal{U}([2N])$ is the label of the other parent of the new individual at time t
- $I_{t,4} \sim$ Bernoulli(s) is Bernoulli-s distributed and determines whether a replacement of a B-individual by a b-individual is rejected ($I_{t,4} = 1$) or not ($I_{t,4} = 0$)
- $I_{t,5} \sim$ Bernoulli(r_1) indicates if a recombination between SL and N1 takes place (and the event $I_{t,5} = 1$ implies a recombination)
- $I_{t,6} \sim$ Bernoulli(r_2) indicates if a recombination between N1 and N2 takes place (again $I_{t,6} = 1$ implies a recombination)

Note that the $I_{t,j}$ are all independent of each other and that $I_{t,1}, I_{t,2}, I_{t,3}$ refer to the labels of individuals from the previous time unit $t-1$.
In the end, we want to describe the ancestral relationships of all sampled neutral loci and denote with (i,j) the j-th neutral locus from the i-th individual. Analogously to [36] we introduce variables which indicate ancestry and type of the allele at a neutral locus at a certain time during the sweep:

- $A_t^u(i,j) \in [2N]$ is the label of the individual living at time u from which the j-th locus of the i-th individual at time t originates, $0 \leqslant u \leqslant t-1$,
- $B_t(i) \in \{0,1\}$ is the allelic type at the selected locus of the i-th individual at time t: a 1 stands for type B at the SL, a 0 for b.

With this notation, $B_t(A_\tau^t(i,j))$ specifies the type at the SL of that individual living at time t from which the locus (i,j), sampled at time τ, the end of the sweep, originates.
We can now define a process in discrete time $M = (M_t)_{t \geqslant 0}$, where each array

$$M_t \in \left([2N]^t \times \{0,1\}\right)^{4N}$$

contains the information about the ancestry up to time t of both neutral loci of all $2N$ individuals (where time is running forward, starting with $t = 0$) and the type at the SL to which the neutral loci are connected to at that time:

$$\begin{aligned} M_t &= (M_t(1,1), M_t(1,2), \ldots, M_t(2N,1), M_t(2N,2)), \text{ with} \\ M_t(i,j) &= (A_t^0(i,j), \ldots, A_t^{t-1}(i,j), B_t(i)), \ t > 0, \text{ and} \\ M_0(i,j) &= (B_0(i)), \ i \in [2N], \ j \in \{1,2\}. \end{aligned}$$

To illustrate this notation, we list the possible event types in one step of the population model and give the adjustments for the process M. Note that across all the possible events, the construction of M_t from M_{t-1} only differs in the coordinates which relate to the new offspring replacing the killed individual $I_{t,1}$. That is, independent of the event which happens, we set for all $i \neq I_{t,1}$:

$$\begin{aligned} &B_t(i) := B_{t-1}(i), \quad A_t^{t-1}(i,j) := i, \\ &A_t^u(i,j) := A_{t-1}^u(i,j), \quad 0 \leqslant u \leqslant t-2,\ j = 1,2. \end{aligned} \tag{2.5}$$

The entries describing the ancestry of the offspring are constructed as follows:

1. a reproduction event with no recombination happens: $I_{t,5} = 0 = I_{t,6}$, $I_{t,1} \neq I_{t,2}$ and the event is not rejected, i.e. either $I_{t,4} = 0$ or $(B_{t-1}(I_{t,1}), B_{t-1}(I_{t,2})) \neq (1,0)$.

$$\begin{aligned} &B_t(I_{t,1}) := B_{t-1}(I_{t,2}), \quad A_t^{t-1}(I_{t,1},j) := I_{t,2}, \\ &A_t^u(I_{t,1},j) := A_{t-1}^u(I_{t,2},j), \quad 0 \leqslant u \leqslant t-2,\ j = 1,2. \end{aligned}$$

2. a proposed replacement is rejected: $I_{t,4} = 1$ and $(B_{t-1}(I_{t,1}), B_{t-1}(I_{t,2})) = (1,0)$. Then we adjust M as stated in (2.5), for all $i \in [2N]$.

3. one recombination happens between SL and N1: $I_{t,5} = 1$, $I_{t,6} = 0$.

$$\begin{aligned} &B_t(I_{t,1}) := B_{t-1}(I_{t,2}), \quad A_t^{t-1}(I_{t,1},j) := (I_{t,3},j), \\ &A_t^u(I_{t,1},j) := A_{t-1}^u(I_{t,3},j), \quad 0 \leqslant u \leqslant t-2,\ j = 1,2. \end{aligned}$$

4. one recombination happens between N1 and N2: $I_{t,5} = 0$, $I_{t,6} = 1$.

$$\begin{aligned} &B_t(I_{t,1}) := B_{t-1}(I_{t,2}), \\ &A_t^{t-1}(I_{t,1},1) := (I_{t,2},1), \quad A_t^u(I_{t,1},1) := A_{t-1}^u(I_{t,2},1), \quad 0 \leqslant u \leqslant t-2, \\ &A_t^{t-1}(I_{t,1},2) := (I_{t,3},2), \quad A_t^u(I_{t,1},2) := A_{t-1}^u(I_{t,3},2), \quad 0 \leqslant u \leqslant t-2. \end{aligned}$$

5. two recombinations happen: $I_{t,5} = I_{t,6} = 1$. This means that the allele at N2 is again inherited from the first parent as in this case we have a double crossover of the chromosomes.

$$\begin{aligned} &B_t(I_{t,1}) := B_{t-1}(I_{t,2}), \\ &A_t^{t-1}(I_{t,1},1) := (I_{t,3},1), \quad A_t^u(I_{t,1},1) := A_{t-1}^u(I_{t,3},1), \quad \text{for } u = 0,\dots,t-2, \\ &A_t^{t-1}(I_{t,1},2) := (I_{t,2},2), \quad A_t^u(I_{t,1},2) := A_{t-1}^u(I_{t,2},2), \quad \text{for } u = 0,\dots,t-2. \end{aligned}$$

As an example, the event described in 4. can be illustrated as follows:

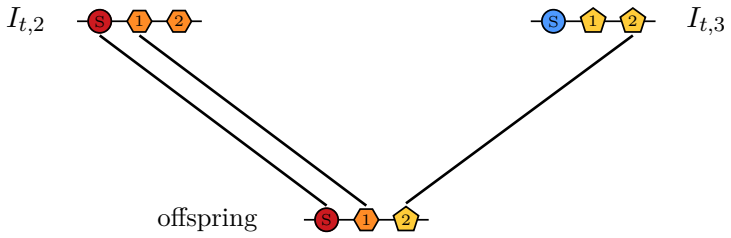


Figure 2.1.3: Event 4, a recombination between N1 and N2 leads to a new composition of alleles in the offspring.

While the offspring copies the type at the SL and the allele at N1 from the parent, $I_{t,2}$, the ancestral line of the allele at N2 now is identical to the line of the second neutral locus of the second parent, $I_{t,3}$.

2.2 The distribution of the partition of a sample

In this section we will state the main results of this thesis. In general, our aim is to approximate the partition Θ describing the ancestral relationships of the set of all sampled neutral loci,

$$[n;1,2] := \{(i,1),(i,2),\ i \in [n] := \{1,\dots,n\}\}. \tag{2.6}$$

Definition 2.4. *With our notation, the partition* $\Theta = \Theta([n;1,2])$ *is defined through*

$$(i,j_i) \sim_\Theta (m,j_m) \quad :\Leftrightarrow \quad A^0_\tau(i,j_i) = A^0_\tau(m,j_m), \quad \textit{for} \ \ i,m \in [n],\ j_i,j_m \in \{1,2\}.$$

That is, two neutral loci (i,j_i) *and* (m,j_m) *are in the same block (equivalence class) of that partition if and only if they descend from the same individual at the beginning of the sweep. As we want to distinguish the loci which are descendants of the first mutant we mark the block* $\{(i,j_i) \mid B_0(A^0_\tau(i,j_i)) = 1\}$ *by a* $*$.

Consider Figure 2.2.1 as an exemplary partition of a group of six individuals sampled at the end of a selective sweep.
A block of the partition which contains only one element (one locus $\{(i,j)\}$) is called a singleton. We call a block which consists of exactly two elements a double-singleton. The two loci of the double-singletons which appear in the below stated result will necessarily come from the same individual.
Before stating the main result we introduce the notion of a marked $\bar{p}$-partition (inspired by

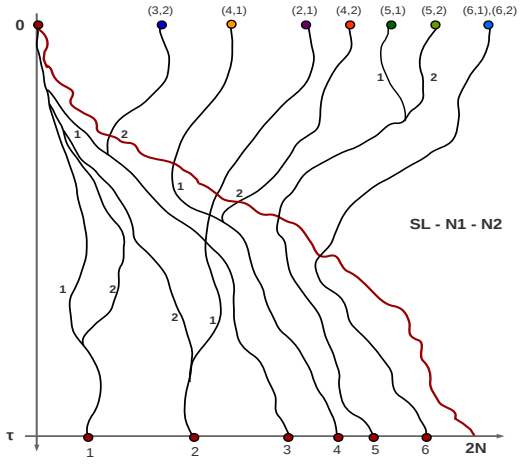


Figure 2.2.1: Partition for 6 individuals sampled at the end of the selective sweep: $\{\{(1,1),(1,2),(2,2),(3,1)\}^*, \{(2,1)\}, \{(3,2)\}, \{(4,1)\}, \{(4,2)\}, \{(5,1)\}, \{(5,2)\}, \{(6,1),(6,2)\}\}$

the definition of a p-partition in [36]) for a d-dimensional vector $\bar{p} \in [0,1]^d$, for some finite d. In this work, we are interested in the case $d = 5$.

Definition 2.5. *Let* $\bar{q} = (q_0, q_1, \ldots, q_4)$, *with* $0 \leqslant q_k \leqslant 1$ *for all* $k = 0, \ldots, 4$ *and* $|\bar{q}| = \sum_{k=0}^{4} q_k = 1$. *Let* $\xi_1, \ldots, \xi_n$ *be independent and identically distributed random variables with*

$$\mathbb{P}(\xi_i = k) = q_k, \quad \textit{for} \quad k = 0, \ldots, 4, \ \forall i \in [n].$$

We call a marked $\bar{q}$*-partition of* $[n; 1, 2]$ *the random partition where*

$(i,1)$ *and* $(i,2)$ *are both in the marked block if* $\xi_i = 0$,
$(i,1)$ *is in the marked block whereas* $(i,2)$ *is a singleton if* $\xi_i = 1$,
$(i,2)$ *is in the marked block whereas* $(i,1)$ *is a singleton if* $\xi_i = 2$,
$\{(i,1),(i,2)\}$ *is an unmarked double-singleton if* $\xi_i = 3$ *and*
$\{(i,1)\}$ *and* $\{(i,2)\}$ *are both unmarked singletons if* $\xi_i = 4$.

We denote with $Q_{\bar{q}}$ *the distribution of a marked* $\bar{q}$*-partition of* $[n; 1, 2]$.

Note that in a $\bar{q}$-partition the only unmarked blocks are singletons $\{(i, j_i)\}$ or double-singletons $\{(i,1),(i,2)\}$ where the two elements are the two neutral loci of one individual $i \in [n]$. There are no unmarked blocks with loci from different individuals.
Formally, the main result of this paper then reads as follows:

Theorem 2.6. *Fix $n \in \mathbb{N}$ and suppose that Assumption 2.2 holds true. Let*

$$p_1 = 1 - \exp\Big(-\frac{r_1}{s}\log(N)\Big), \quad p_2 = 1 - \exp\Big(-\frac{r_2}{s}\log(N)\Big). \tag{2.7}$$

Define $\bar{q}_{\mathrm{G1}} = (q_0, q_1, \ldots, q_4)$ with

$$\begin{aligned} q_0 &= (1-p_1)(1-p_2)(1-p_1p_2), \quad q_1 = (1-p_1)p_2[1+p_1(1-p_2)], \\ q_2 &= p_1p_2(1-p_1)(1-p_2), \quad q_3 = (1-p_2)^2p_1, \\ q_4 &= p_1p_2[1+p_1(1-p_2)]. \end{aligned} \tag{2.8}$$

For the alignment (G1) *there exists a positive constant C such that*

$$|\mathbb{P}(\Theta = \pi) - Q_{\bar{q}_{\mathrm{G1}}}(\pi)| \leqslant C/(\log N),$$

for all $\pi \in \mathcal{P}^*_{[n;1,2]}$, the set of marked partitions of $[n;1,2]$, and the true partition Θ as defined in Definition 2.4.

The result can be interpreted as follows: up to an error of order $1/\log(N)$, the typical partition of a sample of individuals taken at the end of a selective sweep has the same structure as a $\bar{q}_{\mathrm{G1}}$-partition. In particular, this implies that the ancestral relationships of the two neutral loci of a single individual arise nearly independently of all other sampled loci. Furthermore, the only block in the partition which contains loci of different individuals will be the marked block. Hence, no neutral loci will find a common ancestor outside the mutant's family. For a better perception of this result, we have displayed a possible structure of a sample in Figure 2.2.2. In Section 2.3.2, we will discuss in detail the partition structure and also the order of approximation obtained in this theorem. Further, we give an intuition on the structure of the probabilities from $\bar{q}_{\mathrm{G1}}$.

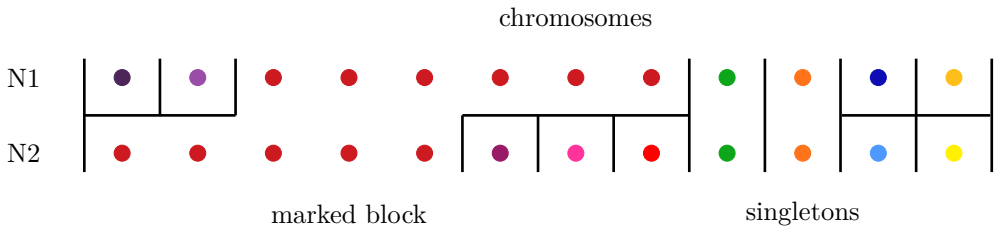


Figure 2.2.2: An exemplary partition of a sample taken at the end of a selective sweep.

For the geometric alignment (G2) we get a simpler structure of the sample.

Theorem 2.7. *Fix $n \in \mathbb{N}$ and suppose that Assumption 2.2 holds true. Let p_1, p_2 be as in*

(2.18) *in Theorem 2.6 and define* $\bar{q}_{\mathrm{G2}} = (q_0, q_1, \ldots, q_4)$ *with*

$$\begin{aligned} q_0 &= (1-p_1)(1-p_2), \quad q_1 = (1-p_1)p_2, \\ q_2 &= p_1(1-p_2), \quad q_3 = 0, \quad q_4 = p_1 p_2. \end{aligned} \tag{2.9}$$

In the case that we consider the alignment (G2) *there exists a positive constant* C *such that*

$$|\mathbb{P}(\Theta = \pi) - Q_{\bar{q}_{\mathrm{G2}}}(\pi)| \leqslant C/(\log N),$$

for all $\pi \in \mathcal{P}^*_{[n;1,2]}$, *the set of marked partitions of* $[n; 1, 2]$.

Note that this Theorem shows that (up to the approximation order) an individual's two neutral loci aligned as in (G2) behave independently in the sense that whether one of them escapes the sweep or not has no influence on the ancestry of the other locus. Furthermore, there are no double-singletons $\{(i,1),(i,2)\}$ as the probability that both loci migrate at the same time is negligible, in the same way as the probability that two escaped lines find a common ancestor in the b-population.

Unless stated otherwise, we will from now on only consider the alignment (G1) and only briefly go back to (G2) when giving the proof of the above Theorem in Section 2.3.3.

2.3 Preparatory results and proof of the main theorems

The main theorem comprises three kinds of assertions: first, there are no other than the five different block types for the two neutral loci of all sampled individuals. For this, we have to show that events leading to unmarked blocks with loci from different individuals have probabilities of order less than or equal to $1/\log(N)$.
The second statement of the Theorem is that the claimed marking probabilities are the correct probabilities for all respective block types. For this, we need to consider all sequences of events which would lead to a specific ancestral relationship and determine their probability. Fortunately, we can exclude many events beforehand by applying results from Schweinsberg and Durrett, [36]. Lastly, we need to show that we can label the individuals independently from each other, that is, that the partition approximately has the structure of a marked $\bar{q}_{\mathrm{G1}}$-partition.

2.3.1 Notation and auxiliary results

We introduce some characteristic times which help us to describe important events during the sweep:

Definition 2.8. *Define the following times of events:*

(i) Denote the first time the j-th locus of the i-th individual migrates into the b-population with

$$R(i,j) := \sup\{t \geqslant 0 : B_t(A^t_\tau(i,j)) = 0\}. \tag{2.10}$$

(ii) The first time that individuals i and m find a common ancestor with respect to their j_i and j_m-locus, respectively, is defined as

$$G^{j_i,j_m}(i,m) := \sup\{t : A^t_\tau(i,j_i) = A^t_\tau(m,j_m)\}. \tag{2.11}$$

Whenever we address the same neutral locus j in the individuals i and m, we write $G^j(i,m)$.

(iii) Define the first time that the genealogies of the two neutral loci of lineage i go back to different B-individuals

$$R^{rec}_B(i) := \sup\{t \geqslant 0 : A^t_\tau(i,1) \neq A^t_\tau(i,2) \text{ and } B_t(A^t_\tau(i,1)) = 1 = B_t(A^t_\tau(i,2))\}. \tag{2.12}$$

(iv) Similarly, the first time that the two neutral loci of lineage i split within the b-population and go back to different b-individuals is defined through

$$\begin{aligned} R^{rec}_b(i) := \sup\{t \geqslant 0 :\ & A^t_\tau(i,1) \neq A^t_\tau(i,2) \text{ and } B_t(A^t_\tau(i,1)) = 0 = B_t(A^t_\tau(i,2)), \\ & A^{t+1}_\tau(i,1) = A^{t+1}_\tau(i,2)\}. \end{aligned} \tag{2.13}$$

(v) Last, we define the first time that a double-recombination in one reproduction event leads to the escape of the first neutral locus into the b-population while the second neutral locus stays in the B-population:

$$\begin{aligned} R^{2rec}_{bB}(i) := \sup\{t \geqslant 0 :\ & B_t(A^t_\tau(i,1)) = 0, B_t(A^t_\tau(i,2)) = 1, \text{ and} \\ & A^{t+1}_\tau(i,1) = A^{t+1}_\tau(i,2), B_{t+1}(A^{t+1}_\tau(i,1)) = 1\}. \end{aligned} \tag{2.14}$$

We here use the usual convention that $\sup \varnothing = -\infty$.

Furthermore, we will use the following notation:

$$r(j) := r_1 + \mathbf{1}_{\{j=2\}} \cdot (r_2 - r_1 r_2), \quad r^*(j) := r_1 + \mathbf{1}_{\{j=2\}} \cdot (r_2 - 2 \cdot r_1 r_2), \quad j = 1,2. \tag{2.15}$$

Here, $r(2)$ is the probability to have a recombination in front of either neutral locus, whereas $r^*(2)$ gives the probability to have exactly one recombination in front of the first or in front of the second neutral locus. Note that $r(1) = r^*(1) = r_1$.

First of all we state the proposition which tells us that a split within the B-population of some sampled lineage i will most likely occur before any neutral locus of some other individual m recombines into the b-population.

Proposition 2.9. *For all $i, m \in [n], i \neq m$ and all $j = 1, 2$, the probability to see a locus migrating into the b-population before (seen backwards in time) the two neutral loci of another individual split within the B-population, can be bounded as follows:*

$$\mathbb{P}(R(m,j) > R_B^{rec}(i) \geqslant 0) \leqslant \frac{C}{\log(N)}. \tag{2.16}$$

More precisely, it holds that

$$\begin{aligned} &\mathbb{P}(R(m,j) \geqslant \tau_N) \leqslant \frac{C}{\log(N)}, \text{ and} \\ &\mathbb{P}(0 \leqslant R_B^{rec}(i) < \tau_N) \leqslant \frac{C}{\log(N)}. \end{aligned} \tag{2.17}$$

This proposition is a key result as it allows us to consider the event of a separation within the B-population separately of events which influence the escape of any locus into the b-population. All calculations which deal with the escape of neutral loci into the b-population when following the ancestral lines backwards in time will then only assume that either $R_B^{\text{rec}}(i) \geqslant \tau_N$ or $R_B^{\text{rec}}(i) = -\infty$ and will not distinguish between the concrete event time of $R_B^{\text{rec}}(i)$ in the former case. Precisely, we will consider the time intervals $[0, \tau_N)$ and $[\tau_N, \tau]$ separately and hence only have to consider those times where the Markov chain X is less than or equal to N in all calculations concerning migration events. The separation of the two time intervals will again be of great use when we prove the multinomial character of the partition of a sample.

The following propositions are concerned with the structure of possible partition elements, that is, they bound the probabilities of events during the sweep which would lead to other than the claimed block types. With the first proposition we bound the probability that a double recombination in one reproduction event leads to the escape of the first neutral locus.

Proposition 2.10. *The probability to witness a double recombination during the sweep can be bounded as follows:*

$$\mathbb{P}\Big(R_{bB}^{2rec}(i) \geqslant 0\Big) \leqslant C/\log(N),$$

for all $i \in [n]$.

The Proposition below states that we will indeed not see a back recombination into the the B-population. Hence, once a lineage or a locus migrates into the b-population it escapes the sweep. Here, the proof is the same as in [36] and will not be given here.

Proposition 2.11 (cf. Prop. 2.1, [36])**.**

$$\mathbb{P}\Big(B_t(A^t_\tau(i,j)) = 1 \text{ for some } t : t \leqslant R(i,j)\Big) \leqslant C(j)/(\log(N))^2.$$

where $C(j)$ *fulfills* $r(j)^2/s^2 \leqslant C(j)/(\log(N))^2$.

This proposition combined with the statement from Proposition 2.10 implies that the only way to have an unmarked singleton $\{(i,1)\}$ while $(i,2)$ is in the marked block, is a split of the two neutral loci within the B-population prior (seen backwards) to a migration of $(i,1)$ into the b-population and combined with the event that the ancestral line of $(i,2)$ never escapes the sweep.
Further, note that for $i = m$, the probability of the event $\{R(m,j) > R^{\text{rec}}_B(i) \geqslant 0\}$ can be bounded by $c/(\log(N))^2$ by the definition of the time $R^{\text{rec}}_B(i)$ and the above Proposition 2.11.
The following result bounds the probability that a coalescence which involves an individual of type b can be witnessed during the sweep.

Proposition 2.12 (cf. Prop. 2.3, [36])**.**

$$\mathbb{P}\Big(G^{j_i,j_m}(i,m) \geqslant 0 \text{ and } B_{G^{j_i}(i,m)+1}(A^{G^{j_i}(i,m)+1}_\tau(i,j_i)) = 0\Big) \leqslant C(\log N)/N.$$

The proofs for this and the following proposition can be taken from [36] and are not repeated here. We only need to adapt the recombination and coalescence probabilities to match the three-locus model (see Lemmas 2.25 and 2.26).
In this next proposition, we bound the probability of the event that a coalescence precedes a migration into the b-population which then would lead to a bigger family of a wild-type b-individual.

Proposition 2.13 (cf. Prop. 2.4, [36])**.**

$$\mathbb{P}(0 \leqslant R(i,j_i) \leqslant G^{j_i,j_m}(i,m)) \leqslant C/(\log N)\ .$$

Before we formulate the auxiliary propositions which concern the success probabilities from $\bar{q}_{\text{G1}}$ in (2.8), we define two quantities which appear in the proofs and also the statements of those propositions:

$$q^{(j)} = 1 - \exp\Big(-\frac{r_j}{s}\sum_{k=1}^{N}\frac{1}{k}\Big), \quad j = 1,2. \tag{2.18}$$

In the main result, the $q^{(j)}$ are approximated by the probabilities p_1 and p_2 from (2.7) (as we can approximate the harmonic sum by the logarithm).
Let us now consider the separation of the two neutral loci of one individual within the B-population. Here we know from Proposition 2.9 that we only need to consider times $t \in [\tau_N, \tau]$.

Proposition 2.14. *Let*

$$q_J^{BB} := 1 - \exp\Big(- \frac{r_2}{s} \sum_{k=1}^{2N-J} \frac{1}{k} \Big), \quad \textit{and} \;\; q^{BB} := q_1^{BB}. \tag{2.19}$$

With τ_J as defined in (2.3) *we have for any $J = 1, 2, \dots, 2N-1$:*

$$\mathbb{P}(R_B^{rec}(i) \geqslant \tau_J) = q_J^{BB} + \mathcal{O}\Big(\frac{1}{\log(N)}\Big).$$

In particular, with p_2 from (2.7) *in Theorem 2.6, we get*

$$\begin{aligned} &q^{BB} = q_N^{BB} + \mathcal{O}(1/\log(N)) \quad \textit{and} \\ &q^{BB} = p_2 + \mathcal{O}(1/\log(N)), \quad \textit{that is,} \quad \mathbb{P}(R_B^{rec}(i) \geqslant \tau_N) = p_2 + \mathcal{O}(1/\log(N)). \end{aligned}$$

We here introduced the notation q^{BB} in order to better understand which effect adds in which way to the marking probabilities. In the end, we will of course replace q^{BB} with p_2 thanks to the last equation in the above Proposition 2.14.
Next, we consider the probability that a locus recombines into the b-population after a certain time τ_J for some $J \in [N]$. For $J = 1$ this corresponds to the escape probability if we can exclude a back-recombination into the B background. Note that this proposition does not give a joint statement for the two loci and thus the proof can be taken from [36] when replacing their recombination probability r with $r^*(j)$ as reasoned in the next section. Since in addition we will later recapitulate the main ideas during the proofs of other results, we refrain from giving a proof of this proposition.

Proposition 2.15 (cf. Prop. 2.2, [36])**.** *Let $q_J^{(j)} = 1 - \exp\Big(- \frac{r^*(j)}{s} \sum_{k=J+1}^{2N} \frac{1}{k} \Big)$ and recall the stopping time τ_J from* (2.3)*. Then*

$$\mathbb{P}(R(i,j) \geqslant \tau_J) = q_J^{(j)} + \mathcal{O}\Big(\frac{1}{(\log N)^2} + \frac{1}{(\log N)\sqrt{J}}\Big).$$

We denote the possible ancestral relations which the two neutral loci of one individual can take at the end of the sweep as follows:

$$\begin{aligned} &\mathcal{A}_0 = (i,1), (i,2) \text{ are both in the marked block,} \\ &\mathcal{A}_1 = (i,1) \text{ in the marked block, } (i,2) \text{ descends from a b-individual,} \\ &\mathcal{A}_2 = (i,2) \text{ in the marked block, } (i,1) \text{ originates from a b-individual,} \\ &\mathcal{A}_3 = (i,1) \text{ and } (i,2) \text{ are descended from the same b-individual,} \\ &\mathcal{A}_4 = (i,1) \text{ and } (i,2) \text{ originate from two different b-individuals.} \end{aligned} \tag{2.20}$$

Note that the formulation here is different from the block types listed in the Definition 2.5,

as we do not specify whether an escaped neutral locus is a singleton or not. With this, we can now state the probabilities for each of the $\mathcal{A}_i$:

Proposition 2.16. *Recall $q^{(1)}, q^{(2)}$ from (2.18) and q^{BB} from (2.19). We have the following marking probabilities:*

$$\begin{aligned}
\mathbb{P}(\mathcal{A}_1) &= (1-q^{(1)})[q^{(1)}q^{BB}(1-q^{(2)})+q^{(2)}] + \mathcal{O}(1/\log(N)),\\
\mathbb{P}(\mathcal{A}_2) &= q^{BB}q^{(1)}(1-q^{(1)})(1-q^{(2)}) + \mathcal{O}(1/\log(N)),\\
\mathbb{P}(\mathcal{A}_3) &= (1-q^{BB})(1-q^{(2)})q^{(1)} + \mathcal{O}(1/\log(N)),\\
\mathbb{P}(\mathcal{A}_4) &= q^{(1)}[q^{(1)}q^{BB}(1-q^{(2)})+q^{(2)}] + \mathcal{O}(1/\log(N)),
\end{aligned} \tag{2.21}$$

and $\mathbb{P}(\mathcal{A}_0) = 1 - \sum_{j=1}^{4}\mathbb{P}(\mathcal{A}_j) + \mathcal{O}(1/\log(N))$.

As the proof of this proposition is rather long we devote the whole Section 2.4.3 to it.
With the above statement we have shown that the probabilities stated in the definition of $\bar{q}_{\mathrm{G1}}$ indeed hold for one individual. What is left to show is that the probabilities for one individual still hold true when considering jointly all $n > 1$ sampled individuals. This is covered by the following Propositions which show the approximate multinomial structure of the sample. To emphasize this again: for proving the main result we need to make sure that we can mimic a sample by marking each lineage approximately independent of the others in order to determine which of its loci escaped the sweep and which not. This is analogous but more involved than the statement of Proposition 2.6 from [36] because we here have five different possible states for each individual and hence need the multinomial rather than the binomial approximation.
The idea is to make use of Proposition 2.9 and give separate statements on the one hand for the number of individuals who experience a split within the B-population at some time t with $\tau_N \leqslant t \leqslant \tau$, and on the other hand for some sample taken at time τ_N. We will show that the latter is approximately distributed as a multinomial random variable with four different categories. If we apply this multinomial statement to a sample from time τ_N which consists of $2k$ individuals, half of them with only locus N1 and half of them with only locus N2 ancestral to a sample from time τ then we can deduce probabilities for the multinomial distribution concerning those individuals whose neutral loci separated within the B-population. Combining both results for the first phase of the sweep with the result on the BB-split for the second phase will then lead to an overall multinomial statement with the correct probabilities as given in the vector $\bar{q}_{\mathrm{G1}}$. The statement of Proposition 2.9 is again crucial in order to be able to consider the time intervals separately.
Before we state the result on the binomial behavior of the number of individuals whose neutral loci split toward the end of the sweep we need a statement which guarantees that a coalescence prior to such a split is unlikely, similar as the statement from Proposition 2.13.

Proposition 2.17. *For any $i, m \in [n], j = 1, 2$ we have*

$$\mathbb{P}(0 \leqslant R_B^{rec}(i) \leqslant G^j(i,m)) \leqslant C/\log(N),$$

for some constant C.

Further, we will need the following statement which is similar to Proposition 2.5 in [36] but considers a different time interval. The proof is however identical and therefore omitted here.

Proposition 2.18 (cf. Proposition 2.5, [36])**.** *We have the following bound on the probability that two sampled individuals will coalesce with respect to any of their neutral loci after time τ_N:*

$$\mathbb{P}(G^{j_i,j_m}(i,m) \geqslant \tau_N) \leqslant C\log(N)/N.$$

Now we can turn to the number of individuals whose neutral loci separate within the B-population.

Proposition 2.19. *Define for $\tau_N \leqslant t \leqslant \tau$*

$$K_t^{BB} := \#\{i \in [n] : R_B^{rec}(i) \geqslant t\}. \tag{2.22}$$

Then for any $J \geqslant N$ it holds that

$$\left|\mathbb{P}(K_{\tau_J}^{BB} = d) - \binom{n}{d}(q_J^{BB})^d(1-q_J^{BB})^{n-d}\right| \leqslant C/\log(N),$$

for some constant C and q_J^{BB} as in (2.19)*. In particular,*

$$\left|\mathbb{P}(K_{\tau_N}^{BB} = d) - \binom{n}{d}(q^{(2)})^d(1-q^{(2)})^{n-d}\right| \leqslant C/\log(N),$$

with $q^{(2)}$ from (2.18)*.*

Note that Proposition 2.9 tells us that with high probability, $K_{\tau_N}^{BB}$ is indeed the total number of sampled individuals whose neutral loci separated within the B-population during the whole sweep.
Define the event that the two neutral loci of an individual i split within the B-population, that is,

$$\mathcal{S}_i := \{R_B^{\mathrm{rec}}(i) \geqslant \tau_N\}, \tag{2.23}$$

and similarly to the definition of the ancestral relations $\mathcal{A}_1, \ldots, \mathcal{A}_5$ from (2.20) define further

the following events for an individual i which is sampled at time t:

$$\begin{aligned}
\tilde{\mathcal{A}}_0^t(i) &= (i,1),(i,2) \text{ never leave the B-population during times 0 and } t,\\
\tilde{\mathcal{A}}_1^t(i) &= \text{all ancestors of } (i,1) \text{ are of type B, whereas } (i,2) \text{ descends from a}\\
&\quad \text{b-individual at some time } t',\ 0 \leqslant t' \leqslant t,\\
\tilde{\mathcal{A}}_2^t(i) &= \text{all ancestors of } (i,2) \text{ are of type B, whereas } (i,1) \text{ descends from a}\\
&\quad \text{b-individual at some time } t',\ 0 \leqslant t' \leqslant t,\\
\tilde{\mathcal{A}}_3^t(i) &= (i,1) \text{ and } (i,2) \text{ have a common ancestor of type b at some time } t',\\
&\quad 0 \leqslant t' \leqslant t, \text{ and do not find different ancestors within the b-population,}\\
\tilde{\mathcal{A}}_4^t(i) &= \{(i,1)\} \text{ and } \{(i,2)\} \text{ originate from two different b-individuals}\\
&\quad \text{at some time in the past } t',\ 0 \leqslant t' \leqslant t.
\end{aligned} \tag{2.24}$$

The difference to (2.20) is that we only consider times t' with $0 \leqslant t' \leqslant t$ and do not make any statement about events which may happen after that specific time t' where a $\tilde{\mathcal{A}}$ is fulfilled. We will now formulate a result for the multinomial character of a sample of n^{BB} individuals taken from the population at time τ_N. Note that we will later choose this number n^{BB} depending on the number of individuals whose neutral loci originate from two different individuals living at time τ_N, that is, n^{BB} can be interpreted as a placeholder for $n - K_{\tau_N}^{BB}$ here. The proposition however makes a statement for the ancestry of all $2 \cdot n^{BB}$ neutral loci from the sample drawn at time τ_N.

Proposition 2.20. *Define for fixed sample size $n^{BB} \in \mathbb{N}$ the random vector*

$$D := (D^0, D^1, D^2, D^3, D^4), \tag{2.25}$$

where D^j equals the number of individuals from τ_N whose neutral loci fulfill the relationship $\tilde{\mathcal{A}}_j^{\tau_N}(i)$,

$$D^j = \#\{i \in [n^{BB}] : \tilde{\mathcal{A}}_j^{\tau_N}(i)\},\ j = 0, \ldots, 4.$$

Let

$$\bar{q}_{\mathrm{G1}}^D := ((1-q^{(1)})(1-q^{(2)}),\ (1-q^{(1)})q^{(2)},\ 0,\ q^{(1)}(1-q^{(2)}),\ q^{(1)}q^{(2)}), \tag{2.26}$$

with $q^{(1)}, q^{(2)}$ as in (2.18) and further

$$\tilde{D} = (\tilde{D}^0, \tilde{D}^1, \tilde{D}^2, \tilde{D}^3, \tilde{D}^4) \sim \mathit{Mult}(n^{BB}, \bar{q}_{\mathrm{G1}}^D). \tag{2.27}$$

Then, for any $d = (d_0, d_1, d_2, d_3, d_4)$ with $d_0, \ldots, d_4 \in [n^{BB}]$ and $|d| = \sum_{j=0}^4 d_j = n^{BB}$ we

have

$$\left|\mathbb{P}(D = d) - \mathbb{P}(\tilde{D} = d)\right| \leqslant C/\log(N),$$

for some constant C.

Note that we did not indicate the dependence of D on the parameter n^{BB} in the notation. We will clarify this dependence by writing $D(n^{BB})$ whenever there is a possible ambiguity. Here, $\tilde{D}^2 = 0$ almost surely reflects the fact, that with high probability we will not see a singleton $(i, 1)$ while $(i, 2)$ is in the marked block as we only consider times between 0 and τ_N and thus all events which would cause such a relationship have only small probability.

Proposition 2.20 implies that, approximately, each individual sampled at τ_N takes one specific ancestral relation for its two neutral loci, independent of the others. This allows us to conclude a similar multinomial structure of the random vector which describes the relation of the neutral loci if they separated within the B-population at some time $t > \tau_N$. We will abuse notation here in the sense that we still use the formulation of the relationships from (2.24) although we assume that $(i, 1)$ and $(i, 2)$ reside in different individuals at the time τ_N where the sample is taken. We can make this slightly more rigorous by using the notation $\tilde{\mathcal{A}}_j^{\tau_N}(A_\tau^{\tau_N}(i, 1), A_\tau^{\tau_N}(i, 2))$ and replacing (i, j_i) by $(A_\tau^{\tau_N}(i, j_i), j_i)$ in the definition (2.24) for every $j_i = 1, 2$.

Corollary 2.21. *Define for fixed sample size* $K^{BB} \in \mathbb{N}$ *the random vector*

$$\bar{D} := (\bar{D}^0, \bar{D}^1, \bar{D}^2, \bar{D}^3, \bar{D}^4), \tag{2.28}$$

where the coordinate $\bar{D}^j$ *equals the number of individuals from* τ_N *whose neutral loci fulfill the relationship* $\tilde{\mathcal{A}}_j^{\tau_N}(A_\tau^{\tau_N}(i, 1), A_\tau^{\tau_N}(i, 2))$, $j = 0, \ldots, 4$, *assuming that all* $2K^{BB}$ *neutral loci reside in different individuals at time* τ_N. *Let*

$$\begin{aligned}\bar{q}_{\mathrm{G1}}^{\bar{D}} := ((1 - q^{(1)})^2(1 - q^{(2)}),\ (1 - q^{(1)})[q^{(1)} + q^{(2)} - q^{(1)}q^{(2)}],\ q^{(1)}(1 - q^{(1)})(1 - q^{(2)}),\\ 0,\ q^{(1)}[q^{(1)} + q^{(2)} - q^{(1)}q^{(2)}]),\end{aligned} \tag{2.29}$$

with $q^{(1)}, q^{(2)}$ *as in* (2.18) *and*

$$\tilde{\bar{D}} = (\tilde{\bar{D}}^0, \tilde{\bar{D}}^1, \tilde{\bar{D}}^2, \tilde{\bar{D}}^3, \tilde{\bar{D}}^4) \sim \mathit{Mult}(K^{BB}; \bar{q}_{\mathrm{G1}}^{\bar{D}}). \tag{2.30}$$

Then, for any $d = (d_0, d_1, d_2, d_3, d_4)$ *with* $d_0, \ldots, d_4 \in [K^{BB}]$ *and* $|d| = \sum_{j=0}^4 d_j = K^{BB}$ *we have*

$$\left|\mathbb{P}(\bar{D} = d) - \mathbb{P}(\tilde{\bar{D}} = d)\right| \leqslant C/\log(N),$$

for some constant C.

Note that $\tilde{\tilde{D}}^3 = 0$ almost surely as, with high probability, we will not see a double-singleton $\{(i,1),(i,2)\}$ if the two loci already separated within the B-population. Corollary 2.21 follows directly from the above Proposition 2.20. We will nevertheless briefly state the idea of the proof in Section 2.4.4.
The result for the whole sample which follows from Propositions 2.19 and 2.20 and the above Corollary 2.21.

Proposition 2.22. *For fixed sample size $n \in \mathbb{N}$ define the process*

$$D^* := (D^{*,0}, D^{*,1}, D^{*,2}, D^{*,3}, D^{*,4}), \tag{2.31}$$

whose coordinate $D^{*,j}$ equals the number of individuals at the end of the sweep whose loci fulfill the relationship $\tilde{\mathcal{A}}_j$, for $j = 0, \dots, 4$. Let

$$\tilde{D}^* = (\tilde{D}^{*,0}, \tilde{D}^{*,1}, \tilde{D}^{*,2}, \tilde{D}^{*,3}, \tilde{D}^{*,4}) \sim \mathrm{Mult}(n; \bar{q}_{\mathrm{G1}}), \tag{2.32}$$

with $\bar{q}_{\mathrm{G1}} = (q_0, q_1, \dots, q_4)$ as in (2.8). Then, for any $d = (d_0, d_1, d_2, d_3, d_4)$ with $d_0, \dots, d_4 \in [n]$ and $|d| = \sum_{j=0}^4 d_j = n$ we have

$$\left|\mathbb{P}(D^* = (d_0, d_1, d_2, d_3, d_4)) - \mathbb{P}(\tilde{D}^* = (d_0, d_1, d_2, d_3, d_4))\right| \leqslant C/\log(N),$$

for some constant C.

Before we apply all of these statements in order to prove the main Theorems 2.6 and 2.7, we will point out several aspects of the main results and also of the above preparatory statements.

2.3.2 Discussion of the main result

As the true partition can be approximated by a marked $\bar{q}_{\mathrm{G1}}$-partition, which comprises the multinomial structure from D^*, we know that we can mimic a sample by taking n individuals and deciding independently for each individual in which relationship the alleles at its two neutral loci were at the beginning of the sweep. Moreover, the statements of Propositions 2.19, 2.20 and Corollary 2.21 allow an easy construction of a typical sample taken after a selective sweep. Consider three marking processes as follows:

- a mark $\mathcal{M}_0$ stands for a recombination within the B-population between the neutral loci N1 and N2 at a time $t \in (\tau_N, \tau)$,
- a mark $\mathcal{M}_1$ stands for a recombination between SL and N1 and an escape of the locus N1 into the b-population,

- a mark $\mathcal{M}_2$ stands for a recombination between N1 and N2 and an escape of the locus N2 into the b-population.

First, we mark each individual with a mark $\mathcal{M}_0$ with probability p_2 from (2.7), independently of all others in the sample. Subsequently, we perform two independent rounds of marking, where the probabilities for the marks depend on the outcome of the first round. In detail: all individuals with an $\mathcal{M}_0$-mark are marked independently of each other with $\mathcal{M}_1$ with probability p_1, as in (2.7), and independent of the outcome there, with a mark $\mathcal{M}_2$ with probability $1-(1-p_1)(1-p_2) = p_1(1-p_2)+p_2$. On the other hand, all individuals which did not receive an $\mathcal{M}_0$-mark get an $\mathcal{M}_1$-mark with probability p_1, and again independent of the outcome and independent of each other, they get an $\mathcal{M}_2$-mark with probability p_2. For each individual, the vector $(m_0, m_1, m_2) \in \{0,1\}^3$ which indicates the absence (0) or presence (1) of a mark $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively, then fully determines its ancestry. As an example, the vector $(0,1,1)$ corresponds to an individual whose alleles are inherited by two different b-individuals, $(1,1,0)$ means that N2 is copied from the mutant while N1 originates from some b-individual.

The interesting aspect of this marking model is that conditional on the outcome of a first binomial experiment, we conduct two independent binomial trials in order to model the escape or captivity of each neutral locus separately. Thus, from the time τ_N on up to the beginning of the sweep, we can decide independently for each neutral locus of each individual whether it originates from the mutant or not, even in the case of the geometric alignment (G1).

This simple structure suggests that the result of the Theorems 2.6 and 2.7 can actually be transfered to the multiple locus case. The calculations for the ancestral relation of the neutral loci of one individual should be of the same nature, only with more cases to distinguish.

Let us now focus on the block types and the order of approximation of our result. Projecting our result to either the first or the second neutral locus gives again the result from Theorem 1.1, [36]. In the second approximation by Schweinsberg and Durrett which results in their Theorem 1.2, the authors show that in the case of one neutral locus, a typical partition may in addition consist of bigger families different from the mutant's family. This statement then holds true up to terms of order $1/(\log N)^2$. From the results obtained here, it is not clear if the methods which were used to obtain this second approximation will allow us to make a similar statement with the same accuracy in the case of two neutral loci in geometry (G1). This will in particular depend on whether it is possible to improve the result of Proposition 2.14, the probability of a split within the B-population, using the methods from the second part of [36].

We will now compare our results to the approximate partition derived by Pfaffelhuber and Studeny in [31] where the authors generalized the work from Etheridge, Pfaffelhuber and Wakolbinger [17] towards the two-locus case. In [31] the authors considered first the diffusion limit of the population process, resulting in (1.20), and subsequently studied large selection

coefficients $\alpha \propto sN$. Whereas this diffusion approximation adds additional uncertainty on the error made for large N in comparison to the underlying discrete model, the results presented here in Chapter 2 enable a clear interpretation of the error term. Theorem 1 in [31] says that up to an error of order $1/(\log \alpha)^2$, the partition of a sample will consist of blocks of the above type and again bigger families different from the mutant's family, as in Theorem 1.2 from [36]. This additional block type is non-negligible in their approximation as the partition of the sample results from a marked Yule tree where recombination marks may hit already coalesced lineages. The phenomenon of a separation of the two neutral loci within the B-population is modeled by generating a refined starting partition for the Yule tree construction: with a certain probability the two neutral loci of one individual split towards the very end of the sweep. This probability indeed equals our probability p_2 (q^{BB}). Furthermore, the probabilities for the different recombination marks and thus block types from Table 1 in [31] show a similar pattern as the above claimed $\bar{q}$-structure, in particular when focusing on events which happen between time 0 and τ_N. Thus our results are consistent with the ones obtained from the diffusion approximation. However, in [31] the impact of the marks distributed according to these probabilities depends on the realization of the Yule tree. This makes it difficult to derive an explicit sampling formula (which is thus omitted in [31]). As the authors point out, however, their construction still lends itself to simulations with fast algorithms, though more complicated than the above proposed marking model with marks $\mathcal{M}_0, \mathcal{M}_1$ and $\mathcal{M}_2$.

2.3.3 Proofs of the main results

The proof of the main Theorem 2.6 uses the propositions from Section 2.3.1 which deal with the three different aspects of the theorem, as described in the beginning of Section 2.3: the exclusion of improbable block types, the marking probabilities and the independent evolution of the individuals. Although the ingredients are all at hand, the formal statement of the proof is rather cumbersome. As we considered the time intervals $[0, \tau_N)$ and $[\tau_N, \tau]$ separately, we need to rigorously formulate a suitable resampling mechanism for time τ_N in order to guarantee that we can combine the results for both time intervals to a joint statement concerning the whole sweep. We will use a similar idea as Schweinsberg and Durrett in their second approximation, as stated in Section 2.3, [36].

Proof of Theorem 2.6. We proceed in four steps. First define the marked partition Θ_1 on $[n; 1, 2]$ as follows:

$$(i, j_i) \sim_{\Theta_1} (m, j_m) \quad :\Leftrightarrow \quad R(i, j_i) < \tau_N,\ R(m, j_m) < \tau_N \ \text{ and } \ A^0_\tau(i, j_i) = A^0_\tau(m, j_m);$$
$$\text{mark the block } \ \{(i, j_i) : R(i, j_i) < \tau_N \ \text{ and } \ B_0(A^0_\tau(i, j_i)) = 1\}.$$

We have $\Theta = \Theta_1$ unless there exists some $(i, j_i) \neq (m, j_m) \in [n; 1, 2]$ with either

$$\begin{aligned} &R(i, j_i) \geqslant \tau_N \quad \text{and} \quad (i, j_i) \sim_\Theta (m, j_m) \quad \text{or} \\ &R(i, j_i) \geqslant \tau_N \quad \text{and} \quad B_0(A^0_\tau(i, j_i)) = 1. \end{aligned}$$

By Proposition 2.9 these events are bounded even in the case $i = m$ such that we obtain

$$\left|\mathbb{P}(\Theta_1 = \pi) - \mathbb{P}(\Theta = \pi)\right| \leqslant C/\log(N)$$

for every partition $\pi \in \mathcal{P}^*_{[n;1,2]}$. (Note that for $i \neq m$ we have a bound of order at least $1/(\log N)^2$.) For the second step consider the population at time τ_N and let

$$n^{BB} = n + K^{BB}_{\tau_N} \leqslant 2 \cdot n \tag{2.33}$$

be the (random) number of individuals at time τ_N which corresponds to the number of individuals carrying genetic material which is ancestral to the sample from time τ. Precisely, there are $n - K^{BB}_{\tau_N}$ individuals at time τ_N where both neutral loci are the ancestors of both neutral loci of a single sampled individual from τ and there are $2 \cdot K^{BB}_{\tau_N}$ individuals who carry the genetic ancestor of either the first or the second locus of an individual sampled at time τ. For each possible value $\ell \leqslant 2n$ of n^{BB} we define a new marked partition Ψ_ℓ on $[\ell; 1, 2]$ as follows: first let

$$\sigma_\ell : [\ell] \to \{i : B_{\tau_N}(i) = 1\}$$

be an injective random map where each of the $\binom{N}{\ell} \cdot \ell!$ possibilities are equally likely. In that way, $\sigma_\ell(1), \sigma_\ell(2), \ldots, \sigma_\ell(\ell)$ is a random sample from all N B-individuals living at time τ_N. With this define

$$\begin{aligned} &(i, j_i) \sim_{\Psi_\ell} (m, j_m) \quad :\Leftrightarrow \quad A^0_{\tau_N}(\sigma_\ell(i), j_i) = A^0_{\tau_N}(\sigma_\ell(m), j_m); \\ &\text{mark the block } \{(i, j_i) : B_0(A^0_{\tau_N}(\sigma_\ell(i), j_i)) = 1\}. \end{aligned}$$

Note that the partition Ψ_ℓ is defined with respect to time τ_N rather than τ.
We now define two maps which ensure a random choice of the labels of the ancestors living at time τ_N of sampled loci from time τ. First, let f^1 be a random bijective map

$$f^1 : [n] \to [n],$$

where again all $n!$ possibilities for f^1 have equal probability. Then, suppose $K^{BB}_{\tau_N} = k$ and define for each possible $k \in \{0, 1, 2 \ldots, n\}$ a surjective map f^2_k which assigns a name $i \in [n+k]$ to a locus (i, j_i) in that way that for exactly $n - k$ individuals the two neutral loci get the

same label, and for all others, the two neutral loci get different labels in $[n+k]$:

$$f_k^2 : [n;1,2] \to [n+k], \quad f_k^2(i,j_i) = \begin{cases} i & \text{if } \ i \leqslant n-k, \\ i & \text{if } \ i > n-k \ \text{ and } \ j_i = 1, \\ i+k & \text{if } \ i > n-k \ \text{ and } \ j_i = 2. \end{cases}$$

In the following we will consider $f^2_{K^{BB}_{\tau_N}}(f^1(i), j_i)$. Define a new marked partition using Ψ_ℓ, f^1 and f_k^2 as follows: let Θ_2 be a random partition on $[n;1,2]$ with

$$\begin{aligned} (i,j_i) \sim_{\Theta_2} (m,j_m) \quad :\Leftrightarrow \quad & R(i,j_i) < \tau_N, \ R(m,j_m) < \tau_N \ \text{ and} \\ & (f^2_{K^{BB}_{\tau_N}}(f^1(i),j_i),j_i) \sim_{\Psi_{n^{BB}}} (f^2_{K^{BB}_{\tau_N}}(f^1(m),j_m),j_m); \\ \text{mark the block } \ \{(i,j_i) : R(i,j_i) < \tau_N \ & \text{ and } \ f^2_{K^{BB}_{\tau_N}}(f^1(i),j_i) \ \text{ is in the} \\ & \text{marked block of } \ \Psi_{n^{BB}}\}. \end{aligned}$$

For a comparison of Θ_1 and Θ_2 consider the following arguments: we have

$$\begin{aligned} & (f^2_{K^{BB}_{\tau_N}}(f^1(i),j_i),j_i) \sim_{\Psi_{n^{BB}}} (f^2_{K^{BB}_{\tau_N}}(f^1(m),j_m),j_m) \\ \Leftrightarrow \quad & A^0_{\tau_N}\Big(\sigma_{n^{BB}}(f^2_{K^{BB}_{\tau_N}}(f^1(i),j_i)),\, j_i\Big) = A^0_{\tau_N}\Big(\sigma_{n^{BB}}(f^2_{K^{BB}_{\tau_N}}(f^1(m),j_m)),\, j_m\Big). \end{aligned}$$

The criterion for the equivalence with respect to Θ as given in Definition 2.4 can be rewritten as follows:

$$A^0_\tau(i,j_i) = A^0_\tau(m,j_m) \quad \Leftrightarrow \quad A^0_{\tau_N}(A^{\tau_N}_\tau(i,j_i),j_i) = A^0_{\tau_N}(A^{\tau_N}_\tau(m,j_m),j_m).$$

By Proposition 2.18 it is unlikely that any two sampled individuals will have found a common ancestor at some time $t \geqslant \tau_N$. Therefore, we know that with probability $1 - \mathcal{O}(1/\log(N))$ the equality $A^{\tau_N}_\tau(i,j_i) = A^{\tau_N}_\tau(m,j_m)$ implies $i = m$. Further, Proposition 2.9 states that with high probability no neutral locus from the sample will have found an ancestor outside the mutant population, again up to terms of order $1/\log(N)$. That ensures that the considered resampling by f^1 and $f^2_{K^{BB}_{\tau_N}}$ indeed only needs to consider the current B-population of N individuals at time τ_N. In addition, Proposition 2.18 states that we will have n^{BB} different individuals at time τ_N who are ancestral to the sample, $n - K^{BB}_{\tau_N}$ of them with two neutral loci and $2 \cdot K^{BB}_{\tau_N}$ with one neutral locus.

Since τ_N is a stopping time and the process which describes the evolution of the population is strong Markov, we know that events happening in $(\tau_N, \tau]$ are independent of events in $[0, \tau_N)$, given the current state at time τ_N. The last argument is that by construction of the Moran model all individuals within one background, B or b, are exchangeable. From these

observations it follows that

$$\left|\mathbb{P}(\Theta_2 = \pi) - \mathbb{P}(\Theta_1 = \pi)\right| \leqslant C/\log(N).$$

The next step of the proof is to compare the partition $\Psi_{n^{BB}}$ with a marked partition defined entirely through the characteristic times from Definition 2.8 with respect to a sample of size n^{BB} from time τ_N. Define the marked partition $\Upsilon_{n^{BB}}$ on $[n^{BB}; 1, 2]$ as follows:

- For $i, m \in [n^{BB}], i \neq m$ and $j_i, j_m \in \{1, 2\}$,

$$(i, j_i) \sim_{\Upsilon_{n^{BB}}} (m, j_m) \quad :\Leftrightarrow \quad R(i, j_i) = R(m, j_m) = -\infty.$$

- For $i \in [n^{BB}]$,

$$(i, 1) \sim_{\Upsilon_{n^{BB}}} (i, 2) \quad :\Leftrightarrow \quad R(i, 1) = R(i, 2) \text{ and } R_b^{\text{rec}}(i) = -\infty.$$

- Mark the block containing $\{(i, j_i) : R(i, j_i) = -\infty\}$.

The partition $\Upsilon_{n^{BB}}$ then consists of one marked block which contains all those neutral loci which were never connected to a b-individual at any point in the past, and otherwise only singletons with one neutral locus or double-singletons composed of the two neutral loci of one individual. For a comparison of $\Upsilon_{n^{BB}}$ and $\Psi_{n^{BB}}$ consider the following for the time interval $[0, \tau_N)$:

- Proposition 2.12 implies that the probability of a coalescence within the b-population is small enough to be ignored.

- Proposition 2.13 says that we can ignore the event that two lineages recombine into the b-population after they have coalesced.

Hence, with high probability, we will not see unmarked blocks with loci from different individuals in $\Psi_{n^{BB}}$. By definition, the same holds true for $\Upsilon_{n^{BB}}$ and thus with high probability, namely $1 - \mathcal{O}(1/\log(N))$, the block types of both partitions coincide.
Further, Proposition 2.11 says that with high probability there is no back-recombination of any locus into the B-population. Therefore, we only have to consider the event that a locus (i, j_i) migrates at all into the b-population (alone or jointly with its counterpart $(i, j_i'), j_i' \neq j_i$). This corresponds exactly to the above definition of $\Upsilon_{n^{BB}}$ and hence we know that the latter produces the same blocks as $\Psi_{n^{BB}}$. For any marked partition $\pi \in \mathcal{P}^*_{[n^{BB};1,2]}$ we therefore get

$$\left|\mathbb{P}(\Upsilon_{n^{BB}} = \pi) - \mathbb{P}(\Psi_{n^{BB}} = \pi)\right| \leqslant C/\log(N). \tag{2.34}$$

Further, from Proposition 2.20 it follows that with $\bar{q}^D_{\mathrm{G1}}$ as in (2.26)

$$\left|\mathbb{P}(\Upsilon_{n^{BB}} = \pi) - Q_{\bar{q}^D_{\mathrm{G1}}}(\pi)\right| \leqslant C/\log(N),$$

where $Q_{\bar{q}^D_{\mathrm{G1}}}$ is the distribution of a marked $\bar{q}^D_{\mathrm{G1}}$-partition on $[n^{BB}; 1, 2]$ and hence by the Triangle inequality

$$\left|\mathbb{P}(\Psi_{n^{BB}} = \pi) - Q_{\bar{q}^D_{\mathrm{G1}}}(\pi)\right| \leqslant C/\log(N).$$

It remains to draw the connection to Θ_2. For this, we need to define another random marked partition which mimics the separation within the B-population. We pursue a similar idea as in Definition 2.9 in [36]. Let $Q^{q^{BB}}_{\bar{q}^D_{\mathrm{G1}}}$ be the distribution of a random marked partition Π^* on the set $[n; 1, 2]$ obtained as follows: let $\zeta_1, \ldots, \zeta_n \sim_{\mathrm{iid}} \mathrm{Bernoulli}(q^{BB})$ with q^{BB} from (2.19). Further consider a random marked $\bar{q}^D_{\mathrm{G1}}$-partition Π^D and a random marked $\bar{q}^{\bar{D}}_{\mathrm{G1}}$-partition $\Pi^{\bar{D}}$, with $\bar{q}^D_{\mathrm{G1}}$ defined in (2.26) and $\bar{q}^{\bar{D}}_{\mathrm{G1}}$ as in (2.29). Then define

$$\begin{aligned}(i, j_i) \sim_{\Pi^*} (m, j_m) \quad :\Leftrightarrow \quad & (i, j_i) \sim_{\Pi^D} (m, j_m) \text{ and } \zeta_i = 0, \text{ or} \\ & (i, j_i) \sim_{\Pi^{\bar{D}}} (m, j_m) \text{ and } \zeta_i = 1; \\ \text{mark the block } \{(i, j_i) : & (i, j_i) \text{ is in the marked block of } \Pi^D \text{ or } \Pi^{\bar{D}}\}.\end{aligned}$$

By Propositions 2.19 and 2.20, Corollary 2.21 and the construction of Θ_2 it now follows that

$$\left|\mathbb{P}(\Theta_2 = \pi) - Q^{q^{BB}}_{\bar{q}^D_{\mathrm{G1}}}(\pi)\right| \leqslant C/\log(N)$$

and finally by Proposition 2.22 we get the statement

$$\left|Q^{q^{BB}}_{\bar{q}^D_{\mathrm{G1}}}(\pi) - Q_{\bar{q}_{\mathrm{G1}}}(\pi)\right| \leqslant C/\log(N).$$

Applying the Triangle inequality finishes the proof. □

Proof of Theorem 2.7. The proof of this theorem essentially uses the same ingredients as the proof of Theorem 2.6. Here however, the construction of the discrete model implies that a recombination which separates one neutral locus from the selected locus does not change the ancestry of the other neutral locus of that individual in this particular step. In addition, the probability for two recombinations in the same reproduction event is again bounded by $1/(\log N)^2$. Hence, for the events happening before time τ_N, in the interval $[0, \tau_N)$, it is irrelevant whether the neutral loci separated within the B-population before. The statement of this theorem thus follows similarly as Corollary 2.21 only with simpler recombination probabilities as in the alignment (G2) the second neutral locus only escapes due to a recombination between the SL and N2. This indeed gives us the entries of the vector

$\bar{p}_{G2}$ as stated in (2.9). □

2.4 Proofs of the auxiliary statements

First of all note that the background Markov chain X defined in (2.1) is the same here as it is in Schweinsberg and Durrett, [36]. This implies that all results concerning this walk, such as the expected number of up-jumps, holds and down-jumps, still hold true when following two or more neutral loci back in time. This will be the argument for referring to [36] for the proofs of the Propositions 2.15, 2.11, 2.12 and 2.13. In the following part we will restate some results from [36] concerning the properties of the Markov chain X.

2.4.1 Properties of the Random Walk

By definition of the Moran model, we have the following transition probabilities with respect to the unconditioned measure (cf. Equations (3.1) to (3.3) in [36]).

$$\begin{aligned}
\mathbb{P}'(X_t = k+1 | X_{t-1} = k) &= \frac{2N-k}{2N}\frac{k}{2N}, \\
\mathbb{P}'(X_t = k-1 | X_{t-1} = k) &= \frac{k}{2N}\frac{2N-k}{2N}(1-s), \\
\mathbb{P}'(X_t = k | X_{t-1} = k) &= \frac{(2N-k)^2 + k^2 + (2N-k)ks}{(2N)^2} = 1 - \frac{(2-s)k(2N-k)}{(2N)^2}.
\end{aligned}$$

Define for $k, j \in \{1, \ldots, 2N-1\}$ and τ_j as in (2.3):

$$\begin{aligned}
\text{number of up-jumps: } U_{k,j} &= \#\{t \geqslant \tau_j \mid X_t = k \text{ and } X_{t+1} = k+1\}, \\
\text{number of down-jumps: } D_{k,j} &= \#\{t \geqslant \tau_j \mid X_t = k \text{ and } X_{t+1} = k-1\}, \\
\text{number of holds: } H_{k,j} &= \#\{t \geqslant \tau_j \mid X_t = k \text{ and } X_{t+1} = k\}, \\
\text{with } J_k &\equiv J_{k,1} \text{ for } J = U, D, H.
\end{aligned}$$

Then Lemma 3.2 from [36] states the following:

Lemma 2.23 (Lemma 3.2, [36]). *Suppose $j, k \in [2N-1]$. Define*

$$p(a,b,k) = \mathbb{P}'(\inf\{s > t : X_s = b\} < \inf\{s > t : X_s = a\} \mid X_t = k), \tag{2.35}$$

with $\mathbb{P}'$ the unconditioned measure. Let $q_0 = 1$ and

$$q_k = \frac{p(k, 2N, k+1)}{p(0, 2N, k+1)} = \frac{s}{(1-(1-s)^{2N-k})}\frac{(1-(1-s)^{2N})}{(1-(1-s)^{k+1})} \geqslant s.$$

Further define $r_{0,j} = 0$ *and* $r_{k,j} = 1$ *for* $j \leqslant k$. *If* $j > k$, *let*

$$r_{k,j} = 1 - \frac{p(k, 2N, j)}{p(0, 2N, j)} = 1 - \frac{(1-(1-s)^{j-k})}{(1-(1-s)^{2N-k})} \frac{(1-(1-s)^{2N})}{(1-(1-s)^{j})} \leqslant (1-s)^{j-k}.$$

and let

$$\beta_k := \frac{k(2N-k)}{k^2 + (2N-k)^2 + sk(2N-k)}. \tag{2.36}$$

Then

$$\begin{aligned} \mathbb{E}(U_{k,j}) &= \frac{r_{k,j}}{q_k}, \\ \mathbb{E}(D_{k,j}) &= \frac{1}{q_{k-1}} - 1\,, \ \textit{for } k > j, \quad \mathbb{E}(D_{k,j}) = \frac{r_{k-1,j}}{q_{k-1}}\,, \ \textit{for } k \leqslant j, \\ \mathbb{E}(H_{k,j}) &= \mathbb{E}(U_{k,j} + D_{k,j}) \cdot \frac{1}{2-s} \cdot \frac{k^2 + (2N-k)^2 + sk(2N-k)}{k(2N-k)} \\ &\leqslant \min\{(1-s)^{j-k}, 1\} \cdot \frac{1}{s\beta_k}. \end{aligned} \tag{2.37}$$

Further, the following relations hold in case of fixation:

$$D_{k,j} = U_{k-1,j}, \ \textit{if } k \leqslant j, \quad D_{k,j} = U_{k-1,j} - 1, \ \textit{if } k > j. \tag{2.38}$$

We will state here only the idea on how to show the statements concerning $U_{k,j}$ and $D_{k,j}$, as we will pursue a similar idea in the proof of Lemma 3.11 (see Appendix B), and refer to [36] for the remaining ideas of the proof.

Proof. The essential idea is to firstly show that the number of up-jumps is in fact geometrically distributed and secondly, to find relations between the number of up-jumps and the number of down-jumps and holds.
Consider the variable $U_{k,j}$. As we condition on fixation we know that in the case $j \leqslant k$, X will necessarily jump from k to $k+1$ for some $t \geqslant \tau_j$. Once it has reached the state $k+1$ we can determine the probabilities for the complementary events that X will go back to k or that X will reach its final state $2N$ without returning to k. We can deduce that $U_{k,j}$ is indeed geometrically distributed and that the probability of the latter event corresponds to the success probability and hence parameter of the distribution. We can express this probability of no return to k by using the definition of $p(a, b, k)$ from (2.35) with respect to the unconditioned measure $\mathbb{P}'$ and get with the identity

$$\mathbb{P}(X_\tau = 2N) = p(0, 2N, 1) = \frac{s}{1-(1-s)^{2N}}$$

and using Bayes' formula the equality

$$\mathbb{P}(X_s > k \ \forall s > t \mid X_t = k+1) = \frac{p(k,2N,k+1)}{p(0,2N,k+1)} = q_k.$$

For $j > k$, $U_{k,j} = 0$ with probability

$$\mathbb{P}(X_t > k \ \forall t > \tau_j) = \frac{p(k,2N,j)}{p(0,2N,j)} = 1 - r_{k,j}$$

and $U_{k,j}$ is only positive on the event $\{T_{k,j} \geqslant 1\}$. Conditional on this event, which has probability $r_{k,j}$, the same calculations as above lead to the conclusion that again $U_{k,j} \sim$ Geometric(q_k) and both observations together result in the claim for $\mathbb{E}(U_{k,j})$.
The results for $D_{k,j}$ and in particular (2.38) follow easily as an up-jump from k to $k+1$ can only occur after a down-jump from $k+1$ to k, except for the first up-jump starting in k. □

Applying Lemma 2.23 we can easily calculate expressions for the following expected values which will be needed later:

Corollary 2.24.

$$\mathbb{E}(U_{2N-k-1,2N-l}) = \frac{r_{2N-k-1,2N-l}}{q_{2N-k-1}}$$
$$= \begin{cases} \frac{(1-(1-s)^{k+1})(1-(1-s)^{2N-k})}{s(1-(1-s)^{2N})}, & k \leqslant l-1 \\ \frac{(1-(1-s)^{k+1})(1-(1-s)^{2N-k})}{s(1-(1-s)^{2N})} - \frac{(1-(1-s)^{k-l+1})(1-(1-s)^{2N-k})}{s(1-(1-s)^{2N-l})}, & 2N-1 > k > l-1 \\ 0, & k = 2N-1 \end{cases}$$

$$\mathbb{E}(H_{2N-k,2N-l}) = \mathbb{E}(U_{2N-k,2N-l} + D_{2N-k,2N-l}) \cdot \tfrac{1}{2-s} \cdot \tfrac{1}{\beta_k}$$
$$= \frac{1}{\beta_k} \cdot \begin{cases} \frac{(1-(1-s)^{k})(1-(1-s)^{2N-k})}{s(1-(1-s)^{2N})}, & k \leqslant l \\ \frac{(1-(1-s)^{k})(1-(1-s)^{2N-k})}{s(1-(1-s)^{2N})} - \frac{(1-(1-s)^{k-l})(1-(1-s)^{2N-k})}{s(1-(1-s)^{2N-l})}, & k > l. \end{cases}$$

Further, the following bounds hold for $k \geqslant l$:

$$\mathbb{E}(U_{2N-k-l,2N-l}) \leqslant \frac{(1-s)^{k-l+1}}{s}, \quad \mathbb{E}(H_{2N-k,2N-l}) \leqslant \frac{(1-s)^{k-l}}{s\beta_k}.$$

These expressions come in handy when we use the symmetry of the Markov chain X conditioned on fixation, as it was mentioned in [36] in the proof of their Proposition 2.1.
We will now define the probabilities that specific lines coalesce in one step, and that one line experiences a particular type of a recombination. Note that in some cases, when the individuals have the same background, we calculate the values of these probabilities without asking

for $I_{t,1}$ and $I_{t,2}$ (or $I_{t,2}$ and $I_{t,3}$) to be different individuals. However, we can satisfy ourselves that the error thus made is small enough as it is not likely to draw the same individual twice for N large.
Define for $k \in [2N-1]$ and $l, i, m \in [2N]$ such that $|k-l| \leqslant 1$ and $i \neq m$ the probabilities for a coalescence of individuals i and m with respect to their j-th locus conditional on the transition of the Markov chain and their background:

$$\begin{aligned} p_{BB}^{c,j}(k,l) &:= \mathbb{P}(A_t^{t-1}(i,j) = A_t^{t-1}(m,j) \mid X_{t-1} = k, X_t = l, B_t(i) = 1 = B_t(m)), \\ p_{bb}^{c,j}(k,l) &:= \mathbb{P}(A_t^{t-1}(i,j) = A_t^{t-1}(m,j) \mid X_{t-1} = k, X_t = l, B_t(i) = 0 = B_t(m)), \\ p_{Bb}^{c,j}(k,l) &:= \mathbb{P}(A_t^{t-1}(i,j) = A_t^{t-1}(m,j) \mid X_{t-1} = k, X_t = l, B_t(i) = 1, B_t(m) = 0). \end{aligned} \tag{2.39}$$

Lemma 2.25 (cf. Lemma 4.1, [36])**.** *We have*

$$\begin{aligned} &p_{BB}^{c,j}(k,k-1) = p_{bb}^{c,j}(k,k+1) = 0, \\ &p_{BB}^{c,j}(k,k+1) = \frac{2}{k(k+1)}\Big[1 - \frac{r^*(j)(2N-k)}{2N}\Big], \quad p_{BB}^{c,j}(k,k) = \frac{2\beta_k}{k(2N-k)}\Big[1 - \frac{r^*(j)(2N-k)}{2N}\Big], \\ &p_{bb}^{c,j}(k,k-1) = \frac{2}{(2N-k)(2N-k+1)}\Big[1 - \frac{r^*(j)k}{2N}\Big], \quad p_{bb}^{c,j}(k,k) = \frac{2\beta_k}{k(2N-k)}\Big[1 - \frac{r^*(j)k}{2N}\Big], \\ &p_{Bb}^{c,j}(k,k+1) = \frac{r^*(j)}{2N(k+1)}, \quad p_{Bb}^{c,j}(k,k) = \frac{r^*(j)\beta_k}{k(2N-k)}, \quad p_{Bb}^{c,j}(k,k-1) = \frac{r^*(j)}{2N(2N-k+1)}. \end{aligned}$$

Proof. The derivations are straightforward and we refer to [36]. We only have to replace their recombination probability r with our $r^*(j)$ from (2.15). Here we need $r^*(j)$ for $j = 2$ as a double recombination would cause the second neutral locus to have the same parent as the selected locus. This is favorable in the coalescence events within the same background but would prevent a coalescence of lineages with respect to N2 when we consider mixed backgrounds. □

We further calculate the probability of the event that a lineage (or one locus of a lineage) changes background due to a recombination. For $k \in [2N-1]$, $j = 1, 2$ and $l, i \in [2N]$ such that $|k-l| \leqslant 1$ let

$$\begin{aligned} p_B^{r^*(j)}(k,l) &:= \mathbb{P}(B_{t-1}(A_t^{t-1}(i,j)) = 0 \mid X_{t-1} = k, X_t = l, B_t(i) = 1), \\ p_b^{r^*(j)}(k,l) &:= \mathbb{P}(B_{t-1}(A_t^{t-1}(i,j)) = 1 \mid X_{t-1} = k, X_t = l, B_t(i) = 0), \\ p_B^{r_2(1-r_1)}(k,l) &:= \mathbb{P}(B_{t-1}(A_t^{t-1}(i,2)) = 0, B_{t-1}(A_t^{t-1}(i,1)) = 1 \mid X_{t-1} = k, X_t = l, B_t(i) = 1), \\ p_B^{r_1 r_2}(k,l) &:= \mathbb{P}(B_{t-1}(A_t^{t-1}(i,1)) = 0, B_{t-1}(A_t^{t-1}(i,2)) = 1 \mid X_{t-1} = k, X_t = l, B_t(i) = 1). \end{aligned}$$

Lemma 2.26 (cf. Lemma 3.3, [36]). *For $r \in \{r^*(j), r_2(1-r_1), r_1 r_2\}$ we have*

$$\begin{aligned}
&p_B^r(k,k-1) = p_b^r(k,k+1) = 0,\\
&p_B^r(k,k+1) = \frac{r(2N-k)}{2N(k+1)}, \quad p_b^r(k,k-1) = \frac{rk}{2N(2N-k+1)},\\
&p_B^r(k,k) = p_b^r(k,k) = \frac{r\beta_k}{2N}.
\end{aligned}$$

Proof. Again, see [36]. Note that N2 also changes background if a recombination between SL and N1 causes the migration. However, when two recombinations happen at the same time, N2 stays with the SL in the former background. Therefore, we have to use the recombination probability $r^*(j)$ from (2.15). □

In the setting of two neutral loci we also need to consider the following event: a recombination between the two neutral loci N1 and N2, where both parents are of type b.
Define for any $i \in [2N]$, $t \in \{1, \dots, \tau\}$

$$p_{bb}^{r_2}(k,l) := \mathbb{P}(I_{t,1} = i, B_{t-1}(I_{t,2}) = 0 = B_{t-1}(I_{t,3}), I_{t,6} = 1 \mid X_{t-1} = k, X_t = l, B_t(i) = 0).$$

This is exactly the probability for a recombination event between N1 and N2 ($I_{t,6} = 1$) where parent and second parent have the b-type at the SL ($B_{t-1}(I_{t,2}) = 0 = B_{t-1}(I_{t,3})$) and where a particular individual (the i-th individual), conditioned to have the b-type in the next generation t, is the newborn ($I_{t,1} = i$). Note that this probability is different from zero only if $l = k, k-1$ as otherwise the parent cannot be of type b.

Lemma 2.27. *We have for $l, k \in [2N-1]$ and $|k-l| \leqslant 1$*

$$p_{bb}^{r_2}(k,l) = \begin{cases} \frac{r_2}{2N}\frac{(2N-k)^2}{k^2+(2N-k)^2+sk(2N-k)} = \frac{r_2}{2N}\frac{(2N-k)\beta_k}{k} & l = k \\ \frac{r_2}{2N} \cdot \frac{2N-k}{2N-(k-1)} & l = k-1 \\ 0 & \textit{else.} \end{cases} \tag{2.40}$$

We here give a very detailed proof in order to give an example of how to calculate the probabilities stated in Lemma 2.25 and Lemma 2.26.

Proof. First of all note that by construction of the model, the outcome of $I_{t,6}$ is independent of all other events. In addition, the value of $B_{t-1}(I_{t,3})$ here only depends on the number of B individuals in generation $t-1$. Thus,

$$\begin{aligned}
p_{bb}^{r_2}(k,l) = \mathbb{P}(I_{t,6} = 1) \cdot \mathbb{P}(B_{t-1}(I_{t,3}) = 0 \mid X_{t-1} = k, X_t = l)\\
\cdot \mathbb{P}(I_{t,1} = i, B_{t-1}(I_{t,2}) = 0 \mid X_{t-1} = k, X_t = l, B_t(i) = 0)
\end{aligned}$$

$$= r_2 \cdot \frac{2N-k}{2N} \cdot \mathbb{P}(I_{t,1} = i, B_{t-1}(I_{t,2}) = 0 \mid X_{t-1} = k, X_t = l, B_t(i) = 0).$$

With Bayes' formula we can rewrite the last factor:

$$\begin{aligned}
&\mathbb{P}(I_{t,1} = i, B_{t-1}(I_{t,2}) = 0 \mid I_{t,1} = i, X_{t-1} = k, X_t = l, B_t(i) = 0) \\
&= \frac{\mathbb{P}(B_t(i) = 0, B_{t-1}(I_{t,2}) = 0 \mid I_{t,1} = i, X_{t-1} = k, X_t = l) \cdot \mathbb{P}(I_{t,1} = i \mid X_{t-1} = k, X_t = l)}{\mathbb{P}(B_t(i) = 0 \mid X_{t-1} = k, X_t = l)} \\
&= \frac{\mathbb{P}(B_t(i) = 0, B_{t-1}(I_{t,2}) = 0 \mid I_{t,1} = i, X_{t-1} = k, X_t = l) \cdot \frac{1}{2N}}{\frac{2N-l}{2N}}.
\end{aligned}$$

Distinguishing between the cases $l = k$ and $l = k-1$ yields the following:

$$\mathbb{P}(B_t(i) = 0, B_{t-1}(I_{t,2}) = 0 \mid I_{t,1} = i, X_{t-1} = k, X_t = k) = \frac{(2N-k)^2}{k^2 + (2N-k)^2 + sk(2N-k)}$$
$$\mathbb{P}(B_t(i) = 0, B_{t-1}(I_{t,2}) = 0 \mid I_{t,1} = i, X_{t-1} = k, X_t = k-1) = 1.$$

Recalling the definition of β_k in (2.36) this implies,

$$\mathbb{P}(I_{t,1} = i, B_{t-1}(I_{t,2}) = 0 \mid I_{t,1} = i, X_{t-1} = k, X_t = k, B_t(i) = 0) = \frac{\beta_k \frac{2N-k}{k} \cdot \frac{1}{2N}}{\frac{2N-k}{2N}} = \frac{\beta_k}{k},$$
$$\mathbb{P}(I_{t,1} = i, B_{t-1}(I_{t,2}) = 0 \mid I_{t,1} = i, X_{t-1} = k, X_t = k-1, B_t(i) = 0) = \frac{1 \cdot \frac{1}{2N}}{\frac{2N-(k-1)}{2N}} = \frac{1}{2N-(k-1)},$$

and putting all together gives the result. □

Analogously to the above, we here also need to calculate the probability that the two neutral loci split within the B-population:

$$p_{BB}^{r_2}(k,l) := \mathbb{P}(I_{t,1} = i, B_{t-1}(I_{t,2}) = 1 = B_{t-1}(I_{t,3}), I_{t,6} = 1 \mid X_{t-1} = k, X_t = l, B_t(i) = 1).$$

Lemma 2.28. *For $l, k \in [2N-1]$ and $|k-l| \leqslant 1$,*

$$p_{BB}^{r_2}(k,l) = \begin{cases} \frac{r_2}{2N} \cdot \beta_k \cdot \frac{k}{2N-k} & l = k \\ \frac{r_2}{2N} \cdot \frac{k}{k+1} & l = k+1 \\ 0 & \textit{else.} \end{cases} \tag{2.41}$$

Proof. The proof is very similar to the above, we will therefore leave out the details.

$$\begin{aligned}
p_{BB}^{r_2}(k,l) &= \mathbb{P}(I_{t,6} = 1) \cdot \mathbb{P}(B_{t-1}(I_{t,3}) = 1 \mid X_{t-1} = k, X_t = l) \\
&\cdot \frac{\mathbb{P}(B_t(i) = 1, B_{t-1}(I_{t,2}) = 1 \mid I_{t,1} = i, X_{t-1} = k, X_t = l) \cdot \mathbb{P}(I_{t,1} = i \mid X_{t-1} = k, X_t = l)}{\mathbb{P}(B_t(i) = 1 \mid X_{t-1} = k, X_t = l)}
\end{aligned}$$

$$= r_2 \cdot \frac{k}{2N} \cdot \frac{\mathbb{P}(B_t(i) = 1, B_{t-1}(I_{t,2}) = 1 \mid I_{t,1} = i, X_{t-1} = k, X_t = l) \cdot \frac{1}{2N}}{\frac{l}{2N}}$$
$$= \frac{r_2}{2N}\frac{k}{l} \cdot \mathbb{P}(B_t(i) = 1, B_{t-1}(I_{t,2}) = 1 \mid I_{t,1} = i, X_{t-1} = k, X_t = l).$$

A distinction of cases results in

$$\mathbb{P}(B_t(i) = 1, B_{t-1}(I_{t,2}) = 1 \mid I_{t,1} = i, X_{t-1} = k, X_t = k) = \frac{k^2}{k^2 + (2N-k)^2 + sk(2N-k)},$$
$$\mathbb{P}(B_t(i) = 1, B_{t-1}(I_{t,2}) = 1 \mid I_{t,1} = i, X_{t-1} = k, X_t = k+1) = 1,$$

and using (2.36) and plugging in ends the proof of the lemma. □

2.4.2 Proofs of the auxiliary propositions

With the above lemmas and the remarks concerning the background Markov chain, we can cite [36] as a reference for the proofs of Propositions 2.11, 2.12, 2.13 and 2.15. We just have to insert our probabilities for coalescence and recombination from Lemma 2.25 and 2.26 into the calculations from [36] in order to receive the same bounds. All remaining ideas and calculations can be transfered straightforwardly. Instead of restating the proofs of those statements here, we recapitulate many of the ideas and calculations used in [36] in the following proofs of Propositions 2.9 and 2.10 and in the next section which deals with the proof of Proposition 2.16.

Throughout this section, we will denote by c, c', C, C' constants which are independent of N and which may vary from line to line.

In order to prove Proposition 2.9, we first restate (without proof) a bound on the probability that the Markov chain X is in a certain state at recombination events of interest.

Lemma 2.29 (cf. Lemma 3.4, [36]). *We have for $i \in [n], j = 1, 2$ and $k \in [2N]$*

$$\mathbb{P}(X_{R(i,j)} = k) \leqslant \frac{r^*(j)}{sk}, \tag{2.42}$$

Further we state a result from [36], which was obtained in the proof of their Proposition 2.1. For $\mathcal{T} = R(i,j), G^j(i,m)$ or $\mathcal{T} = \tau_J, \tau_J^*$ with $i, m \in [n], j = 1, 2, J \in [2N]$, we have

$$\begin{aligned}
\mathbb{E}[\#\{t < \mathcal{T} : X_t = k \text{ and } X_{t+1} = k+1 \mid X_{\mathcal{T}} = j] &= \mathbb{E}[U_{2N-k-1,2N-j}], \\
\mathbb{E}[\#\{t < \mathcal{T} : X_t = k \text{ and } X_{t+1} = k-1 \mid X_{\mathcal{T}} = j] &= \mathbb{E}[D_{2N-k+1,2N-j}], \\
\mathbb{E}[\#\{t < \mathcal{T} : X_t = k \text{ and } X_{t+1} = k \mid X_{\mathcal{T}} = j] &= \mathbb{E}[H_{2N-k,2N-j}].
\end{aligned} \tag{2.43}$$

These relations, which are explained by a symmetry argument and the strong Markov property of X, will be used in many proofs throughout this whole section.

We start with giving a proof of Proposition 2.9 following similar ideas as the ones which are used in the proof of Proposition 2.4 in [36] in order to reach the statement given in Equation (2.16).

Proof of Proposition 2.9. We start with the first statement:

$$\begin{aligned}\mathbb{P}(R(m,j) > R_B^{\mathrm{rec}}(i) \geqslant 0) &= \sum_{k=1}^{2N-1}\sum_{l=1}^{2N-1} \mathbb{P}(R(m,j) > R_B^{\mathrm{rec}}(i) \geqslant 0, X_{R(m,j)} = k, X_{R_B^{\mathrm{rec}}(i)} = l)\\ &= \sum_{k=1}^{2N-1}\sum_{l=1}^{2N-1} \mathbb{P}(R(m,j) > R_B^{\mathrm{rec}}(i) \geqslant 0, X_{R_B^{\mathrm{rec}}(i)} = l \mid X_{R(m,j)} = k)\cdot \mathbb{P}(X_{R(m,j)} = k).\end{aligned}$$

In order to bound the first probability, we apply the symmetry argument from (2.43) and do not take into account that $R_B^{\mathrm{rec}}(i)$ is actually the *first* time a BB-split happens.

$$\begin{aligned}&\mathbb{P}(R(m,j) > R_B^{\mathrm{rec}}(i) \geqslant 0, X_{R_B^{\mathrm{rec}}(i)} = l \mid X_{R(m,j)} = k)\\ &\leqslant p_{BB}^{r_2}(l,l)\cdot \mathbb{E}(H_{2N-l,2N-k}) + p_{BB}^{r_2}(l,l+1)\cdot \mathbb{E}(U_{2N-l-1,2N-k})\\ &\leqslant \frac{r_2}{2N}\frac{l}{2N-l}\beta_l \cdot \frac{\min\{(1-s)^{l-k},1\}}{s\beta_l} + \frac{r_2}{2N}\frac{l}{l+1}\frac{\min\{(1-s)^{l-k+1},1\}}{s}\\ &\leqslant \frac{r_2}{s}\min\{(1-s)^{l-k},1\}\cdot\frac{l+2N-l}{2N(2N-l)} = \frac{r_2}{s}\cdot\frac{\min\{(1-s)^{l-k},1\}}{2N-l}.\end{aligned}$$

Together with (2.42) from Lemma 2.29,

$$\begin{aligned}\mathbb{P}(R(m,j) > R_B^{\mathrm{rec}}(i) \geqslant 0) &\leqslant \sum_{k=1}^{2N-1}\sum_{l=1}^{2N-1}\frac{r_2}{s}\cdot\frac{\min\{(1-s)^{l-k},1\}}{2N-l}\cdot\frac{r^*(j)}{sk}\\ &= \frac{r^*(j)\cdot r_2}{s^2}\sum_{k=1}^{2N-1}\frac{1}{k}\cdot\Big\{\sum_{l=1}^{k}\frac{1}{2N-l} + \sum_{l=k+1}^{2N-1}\frac{(1-s)^{l-k}}{2N-l}\Big\}\\ &\leqslant \frac{r^*(j)\cdot r_2}{s^2}\sum_{k=1}^{2N-1}\frac{1}{k}\cdot\Big\{\frac{k}{2N-k} + \sum_{l=0}^{\infty}(1-s)^l\Big\} \leqslant c\cdot r^*(j)\cdot r_2\cdot\sum_{k=1}^{2N-1}\frac{1}{k} \leqslant \frac{C}{\log N},\end{aligned}$$

as both recombination probabilities can be bounded by a constant over $\log(N)$.
The first line of (2.17) follows from Proposition 2.15 as $\sum_{k=N+1}^{2N}\frac{1}{k} \leqslant 1$ for all N. This implies

$$\begin{aligned}\mathbb{P}(R(m,j) \geqslant \tau_N) &= 1 - \exp\Big(-\frac{r^*(j)}{s}\sum_{k=N+1}^{2N}\frac{1}{k}\Big) + \mathcal{O}\Big(\tfrac{C}{\log(N)}\Big)\\ &= 1 - \sum_{l=0}^{\infty}\frac{1}{l!}\underbrace{\Big(-\frac{r^*(j)}{s}\Big)^l}_{=\mathcal{O}(1/(\log N)^l)}\underbrace{\Big[\sum_{k=N+1}^{2N}\frac{1}{k}\Big]^l}_{\leqslant 1} = 1 - \Big[1 - \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)\Big] \leqslant \frac{C}{\log(N)}.\end{aligned}$$

For the second bound from (2.17) we sum over the possible states of the Markov chain X at

the time of the separation and translate the event $\{R_B^{\mathrm{rec}}(i) = t\}$ for some t into the variables which describe the ancestry of the loci during the evolution of the population:

$$\begin{aligned}
&\mathbb{P}(0 \leqslant R_B^{\mathrm{rec}}(i) < \tau_N) = \sum_{l=1}^{N-1} \mathbb{P}(0 \leqslant R_B^{\mathrm{rec}}(i) < \tau_N, X_{R_B^{\mathrm{rec}}(i)} = l) \\
&= \sum_{l=1}^{N-1} \sum_{|k-l|\leqslant 1} \sum_{t=0}^{\infty} \mathbb{P}(A_\tau^t(i,1) \neq A_\tau^t(i,2), B_t(A_\tau^t(i,1)) = 1 = B_t(A_\tau^t(i,2)), \\
&\qquad \forall t' > t : A_\tau^{t'}(i,1) = A_\tau^{t'}(i,2), B_{t'}(A_\tau^{t'}(i,1)) = 1 = B_{t'}(A_\tau^{t'}(i,2)), t \leqslant \tau_N, X_t = l, X_{t+1} = k) \\
&= \sum_{l=1}^{N-1} \sum_{|k-l|\leqslant 1} \sum_{t=0}^{\infty} \mathbb{P}(t \leqslant \tau_N, X_t = l, X_{t+1} = k) \cdot \mathbb{P}(A_\tau^t(i,1) \neq A_\tau^t(i,2), B_t(A_\tau^t(i,1)) = 1, \\
&\qquad B_t(A_\tau^t(i,2)) = 1 \mid \forall t' > t : A_\tau^{t'}(i,1) = A_\tau^{t'}(i,2), B_{t'}(A_\tau^{t'}(i,1)) = 1 = B_{t'}(A_\tau^{t'}(i,2)), \\
&\qquad t \leqslant \tau_N, X_t = l, X_{t+1} = k) \\
&\qquad \cdot \underbrace{\mathbb{P}(\forall t' > t : A_\tau^{t'}(i,1) = A_\tau^{t'}(i,2), B_{t'}(A_\tau^{t'}(i,1)) = 1 = B_{t'}(A_\tau^{t'}(i,2)) \mid t \leqslant \tau_N, X_t = l, X_{t+1} = k)}_{\leqslant 1} \\
&\leqslant \sum_{l=1}^{N-1} \Big[p_{BB}^{r_2}(l,l) \sum_{t=0}^{\infty} \mathbb{P}(t \leqslant \tau_N, X_t = l, X_{t+1} = l) \\
&\qquad + p_{BB}^{r_2}(l,l+1) \sum_{t=0}^{\infty} \mathbb{P}(t \leqslant \tau_N, X_t = l, X_{t+1} = l+1) \Big] \\
&\leqslant \sum_{l=1}^{N-1} \Big[p_{BB}^{r_2}(l,l) \cdot \mathbb{E}(H_l) + p_{BB}^{r_2}(l,l+1) \cdot \mathbb{E}(U_l) \Big] \leqslant \sum_{l=1}^{N-1} \Big[\frac{r_2 \beta_l l}{2N(2N-l)} \cdot \frac{1}{\beta_l s} + \frac{r_2}{2N} \cdot \frac{1}{s} \Big] \\
&= \frac{r_2}{s} \sum_{l=1}^{N-1} \frac{l + 2N - l}{2N(2N-l)} = \frac{r_2}{s} \sum_{l=1}^{N-1} \frac{1}{2N-l} \leqslant \frac{r_2}{s} \frac{N-1}{2N-(N-1)} \leqslant \frac{r_2}{s} \leqslant \frac{C}{\log(N)}.
\end{aligned}$$

Note that we also could have derived the first inequality in (2.17) in the same way. □

Proof of Proposition 2.10. The idea of the proof here is very similar to the one for the second line of (2.17) (see proof above).

$$\begin{aligned}
&\mathbb{P}(R_{bB}^{2\mathrm{rec}}(i) \geqslant 0) = \sum_{l=1}^{2N} \sum_{t=0}^{\infty} \mathbb{P}(R_{bB}^{2\mathrm{rec}}(i) = t, X_t = l, t < \tau) \\
&= \sum_{l=1}^{2N} \sum_{t=0}^{\infty} \mathbb{P}(t < \tau, X_t = l, \forall t' > t : A_\tau^{t'}(i,1) = A_\tau^{t'}(i,2) \text{ and } B_{t'}(A_\tau^{t'}(i,1)) = 1, \\
&\qquad B_t(A_\tau^t(i,1)) = 0, B_t(A_\tau^t(i,2)) = 1) \\
&= \sum_{l=1}^{2N} \sum_{t=0}^{\infty} \sum_{k=l,l+1} \mathbb{P}(B_t(A_\tau^t(i,1)) = 0, B_t(A_\tau^t(i,2)) = 1 \mid t < \tau, X_t = l, X_{t+1} = k, \\
&\qquad \forall t' > t : A_\tau^{t'}(i,1) = A_\tau^{t'}(i,2) \text{ and } B_{t'}(A_\tau^{t'}(i,1)) = 1) \cdot \mathbb{P}(t < \tau, X_t = l, X_{t+1} = k)
\end{aligned}$$

$$\cdot \underbrace{\mathbb{P}(\forall t' > t : A_\tau^{t'}(i,1) = A_\tau^{t'}(i,2) \text{ and } B_{t'}(A_\tau^{t'}(i,1)) = 1 \mid t < \tau, X_t = l, X_{t+1} = k)}_{\leqslant 1}$$

$$\leqslant \sum_{l=1}^{2N} \Big[p_B^{r_1 r_2}(l,l) \sum_{t=0}^{\infty} \mathbb{P}(t < \tau, X_t = l = X_{t+1}) + p_B^{r_1 r_2}(l,l+1) \sum_{t=0}^{\infty} \mathbb{P}(t < \tau, X_t = l, X_{t+1} = l+1) \Big]$$

$$= \sum_{l=1}^{2N} \Big[p_B^{r_1 r_2}(l,l) \cdot \mathbb{E}(H_l) + p_B^{r_1 r_2}(l,l+1) \cdot \mathbb{E}(U_l) \Big]$$

$$\leqslant \sum_{l=1}^{2N} \Big[\frac{r_1 r_2 \beta_l}{2N} \cdot \frac{1}{\beta_l} + \frac{r_1 r_2 (2N-l)}{2N(l+1)} \cdot \frac{1}{s} \Big] \leqslant r_1 r_2 + \frac{r_1 r_2}{s} \sum_{l=1}^{2N} \frac{1}{l} = \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big).$$

□

2.4.3 Calculation of the success probabilities for the different block types

Before we now come to the proofs regarding the marking probabilities we will first state the following Lemma which is used for many essential simplifications throughout the whole section.

Lemma 2.30. *For $r, s \in (0,1)$, $r \leqslant \tilde{r}/\log(2N)$, $r < s$ and $M \in [2N]$ we have*

$$\frac{r}{s} \sum_{l=1}^{M} \frac{1}{l+1} \exp\Big(-\frac{r}{s} \sum_{k=l+1}^{M} \frac{1}{k} \Big) = 1 - \exp\Big(-\frac{r}{s} \cdot \sum_{k=1}^{M} \frac{1}{k} \Big) + \mathcal{O}\Big(\frac{1}{\log(N)}\Big).$$

A proof of this lemma can be found in the Appendix A. Here we continue with proving the result for the probability of the event $\{R_B^{\text{rec}}(i) \geqslant \tau_J\}$. The proof follows the steps of the proof of Proposition 2.2 in [36]. As we will repeat the same ideas in different contexts throughout this whole section, we will briefly summarize the idea beforehand.
The aim is to show that we can approximate the probability $\mathbb{P}(\mathcal{C})$ of an event of interest $\mathcal{C}$ by a value $\exp(-\mathbb{E}[\tilde{\eta}_{\mathcal{C}}^X])$.
In a first step we consider $\mathcal{C}$ conditional on the Markov chain X and perform a Poisson approximation (see Lemma A.2 and the comment in the proof below) such that we can express the desired probability through some exponential $\exp(-\eta_{\mathcal{C}}^X)$ with $\eta_{\mathcal{C}}^X$ chosen such that the error made by the approximation is small. The expression $\tilde{\eta}_{\mathcal{C}}^X$ is chosen such that it is close to the argument $\eta_{\mathcal{C}}^X$ in the sense that the expectation of their difference is small. With the Triangle inequality and the Mean Value theorem applied to the exponential function we first get

$$\begin{aligned} |\mathbb{P}(\mathcal{C} \mid X) - \exp(-\tilde{\eta}_{\mathcal{C}}^X)| &\leqslant |\mathbb{P}(\mathcal{C} \mid X) - \exp(-\eta_{\mathcal{C}}^X)| + |\exp(-\eta_{\mathcal{C}}^X) - \exp(-\tilde{\eta}_{\mathcal{C}}^X)| \\ &\leqslant |\mathbb{P}(\mathcal{C} \mid X) - \exp(-\eta_{\mathcal{C}}^X)| + |\eta_{\mathcal{C}}^X - \tilde{\eta_{\mathcal{C}}}^X|. \end{aligned}$$

Then the aimed approximation fulfills the following inequality by taking expectations in the

above and applying the Jensen inequality:

$$\begin{aligned}|\mathbb{P}(\mathcal{C}) - \exp(-\mathbb{E}[\tilde{\eta}_{\mathcal{C}}^X])| &\leqslant |\mathbb{P}(\mathcal{C}) - \mathbb{E}\big[\exp(-\tilde{\eta}_{\mathcal{C}}^X)\big]| + |\mathbb{E}\big[\exp(-\tilde{\eta}_{\mathcal{C}}^X)\big] - \exp(-\mathbb{E}[\tilde{\eta}^X])| \\ &\leqslant \mathbb{E}|\mathbb{P}(\mathcal{C} \mid X) - \exp(-\eta_{\mathcal{C}}^X)| + \mathbb{E}|\eta_{\mathcal{C}}^X - \tilde{\eta}_{\mathcal{C}}^X| + \mathbb{E}|\tilde{\eta}_{\mathcal{C}}^X - \mathbb{E}[\tilde{\eta}_{\mathcal{C}}^X]|.\end{aligned} \tag{2.44}$$

We aim at bounding the right-hand side: the first term can be small if the $\eta_{\mathcal{C}}^X$ obtained by the Poisson approximation has suitable properties. The second term then depends on the choice of $\tilde{\eta}_{\mathcal{C}}^X$ and hence can be made small with the right pick. The last term is equal to the square root of the variance of $\tilde{\eta}_{\mathcal{C}}^X$. In the next proof we will see that the most evolved part is actually to prove a sufficient bound for the latter.

Proof of Proposition 2.14. We will follow the above described path. Let $\theta_t^{BB} := p_{BB}^{r_2}(X_{t-1}, X_t)$. Then $R_B^{\mathrm{rec}}(i) < \tau_J$ for some $i \in [n]$ if in all steps from τ up to τ_J, this line was never the newborn of two B-individuals when the reproduction event included a recombination between the two neutral loci. Conditional on X, the complementary event $\mathbb{P}(R_B^{\mathrm{rec}}(i) \geqslant \tau_J) = 1 - \mathbb{P}(R_B^{\mathrm{rec}}(i) < \tau_J)$, then can be written as follows:

$$\mathbb{P}(R_B^{\mathrm{rec}}(i) \geqslant \tau_J \mid X) = 1 - \prod_{t=\tau_J+1}^{\tau} (1 - \theta_t^{BB}). \tag{2.45}$$

Theorem 3.6.1 in [15], a limit theorem in the context of Poisson Convergence, now states that for $\eta_J^{BB} := \sum_{t=\tau_J+1}^{\tau} \theta_t^{BB}$ the following holds true:

$$\Big|1 - \prod_{t=\tau_J+1}^{\tau} (1 - \theta_t^{BB}) - (1 - \exp(-\eta_J^{BB}))\Big| \leqslant \sum_{t=\tau_J+1}^{\tau} (\theta_t^{BB})^2,$$

We will give a short proof of this statement in Lemma A.2. After taking expectations, the above reads

$$\Big|\mathbb{P}(R_B^{\mathrm{rec}}(i) \geqslant \tau_J) - (1 - \mathbb{E}[\exp(-\eta_J^{BB})])\Big| \leqslant \mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau} (\theta_t^{BB})^2\Big].$$

In the next steps we follow the recipe from page 65: first, we bound the expectation on the right-hand side, then we calculate $\mathbb{E}[\tilde{\eta}_J^{BB}]$ for $\tilde{\eta}_J^{BB} := \sum_{t=\tau_J+1}^{\tau} \theta_t^{BB} \mathbf{1}_{\{X_{t-1} \geqslant J\}}$ and show that $\mathbb{E}[\tilde{\eta}_J^{BB} - \eta_J^{BB}] \leqslant c/\log(N)$. Last, we bound the variance of $\tilde{\eta}_J^{BB}$ in order to approximate $\mathbb{E}[\exp(-\tilde{\eta}_J^{BB})]$ by $\exp(-\mathbb{E}[\tilde{\eta}_J^{BB}])$.

$$\begin{aligned}\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau} (\theta_t^{BB})^2\Big] &\leqslant \sum_{l=1}^{2N-1} \Big\{(p_{BB}^{r_2}(l,l))^2 \mathbb{E}(H_l) + (p_{BB}^{r_2}(l,l+1))^2 \mathbb{E}(U_l)\Big\} \\ &\leqslant \sum_{l=1}^{2N-1} \Big\{\frac{r_2^2 l^2}{(2N)^2} \frac{1}{s} \Big[\frac{\beta_l}{(2N-l)^2} + \frac{1}{(l+1)^2}\Big]\Big\}\end{aligned}$$

$$= \frac{r_2^2}{s}\Big\{ \sum_{l=1}^{2N-1} \frac{l^2}{(2N)^2} \frac{l}{(l^2+(2N-l)^2+sl(2N-l))(2N-l)} + \sum_{l=1}^{2N-1} \frac{l^2}{(2N)^2(l+1)^2} \Big\}$$
$$\leqslant \frac{r_2^2}{s}\Big\{ \sum_{l=1}^{2N-1} \frac{1}{2N} + \sum_{l=1}^{2N-1} \frac{1}{(2N)^2} \Big\} \leqslant \frac{C}{(\log(N))^2}, \tag{2.46}$$

and thus,

$$\Big|\mathbb{P}(R_B^{\mathrm{rec}}(i) \geqslant \tau_J) - (1 - \mathbb{E}[\exp(-\eta_J^{BB})])\Big| \leqslant \frac{C}{(\log(N))^2}. \tag{2.47}$$

For $\tilde{\eta}_J^{BB}$ defined as above, we have

$$\mathbb{E}[\eta_J^{BB} - \tilde{\eta}_J^{BB}] \leqslant \mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau} \theta_t^{BB} \mathbf{1}_{\{X_{t-1}<J\}}\Big] = \sum_{l=1}^{J-1} \Big[p_{BB}^{r_2}(l,l)\mathbb{E}(H_{l,J}) + p_{BB}^{r_2}(l,l+1)\mathbb{E}(U_{l,J})\Big]$$
$$\leqslant \sum_{l=1}^{J-1} \Big[\frac{r_2 l}{2N(2N-l)} \frac{(1-s)^{J-l}}{s} + \frac{r_2}{2N} \frac{(1-s)^{J-l}}{s}\Big]$$
$$\leqslant \frac{r_2}{s} \sum_{l=1}^{N} \frac{(1-s)^{J-l}}{2N} + \frac{r_2}{s} \sum_{l=N+1}^{J-1} \frac{(1-s)^{J-l}}{2N-l} + \frac{c}{N\log(N)}$$
$$\leqslant \frac{c}{N\log(N)} + \frac{c'\mathbf{1}_{\{J>N\}}}{(2N-J+1)\log(N)} \leqslant \frac{C}{\log(N)}, \tag{2.48}$$

and hence we can work with $\tilde{\eta}_J^{BB}$ instead of η_J^{BB}. Let us calculate the expectation of $\tilde{\eta}_J^{BB}$:

$$\mathbb{E}[\tilde{\eta}_J^{BB}] = \sum_{l=J}^{2N-1} \Big[p_{BB}^{r_2}(l,l)\mathbb{E}(H_l) + p_{BB}^{r_2}(l,l+1)\mathbb{E}(U_l)\Big] \tag{2.49}$$
$$= \sum_{l=J}^{2N-1} \Big[\frac{r_2\beta_l l}{2N(2N-l)} \frac{(1-(1-s)^{2N-l})(1-(1-s)^l)}{s\beta_l(1-(1-s)^{2N})}$$
$$+ \frac{r_2 l}{2N(l+1)} \frac{(1-(1-s)^{2N-l})(1-(1-s)^{l+1})}{s(1-(1-s)^{2N})}\Big]$$
$$= \frac{r_2}{s} \sum_{l=J}^{2N-1} \frac{(1-(1-s)^{2N-l})}{(1-(1-s)^{2N})} \Big[\frac{l(1-(1-s)^l)}{2N(2N-l)} + \frac{(1-(1-s)^{l+1})}{2N}\Big] + \mathcal{O}\Big(\tfrac{1}{N\log(N)}\Big).$$

As in [36] we can approximate the above expression in three steps:

$$\begin{aligned}
(1)\quad & \frac{r_2}{s}\sum_{l=J}^{2N-1}\Big[1-\frac{(1-(1-s)^{2N-l})}{(1-(1-s)^{2N})}\Big]\cdot\Big[\frac{l(1-(1-s)^l)}{2N(2N-l)}+\frac{(1-(1-s)^{l+1})}{2N}\Big]\\
& \leqslant \frac{r_2}{s}\sum_{l=J}^{2N-1}(1-s)^{2N-l}\frac{2}{2N-l}\leqslant\frac{c}{\log(N)},\\
& \text{implying that we can replace the fraction by 1,}\\
(2)\quad & \frac{r_2}{s}\sum_{l=J}^{2N-1}\Big[\frac{(1-(1-s)^{l+1})}{2N}-\frac{(1-(1-s)^l)}{2N}\Big]=\frac{r_2}{s}\sum_{l=J}^{2N-1}\frac{s(1-s)^l}{2N}\leqslant\frac{c}{N\log(N)},\\
& \text{implying that we can substitute } 1-(1-s)^{l+1} \text{ by } 1-(1-s)^l,\\
(3)\quad & \frac{r_2}{s}\sum_{l=J}^{2N-1}[1-(1-(1-s)^l)]\frac{l+2N-l}{2N(2N-l)}=\frac{r_2}{s}\sum_{l=J}^{2N-1}(1-s)^l\frac{1}{2N-l}\leqslant\frac{c}{\log(N)},\\
& \text{implying that we can substitute } 1-(1-s)^l \text{ by } 1,
\end{aligned}\tag{2.50}$$

and make an error of at most order $1/\log(N)$. Hence,

$$\mathbb{E}[\tilde{\eta}_J^{BB}]=\frac{r_2}{s}\sum_{l=J}^{2N-1}\frac{1}{2N-l}+\mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)=\frac{r_2}{s}\sum_{l=1}^{2N-J}\frac{1}{l}+\mathcal{O}\Big(\tfrac{1}{\log(N)}\Big).\tag{2.51}$$

With the Mean Value theorem and Jensen's Inequality we get as in (2.44)

$$\begin{aligned}
\big|\mathbb{E}(\exp(-\eta_J^{BB}))-\exp(-\mathbb{E}(\tilde{\eta}_J^{BB}))\big| &\leqslant \mathbb{E}\big|\exp(-\eta_J^{BB})-\exp(-\mathbb{E}(\tilde{\eta}_J^{BB}))\big|\leqslant\mathbb{E}\big|\eta_J^{BB}-\mathbb{E}(\tilde{\eta}_J^{BB})\big|\\
&=\mathbb{E}\big|\eta_J^{BB}-\tilde{\eta}_J^{BB}+\tilde{\eta}_J^{BB}-\mathbb{E}(\tilde{\eta}_J^{BB})\big|\leqslant\mathbb{E}\big|\eta_J^{BB}-\tilde{\eta}_J^{BB}\big|+(\mathrm{Var}(\tilde{\eta}_J^{BB}))^{1/2}\\
&\leqslant\frac{C}{\log(N)}+(\mathrm{Var}(\tilde{\eta}_J^{BB}))^{1/2}
\end{aligned}$$

and the last step is therefore to bound the variance of $\tilde{\eta}_J^{BB}$. By (2.49) and (3.13) from [36] we have

$$\mathrm{Var}(\tilde{\eta}_J^{BB})\leqslant 4r_2^2\max\Big\{\mathrm{Var}\Big[\sum_{l=J}^{2N-1}\frac{1}{r_2}p_{BB}^{r_2}(l,l)H_l\Big],\mathrm{Var}\Big[\sum_{l=J}^{2N-1}\frac{1}{r_2}p_{BB}^{r_2}(l,l+1)U_l\Big]\Big\},\tag{2.52}$$

and thus it suffices to bound each of the inner variances by a constant. Define

$$\begin{aligned}
a_l^{BB}&:=\frac{1}{r_2}p_{BB}^{r_2}(l,l+1)=\frac{l}{2N(l+1)}<\frac{1}{2N},\\
b_l^{BB}&:=\frac{1}{r_2}p_{BB}^{r_2}(l,l)=\frac{\beta_l l}{2N(2N-l)}\leqslant\frac{l^2}{4N^3}.
\end{aligned}$$

Then, (3.17) in [36] gives us a bound on the covariance of the number of up-jumps and similar

as in (3.18), [36], we deduce

$$\begin{aligned}\mathrm{Var}\Big[\sum_{l=J}^{2N-1}\frac{1}{r_2}p_{BB}^{r_2}(l,l+1)U_l\Big] &= \sum_{l=J}^{2N-1}\sum_{k=J}^{2N-1}a_l^{BB}a_k^{BB}\mathrm{Cov}(U_l,U_k)\\ &\leqslant \frac{s\sqrt{2}}{s^2}\sum_{l=J}^{2N-1}\sum_{k=l}^{2N-1}\frac{1}{(2N)^2}(1-s)^{(k-l)/2}\leqslant\frac{c}{N}.\end{aligned}$$

The bound on the covariance of H_l in [36] reads

$$\mathrm{Cov}(H_l,H_k)\leqslant c\frac{(1-p_l)(1-p_k)}{p_lp_k}(1-s)^{(k-l)/2}, \tag{2.53}$$

for $l\leqslant k$ and with $p_l=\mathbb{P}(X_t\neq X_{t-1}\mid X_{t-1}=l)=l(2N-l)(2-s)/(2N)^2$. As

$$\frac{b_l^{BB}(1-p_l)}{p_l}=\frac{\beta_l l}{2N(2N-l)}\frac{l(2N-l)}{(2N)^2\beta_l}\frac{(2N)^2}{l(2N-l)(2-s)}=\frac{l}{2N(2N-l)(2-s)}\leqslant\frac{l}{2N(2N-l)}$$

this fraction increases with l and we derive similar as in (3.19),[36]:

$$\begin{aligned}\mathrm{Var}\Big[\sum_{l=J}^{2N-1}\frac{1}{r_2}p_{BB}^{r_2}(l,l)H_l\Big] &= \sum_{l=J}^{2N-1}\sum_{k=J}^{2N-1}b_l^{BB}b_k^{BB}\mathrm{Cov}(H_l,H_k)\\ &\leqslant 2c\sum_{l=J}^{2N-1}\sum_{k=l}^{2N-1}\frac{b_l^{BB}(1-p_l)}{p_l}\frac{b_k^{BB}(1-p_k)}{p_k}(1-s)^{(k-l)/2}\\ &\leqslant 2c\sum_{l=J}^{2N-1}\sum_{k=l}^{2N-1}\frac{l}{2N(2N-l)}\frac{k}{2N(2N-k)}(1-s)^{(k-l)/2}\\ &= 2c\sum_{l=J}^{2N-1}\frac{l}{2N(2N-l)(1-s)^{l/2}}\sum_{k=l}^{2N-1}\frac{k}{2N(2N-k)}(1-s)^{k/2}.\end{aligned} \tag{2.54}$$

The straightforward attempt to bound this double sum leads to an upper bound of order $\mathcal{O}(\log(N))$ instead of $\mathcal{O}(1)$. We therefore need to do more evolved calculations. Let $g:=(1-s)^{1/2}$, then $g<1$ as $0<s<1$. Consider first the inner sum over the k. We will treat it as a lower sum in the context of Riemann integration (see the proof of Lemma 2.30 in the Appendix for more details concerning this topic) and bound it by an integral.

$$\sum_{k=l}^{2N-1}\frac{kg^k}{2N(2N-k)}\leqslant g^{2N-1}+\sum_{k=l}^{2N-2}\frac{g^k}{2N-k}\leqslant g^{2N-1}+\int_l^{2N-1}\frac{g^{x-1}}{2N-x}\,\mathrm{d}x.$$

The integral then can be calculated using substitution and integration by parts:

$$\int_l^{2N-1}\frac{g^{x-1}}{2N-x}\,\mathrm{d}x=g^{2N-1}\int_1^{2N-l}\frac{g^{-y}}{y}\,\mathrm{d}y=g^{2N-1}\Big\{\Big[\frac{-g^{-y}}{y\log g}\Big]_1^{2N-l}-\int_1^{2N-l}\frac{g^{-y}}{y^2\log g}\,\mathrm{d}y\Big\}.$$

Keep in mind that $g < 1$ and thus $-\log(g) > 0$. For $1 \leqslant y \leqslant 2N - l$, the function $f : y \mapsto g^{-y}/y^2$ is always bounded by $g^{-2N+l}/(2N-l)^2$ as, on the one hand, it is monotonously increasing for all $y > 2/(-\log g)$:

$$f'(y) = \frac{-\log(g)g^{-y}}{y^2} - \frac{2g^{-y}}{y^3} = -\frac{g^{-y}}{y^2}\Big(\log(g) + \frac{2}{y}\Big) > 0 \quad \Leftrightarrow \quad y > \frac{2}{-\log(g)},$$

and on the other hand, $f'(y) < 0$ for all $1 \leqslant y \leqslant 2/(-\log g)$ with $f(1) = 1/g$. Consequently,

$$\begin{aligned}
&\Big|\sum_{k=l}^{2N-1} \frac{k}{2N(2N-k)} g^k\Big| \leqslant g^{2N-1}\Big|1 + \Big[\frac{-g^{-y}}{y\log g}\Big]_1^{2N-l} + (2N-l)\frac{g^{-2N+l}}{(2N-l)^2\log g}\Big| \\
&= g^{2N-1}\Big\{1 + \frac{g^{-2N+l}}{(2N-l)\log g} + \frac{1}{g\log g} + \frac{g^{-2N+l}}{(2N-l)\log g}\Big\} \leqslant \frac{2}{\log(g)}\Big\{\frac{g^{l-1}}{2N-l} + g^{2N-2}\Big\}.
\end{aligned}$$

Inserting this in the sum over all l from (2.54) gives us the following:

$$\begin{aligned}
&\sum_{l=J}^{2N-1} \frac{l}{2N(2N-l)g^l} \frac{2}{\log(g)}\Big\{\frac{g^{l-1}}{2N-l} + g^{2N-2}\Big\} \leqslant \sum_{l=J}^{2N-1}\Big\{\frac{2}{(2N-l)^2 g\log(g)} + \frac{2g^{2N-l-2}}{(2N-l)\log(g)}\Big\} \\
&\leqslant \frac{2}{g\log(g)} \sum_{l=1}^{2N-J} \frac{1}{l^2} + \frac{2g^{-2}}{\log(g)} \sum_{l=1}^{2N-J} g^l \leqslant C,
\end{aligned}$$

for some constant C as we have $1/l^2$ in the first sum and $g < 1$ in the second. As a result, the variance in (2.54) can be bounded by a constant:

$$\mathrm{Var}\Big[\sum_{l=J}^{2N-1} \frac{1}{r_2} p_{BB}^{r_2}(l,l) H_l\Big] \leqslant 2c \sum_{l=J}^{2N-1} \frac{l}{2N(2N-l)g^l} \sum_{k=l}^{2N-1} \frac{kg^k}{2N(2N-k)} \leqslant C'.$$

With the factor r_2^2 in front of the maximum in (2.52), we therefore get

$$\mathrm{Var}(\tilde{\eta}_J^{BB}) \leqslant \frac{c}{(\log N)^2},$$

and hence, together with (2.48):

$$\big|\mathbb{E}(\exp(-\eta_J^{BB})) - \exp(-\mathbb{E}(\tilde{\eta}_J^{BB}))\big| \leqslant \frac{c}{\log(N)}.$$

Using the above and the result from (2.51) in (2.47) ends the proof of the first statement. The second equation follows from the first when considering $J = 1$ and using the result from (A.1) in the proof of Lemma 2.30. □

Recall the notation $\mathcal{S}_i$ from (2.23) which denoted the event that the two neutral loci of individual i separated within the B-population, with complement $\mathcal{S}_i^c$. Due to Proposition 2.9,

we do not have to worry about any chronological overlapping of a migration event and $\mathcal{S}_i$. From now on,

$$\mathbb{P}^{\mathcal{S}_i}(\cdot) = \mathbb{P}(\cdot \mid \mathcal{S}_i), \quad \text{and} \quad \mathbb{P}^{\mathcal{S}_i^c}(\cdot) = \mathbb{P}(\cdot \mid \mathcal{S}_i^c).$$

and all calculations will consider only times $t \leqslant \tau_N$. We will describe for all of the five potential block types the possible (and probable) events leading to the particular ancestral relation of the two neutral loci of an individual i. A key argument for the exclusion of many event types will be that by Proposition 2.11, a recombination from the b- back into the B-population only happens with a probability smaller than $c/(\log(N))^2$. Hence, once the ancestral line of some locus goes back to a b-individual, it will never connect to a B-individual again.
Likewise important is that, as a consequence of Proposition 2.13, once the event $\mathcal{S}_i$ holds true, the lineages of $(i,1)$ and $(i,2)$ will never coalesce again, except if they both stay in the mutant population during the whole sweep.
In the following list, which refers to the ancestral relationships as defined right before Proposition 2.16, we will ignore events which have probability $\leqslant C/\log(N)$ for some constant C.

1. $\mathcal{A}_0$: this happens only if $R(i,1) = R(i,2) = -\infty$, independent of whether $\mathcal{S}_i$ or $\mathcal{S}_i^c$ holds true. However, the probabilities differ depending on the value of $R_B^{\mathrm{rec}}(i)$.

2. $\mathcal{A}_1$: again, independent of $\mathcal{S}_i$ or $\mathcal{S}_i^c$, we need $R(i,1) = -\infty$ and $R(i,2) \geqslant 0$, but again, the probabilities differ depending on the value of $R_B^{\mathrm{rec}}(i)$.

3. $\mathcal{A}_2$: by Propositions 2.10 and 2.11, the event $\mathcal{S}_i$ is necessary for this relation of the two neutral loci. Further, we need that $R(i,2) = -\infty$ and $R(i,1) \geqslant 0$.

4. $\mathcal{A}_3$: in contrast to $\mathcal{A}_2$ we here, by Proposition 2.13, necessarily need $\mathcal{S}_i^c$ to be true. In addition it is necessary that $R(i,1) = R(i,2) \geqslant 0$ and $R_b^{\mathrm{rec}}(i) = -\infty$.

5. $\mathcal{A}_4$: either $(i,1)$ and $(i,2)$ recombine into the b-population at different times, $R(i,1) > R(i,2) \geqslant 0$ (only in case of $\mathcal{S}_i$) or vices versa, or they recombine together, that is, $R(i,1) = R(i,2) \geqslant 0$, and in addition $R_b^{\mathrm{rec}}(i) \geqslant 0$.

We will focus on calculating the probability of $\mathcal{A}_4$ (corresponds to *(i)* of Proposition 2.16), and deduce the other statements through easy adaption.

Probability of the ancestral relation $\mathcal{A}_4$.

Let us first consider $\mathcal{A}_4$ conditional on the event that the two loci of individual i are still connected to the same individual at time τ_N. By Propositions 2.10 and 2.13, there are now exactly two ways for the N1 and N2 locus to appear as two unmarked singletons in the partition of the sample. Note again that we only consider the time span $[0, \tau_N]$ here.

e1) $R(i,2) > R(i,1) \geqslant 0$: first $(i,2)$ recombines into the b-population, then $(i,1)$ recombines into the b-population (and connects to a different individual than $(i,2)$). This event will be called $[2,1]^{rec}_{b,i}$.

e2) $R(i,2) = R(i,1) > R^{\mathrm{rec}}_b(i) \geqslant 0$: the tuple $(i,1),(i,2)$ recombines into the b-population, afterwards a second recombination splits them and $(i,2)$ is attached to another b-individual. This event is denoted by $[12,2]^{rec}_{b,i}$.

As we know from Propositions 2.11 and 2.12 the events which would annihilate the singleton status after an event of type e1) or e2) are of order at most $1/\log(N)^2$ and hence we do not need to account for them. All in all,

$$\mathbb{P}(\mathcal{A}_4 \mid \mathcal{S}_i^c) = \mathbb{P}^{\mathcal{S}_i^c}([2,1]^{rec}_{b,i}) + \mathbb{P}^{\mathcal{S}_i^c}([12,2]^{rec}_{b,i}) + \mathcal{O}(1/\log(N)).$$

If the two loci of an individual i split within the B-population, that is, if $\mathcal{S}_i$ is true, then the chronological order of the migration of the loci into the b-population is not important. Thus, the event $[2,1]^{rec}_{b,i}$ is again one possibility, in the same way as the following:

e3) First $(i,1)$ recombines into the b-population, then $(i,2)$ recombines into the b-population (and connects to a different individual than $(i,1)$). This event will be called $[1,2]^{rec}_{b,i}$.

In this case, we get

$$\mathbb{P}(\mathcal{A}_4 \mid \mathcal{S}_i) = \mathbb{P}^{\mathcal{S}_i}([2,1]^{rec}_{b,i}) + \mathbb{P}^{\mathcal{S}_i}([1,2]^{rec}_{b,i}) + \mathcal{O}(1/\log(N)),$$

and hence in the end we have

$$\begin{aligned}\mathbb{P}(\mathcal{A}_4) &= \mathbb{P}^{\mathcal{S}_i}(\mathcal{A}_4)\cdot\mathbb{P}(\mathcal{S}_i) + \mathbb{P}^{\mathcal{S}_i^c}(\mathcal{A}_4)\cdot\mathbb{P}(\mathcal{S}_i^c)\\ &= \big(\mathbb{P}^{\mathcal{S}_i}([2,1]^{rec}_{b,i}) + \mathbb{P}^{\mathcal{S}_i}([1,2]^{rec}_{b,i})\big)\cdot\mathbb{P}(\mathcal{S}_i)\\ &\quad + \big(\mathbb{P}^{\mathcal{S}_i^c}([2,1]^{rec}_{b,i}) + \mathbb{P}^{\mathcal{S}_i^c}([12,2]^{rec}_{b,i})\big)\cdot\mathbb{P}(\mathcal{S}_i^c) + \mathcal{O}(1/\log(N)).\end{aligned} \tag{2.55}$$

We start with the case of $\mathcal{S}_i^c$, and first consider the event from e2), then e1) and finally add them up. In the following calculations we will assume that i was sampled at τ_N and hence, by definition its two neutral loci are still linked to each other at that time, which enables us to use the measure $\mathbb{P}$ instead of $\mathbb{P}^{\mathcal{S}_i^c}$.

Lemma 2.31. *With the notation from* e2) *we get*

$$\begin{aligned}\mathbb{P}([12,2]^{rec}_{b,i}) = \sum_{l=1}^{N}\Big\{\frac{r_1(1-r_2)}{s}\cdot\exp\Big(-\frac{r(2)}{s}\sum_{m=l+1}^{N}\frac{1}{m}\Big)(1-(1-s)^{l+1})\cdot\frac{1}{l+1}\Big\}\\ \cdot\Big\{1-\exp\Big(-\frac{r_2}{s}\cdot\sum_{k=1}^{l}\frac{1}{k}\Big)\Big\} + \mathcal{O}\Big(\frac{1}{\log N}\Big).\end{aligned}$$

Proof. Recall the definition of $R_b^{\mathrm{rec}}(i)$ from (2.13) and note that by Proposition 2.12 we have that

$$\mathbb{P}(R_b^{\mathrm{rec}}(i) \geqslant 0) = \mathbb{P}(R(i,1) = R(i,2) \geqslant 0, R_b^{\mathrm{rec}}(i) \geqslant 0) + \mathcal{O}\Big(\tfrac{\log(N)}{N}\Big).$$

With this, we get

$$\begin{aligned}\mathbb{P}([12,2]_{b,i}^{rec}) &= \mathbb{P}(R(i,1) = R(i,2) \geqslant 0, R_b^{\mathrm{rec}}(i) \geqslant 0) + \mathcal{O}\Big(\tfrac{\log(N)}{N}\Big) \\ &= \sum_{l=1}^{2N} \mathbb{P}(R_b^{\mathrm{rec}}(i) \geqslant 0 \mid R(i,1) = R(i,2) \geqslant 0, X_{R(i,2)} = l) \\ &\qquad \cdot \mathbb{P}(R(i,1) = R(i,2) \geqslant 0, X_{R(i,2)} = l) + \mathcal{O}\Big(\tfrac{\log(N)}{N}\Big).\end{aligned}$$

We will abbreviate the event considered in the second probability by

$$\mathcal{R}(i;l) := \{R(i,1) = R(i,2) \geqslant 0, X_{R(i,2)} = l\}, \quad l \in [N]. \tag{2.56}$$

The following Lemma states the result of the calculation of $\mathbb{P}(R_b^{\mathrm{rec}}(i) \geqslant 0 \mid \mathcal{R}(i;l))$.

Lemma 2.32. *Recall the notation from Lemma 2.27, the times τ_l and τ_l^* defined in* (2.3) *and* (2.4)*, respectively, and define*

$$\theta_t^{bb} := p_{bb}^{r_2}(X_t, X_{t+1}), \quad \textit{and} \quad \eta_l := \sum_{t'=1}^{\tau_l - 1} \theta_{t'}^{bb}, \ \eta_l^* := \sum_{t'=1}^{\tau_l^* - 1} \theta_{t'}^{bb}, \ \tilde{\eta}_l := \sum_{t'=1}^{\tau_l^* - 1} \theta_{t'}^{bb} \mathbf{1}_{\{X_{t'} \leqslant l\}}.$$

Then we have the following statements, where C is some constant which may vary from line to line:

1. $\mathbb{P}(R_b^{rec}(i) \geqslant 0 \mid \mathcal{R}(i;l)) \geqslant 1 - \mathbb{E}\Big(\mathrm{e}^{-\eta_l}\Big) + \dfrac{C}{(\log(N))^2}$,
   $\mathbb{P}(R_b^{rec}(i) \geqslant 0 \mid \mathcal{R}(i;l)) \leqslant 1 - \mathbb{E}\Big(\mathrm{e}^{-\eta_l^*}\Big) + \dfrac{C}{(\log(N))^2}$.
2. $\mathbb{E}(\tilde{\eta}_l - \eta_l) \leqslant \dfrac{C}{l \log(N)}$, $\mathbb{E}(\eta_l^* - \tilde{\eta}_l) \leqslant \dfrac{C}{l \log(N)}$.
3. $\mathbb{E}(\tilde{\eta}_l) = \dfrac{r_2}{s} \displaystyle\sum_{k=1}^{l} \frac{1}{k} + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)$.
4. $\Big|\mathbb{E}(\mathrm{e}^{-\tilde{\eta}_l}) - \mathrm{e}^{-\mathbb{E}(\tilde{\eta}_l)}\Big| \leqslant C'/\log(N)$.

From this it follows that

$$\mathbb{P}(R_b^{rec}(i) \geqslant 0 \mid \mathcal{R}(i;l)) = 1 - \exp\Big(-\frac{r_2}{s}\sum_{k=1}^{l}\frac{1}{k}\Big) + \mathcal{O}(1/\log(N)).$$

The proof follows the recipe described on page 65 and uses essentially the same methods as were presented throughout the proof of Proposition 2.14. However, the proof here is more delicate as we have the additional information that $X_{R(i,2)} = l$. This makes it necessary to perform two Poisson approximations, one as a lower and one as an upper bound, as stated in 1. of Lemma 2.32, as we only know that we are in state l but not that it is the first time. Hence we make the detour via the bounds with the first and last time the state l was visited by X. Claim 2. however says that both approximations can again be approximated using the same exponent $\tilde{\eta}$.

Proof. With $p_{bb}^{r_2}(k,k')$ from (2.40) we have a lower and upper bound as follows:

$$1 - \prod_{t=1}^{\tau_l}(1-\theta_{t-1}^{bb}) \leqslant \mathbb{P}(R_b^{\mathrm{rec}}(i) \geqslant 0 \mid \mathcal{R}(i;l), X) \leqslant 1 - \prod_{t=1}^{\tau_l^*}(1-\theta_{t-1}^{bb}).$$

After taking expectations, the above reads as follows

$$1 - \mathbb{E}\Big(\prod_{t=1}^{\tau_l}(1-\theta_{t-1}^{bb})\Big) \leqslant \mathbb{P}(R_b^{\mathrm{rec}}(i) \geqslant 0 \mid \mathcal{R}(i;l)) \leqslant 1 - \mathbb{E}\Big(\prod_{t=1}^{\tau_l^*}(1-\theta_{t-1}^{bb})\Big), \tag{2.57}$$

1. Consider first a Poisson approximation of the lower bound. Applying the limit theorem from [15] which is stated in Lemma A.2 yields

$$\Big|1 - \prod_{t=1}^{\tau_l}(1-\theta_{t-1}^{bb}) - \Big(1-\mathrm{e}^{-\eta_l}\Big)\Big| \leqslant \sum_{t'=1}^{\tau_l-1}(\theta_{t'}^{bb})^2.$$

Taking expectations and using the results from Lemma 2.27 as well as (2.37) from Lemma 2.23, this inequality gives us

$$\begin{aligned}
&\Big|1 - \mathbb{E}\Big(\prod_{t=1}^{\tau_l}(1-\theta_{t-1}^{bb})\Big) - \Big(1-\mathbb{E}\Big(\mathrm{e}^{-\eta_l}\Big)\Big)\Big| \leqslant \mathbb{E}\Big(\sum_{t'=1}^{\tau_l-1}(\theta_{t'}^{bb})^2\Big) \leqslant \mathbb{E}\Big(\sum_{t'=1}^{\tau}(\theta_{t'}^{bb})^2\Big) \\
&= \sum_{k=1}^{2N-1}\mathbb{E}(D_k \cdot p_{bb}^{r_2}(k,k-1)^2 + H_k \cdot p_{bb}^{r_2}(k,k)^2) \\
&= \sum_{k=1}^{2N-1}\mathbb{E}(D_k)\cdot\frac{(r_2)^2}{(2N)^2}\frac{(2N-k)^2}{(2N-k+1)^2} + \mathbb{E}(H_k)\cdot\frac{(r_2)^2}{(2N)^2}\frac{(2N-k)^2\beta_k^2}{k^2}\Big)
\end{aligned}$$

$$\leqslant \sum_{k=1}^{2N-1} \frac{(r_2)^2}{(2N)^2}\frac{1}{s}\Big(\underbrace{\frac{(2N-k)^2}{(2N-k+1)^2}}_{\leqslant 1} + \underbrace{\beta_k \frac{(2N-k)^2}{k^2}}_{\leqslant (2N)^2\frac{1}{k^2}}\Big)$$

$$\leqslant \frac{r_2^2}{s}\underbrace{\sum_{k=1}^{2N-1}\Big(\frac{1}{(2N)^2}+\frac{1}{k^2}\Big)}_{\leqslant c} \leqslant \frac{C}{(\log(N))^2}. \tag{2.58}$$

For the upper bound note that we can again use the sum from 1 to τ as a bound.

$$\Big|1-\mathbb{E}\Big(\prod_{t=1}^{\tau_l^*}(1-\theta_{t-1}^{bb})\Big)-\Big(1-\mathbb{E}\Big(\mathrm{e}^{-\eta_l^*}\Big)\Big)\Big| \leqslant \sum_{t'=1}^{\tau_l^*-1}\mathbb{E}\Big((\theta_{t'}^{bb})^2\Big) \leqslant \sum_{t'=1}^{\tau}\mathbb{E}\Big((\theta_{t'}^{bb})^2\Big) \leqslant \frac{C}{(\log N)^2}.$$

Both lower and upper bound in (2.57) can therefore be approximated by $1-\mathbb{E}\Big(\mathrm{e}^{-\mu}\Big)$ with $\mu=\eta_l$ or $\mu=\eta_l^*$, respectively.

2. Next, we show that both η_l and η_l^* can be approximated by $\tilde{\eta}_l$. Indeed, on the one hand we have

$$\begin{aligned}
\mathbb{E}(\tilde{\eta}_l-\eta_l) &= \mathbb{E}\Big(\sum_{t'=\tau_l}^{\tau_l^*-1}\theta_{t'}^{bb}\mathbf{1}_{\{X_{t'}\leqslant l\}}\Big) \leqslant \sum_{k=1}^{l}\Big\{p_{bb}^{r_2}(k,k)\mathbb{E}(H_{k,l})+p_{bb}^{r_2}(k,k-1)\mathbb{E}(D_{k,l})\Big\}\\
&\leqslant \sum_{k=1}^{l}\Big\{\frac{r_2}{2N}\frac{(2N-k)\beta_k}{k}\cdot\frac{(1-s)^{l-k}}{s\beta_k}+\frac{r_2}{2N}\frac{2N-k}{2N-k+1}\cdot\frac{(1-s)^{l-k+1}}{s}\Big\}\\
&\leqslant \frac{r_2}{s}\sum_{k=1}^{l}\frac{(1-s)^{l-k}}{k} = \frac{r_2(1-s)^l}{s}\sum_{k=1}^{l}\frac{1}{k}\frac{1}{(1-s)^k} \leqslant \frac{r_2(1-s)^l}{s}\frac{c}{l}\frac{1}{(1-s)^l}\\
&\leqslant \frac{C}{l\log(N)},
\end{aligned}$$

where we used Lemma A.1 (Lemma 3.5 in [36]) in order to bound the sum. On the other hand it holds that

$$\begin{aligned}
\mathbb{E}(\eta_l^*-\tilde{\eta}_l) &= \mathbb{E}\Big(\sum_{t'=1}^{\tau_l^*-1}\theta_{t'}^{bb}\mathbf{1}_{\{X_{t'}>l\}}\Big)\\
&\leqslant \sum_{k=l+1}^{2N}\Big\{p_{bb}^{r_2}(k,k)\mathbb{E}(H_{2N-k,2N-l})+p_{bb}^{r_2}(k,k-1)\mathbb{E}(D_{2N-k+1,2N-l})\Big\}\\
&\leqslant \sum_{k=l+1}^{2N}\Big\{\frac{r_2}{2N}\frac{(2N-k)\beta_k}{k}\cdot\frac{(1-s)^{k-l}}{s\beta_k}+\frac{r_2}{2N}\frac{(1-s)^{k-l}}{s}\Big\}\\
&\leqslant \frac{r_2}{s}\sum_{k=l+1}^{2N}\frac{(1-s)^{k-l}}{k} \leqslant \frac{r_2}{s(l+1)}\sum_{k=0}^{\infty}(1-s)^k \leqslant \frac{C}{l\log(N)}.
\end{aligned}$$

3. In the third step, we calculate the expected value of $\tilde{\eta}$:

$$\begin{aligned}
\mathbb{E}(\tilde{\eta}_l) &= \mathbb{E}\Big(\sum_{t=1}^{\tau_l^*-1} \theta_t^{bb}\mathbf{1}_{\{X_t\leqslant l\}}\Big) = \sum_{k=1}^{l}\Big\{p_{bb}^{r_2}(k,k)\mathbb{E}(H_k) + p_{bb}^{r_2}(k,k-1)\mathbb{E}(D_k)\Big\} \\
&= \frac{r_2}{s}\sum_{k=1}^{l}(1-(1-s)^k)\Big\{\frac{(2N-k)}{2Nk}\frac{1-(1-s)^{2N-k}}{1-(1-s)^{2N}} + \frac{2N-k}{2N(2N-k+1)}\frac{1-(1-s)^{2N-k+1}}{1-(1-s)^{2N}}\Big\} \\
&\quad - \frac{r_2}{2N}\sum_{k=1}^{l}\frac{2N-k}{2N(2N-k+1)} \\
&= \frac{r_2}{s}\sum_{k=1}^{l}(1-(1-s)^k)\frac{(2N-k)}{2N}\Big\{\frac{2N-k+1+k}{k(2N-k+1)}\Big\} + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big) \\
&= \frac{r_2}{s}\sum_{k=1}^{l}\frac{1-(1-s)^k}{k} + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big) = \frac{r_2}{s}\sum_{k=1}^{l}\frac{1}{k} + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big).
\end{aligned}$$

Here we used similar simplifying approximations as in (2.50).

4. In order to show the last inequality, we have to bound the variance of $\tilde{\eta}_l$. Analogously to the proof of Proposition 2.14, we define

$$\begin{aligned}
a_k &:= \frac{1}{r_2}p_{bb}^{r_2}(k,k-1) = \frac{2N-k}{2N(2N-k+1)} \leqslant \frac{1}{2N}, \\
b_k &:= \frac{1}{r_2}p_{bb}^{r_2}(k,k) = \frac{(2N-k)^2}{2N(k^2+(2N-k)^2+sk(2N-k))} \leqslant \frac{(2N-k)^2}{4N^3}.
\end{aligned}$$

Note that the here defined a_k is smaller than the corresponding a_k from Schweinsberg and Durrett. From equation (3.13) of [36] we have

$$\begin{aligned}
\mathrm{Var}(\tilde{\eta}_l) &\leqslant 4r_2^2\max\Big\{\mathrm{Var}\Big(\sum_{k=1}^{l}a_k\underbrace{D_k}_{=U_k-1}\Big), \mathrm{Var}\Big(\sum_{k=1}^{l}b_kH_k\Big)\Big\} \\
&= 4r_2^2\max\Big\{\mathrm{Var}\Big(\sum_{k=1}^{l}a_kU_k\Big), \mathrm{Var}\Big(\sum_{k=1}^{l}b_kH_k\Big)\Big\}.
\end{aligned}$$

Thus, as the a_k are bounded uniformly by $1/2N$, the variance in the first argument can be treated in the exact same way and with the same result as in [36]:

$$\mathrm{Var}\Big(\sum_{k=1}^{l}a_kU_k\Big) \leqslant \mathrm{Var}\Big(\sum_{k=1}^{2N-1}a_kU_k\Big) \leqslant C.$$

For bounding the variance in the second argument, consider the following:

$$\mathrm{Var}\Big(\sum_{k=1}^{l}b_kH_k\Big) = \sum_{k=1}^{l}\sum_{k'=1}^{l}b_kb_{k'}\mathrm{Cov}(H_k,H_{k'}) \leqslant 2\cdot\sum_{k=1}^{l}\sum_{k'=k}^{l}b_kb_{k'}\mathrm{Cov}(H_k,H_{k'})$$

$$\leqslant C \cdot \sum_{k=1}^{l} \sum_{k'=k}^{l} b_k b_{k'} \frac{1}{p_k p_{k'}} (1-s)^{(k'-k)/2},$$

where we used (2.53) which states a bound on the covariance of the number of holds from [36]. Here $p_k = \frac{k(2N-k)(2-s)}{(2N)^2}$ thus

$$\frac{b_k}{p_k} \leqslant \frac{(2N-k)^2}{4N^3} \frac{(2N)^2}{k(2N-k)(2-s)} = \frac{2N-k}{N(2-s)k} \leqslant \frac{1}{(2-s)} \cdot \frac{2}{k}.$$

With this we can bound the variance of the holds by some constant:

$$\begin{aligned} \mathrm{Var}\Big(\sum_{k=1}^{l} b_k H_k\Big) &\leqslant C \sum_{k=1}^{l} \sum_{k'=k}^{l} \frac{b_k b_{k'}}{p_k p_{k'}} (1-s)^{(k'-k)/2} \leqslant \frac{4c}{(2-s)^2} \sum_{k=1}^{l} \sum_{k'=k}^{l} \frac{1}{kk'} (1-s)^{(k'-k)/2} \\ &= \frac{4c}{(2-s)^2} \sum_{k=1}^{l} \frac{1}{k} \underbrace{\sum_{k'=0}^{l-k} \frac{1}{k'+k} (1-s)^{(k')/2}}_{\leqslant \frac{1}{k} \cdot c'} \leqslant C' \cdot \sum_{k=1}^{l} \frac{1}{k^2} \leqslant C. \end{aligned}$$

As r_2^2 is bounded by a constant times $1/(\log(N))^2$, statement number three follows from the Mean Value Theorem and Jensen's inequality as described on page 65:

$$\Big|\mathbb{E}(\mathrm{e}^{-\tilde{\eta}_l}) - \mathrm{e}^{-\mathbb{E}(\tilde{\eta}_l)}\Big| \leqslant \mathbb{E}\Big|(\mathrm{e}^{-\tilde{\eta}_l} - \mathrm{e}^{-\mathbb{E}(\tilde{\eta}_l)}\Big| \leqslant \mathbb{E}\Big|(\tilde{\eta}_l - \mathbb{E}(\tilde{\eta}_l)\Big| \leqslant (\mathrm{Var}(\tilde{\eta}_l))^{1/2} \leqslant \frac{C}{\log(N)}. \quad (2.59)$$

Combining the statements from steps 1. to 4. and recalling (2.44) results in

$$\Big|\mathbb{P}(R_b^{\mathrm{rec}}(i) \geqslant 0 \mid \mathcal{R}(i;l)) - \Big(1 - \exp\Big(-\frac{r_2}{s} \sum_{k=1}^{l} \frac{1}{k}\Big)\Big)\Big| \leqslant \frac{C}{\log(N)}.$$

□

It remains to calculate $\mathbb{P}(\mathcal{R}(i;l))$ for an individual i sampled at time τ_N.

Lemma 2.33.

$$\begin{aligned} \mathbb{P}(\mathcal{R}(i;l)) &= \frac{r_1(1-r_2)}{s}\Big(\frac{1-(1-s)^{2N-l}}{1-(1-s)^{2N}}\Big) \exp\Big(-\frac{r(2)}{s} \sum_{m=l+1}^{N} \frac{1}{m}\Big) \\ &\Big[\frac{(2N-l)(1-(1-s)^{l+1})}{2N(l+1)} + \exp\Big(-\frac{r(2)}{s}\frac{1}{l}\Big)\frac{1-(1-s)^l}{2N}\Big] + \mathcal{O}\Big(\frac{1}{(l+1)\cdot \log N}\Big). \end{aligned}$$

Proof. By decomposing the event and conditioning appropriately we get

$$\mathbb{P}(\mathcal{R}(i;l)) = \sum_{t=0}^{\infty} \mathbb{P}(R(i,2) = t \leqslant \tau_N, \text{ and } X_t = l, R(i,1) = R(i,2))$$

$$
\begin{aligned}
&= \sum_{t=0}^{\infty} \sum_{k=l,l+1} \mathbb{P}(t \leqslant \tau_N, X_t = l, X_{t+1} = k, \forall t' > t : B_{t'}(A^{t'}_{\tau_N}(i,2)) = 1 = B_{t'}(A^{t'}_{\tau_N}(i,1)),\\
&\qquad B_t(A^t_{\tau_N}(i,2)) = 0 = B_t(A^t_{\tau_N}(i,1)))\\
&= \sum_{t=0}^{\infty} \sum_{k=l,l+1} \mathbb{P}(B_t(A^t_{\tau_N}(i,2)) = 0 = B_t(A^t_{\tau_N}(i,1)) \mid t \leqslant \tau_N, X_t = l, X_{t+1} = k,\\
&\qquad \forall t' > t : B_{t'}(A^{t'}_{\tau_N}(i,1)) = 1 = B_{t'}(A^{t'}_{\tau_N}(i,2))) \cdot \mathbb{P}(t \leqslant \tau_N, X_t = l, X_{t+1} = k)\\
&\qquad \cdot \mathbb{P}(\forall t' > t : B_{t'}(A^{t'}_{\tau_N}(i,1)) = 1 = B_{t'}(A^{t'}_{\tau_N}(i,2)) \mid t \leqslant \tau_N, X_t = l, X_{t+1} = k)\\
&= \sum_{t=0}^{\infty} \sum_{k=l,l+1} p_B^{r_1(1-r_2)}(l,k) \cdot \mathbb{P}(t \leqslant \tau_N, X_t = l, X_{t+1} = k)\\
&\qquad \cdot \underbrace{\mathbb{P}(\forall t' > t : B_{t'}(A^{t'}_{\tau_N}(i,1)) = 1 = B_{t'}(A^{t'}_{\tau_N}(i,2)) \mid t \leqslant \tau_N, X_t = l, X_{t+1} = k)}_{=:P_k^t.} \qquad (2.60)
\end{aligned}
$$

As before, the idea is to do a Poisson approximation for P_k^t. Thus, we condition on the Markov chain X and get a lower and upper bound as follows

$$
\begin{aligned}
&\prod_{t'=\tau_k}^{\tau_N} \Big(1 - p_B^{r(2)}(X_{t'-1}, X_{t'})\Big)\\
&\leqslant P_k^{X,t} := \mathbb{P}(\forall t' > t : B_{t'}(A^{t'}_{\tau_N}(i,1)) = 1 = B_{t'}(A^{t'}_{\tau_N}(i,2)) \mid t \leqslant \tau_N, X_{t+1} = k, X)\\
&= \mathbb{P}((i,1),(i,2) \text{ do not migrate into the b-population during steps}\\
&\qquad \text{at times } t+1, t+2, \ldots, \tau \mid t \leqslant \tau_N, X_{t+1} = k, X)\\
&\leqslant \prod_{t'=\tau_k^*}^{\tau_N} \Big(1 - p_B^{r(2)}(X_{t'-1}, X_{t'})\Big).
\end{aligned}
$$

As not only the idea but also the calculations are very similar to the ones in the proof of Proposition 2.14 and Lemma 2.32 we will not give many details here. Define

$$
\begin{aligned}
\eta_k^- &:= \sum_{t'=\tau_k}^{\tau_N} p_B^{r(2)}(X_{t'-1}, X_{t'}), \quad \eta_k^+ := \sum_{t'=\tau_k^*}^{\tau_N} p_B^{r(2)}(X_{t'-1}, X_{t'}), \quad \text{and}\\
\tilde{\eta}_k &:= \sum_{t'=\tau_k}^{\tau_N} p_B^{r(2)}(X_{t'-1}, X_{t'}) \mathbf{1}_{\{X_{t'} \geqslant k\}}.
\end{aligned}
$$

Then, the statement for the Poisson approximation implies that

$$
\begin{aligned}
&\Big|\mathbb{E}\Big[\prod_{t'=\tau_k}^{\tau_N} \Big(1 - p_B^{r(2)}(X_{t'-1}, X_{t'})\Big)\Big] - \mathbb{E}\big[\exp\left(\eta_k^-\right)\big]\Big| \leqslant \mathbb{E}\Big[\sum_{t'=\tau_k}^{\tau_N} \big(p_B^{r(2)}(X_{t'-1}, X_{t'})\big)^2\Big]\\
&\leqslant \sum_{k=1}^{2N-1} \Big[\mathbb{E}[U_k] \cdot \frac{(r_1+r_2)^2}{(2N)^2} \frac{(2N-k)^2}{(k+1)^2} + \mathbb{E}[H_k] \cdot \beta_k^2 \frac{(r_1+r_2)^2}{(2N)^2}\Big]
\end{aligned}
$$

$$\leqslant \frac{(r_1+r_2)^2}{s}\sum_{k=1}^{2N-1}\Big[\frac{1}{(k+1)^2}+\frac{1}{(2N)^2}\Big] \leqslant \frac{C}{(\log(N))^2}, \tag{2.61}$$

and analogously,

$$\Big|\mathbb{E}\Big[\prod_{t'=\tau_k^*}^{\tau_N}\Big(1-p_B^{r(2)}(X_{t'-1},X_{t'})\Big)\Big]-\mathbb{E}\big[\exp\left(\eta_k^+\right)\big]\Big| \leqslant \frac{C}{(\log(N))^2}.$$

Furthermore, by applying (2.37) from Lemma 2.23 and Corollary 2.24 we first get

$$\begin{aligned}
\mathbb{E}(\eta_k^- - \tilde{\eta}_k) &= \mathbb{E}\Big(\sum_{t'=\tau_k}^{\tau_N} p_B^{r(2)}(X_{t'-1},X_{t'})\mathbf{1}_{\{X_{t'}<k\}}\Big)\\
&\leqslant \sum_{m=1}^{k-1}\Big\{\mathbb{E}(U_{m,k})p_B^{r(2)}(m,m+1)+\mathbb{E}(H_{m,k})p_B^{r(2)}(m,m)\Big\}\\
&\leqslant \sum_{m=1}^{k-1}\frac{(1-s)^{k-m}}{s}\Big(\frac{(r(2)(2N-m)}{2N(m+1)}+\frac{r(2)}{2N}\Big) \leqslant \frac{Cr(2)}{k} \leqslant \frac{C}{k\log(N)},
\end{aligned}$$

and similarly, by using the argument from (2.43),

$$\begin{aligned}
\mathbb{E}(\tilde{\eta}_k - \eta_k^+) &= \mathbb{E}\Big(\sum_{t'=\tau_k}^{\tau_k^*-1} p_B^{r(2)}(X_{t'-1},X_{t'})\mathbf{1}_{\{X_{t'}\geqslant k\}}\Big)\\
&\leqslant \sum_{m=k}^{2N}\Big\{\mathbb{E}(U_{2N-m-1,2N-k})p_B^{r(2)}(m,m+1)+\mathbb{E}(H_{2N-m,2N-k})p_B^{r(2)}(m,m)\Big\}\\
&\leqslant \sum_{m=k}^{2N}\frac{(1-s)^{k-m}}{s}\Big(\frac{(r(2)(2N-m)(1-s)}{2N(m+1)}+\frac{r(2)}{2N}\Big) \leqslant \frac{C}{k\log(N)}.
\end{aligned}$$

Both η_k^- and η_k^+ can therefore be approximated by $\tilde{\eta}_k$. The expectation of $\tilde{\eta}_k$ is obtained using very similar approximations as stated in (2.50) in the proof of Proposition 2.14 and as done in the proof of Lemma 3.8 in [36]. Note however, that we only consider times up to τ_N and that the following holds true for any $l = 1, 2, \ldots, N$:

$$\sum_{t=0}^{\infty}\mathbb{P}(t\leqslant\tau_N, X_t = l = X_{t+1}) = \mathbb{E}\Big[\sum_{t=0}^{\tau_N}\mathbf{1}_{\{X_t=l=X_{t+1}\}}\Big] = \mathbb{E}(H_l)-\mathbb{E}(H_{l,N}). \tag{2.62}$$

By (2.37) we can bound $\mathbb{E}(H_{l,N})$ by $(1-s)^{N-l}/(s\beta_l)$. Both considerations can be transfered

to $\sum_{t=0}^{\infty}\mathbb{P}(t\leqslant\tau_N, X_t=l, X_{t+1}=l+1)$ and the number of up-jumps. We obtain

$$\begin{aligned}\mathbb{E}(\tilde{\eta}_k) &= \sum_{m=k}^{N}\Big[\mathbb{E}\big[U_m-U_{m,N}\big]\cdot\frac{r(2)(2N-m)}{2N(m+1)}+\mathbb{E}\big[H_m-H_{m,N}\big]\cdot\beta_m\frac{r(2)}{2N}\Big]\\ &= \sum_{m=k}^{N}\Big[\mathbb{E}[U_m]\cdot\frac{r(2)(2N-m)}{2N(m+1)}+\mathbb{E}[H_m]\cdot\beta_m\frac{r(2)}{2N}\Big]+\mathcal{O}\Big(\tfrac{1}{k\log N}\Big) \qquad (2.63)\\ &= \frac{r(2)}{s}\sum_{m=k+1}^{N}\frac{1}{m}+\mathcal{O}\Big(\tfrac{1}{N}+\tfrac{1}{k\log N}\Big),\end{aligned}$$

where the error in the second line is due to $\mathbb{E}(H_{m,N})$ and $\mathbb{E}(H_{m,N})$ and follows from the formula of geometric sums and the bound on $r(2)$. Further, we use the result from Lemma 3.10 in [36] (obtained by similar calculations as performed in step 3 of the proof of Lemma 2.32) in order to bound the variance of $\tilde{\eta}_k$,

$$\mathrm{Var}(\tilde{\eta}_k)\leqslant\frac{C}{k(\log N)^2}.$$

Finally, by the same calculation as in (2.59) we obtain

$$P_k := P_k^t = \exp\Big(-\frac{r(2)}{s}\sum_{m=k}^{N}\frac{1}{m}\Big)+\mathcal{O}\Big(\tfrac{1}{(\log N)^2}+\tfrac{1}{\sqrt{k}\log N}\Big).$$

We finish the calculation with the Lemma 2.23 and using again the equality stated in (2.62):

$$\begin{aligned}\mathbb{P}(\mathcal{R}(i;l)) &= P_l\cdot p_B^{r_1(1-r_2)}(l,l)\cdot\sum_{t=0}^{\infty}\mathbb{P}(t\leqslant\tau_N, X_t=l=X_{t+1})\\ &\quad+P_{l+1}\cdot p_B^{r_1(1-r_2)}(l,l+1)\cdot\sum_{t=0}^{\infty}\mathbb{P}(t<\tau, X_t=l, X_{t+1}=l+1)\\ &= \frac{r_1(1-r_2)}{2N}\frac{(1-(1-s)^l)(1-(1-s)^{2N-l})}{s(1-(1-s)^{2N})}\cdot\exp\Big(-\frac{r(2)}{s}\sum_{m=l}^{2N}\frac{1}{m}\Big)\\ &\quad+\frac{r_1(1-r_2)(2N-l)}{2N(l+1)}\frac{(1-(1-s)^{l+1})(1-(1-s)^{2N-l})}{s(1-(1-s)^{2N})}\cdot\exp\Big(-\frac{r(2)}{s}\sum_{m=l+1}^{2N}\frac{1}{m}\Big)\\ &\quad+\mathcal{O}\Big(\tfrac{1}{(l+1)\cdot\log N}\Big)+\mathcal{O}\Big(\tfrac{1}{l\cdot(\log N)^3}+\tfrac{1}{l^{\frac{3}{2}}\cdot(\log N)^2}\Big)\\ &= \frac{r_1(1-r_2)}{s}\cdot\Big(\frac{1-(1-s)^{2N-l}}{1-(1-s)^{2N}}\Big)\cdot\exp\Big(-\frac{r(2)}{s}\sum_{m=l+1}^{2N}\frac{1}{m}\Big)\cdot\Big[\frac{(2N-l)(1-(1-s)^{l+1})}{2N(l+1)}\\ &\quad+\exp\Big(-\frac{r(2)}{s}\frac{1}{l}\Big)\cdot\frac{1-(1-s)^l}{2N}\Big]+\mathcal{O}\Big(\tfrac{1}{(l+1)\cdot\log N}\Big).\end{aligned}$$

□

Combining both results now yields the statement of Lemma 2.31.

$$\begin{aligned}
&\mathbb{P}([12,2]^{rec}_{b,i}) = \sum_{l=1}^{N} \mathbb{P}(\mathcal{R}(i;l)) \cdot \mathbb{P}(R^{\mathrm{rec}}_b(i) \geqslant 0 \mid \mathcal{R}(i;l)) \\
&= \sum_{l=1}^{N} \Big\{ \frac{r_1(1-r_2)}{s} \cdot \frac{1-(1-s)^{2N-l}}{1-(1-s)^{2N}} \cdot \exp\Big(-\frac{r(2)}{s} \sum_{m=l+1}^{N} \frac{1}{m} \Big) \cdot \Big[\frac{(2N-l)(1-(1-s)^{l+1})}{2N(l+1)} \\
&\qquad + \exp\Big(-\frac{r(2)}{s} \frac{1}{l} \Big) \cdot \frac{1-(1-s)^{l}}{2N} \Big] + \mathcal{O}\Big(\tfrac{1}{(l+1)\cdot \log N} \Big) \Big\} \\
&\qquad \cdot \Big\{ 1 - \exp\Big(-\frac{r_2}{s} \sum_{k=1}^{l} \frac{1}{k} \Big) + \mathcal{O}\Big(\tfrac{1}{\log N} \Big) \Big\} \\
&= \sum_{l=1}^{N} \Big\{ \frac{r_1(1-r_2)}{s} \cdot \exp\Big(-\frac{r(2)}{s} \sum_{m=l+1}^{N} \frac{1}{m} \Big) (1-(1-s)^{l+1}) \cdot \frac{2N+1}{2N(l+1)} \\
&\qquad + \mathcal{O}\Big(\tfrac{1}{(l+1)\cdot \log N} \Big) \Big\} \cdot \Big\{ 1 - \exp\Big(-\frac{r_2}{s} \sum_{k=1}^{l} \frac{1}{k} \Big) + \mathcal{O}\Big(\tfrac{1}{\log N} \Big) \Big\} + \mathcal{O}\Big(\tfrac{1}{N \log N} \Big) \\
&= \sum_{l=1}^{N} \Big\{ \frac{r_1(1-r_2)}{s} \exp\Big(-\frac{r(2)}{s} \sum_{m=l+1}^{N} \frac{1}{m} \Big) (1-(1-s)^{l+1}) \frac{1}{l+1} \Big\} \Big\{ 1 - \exp\Big(-\frac{r_2}{s} \sum_{k=1}^{l} \frac{1}{k} \Big) \Big\} \\
&\qquad + \mathcal{O}\Big(\tfrac{1}{\log N} \Big).
\end{aligned}$$

□

Lemma 2.34. *With the notation from* e1) *we get*

$$\begin{aligned}
\mathbb{P}([2,1]^{rec}_{b,i}) = \sum_{l=1}^{N} \Big\{ \frac{r_2(1-r_1)}{s} (1-(1-s)^{l+1}) \cdot \exp\Big(-\frac{r(2)}{s} \sum_{m=l+1}^{N} \frac{1}{m} \Big) \cdot \frac{1}{l+1} \Big\} \\
\cdot \Big\{ \sum_{k=1}^{l-1} \frac{r_1}{s} \cdot \exp\Big(-\frac{r_1}{s} \sum_{n=k+1}^{l} \frac{1}{n} \Big) \cdot (1-(1-s)^{k+1}) \cdot \frac{1}{k+1} \Big\} + \mathcal{O}\Big(\frac{1}{(\log N)} \Big).
\end{aligned}$$

Proof. In terms of the random times $R(i,1)$ and $R(i,2)$, we have

$$\begin{aligned}
\mathbb{P}([2,1]^{rec}_{b,i}) = \sum_{l=1}^{N} &\mathbb{P}(R(i,1) \geqslant 0 \mid R(i,2) > R(i,1), X_{R(i,2)} = l) \\
&\cdot \mathbb{P}(R(i,2) > R(i,1), X_{R(i,2)} = l).
\end{aligned}$$

We now consider separately the two probabilities that make up the product. Revisiting the calculation of $\mathbb{P}(\mathcal{R}(i;l))$ in the proof of Lemma 2.33, we notice that the event $R(i,1) = R(i,2)$ only influences the recombination probability. Therefore, we only need to change $p_B^{r_1(1-r_2)}(l,k)$

into $p_B^{r_2(1-r_1)}(l,k)$ and get

$$\mathbb{P}(R(i,2) > R(i,1), X_{R(i,2)} = l) = \frac{r_2(1-r_1)}{s} \cdot \frac{1-(1-s)^{2N-l}}{1-(1-s)^{2N}} \cdot \exp\Big(-\frac{r(2)}{s}\sum_{m=l+1}^{2N}\frac{1}{m}\Big)$$
$$\cdot \Big[\frac{(2N-l)(1-(1-s)^{l+1})}{2N(l+1)} + \exp\Big(-\frac{r(2)}{s}\frac{1}{l}\Big)\cdot\frac{1-(1-s)^l}{2N}\Big] + \mathcal{O}\Big(\tfrac{1}{(l+1)\cdot\log N}\Big).$$

Now consider the first probability. Similar calculations as for (2.60) lead to

$$\begin{aligned}
&\mathbb{P}(R(i,1) \geqslant 0 \mid R(i,2) > R(i,1), X_{R(i,2)} = l)\\
&= \sum_{t=0}^{\infty}\sum_{k=1}^{N}\sum_{m=k,k+1} p_B^{r_1}(k,m)\mathbb{P}(X_t = k, X_{t+1} = m, t < R(i,2) \mid R(i,2) > R(i,1), X_{R(i,2)} = l)\\
&\underbrace{\mathbb{P}(\forall t' > t : B_{t'}(A^{t'}_{\tau_N}(i,1)) = 1 \mid t < R(i,2), X_t = k, X_{t+1} = m, R(i,2) > R(i,1), X_{R(i,2)} = l)}_{=:P^{t,(1)}_{k;m,l}}.
\end{aligned}$$

The calculation of $P^{t,(1)}_{k;m,l}$ is done as in the proof of Lemma 2.33. However, we here additionally condition on the event $R(i,2) > R(i,1), X_{R(i,2)} = l$ and thus have extra information. Keep in mind that we know neither the number of visits of the state m before $t+1$ nor the number of visits of state l after time $R(i,2)$. We will use a similar idea as in the proof of Lemma 2.32 and bound the probability $P^{t,(1)}_{k;m,l}$ from above and below using at the one hand the time span between the first visit of m and the last visit of l and on the other hand the last visit in m and the first visit in l.

As we will see later, we only need to calculate an explicit value for the case where $m \leqslant l$. In the other case $m > l$ we can bound the above probability by one and still are able to drop the terms in question while making an error of order at most $\mathcal{O}(1/\log N)$. With $\theta_t^{r_1} := p_B^{r_1}(X_t, X_{t+1})$ and assuming $\tau_m^* < \tau_l$ we can write

$$\begin{aligned}
&\prod_{t'=\tau_m}^{\tau_l^*-1}\Big(1-\theta_{t'}^{r_1}\Big) \leqslant P^{X,t,(1)}_{k;m,l} := \mathbb{P}(\forall t' > t : B_{t'}(A^{t'}_{\tau_N}(i,1)) = 1 \mid t < R(i,2), X_t = k,\\
&\qquad X_{t+1} = m, R(i,2) > R(i,1), X_{R(i,2)} = l, X)\\
&= \mathbb{P}((i,1) \text{ does not migrate into the b-population during steps at times } t+1,\\
&\qquad t+2, \dots, \tau_N \mid t < R(i,2), X_t = k, X_{t+1} = m, R(i,2) > R(i,1), X_{R(i,2)} = l, X)\\
&\leqslant \prod_{t'=\tau_m^*}^{\tau_l-1}\Big(1-\theta_{t'}^{r_1}\Big). \qquad (2.64)
\end{aligned}$$

We will see later that the assumption $\tau_m^* < \tau_l$ for the upper bound can be dropped by replacing the range of the product with the minimal distance between a visit in m and a visit in l.

The idea is to do again a Poisson approximation for both products and then to compare their difference. We will first consider the lower bound. Define

$$\eta_{m,l}^{(1),-,*} := \sum_{t'=\tau_m}^{\tau_l^*-1} \theta_{t'}^{r_1}, \quad \text{and} \quad \tilde{\eta}_{m,l}^{(1),-,*} := \sum_{t'=\tau_m}^{\tau_l^*-1} \theta_{t'}^{r_1} \mathbf{1}_{\{m \leqslant X_{t'} \leqslant l\}}.$$

Very similar to the previous proof of Lemma 2.33 and in particular by the same argument as in the proof of Lemma 3.6 in [36] we get

$$\begin{aligned} &\Big| P_{k;m,l}^{X,t,(1)} - \exp(-\eta_{m,l}^{(1),-,*}) \Big| \leqslant \sum_{t'=\tau_m}^{\tau} \Big(\theta_{t'}^{r_1}\Big)^2, \quad \text{and further} \\ &\Big| P_{k;m,l}^{t,(1)} - \mathbb{E}(\exp(-\eta_{m,l}^{(1),-,*})) \Big| \leqslant \mathbb{E}\Big(\sum_{t'=1}^{\tau} \Big(\theta_{t'}^{r_1}\Big)^2 \Big) \leqslant \frac{C}{(\log N)^2}, \end{aligned} \tag{2.65}$$

for some constant C. We will now show that the expectation of the difference between $\eta_{m,l}^{(1),-,*}$ and $\tilde{\eta}_{m,l}^{(1),-,*}$ is small enough.

$$\begin{aligned} &\mathbb{E}\Big(\eta_{m,l}^{(1),-,*} - \tilde{\eta}_{m,l}^{(1),-,*}\Big) = \mathbb{E}\Big(\sum_{t'=\tau_m}^{\tau_l^*-1} \theta_{t'}^{r_1} \cdot \Big(\mathbf{1}_{\{X_{t'}<m\}} + \mathbf{1}_{\{X_{t'}>l\}} \Big) \Big) \\ \leqslant &\sum_{k=1}^{m-1} \Big(\mathbb{E}(U_{k,m}) \cdot p_B^{r_1}(k,k+1) + \mathbb{E}(H_{k,m}) \cdot p_B^{r_1}(k,k) \Big) \\ &\quad + \sum_{k=l+1}^{2N} \Big(\mathbb{E}(U_{2N-k-1,2N-l}) \cdot p_B^{r_1}(k,k+1) + \mathbb{E}(H_{2N-k,2N-l}) \cdot p_B^{r_1}(k,k) \Big) \\ \leqslant &\sum_{k=1}^{m-1} \frac{(1-s)^{m-k}}{s} \Big(\frac{r_1(2N-k)}{2N(k+1)} + \frac{r_1}{2N} \Big) + \sum_{k=l+1}^{2N} \frac{(1-s)^{k-l}}{s} \Big(\frac{r_1(2N-k)}{2N(k+1)} + \frac{r_1}{2N} \Big) \\ \leqslant &\frac{C_1 r_1}{m-1} + \frac{C_2 r_1}{2N} + \frac{C_3 r_1}{l+2} \leqslant \frac{C}{(m-1)\log(N)}, \end{aligned}$$

where we applied (2.43) and used $U_{2N-k-1,2N-l}$ and $H_{2N-k,2N-l}$ as upper bounds for the number of up-jumps and holds, respectively, and did not account for the fact that we only need to count the jumps from time τ_m on. To receive the bounds in the last line we used on the one hand Lemma A.1 and on the other hand the formula for a geometric sum. As we have

$$\mathrm{Var}(\tilde{\eta}_{m,l}^{(1),-,*}) \leqslant \mathrm{Var}\Big(\sum_{t'=\tau_m}^{\tau} \theta_{t'}^{r_1} \mathbf{1}_{\{X_{t'} \geqslant m\}} \Big),$$

and the latter is known to be at most of order $1/(m \cdot (\log(N))^2)$ by Lemma 3.10, [36], we can here as well approximate $\mathbb{E}(\exp(-\tilde{\eta}_{m,l}^{(1),-,*})$ by $\exp(-\mathbb{E}(\tilde{\eta}_{m,l}^{(1),-,*}))$ in (2.65) and still have the

order of approximation we need. The last thing to do is therefore to calculate the expected value of $\tilde{\eta}_{m,l}^{(1),-,*}$.

$$\begin{aligned}
&\mathbb{E}(\tilde{\eta}_{m,l}^{(1),-,*}) = \mathbb{E}(\sum_{t'=\tau_m}^{\tau_l^*-1} \theta_{t'}^{r_1} \mathbf{1}_{\{X_{t'}\geqslant m\}} \mathbf{1}_{\{X_{t'}\leqslant l\}}) \\
&= \sum_{k=m}^{l-1} \Big[\mathbb{E}(U_k) p_B^{(1)}(k,k+1) + \mathbb{E}(H_k) p_B^{(1)}(k,k)\Big] \\
&= \sum_{k=m}^{l-1} \Big[\frac{r_1(2N-k)}{2N(k+1)} \frac{(1-(1-s)^{2N-k})(1-(1-s)^{k+1})}{s(1-(1-s)^{2N})} \\
&\quad + \frac{r_1}{2N}\frac{1}{2-s}\Big(\frac{(1-(1-s)^{2N-k})(1-(1-s)^{k+1}) + (1-(1-s)^{2N-k+1})(1-(1-s)^k)}{s(1-(1-s)^{2N})} - 1\Big)\Big] \\
&= \frac{r_1}{s}\sum_{k=m}^{l-1} \frac{(1-(1-s)^{2N-k})}{(1-(1-s)^{2N})}\Big(\frac{(2N-k)(1-(1-s)^{k+1})}{2N(k+1)} + \frac{(1-(1-s)^k)}{2N}\Big) \\
&= \frac{r_1}{s}\frac{2N+1}{2N}\sum_{k=m}^{l-1}\frac{(1-(1-s)^k)}{k+1} + \mathcal{O}\Big(\tfrac{1}{N}\Big) = \frac{r_1}{s}\sum_{k=m}^{l-1}\frac{1}{k} + \mathcal{O}\Big(\tfrac{1}{m\log(N)}\Big).
\end{aligned}$$

The approximations in the last two steps of the above calculations follow the same reasoning as stated in (2.50).

The lower bound in (2.64) can be treated just alike. We will approximate the number of up-jumps and holds between the last hit of m and the first hit of l by the same number of jumps as above, namely those happening between τ_m and τ_l^* starting from a k between m and $l-1$. Define

$$\eta_{m,l}^{(1),*,-} := \sum_{t'=\tau_m^*}^{\tau_l-1} \theta_{t'}^{r_1}, \quad \text{and} \quad \tilde{\eta}_{m,l}^{(1),*,-} := \sum_{t'=\tau_m}^{\tau_l^*-1} \theta_{t'}^{r_1} \mathbf{1}_{\{m\leqslant X_{t'}\leqslant l\}} = \tilde{\eta}_{m,l}^{(1),-,*},$$

and consider the expectation of their difference. Similar calculations as for the upper bound lead to

$$\mathbb{E}\Big(\tilde{\eta}_{m,l}^{(1),*,-} - \eta_{m,l}^{(1),*,-}\Big) = \mathbb{E}\Big(\sum_{t'=\tau_m}^{\tau_m^*-1} \theta_{t'}^{r_1}\mathbf{1}_{\{m\leqslant X_{t'}\leqslant l\}} + \sum_{t'=\tau_l+1}^{\tau_l^*-1} \theta_{t'}^{r_1}\mathbf{1}_{\{m\leqslant X_{t'}\leqslant l\}}\Big) \leqslant \frac{C}{m\log(N)}.$$

From this calculations we see that regardless whether $\tau_m^* < \tau_l$ holds true we can approximate $\tilde{\tilde{\eta}}_{m,l}^{(1),*,-}$ which looks at the minimal number of steps between visits in m and l again with $\tilde{\eta}_{m,l}^{(1),-,*}$ with the same error bound.

This shows us, that both bounds from (2.64) can be approximated by the same expression and therefore they coincide up to terms of order less or equal to $1/(m\log(N))$. For $m \leqslant l$ we

get:

$$P^{t,(1)}_{k;m,l} = \exp\Big(-\frac{r_1}{s}\sum_{n=m}^{l}\frac{1}{n}\Big) + \mathcal{O}\Big(\tfrac{1}{(\log N)^2} + \tfrac{1}{\sqrt{m}\log N}\Big). \tag{2.66}$$

We can finish with applying again (2.43), plugging in the results from above and making similar calculations as in the end of the proof of Lemma 2.33:

$$\begin{aligned}
&\mathbb{P}(R(i,1) \geqslant 0 \mid R(i,2) > R(i,1), X_{R(i,2)} = l) = \sum_{t=0}^{\infty}\sum_{k=1}^{N}\sum_{m=k,k+1} p_B^{r_1}(k,m)\cdot P^{t,(1)}_{k;m,l} \\
&\quad\cdot \underbrace{\mathbb{P}(X_t = k, X_{t+1} = m, t < R(i,2) \mid R(i,2) > R(i,1), X_{R(i,2)} = l)}_{=\mathbb{P}(X_t=k, X_{t+1}=m, t<R(i,2)\mid X_{R(i,2)}=l)} \\
&= \sum_{k=1}^{N}\Big(p_B^{r_1}(k,k)\cdot P^{t,(1)}_{k;k,l}\cdot\mathbb{E}(H_{2N-k,2N-l}) + p_B^{r_1}(k,k+1)\cdot P^{t,(1)}_{k;k+1,l}\cdot\mathbb{E}(U_{2N-k-1,2N-l})\Big) \\
&= \sum_{k=1}^{l-1}\frac{r_1}{s}\cdot\exp\Big(-\frac{r_1}{s}\sum_{n=k+1}^{l}\frac{1}{n}\Big)\cdot(1-(1-s)^{k+1})\cdot\frac{1}{k+1} + \mathcal{O}\Big(\tfrac{1}{(\log N)}\Big),
\end{aligned}$$

Together with the first part of the calculation this yields

$$\begin{aligned}
&\mathbb{P}([2,1]^{rec}_{b,i}) \\
&= \sum_{l=1}^{N}\mathbb{P}(R(i,2) > R(i,1), X_{R(i,2)} = l)\cdot\mathbb{P}(R(i,1)\geqslant 0 \mid R(i,2) > R(i,1), X_{R(i,2)} = l) \\
&= \sum_{l=1}^{N}\Big\{\frac{r_2(1-r_1)}{s}\cdot\exp\Big(-\frac{r(2)}{s}\sum_{m=l+1}^{N}\frac{1}{m}\Big)\cdot\Big[\frac{(2N-l)(1-(1-s)^{l+1})}{2N(l+1)} + \frac{1-(1-s)^{l+1}}{2N}\Big] \\
&\qquad + \mathcal{O}\Big(\tfrac{1}{(l+1)\cdot\log N}\Big)\Big\} \\
&\qquad\cdot\Big\{\sum_{k=1}^{l-1}\frac{r_1}{s}\cdot\exp\Big(-\frac{r_1}{s}\sum_{n=k+1}^{l}\frac{1}{n}\Big)\cdot(1-(1-s)^{k+1})\cdot\frac{1}{k+1} + \mathcal{O}\Big(\tfrac{1}{(\log N)}\Big)\Big\} \\
&= \sum_{l=1}^{N}\Big\{\frac{r_2(1-r_1)}{s}(1-(1-s)^{l+1})\cdot\exp\Big(-\frac{r(2)}{s}\sum_{m=l+1}^{N}\frac{1}{m}\Big)\cdot\frac{1}{l+1}\Big\} \\
&\qquad\cdot\Big\{\sum_{k=1}^{l-1}\frac{r_1}{s}\cdot\exp\Big(-\frac{r_1}{s}\sum_{n=k+1}^{l}\frac{1}{n}\Big)\cdot(1-(1-s)^{k+1})\cdot\frac{1}{k+1}\Big\} + \mathcal{O}\Big(\tfrac{1}{(\log N)}\Big).
\end{aligned}$$

□

It remains to add up both probabilities given in Lemma 2.31 and Lemma 2.34. Recall that

$$r(2) = r_1 + r_2 - r_1 r_2 = r_1 + r_2 + \mathcal{O}(1/(\log N)^2) \tag{2.67}$$

and that we hence can approximate as follows:

$$\exp\Big(-\frac{r(2)}{s}\sum_{m=l+1}^{N}\frac{1}{m}\Big)=\exp\Big(-\frac{r_1+r_2}{s}\sum_{m=l+1}^{N}\frac{1}{m}\Big)+\mathcal{O}(1/\log(N)).$$

With this, the sum is calculated easily using the statement of Lemma 2.30.

$$\begin{aligned}
&\mathbb{P}^{\mathcal{S}_i^c}(\mathcal{A}_4)=\mathbb{P}^{\mathcal{S}_i^c}([2,1]_{b,i}^{rec})+\mathbb{P}^{\mathcal{S}_i^c}([12,2]_{b,i}^{rec})\\
&=\sum_{l=1}^{N}\frac{r_2}{s}\frac{1}{l+1}\exp\Big(-\frac{r(2)}{s}\sum_{m=l+1}^{N}\frac{1}{m}\Big)\cdot\sum_{k=1}^{l-1}\frac{r_1}{s}\frac{1}{k+1}\exp\Big(-\frac{r_1}{s}\sum_{n=k+1}^{l}\frac{1}{n}\Big)\\
&\quad+\sum_{l=1}^{N}\frac{r_1}{s}\frac{1}{l+1}\exp\Big(-\frac{r(2)}{s}\sum_{m=l+1}^{N}\frac{1}{m}\Big)\Big(1-\exp\Big(-\frac{r_2}{s}\cdot\sum_{k=1}^{l}\frac{1}{k}\Big)\Big)+\mathcal{O}\Big(\tfrac{1}{\log N}\Big)\\
&=1-\exp\Big(-\frac{r_1+r_2}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)-\exp\Big(-\frac{r_1}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\cdot\Big(1-\exp\Big(-\frac{r_2}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\Big)\\
&\quad-\exp\Big(-\frac{r_2}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\cdot\Big(1-\exp\Big(-\frac{r_1}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\Big)+\mathcal{O}\Big(\tfrac{1}{\log N}\Big)\\
&=\Big(1-\exp\Big(-\frac{r_1}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\Big)\cdot\Big(1-\exp\Big(-\frac{r_2}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\Big)+\mathcal{O}\Big(\tfrac{1}{\log N}\Big).
\end{aligned}$$

Let us now consider the case $\mathcal{S}_i$. Here, the situation is very simple as Proposition 2.20 tells us, that the two neutral loci of one individual can be treated as two independent individuals after they split within the B-population. We have to keep in mind though, that on the lineage which represents the ancestor of $(i,2)$ also a recombination between the selected and the first neutral locus leads to an escape of N2. A double recombination in one step however is again not conducive. We therefore have to use $r^*(2)$ when we apply Proposition 2.15 here.

$$\begin{aligned}
\mathbb{P}(\mathcal{A}_4\mid\mathcal{S}_i)&=\mathbb{P}^{\mathcal{S}_i}(R(i,1)\geqslant 0,R(i,2)\geqslant 0)+\mathcal{O}(1/\log(N))\\
&=\mathbb{P}^{\mathcal{S}_i}(R(i,1)\geqslant 0)\cdot\mathbb{P}^{\mathcal{S}_i}(R(i,2)\geqslant 0)+\mathcal{O}(1/\log(N))\\
&=\Big(1-\exp\Big(-\frac{r_1}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\Big)\cdot\Big(1-\exp\Big(-\frac{r_1+r_2}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\Big)+\mathcal{O}\Big(\tfrac{1}{\log N}\Big).
\end{aligned}$$

Turning back to Equation (2.55) and using the definition of $q^{(1)},q^{(2)}$ and q^{BB} from Equation (2.18) in Proposition 2.16, we get:

$$\begin{aligned}
\mathbb{P}(\mathcal{A}_4)&=(1-q^{BB})q^{(2)}q^{(1)}+q^{BB}q^{(1)}[1-(1-q^{(1)})(1-q^{(2)})]+\mathcal{O}(1/\log(N))\\
&=q^{(1)}[q^{(1)}q^{BB}(1-q^{(2)})+q^{(2)}]+\mathcal{O}(1/\log(N)).
\end{aligned}\tag{2.68}$$

Probability of the ancestral relation $\mathcal{A}_3$.

We will briefly calculate the probability of a double-singleton, $\mathcal{A}_3$. As mentioned before, this block type only has a non-negligible probability under $\mathbb{P}^{\mathcal{S}_i^c}$. By Proposition 2.12 we have

$$\begin{aligned}
&\mathbb{P}^{\mathcal{S}_i^c}(\mathcal{A}_3)\\
&= \mathbb{P}^{\mathcal{S}_i^c}((i,1),(i,2) \text{ recombine together in the b-population and do not separate}) + \mathcal{O}(\log(N)/N)\\
&= \sum_{l=1}^{N} \mathbb{P}^{\mathcal{S}_i^c}(R(i,1) = R(i,2) \geqslant 0, R_b^{\mathrm{rec}}(i) = -\infty, X_{R(i,1)} = l) + \mathcal{O}(\log(N)/N)\\
&= \sum_{l=1}^{N} \mathbb{P}^{\mathcal{S}_i^c}(R_b^{\mathrm{rec}}(i) = -\infty \mid \mathcal{R}(i;l)) \cdot \mathbb{P}^{\mathcal{S}_i^c}(\mathcal{R}(i;l)) + \mathcal{O}(\log(N)/N),
\end{aligned}$$

with $\mathcal{R}(i;l)$ as in (2.56). We now can plug in the results from the last section, make similar approximations and use in particular (2.67) and the statement of Lemma 2.30 in order to receive the following:

$$\begin{aligned}
\mathbb{P}^{\mathcal{S}_i^c}(\mathcal{A}_3) &= \sum_{l=1}^{N} \Big\{ \exp\Big(-\frac{r_2}{s} \sum_{k=1}^{l} \frac{1}{k} \Big) \exp\Big(\mathcal{O}\Big(\tfrac{1}{N}\Big)\Big) + \mathcal{O}\Big(\tfrac{\log N}{N}\Big) \Big\} \cdot \Big\{ \frac{r_1}{s} \cdot \frac{1-(1-s)^{2N-l}}{1-(1-s)^{2N}}\\
&\cdot \exp\Big(-\frac{r(2)}{s} \sum_{m=l+1}^{N} \frac{1}{m} \Big) \Big[\frac{(2N-l)(1-(1-s)^{l+1})}{2N(l+1)} + \exp\Big(-\frac{r(2)}{s} \frac{1}{l} \Big) \cdot \frac{1-(1-s)^{l}}{2N} \Big] + \mathcal{O}\Big(\tfrac{1}{(\log N)^2}\Big) \Big\}\\
&= \frac{r_1}{s} \sum_{l=1}^{N} \exp\Big(-\frac{r_2}{s} \cdot \sum_{k=1}^{l} \frac{1}{k} \Big) \cdot \frac{1}{l+1} \cdot \exp\Big(-\frac{r(2)}{s} \sum_{m=l+1}^{N} \frac{1}{m} \Big) + \mathcal{O}\Big(\tfrac{1}{\log N}\Big)\\
&= \exp\Big(-\frac{r_2}{s} \cdot \sum_{k=1}^{N} \frac{1}{k} \Big) \cdot \Big(1 - \exp\Big(-\frac{r_1}{s} \sum_{k=1}^{N} \frac{1}{k} \Big) \Big) + \mathcal{O}\Big(\tfrac{1}{\log N}\Big)
\end{aligned}$$

The total probability of the block type described through $\mathcal{A}_3$ then reads

$$\mathbb{P}(\mathcal{A}_3) = (1-q^{BB})(1-q^{(2)})q^{(1)} + \mathcal{O}(1/\log(N)). \tag{2.69}$$

Probability of the ancestral relation $\mathcal{A}_2$.

We continue with a short derivation of the block type describing $\mathcal{A}_2$, which we only need to consider under $\mathbb{P}^{\mathcal{S}_i}$. Similar as for $\mathbb{P}^{\mathcal{S}_i}(\mathcal{A}_4)$ we get by Propositions 2.20 and then 2.15,

$$\begin{aligned}
\mathbb{P}^{\mathcal{S}_i}(\mathcal{A}_2) &= \mathbb{P}^{\mathcal{S}_i}(R(i,1) \geqslant 0) \cdot \mathbb{P}^{\mathcal{S}_i}(R(i,2) = -\infty) + \mathcal{O}(1/\log(N))\\
&= \Big(1 - \exp\Big(-\frac{r_1}{s} \sum_{k=1}^{N} \frac{1}{k} \Big) \Big) \cdot \exp\Big(-\frac{r_1+r_2}{s} \sum_{k=1}^{N} \frac{1}{k} \Big) + \mathcal{O}\Big(\tfrac{1}{\log N}\Big).
\end{aligned}$$

Thus,

$$\mathbb{P}(\mathcal{A}_2) = q^{BB} q^{(1)}(1-q^{(1)})(1-q^{(2)}) + \mathcal{O}(1/\log(N)). \tag{2.70}$$

Probability of the ancestral relation $\mathcal{A}_1$.

For proving the result for $\mathcal{A}_1$ under $\mathbb{P}^{\mathcal{S}_i^c}$ we go a bit more into detail: let i be an individual sampled at time τ_N. Then,

$$\begin{aligned}
&\mathbb{P}^{\mathcal{S}_i^c}(\mathcal{A}_1) = \mathbb{P}(\exists t : (i,2) \text{ solely recombines for the first time into the b-population,} \\
&\qquad (i,1) \text{ stays in B in } (0,\tau_N)) + \mathcal{O}(1/(\log N)^2) \\
&= \sum_{l=1}^{N} \sum_{t=0}^{\infty} \mathbb{P}((i,1) \text{ stays in B in } (0,t) \mid \text{at } t, (i,2) \text{ solely recombines for the first time into} \\
&\qquad \text{the b-population, } (i,1) \text{ is in the B-population in } [t,\tau_N], t \leqslant \tau_N, X_t = l) \\
&\quad \cdot \sum_{k=l,l+1} \mathbb{P}(\text{at } t, (i,2) \text{ solely recombines for the first time into the b-population, } (i,1) \\
&\qquad \text{is in the B-population in} [t,\tau_N], t \leqslant \tau_N, X_t = l, X_{t+1} = k) + \mathcal{O}(1/(\log N)^2).
\end{aligned}$$

We first have a look at the last probability. Similar as before,

$$\begin{aligned}
&\mathbb{P}(\text{at } t, (i,2) \text{ solely recombines for the first time into the b-population,} \\
&\qquad (i,1) \text{ is in the B-population in } [t,\tau_N], t \leqslant \tau_N, X_t = l, X_{t+1} = k) \\
&= \mathbb{P}(B_t(A^t_{\tau_N}(i,2)) = 0, B_t(A^t_{\tau_N}(i,1)) = 1, \forall t' > t : B_{t'}(A^{t'}_{\tau_N}(i,1)) = 1, \\
&\qquad B_{t'}(A^{t'}_{\tau_N}(i,2)) = 1, t \leqslant \tau_N, X_t = l, X_{t+1} = k) + \mathcal{O}(1/(\log N)^2) \\
&= p_B^{r_2(1-r_1)}(l,k) \cdot \mathbb{P}(t \leqslant \tau_N, X_t = l, X_{t+1} = k) \cdot P_k + \mathcal{O}(1/(\log N)^2),
\end{aligned}$$

with P_k as in (2.60).
The first probability can be treated as follows:

$$\begin{aligned}
&\mathbb{P}((i,1) \text{ stays in B in } (0,t) \mid \text{at } t, (i,2) \text{ solely recombines for the first time into the} \\
&\qquad \text{b-population, } (i,1) \text{ is in the B-population in } [t,\tau_N], t \leqslant \tau_N, X_t = l) \\
&= \mathbb{P}((i,1) \text{ stays in B in } (0,t) \mid (i,1) \text{ is in the B-population in } [t,\tau_N], t \leqslant \tau_N, X_t = l) =: P_l^1,
\end{aligned}$$

and we can again bound this from below and above as follows:

$$\prod_{t=1}^{\tau_l^*-1} (1 - p_B^{r_1}(X_t, X_{t+1})) \leqslant P_l^1 \leqslant \prod_{t=1}^{\tau_l-1} (1 - p_B^{r_1}(X_t, X_{t+1})).$$

This inequality however corresponds to (2.64) with $m = 1$ and $\tau_m = \tau_m^*$. The result therefore follows from (2.66):

$$P_l^1 = \exp\Big(-\frac{r_1}{s}\cdot\sum_{k=1}^{l}\frac{1}{k}\Big) + \mathcal{O}(\tfrac{1}{\log N}).$$

With this, we can close the proof with

$$\begin{aligned}\mathbb{P}^{\mathcal{S}_i^c}(\mathcal{A}_1) = \exp\Big(-\frac{r_1}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\cdot\sum_{l=1}^{N}\exp\Big(\frac{r_1}{s}\sum_{k=l+1}^{N}\frac{1}{k}\Big)\cdot\Big\{p_B^{r_2(1-r_1)}(l,l)\cdot P_l\cdot\mathbb{E}(H_l)\\ + p_B^{r_2(1-r_1)}(l,l+1)\cdot P_{l+1}\cdot\mathbb{E}(U_l)\Big\} + \mathcal{O}(1/(\log N)),\end{aligned}$$

where we again used that the terms resulting from the number of holds and up-jumps starting from l after time τ_N can be bounded (see proof of Lemma 2.33). The expression in the curly brackets is known from Lemma 2.33. With the simplifications used before we finally get

$$\begin{aligned}\mathbb{P}^{\mathcal{S}_i^c}(\mathcal{A}_1) &= \exp\Big(-\frac{r_1}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\cdot\sum_{l=1}^{N}\exp\Big(\frac{r_1}{s}\sum_{k=l+1}^{N}\frac{1}{k}\Big)\cdot\Big[\frac{r_2}{s}\frac{1}{l+1}\exp\Big(-\frac{r(2)}{s}\sum_{k=l+1}^{N}\frac{1}{k}\Big)\\ &\quad + \mathcal{O}\Big(\tfrac{1}{(\log N)^2}\Big)\Big] + \mathcal{O}\Big(\tfrac{1}{\log N}\Big)\\ &= \exp\Big(-\frac{r_1}{s}\cdot\sum_{k=1}^{N}\frac{1}{k}\Big)\Big(1-\exp\Big(-\frac{r_2}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\Big) + \mathcal{O}\Big(\tfrac{1}{\log N}\Big).\end{aligned}$$

The probability of $\mathcal{A}_1$ under $\mathbb{P}^{\mathcal{S}_i}$ can be easily calculated by applying again Propositions 2.20 and 2.15:

$$\mathbb{P}^{\mathcal{S}_i}(\mathcal{A}_1) = \exp\Big(-\frac{r_1}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\cdot\Big(1-\exp\Big(-\frac{r_1+r_2}{s}\sum_{k=1}^{N}\frac{1}{k}\Big)\Big) + \mathcal{O}\Big(\tfrac{1}{\log N}\Big).$$

Together,

$$\begin{aligned}\mathbb{P}(\mathcal{A}_1) &= (1-q^{BB})(1-q^{(1)})q^{(2)} + q^{BB}(1-q^{(1)})(1-[(1-q^{(1)})(1-q^{(2)})]) + \mathcal{O}(1/\log(N))\\ &= (1-q^{(1)})[q^{(1)}q^{BB}(1-q^{(2)}) + q^{(2)}] + \mathcal{O}(1/\log(N)).\end{aligned} \tag{2.71}$$

Probability of the ancestral relation $\mathcal{A}_0$.

The probability of $\mathcal{A}_0$ can easily be deduced from the previous calculations and in particular by applying Proposition 2.15:

$$\begin{aligned}\mathbb{P}(\mathcal{A}_0) &= q^{BB}(1-q^{(1)})[(1-q^{(1)})(1-q^{(2)})] + (1-q^{BB})(1-q^{(1)})(1-q^{(2)}) + \mathcal{O}(1/\log(N))\\ &= (1-q^{(1)})(1-q^{(2)})(1-q^{(1)}q^{BB}) + \mathcal{O}(1/\log(N)),\end{aligned} \tag{2.72}$$

and a test shows that indeed, $\sum_{k=0}^{4}\mathbb{P}(\mathcal{A}_k)=1$. Recalling that $q^{BB}=q^{(2)}+\mathcal{O}(1/\log(N))$ (see for example Proposition 2.14) we can check that all derived probabilities correspond to the entries in $\bar{q}_{\mathrm{G1}}$ up to error terms of order at most $\mathcal{O}(1/\log(N))$.

2.4.4 Proof of the multinomial marking property

The proof of Proposition 2.17 is very similar to the proof of Proposition 2.13 (c.f. Proposition 2.4 in [36]) which was omitted in this work.

Proof of Proposition 2.17. As in the proof of Lemma 4.2 in [36] we can derive the following bound for the probability that a coalescence with respect to any locus j happens within the B-population:

$$\begin{aligned}
&\mathbb{P}(X_{G^j(i,m)}=k, B_{G^j(i,m)+1}(A_\tau^{G^j(i,m)}(i,j))=1=B_{G^j(i,m)+1}(A_\tau^{G^j(i,m)}(m,j)))\\
&\leqslant \mathbb{P}(X_{G^j(i,m)}=k \mid B_{G^j(i,m)+1}(A_\tau^{G^j(i,m)}(i,j))=1=B_{G^j(i,m)+1}(A_\tau^{G^j(i,m)}(m,j)))\\
&\leqslant p_{BB}^{c,j}(k,k)\mathbb{E}(H_k)+p_{BB}^{c,j}(k,k+1)\mathbb{E}(U_k)\leqslant \frac{2}{k(2N-k)s}+\frac{2}{k(k+1)s}\leqslant \frac{4N}{sk^2(2N-k)}.
\end{aligned}$$

With this, we can continue as follows by making again use of the symmetry argument from (2.43) for the first inequality:

$$\begin{aligned}
&\mathbb{P}(0\leqslant R_B^{\mathrm{rec}}(i)\leqslant G^j(i,m))=\sum_{k=1}^{2N-1}\sum_{l=1}^{2N-1}\mathbb{P}(0\leqslant R_B^{\mathrm{rec}}(i)\leqslant G^j(i,m), X_{G^j(i,m)}=k, X_{R_B^{\mathrm{rec}}(i)}=l)\\
&=\sum_{k=1}^{2N-1}\mathbb{P}(X_{G^j(i,m)}=k, B_{G^j(i,m)+1}(A_\tau^{G^j(i,m)}(i,j))=1=B_{G^j(i,m)+1}(A_\tau^{G^j(i,m)}(m,j)))\\
&\quad\cdot\sum_{l=1}^{2N-1}\mathbb{P}(0\leqslant R_B^{\mathrm{rec}}(i)\leqslant G^j(i,m), X_{R_B^{\mathrm{rec}}(i)}=l\mid X_{G^j(i,m)}=k, B_{G^j(i,m)+1}(A_\tau^{G^j(i,m)}(i,j))=1,\\
&\qquad B_{G^j(i,m)+1}(A_\tau^{G^j(i,m)}(m,j))=1)+\mathcal{O}(\log(N)/N)\\
&\leqslant\sum_{k=1}^{2N-1}\frac{4N}{sk^2(2N-k)}\sum_{l=1}^{2N-1}\{p_{BB}^{r_2}(k,k)\mathbb{E}(H_{2N-l,2N-k})+p_{BB}^{r_2}(k,k+1)\mathbb{E}(U_{2N-l-1,2N-k})\}\\
&\qquad+\mathcal{O}(\log(N)/N)\\
&\leqslant\sum_{k=1}^{2N-1}\frac{4N}{sk^2(2N-k)}\sum_{l=1}^{2N-1}\{\underbrace{\frac{r_2}{2Ns}\frac{l}{2N-l}\min\{(1-s)^{l-k},1\}+\frac{r_2}{2Ns}\frac{l}{l+1}\min\{(1-s)^{l+1-k},1\}}_{\leqslant\frac{r_2}{2Ns}\min\{(1-s)^{l-k},1\}\frac{l+(2N-l)}{2N-l}}\}\\
&\qquad+\mathcal{O}(\log(N)/N)\\
&\leqslant\sum_{k=1}^{2N-1}\frac{4N}{sk^2(2N-k)}\Big[\sum_{l=1}^{k}\frac{r_2}{(2N-l)s}+\sum_{l=k+1}^{2N-1}\frac{r_2(1-s)^{l-k}}{(2N-l)s}\Big]+\mathcal{O}\Big(\tfrac{\log(N)}{N}\Big)
\end{aligned}$$

$$\leqslant \frac{4r_2}{s^2}\sum_{k=1}^{2N-2}\frac{N}{k^2(2N-k)}\Big[\frac{k}{2N-k}+\frac{1/s}{2N-k-1}\Big]+\mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)$$
$$\leqslant \frac{4r_2}{s^3}\Big\{\sum_{k=1}^{N}\frac{1}{k^2}\frac{N+1}{N-1}+\sum_{k=N+1}^{2N-2}\frac{k+1}{k}\frac{1}{(2N-k-1)^2}\Big\}+\mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)\leqslant\mathcal{O}\Big(\tfrac{1}{\log(N)}\Big).$$

□

Before we come to the binomial and multinomial statements, we state a coupling Lemma which is needed in each of the following proofs and which generalizes Lemma 5.1 from [36].

Lemma 2.35. *Let V_N and V'_N be $\{0,1,\dots,n\}$-valued random variables with*

$$\mathbb{E}(V_N)=\mathbb{E}(V'_N)+\mathcal{O}(\varepsilon(N)), \tag{2.73}$$

where $\varepsilon(N)$ is a function of N with $\varepsilon(N)\to 0$ for $N\to\infty$. Then there exist random variables $\tilde{V}_N$ and $\tilde{V}'_N$ on some probability space such that

$$\begin{aligned}&V_N \text{ and } \tilde{V}_N \text{ have the same distribution,}\\ &V'_N \text{ and } \tilde{V}'_N \text{ have the same distribution, and}\\ &\mathbb{P}(\tilde{V}_N\neq\tilde{V}'_N)\leqslant n\cdot\max\{\mathbb{P}(\tilde{V}_N\geqslant 2),\mathbb{P}(\tilde{V}'_N\geqslant 2)\}+\mathcal{O}(\varepsilon(N)).\end{aligned} \tag{2.74}$$

The proof is straightforward and follows that in [36]. We will however give more details in order to be sure that the generalization indeed holds true.

Proof. By the general idea of coupling we can construct two random variables $\tilde{V}_N$ and $\tilde{V}'_N$ on the same probability space such that their distributions coincide with V_N and V'_N, respectively. Further, the coupling can be done in such a way that

$$\begin{aligned}\mathbb{P}(\tilde{V}_N=\tilde{V}'_N)&=\sum_{k=0}^{n}\mathbb{P}(\tilde{V}_N=\tilde{V}'_N=k)=\sum_{k=2}^{n}\mathbb{P}(\tilde{V}_N=\tilde{V}'_N=k)\\ &\quad+\min\{\mathbb{P}(\tilde{V}_N=0),\mathbb{P}(\tilde{V}'_N=0)\}+\min\{\mathbb{P}(\tilde{V}_N=1),\mathbb{P}(\tilde{V}'_N=1)\}\\ &\geqslant\min\{\mathbb{P}(\tilde{V}_N=0),\mathbb{P}(\tilde{V}'_N=0)\}+\min\{\mathbb{P}(\tilde{V}_N=1),\mathbb{P}(\tilde{V}'_N=1)\},\\ \Leftrightarrow\ \mathbb{P}(\tilde{V}_N\neq\tilde{V}'_N)&\leqslant 1-\min\{\mathbb{P}(\tilde{V}_N=0),\mathbb{P}(\tilde{V}'_N=0)\}-\min\{\mathbb{P}(\tilde{V}_N=1),\mathbb{P}(\tilde{V}'_N=1)\}.\end{aligned} \tag{2.75}$$

By standard calculations we can now derive bounds for the two minima on the right-hand side:

$$\begin{aligned}&\mathbb{E}(\tilde{V}_N)=\sum_{k=0}^{n}k\cdot\mathbb{P}(\tilde{V}_N=k)\geqslant\sum_{k=1}^{n}\mathbb{P}(\tilde{V}_N=k)=1-\mathbb{P}(\tilde{V}_N=0)\\ \Leftrightarrow\quad&\mathbb{P}(\tilde{V}_N=0)\geqslant 1-\mathbb{E}(\tilde{V}_N);\ \text{ and analogously, }\ \mathbb{P}(\tilde{V}'_N=0)\geqslant 1-\mathbb{E}(\tilde{V}'_N),\end{aligned}$$

$$\Rightarrow \quad \min\{\mathbb{P}(\tilde{V}_N = 0), \mathbb{P}(\tilde{V}'_N = 0)\} \geqslant 1 - \max\{\mathbb{E}(\tilde{V}_N), \mathbb{E}(\tilde{V}'_N)\}.$$

In the same way,

$$\begin{aligned}
&\mathbb{E}(\tilde{V}_N) = \sum_{k=0}^{n} k \cdot \mathbb{P}(\tilde{V}_N = k) = \mathbb{P}(\tilde{V}_N = 1) + \sum_{k=2}^{n} k \cdot \mathbb{P}(\tilde{V}_N = k) \leqslant \mathbb{P}(\tilde{V}_N = 1) + n \cdot \mathbb{P}(\tilde{V}_N \geqslant 2) \\
\Leftrightarrow \quad &\mathbb{P}(\tilde{V}_N = 1) \geqslant \mathbb{E}(\tilde{V}_N) - n\mathbb{P}(\tilde{V}_N \geqslant 2); \quad \text{and analogously,} \quad \mathbb{P}(\tilde{V}'_N = 1) \geqslant \mathbb{E}(\tilde{V}'_N) - n\mathbb{P}(\tilde{V}'_N \geqslant 2), \\
\Rightarrow \quad &\min\{\mathbb{P}(\tilde{V}_N = 1), \mathbb{P}(\tilde{V}'_N = 1)\} \geqslant \min\{\mathbb{E}(\tilde{V}_N), \mathbb{E}(\tilde{V}'_N)\} - n \max\{\mathbb{P}(\tilde{V}_N \geqslant 2), \mathbb{P}(\tilde{V}'_N \geqslant 2)\}.
\end{aligned}$$

Both inequalities together with (2.73) give the following:

$$\begin{aligned}
&\min\{\mathbb{P}(\tilde{V}_N = 0), \mathbb{P}(\tilde{V}'_N = 0)\} + \min\{\mathbb{P}(\tilde{V}_N = 1), \mathbb{P}(\tilde{V}'_N = 1)\} \\
&\geqslant 1 - \max\{\mathbb{E}(\tilde{V}_N), \mathbb{E}(\tilde{V}'_N)\} + \min\{\mathbb{E}(\tilde{V}_N), \mathbb{E}(\tilde{V}'_N)\} - n \cdot \max\{\mathbb{P}(\tilde{V}_N \geqslant 2), \mathbb{P}(\tilde{V}'_N \geqslant 2)\} \\
&= 1 - \mathcal{O}(\varepsilon(N)) - n \cdot \max\{\mathbb{P}(\tilde{V}_N \geqslant 2), \mathbb{P}(\tilde{V}'_N \geqslant 2)\}. \qquad (2.76)
\end{aligned}$$

Finally, (2.76) in (2.75) gives the statement of the lemma:

$$\begin{aligned}
\mathbb{P}(\tilde{V}_N \neq \tilde{V}'_N) &\leqslant 1 - [1 - \mathcal{O}(\varepsilon(N)) - n \cdot \max\{\mathbb{P}(\tilde{V}_N \geqslant 2), \mathbb{P}(\tilde{V}'_N \geqslant 2)\}] \\
&= n \cdot \max\{\mathbb{P}(\tilde{V}_N \geqslant 2), \mathbb{P}(\tilde{V}'_N \geqslant 2)\} + \mathcal{O}(\varepsilon(N)).
\end{aligned}$$

□

In the following proofs we will always define certain binomially distributed random variables, show that we have approximate equality of appropriate expectations and that we can bound the right-hand side of (2.74) by terms of order at most $1/\log(N)$ before applying the above lemma.
We will start with proving Proposition 2.20 as it is the most involved and subsequently deduce briefly the proofs of the remaining statements dealing with the multinomial character of the partition of a sample.

Proof of Proposition 2.20. Suppose that we sample n^{BB} individuals at time τ_N and follow back their neutral gene genealogies. Consider for t with $0 \leqslant t \leqslant \tau_N$ the process

$$K^1_t := \#\{i \in [n^{BB}] : R(i,1) \geqslant t\}, \qquad (2.77)$$

with $K^1_{\tau_N} = 0$. Then, by Proposition 2.9, $K^1 := (K^1_t)_{\tau_N \geqslant t \geqslant 0}$ is approximately identical to the process $(K_t)_{\tau \geqslant t \geqslant 0}$ from [36] and therefore we have the statement from their Proposition 2.6, that is

Lemma 2.36 (cf. Prop. 2.6, [36]). *Let* $q_J^{(1)} = 1 - \exp(-\frac{r_1}{s}\sum_{k=J+1}^{N}\frac{1}{k})$. *For* $J \in [N]$, *it holds*

$$\left|\mathbb{P}(K^1_{\tau_J} = d) - \binom{n^{BB}}{d}(q_J^{(1)})^d(1-q_J^{(1)})^{n^{BB}-d}\right| \leqslant \frac{C}{\log N}$$

for $d = 0, 1, \ldots, n^{BB}$.

From the proof of Lemma 5.2, [36], which leads to the above stated Proposition, we know that K^1 can be coupled with a process $\tilde{K}^1$ such that they are almost surely equal in the limit of $N \to \infty$. Precisely we have for any $J \in [N]$ (cf. equation (5.1), [36])

$$\mathbb{P}(K^1_t \neq \tilde{K}^1_t \text{ for some } \tau_N \geqslant t \geqslant \tau_J) \leqslant \frac{C}{\log N},$$

and

$$\tilde{K}^1_{\tau_J} \mid X \sim \mathrm{Bin}(n^{BB}, F^1_J), \quad \text{with} \quad F^1_J = 1 - \prod_{t=\tau_J+1}^{\tau_N}(1-\theta_t^{r_1}), \quad \text{and} \quad \theta_t^{r_1} = p_B^{r_1}(X_{t-1}, X_t).$$

Note that the statement in [36] gives a better result for $J \leqslant CN/\log(N)$ for some constant C. For our purposes, the above (which holds for all $J \in [N]$) is sufficient.
We now define the process which counts the number of individuals whose second neutral locus recombines into the b-population during the first half of the sweep whereas the first neutral locus stays in the B-population,

$$K^2_t := \#\{i \in [n^{BB}] : R(i,2) \geqslant t, R(i,1) = -\infty\}, \quad 0 \leqslant t \leqslant \tau_N, \tag{2.78}$$

and will study its behavior conditional on X and the value K^1_0. Again, $K^2_{\tau_N} = 0$ and K^2_t only increases if a recombination happens that strikes the N2-locus of an individual which was not counted by the process K^1. We will see that with a similar reasoning as in [36] we get the following analogous statement:

Lemma 2.37. *Define for* $\theta_t^{r_2} := p_B^{r_2}(X_{t-1}, X_t)$ *the process*

$$\begin{aligned}&\tilde{K}^2 := (\tilde{K}^2_t)_{\tau_N \geqslant t \geqslant 0} \text{ with } \tilde{K}^2_{\tau_N} = 0 \text{ and}\\ &\tilde{K}^2_{t-1} - \tilde{K}^2_t \mid X, K^1_0, (\tilde{K}^2_u)_{\tau_N \geqslant u \geqslant t} \sim \mathrm{Bin}(n^{BB} - K^1_0 - \tilde{K}^2_t, \theta_t^{r_2}).\end{aligned} \tag{2.79}$$

Then, for $J \in [N]$,

$$\mathbb{P}(K^2_t \neq \tilde{K}^2_t \text{ for some } t \geqslant \tau_J) \leqslant \frac{C}{\log N}.$$

Proof. The proof is very similar to the one of Lemma 5.2 in [36], nevertheless we will be

much more detailed here in order to show that the arguments can indeed be transferred to the somewhat more delicate situation at hand.
First of all, we will check if the expected values of the differences of both processes are identical (under some specific measure) such that we can apply the general coupling Lemma 2.35.
From (2.79) we have

$$\mathbb{E}(\tilde{K}^2_{t-1} - \tilde{K}^2_t \mid X, K^1_0, (\tilde{K}^2_u)_{\tau_N \geqslant u \geqslant t}) = (n^{BB} - K^1_0 - \tilde{K}^2_t)\theta^{r_2}_t.$$

Let us define the set of all individuals i living at time t from which exactly k of the sampled loci $(m, 2)$, $m \in [n^{BB}]$, from time τ_N originate:

$$\begin{aligned} &\mathcal{G}^{2,k}_t := \{i \in [2N] : A^t_{\tau_N}(m, 2) = i \text{ for exactly } k \text{ different } m \in [n^{BB}]\}, \\ &k = 0, 1, \ldots, n^{BB},\ 0 \leqslant t \leqslant \tau_N. \end{aligned} \tag{2.80}$$

In other words, exactly k of the sampled N2-loci are descendants of an individual $i \in \mathcal{G}^{2,k}_t$ from time t. In particular, this implies the following identities:

$$\sum_{k=1}^{n^{BB}} k \cdot |\mathcal{G}^{2,k}_t| = n^{BB}, \quad \bigcup_{k=0}^{n}{\cdot}\, \mathcal{G}^{2,k}_t = [2N],\ 0 \leqslant t \leqslant \tau_N, \quad \text{and} \ \ \mathcal{G}^{2,1}_{\tau_N} = [n^{BB}]. \tag{2.81}$$

By translating the event $K^2_{t-1} - K^2_t = k$ into the evolution defining random variables from page 34 and applying Proposition 2.10 which bounds the probability of a double recombination we obtain

$$\begin{aligned}
&\mathbb{E}(K^2_{t-1} - K^2_t \mid X, K^1_0, (K^2_u)^{\tau_N}_{u=t}) = \sum_{k=0}^{n^{BB}} k \cdot \mathbb{P}(K^2_{t-1} - K^2_t = k \mid X, K^1_0, (K^2_u)^{\tau_N}_{u=t}) \\
&= \sum_{k=0}^{n^{BB}} k \cdot \sum_{i=1}^{2N} \Big(1 - \mathcal{O}\Big(\tfrac{1}{\log N}\Big)\Big) \mathbb{P}\big(I_{t,1} = i, i \in \mathcal{G}^{2,k}_t, B_t(i) = 1, B_t(I_{t,2}) = 1, B_{t-1}(I_{t,3}) = 0, I_{t,6} = 1, \\
&\qquad I_{t,5} = 0, \forall s < t : B_s(A^s_t(i, 1)) = 1 \mid X, K^1_0, (K^2_u)^{\tau_N}_{u=t}\big) \\
&= \mathbb{P}(I_{t,6} = 1)\Big(1 - \mathcal{O}\Big(\tfrac{1}{\log N}\Big)\Big) \cdot \sum_{k=0}^{n^{BB}} k \sum_{i=1}^{2N} \mathbb{P}\big(B_{t-1}(I_{t,3}) = 0 \mid I_{t,1} = i, i \in \mathcal{G}^{2,k}_t, B_t(i) = 1, \\
&\qquad B_t(I_{t,2}) = 1, \underbrace{I_{t,5} = 0, \forall s < t : B_s(A^s_t(i, 1)) = 1}_{\Rightarrow \text{ no constraint on the type of } I_{t,3}}, X, K^1_0, (K^2_u)^{\tau_N}_{u=t}\big) \\
&\cdot \mathbb{P}\big(B_t(I_{t,2}) = 1, I_{t,1} = i \mid i \in \mathcal{G}^{2,k}_t, B_t(i) = 1, I_{t,5} = 0, \forall s < t : B_s(A^s_t(i, 1)) = 1, X, K^1_0, (K^2_u)^{\tau_N}_{u=t}\big) \\
&\cdot \underbrace{\mathbb{P}\big(I_{t,5} = 0 \mid i \in \mathcal{G}^{2,k}_t, B_t(i) = 1, \forall s < t : B_s(A^s_t(i, 1)) = 1, X, K^1_0, (K^2_u)^{\tau_N}_{u=t}\big)}_{=1-\mathcal{O}(1/\log(N))} \\
&\cdot \mathbb{P}\big(i \in \mathcal{G}^{2,k}_t, B_t(i) = 1, \forall s < t : B_s(A^s_t(i, 1)) = 1 \mid X, K^1_0, (K^2_u)^{\tau_N}_{u=t}\big)
\end{aligned}$$

$$= r_2 \frac{2N - X_{t-1}}{2N}\big(1 - \mathcal{O}\big(\tfrac{1}{\log N}\big)\big) \sum_{k=0}^{n^{BB}} k \sum_{i=1}^{2N} \mathbb{P}\big(i \in \mathcal{G}_t^{2,k}, B_t(i) = 1, \forall s < t : B_s(A_t^s(i,1)) = 1$$
$$\mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}) \cdot \mathbb{P}\big(B_t(I_{t,2}) = 1, I_{t,1} = i \mid i \in \mathcal{G}_t^{2,k}, B_t(i) = 1, I_{t,5} = 0,$$
$$\forall s < t : B_s(A_t^s(i,1)) = 1, X, K_0^1, (K_u^2)_{u=t}^{\tau_N}\big).$$

Consider first the conditional probability that the parent is of type B and the offspring will replace a specific individual i. This probability depends on the values of the process X and in particular its type of transition from time $t-1$ to t:

$$\mathbb{P}\big(B_t(I_{t,2}) = 1, I_{t,1} = i \mid i \in \mathcal{G}_t^{2,k}, B_t(i) = 1, \underbrace{I_{t,5} = 0, \forall s < t : B_s(A_t^s(i,1)) = 1}_{\text{no constraint on the choice of } I_{t,1}}, X, K_0^1, (K_u^2)_{u=t}^{\tau_N}\big)$$
$$= \begin{cases} \frac{X_t \cdot 1}{X_t^2 + (2N - X_t)^2 + s X_t (2N - X_t)} = \frac{\beta_{X_t}}{2N - X_t}, & X_t = X_{t-1}, \\ \frac{1}{X_t}, & X_t = X_{t-1} + 1. \end{cases} \tag{2.82}$$

This can be seen as follows: if there is a hold, then there are $X_t^2 + (2N - X_t)^2 + s X_t (2N - X_t)$ possible combinations for $I_{t,1}$ and $I_{t,2}$. The parent is of type B if we draw one of the X_t individuals, the offspring however needs to be one particular individual i, so there is only one way to choose it. If the chain jumps by one in the t-th step, then the parent is necessarily of type B, as well as the offspring. Here we have to consider the probability that the new offspring is indeed the i-th individual, so one specific of all B-individuals in the next generation t. Note that this probability is independent of k and i and thus can be taken out of the sum. The statements in (2.81) let us treat the remaining part:

$$\sum_{i=1}^{2N} \sum_{k=0}^{n^{BB}} k \cdot \mathbb{P}\big(i \in \mathcal{G}_t^{2,k}, B_t(i) = 1, \forall s < t : B_s(A_t^s(i,1)) = 1 \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}\big)$$
$$= \mathbb{E}\Big[\sum_{k=1}^{n^{BB}} k \cdot \sum_{i=1}^{2N} \mathbf{1}_{\{i \in \mathcal{G}_t^{2,k}, B_t(i)=1, \forall s<t : B_s(A_t^s(i,1))=1\}} \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N} \Big]$$
$$= \mathbb{E}\Big[\sum_{k=1}^{n^{BB}} k \cdot \big|\{i \in [2N] : i \in \mathcal{G}_t^{2,k}, B_t(i) = 1, \forall s < t : B_s(A_t^s(i,1)) = 1\}\big| \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N} \Big]$$
$$= \mathbb{E}\Big[n^{BB} - K_0^1 - K_t^2 \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N} \Big]$$
$$+ \mathbb{E}\Big[\big|\{m \in [n^{BB}] : A_{\tau_N}^t(m,2) \in \mathcal{G}_t^{2,k}, B_t(A_{\tau_N}^t(m,2)) = 1 = B_t(A_{\tau_N}^t(m,1)),$$
$$B_s(A_{\tau_N}^s(i,1)) = 0 \text{ for some } t \leqslant s \leqslant \tau_N\}\big| \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N} \Big]$$
$$= n^{BB} - K_0^1 - K_t^2 + \mathcal{O}\big(\tfrac{1}{(\log N)^2}\big). \tag{2.83}$$

as the second expectation can be bounded by Proposition 2.11 and hence only those from the

set $\mathcal{G}_t^{2,k}$ count (k-fold) who will never leave the B-population with respect to their N1-locus and who currently reside in the B-population.
Recall the value of $\theta_t^{r_2} = p_B^{r_2}(X_{t-1}, X_t)$ given in Lemma 2.26. We see that it is exactly the result from (2.82) multiplied by $r_2 \cdot \frac{2N - X_{t-1}}{2N}$. This together with the result from (2.83) gives

$$\mathbb{E}(K_{t-1}^2 - K_t^2 \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}) = \theta_t^{r_2}\big[n^{BB} - K_0^1 - K_t^2\big]\Big(1 - \mathcal{O}\Big(\tfrac{1}{\log N}\Big)\Big)$$

We notice that, up to terms of order at most $\theta_t^{r_2}/(\log N)$, the above calculated expectations are equal on the event that $\{K_t^2 = \tilde{K}_t^2\}$ and apply Lemma 2.35 in order to conclude the statement of Lemma 2.37. Note, that we use Lemma 2.35 in the sense that the inequality holds for each specific configuration of $\{X, K_0^1, (K_u^2)_{u=t}^{\tau_N}\}$.

$$\begin{aligned}
&\mathbb{P}(K_t^2 \neq \tilde{K}_t^2 \text{ for some } \tau_N \geqslant t \geqslant \tau_J \mid X, K_0^1) = \sum_{t=\tau_J}^{\tau_N} \mathbb{P}(K_t^2 \neq \tilde{K}_t^2 \text{ and } K_s^2 = \tilde{K}_s^2 \ \forall s > t \mid X, K_0^1) \\
&= \sum_{t=\tau_J+1}^{\tau_N} \mathbb{E}\big[\mathbb{P}(K_{t-1}^2 - K_t^2 \neq \tilde{K}_{t-1}^2 - \tilde{K}_t^2, K_s^2 = \tilde{K}_s^2 \ \forall s \geqslant t \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}, (\tilde{K}_u^2)_{u=t}^{\tau_N}) \mid X, K_0^1\big] \\
&= \sum_{t=\tau_J+1}^{\tau_N} \mathbb{E}\Big[\mathbb{P}(K_{t-1}^2 - K_t^2 \neq \tilde{K}_{t-1}^2 - \tilde{K}_t^2 \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}, (\tilde{K}_u^2)_{u=t}^{\tau_N}, \\
&\qquad K_s^2 = \tilde{K}_s^2 \ \forall s \geqslant t) \cdot \mathbb{P}(K_s^2 = \tilde{K}_s^2 \ \forall s \geqslant t \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}, (\tilde{K}_u^2)_{u=t}^{\tau_N}) \mid X, K_0^1\Big] \\
&\leqslant \sum_{t=\tau_J+1}^{\tau_N} \mathbb{E}\Big[\Big(n^{BB} \cdot \max\Big\{\mathbb{P}(K_{t-1}^2 - K_t^2 \geqslant 2 \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}, (\tilde{K}_u^2)_{u=t}^{\tau_N}, K_s^2 = \tilde{K}_s^2 \ \forall s \geqslant t), \\
&\qquad \mathbb{P}(\tilde{K}_{t-1}^2 - \tilde{K}_t^2 \geqslant 2 \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}, (\tilde{K}_u^2)_{u=t}^{\tau_N}, K_s^2 = \tilde{K}_s^2 \ \forall s \geqslant t)\Big\} + \mathcal{O}\Big(\tfrac{\theta_t^{r_2}}{\log N}\Big)\Big) \\
&\qquad \cdot \mathbb{P}(K_s^2 = \tilde{K}_s^2 \ \forall s \geqslant t \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}, (\tilde{K}_u^2)_{u=t}^{\tau_N}) \mid X, K_0^1\Big] \\
&\leqslant \sum_{t=\tau_J+1}^{\tau_N} \mathbb{E}\Big[n^{BB} \cdot \big[\mathbb{P}(K_{t-1}^2 - K_t^2 \geqslant 2, K_s^2 = \tilde{K}_s^2 \ \forall s \geqslant t \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}, (\tilde{K}_u^2)_{u=t}^{\tau_N}) \\
&\qquad + \mathbb{P}(\tilde{K}_{t-1}^2 - \tilde{K}_t^2 \geqslant 2, K_s^2 = \tilde{K}_s^2 \ \forall s \geqslant t \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}, (\tilde{K}_u^2)_{u=t}^{\tau_N})\big] \mid X, K_0^1\Big] \\
&\qquad + \mathcal{O}\Big(\sum_{t=\tau_J+1}^{\tau_N} \tfrac{\theta_t^{r_2}}{\log N}\Big) \\
&\leqslant \sum_{t=\tau_J+1}^{\tau_N} \mathbb{E}\Big[n^{BB} \cdot \big[\mathbb{P}(K_{t-1}^2 - K_t^2 \geqslant 2 \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}, (\tilde{K}_u^2)_{u=t}^{\tau_N}) \\
&\qquad + \mathbb{P}(\tilde{K}_{t-1}^2 - \tilde{K}_t^2 \geqslant 2 \mid X, K_0^1, (K_u^2)_{u=t}^{\tau_N}, (\tilde{K}_u^2)_{u=t}^{\tau_N})\big] \mid X, K_0^1\Big] + \mathcal{O}\Big(\sum_{t=\tau_J+1}^{\tau_N} \tfrac{\theta_t^{r_2}}{\log N}\Big) \\
&= n^{BB} \sum_{t=\tau_J+1}^{\tau_N} \Big[\mathbb{P}(K_{t-1}^2 - K_t^2 \geqslant 2 \mid X, K_0^1) + \mathbb{P}(\tilde{K}_{t-1}^2 - \tilde{K}_t^2 \geqslant 2 \mid X, K_0^1)\Big] + \mathcal{O}\Big(\sum_{t=\tau_J+1}^{\tau_N} \tfrac{\theta_t^{r_2}}{\log N}\Big).
\end{aligned} \tag{2.84}$$

From this, we can get the statement of the lemma by calculating the expectation from the above sum. First, as mentioned above, the error term with the sum over the $\theta_t^{r_2}$ is bounded by $C/\log(N)$ from the calculations in (2.63). Next, consider the part which deals with the process K^2:

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N} \mathbb{P}(K_{t-1}^2 - K_t^2 \geqslant 2 \mid X, K_0^1)\Big] = \mathbb{E}\Big[\sum_{t=1}^{\infty} \mathbf{1}_{\{\tau_J+1\leqslant t\leqslant \tau_N\}} \mathbb{E}(\mathbf{1}_{\{K_{t-1}^2 - K_t^2 \geqslant 2\}} \mid X, K_0^1)\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\sum_{t=1}^{\infty} \mathbf{1}_{\{K_{t-1}^2 - K_t^2 \geqslant 2,\ \tau_J+1\leqslant t\leqslant \tau_N\}} \mid X, K_0^1\Big]\Big] \leqslant \mathbb{E}\Big[\tfrac{n^{BB}-K_0^1}{2} \mathbf{1}_{\{K_{t-1}^2 - K_t^2 \geqslant 2, \text{ for some } \tau_J+1\leqslant t\leqslant \tau_N\}}\Big] \\
&\leqslant \frac{n^{BB}}{2} \mathbb{P}(K_{t-1}^2 - K_t^2 \geqslant 2, \text{ for some } \tau_J + 1 \leqslant t \leqslant \tau_N) \\
&\leqslant \frac{n^{BB}}{2} \mathbb{P}(\exists\, i, m \text{ s.t. } G^2(i,m) \geqslant R(i,2) \geqslant \tau_J) \leqslant C/\log(N), \qquad (2.85)
\end{aligned}$$

due to Proposition 2.13. Last, by definition we have for the process $\tilde{K}^2$ that

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N} \mathbb{P}(\tilde{K}_{t-1}^2 - \tilde{K}_t^2 \geqslant 2 \mid X, K_0^1)\Big] \\
&= \mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N} \mathbb{E}\big[\mathbb{P}(\tilde{K}_{t-1}^2 - \tilde{K}_t^2 \geqslant 2 \mid X, K_0^1, (\tilde{K}_u^2)_{u=t}^{\tau_N}) \mid X, K_0^1\big]\Big] \\
&\leqslant \mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N} \mathbb{E}\Big[\binom{n^{BB} - K_0^1 - \tilde{K}_t^2}{2} \cdot (\theta_t^{r_2})^2 \mid X, K_0^1\Big]\Big] \leqslant \binom{n^{BB}}{2} \mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N} (\theta_t^{r_2})^2\Big] \leqslant \frac{C}{(\log N)^2},
\end{aligned}$$

by similar calculations as for (2.61). From (2.84) we then get the statement of the Lemma:

$$\begin{aligned}
\mathbb{P}(K_t^2 \neq \tilde{K}_t^2 \text{ for some } t \geqslant \tau_J) &= \mathbb{E}\Big[\mathbb{P}(K_t^2 \neq \tilde{K}_t^2 \text{ for some } t \geqslant \tau_J \mid X, K_0^1)\Big] \\
&\leqslant n^{BB} \cdot \Big(\frac{1}{\log N} + \frac{1}{(\log(N))^2}\Big) + \mathcal{O}\Big(\tfrac{1}{\log N}\Big) \leqslant \frac{C}{\log N}.
\end{aligned}$$

□

Lemma 2.38.

$$\tilde{K}_{\tau_J}^2 \mid X, K_0^1 \sim \mathrm{Bin}(n^{BB} - K_0^1, F_J^2),$$

with $F_J^2 := 1 - \prod_{t=\tau_J+1}^{\tau_N}(1 - \theta_t^{r_2})$.

Proof. Consider the process $\bar{K}_t^2 := n^{BB} - K_0^1 - \tilde{K}_{\tau-t}^2$. Then, $\bar{K}_0^2 = n^{BB} - K_0^1$ and further, $\bar{K}_1^2 \sim \mathrm{Bin}(\bar{K}_0^2, 1 - \theta_\tau^{r_2})$, as only those who were not counted by $\tilde{K}_{\tau-1}^2$ will be a "success" for

the process $\bar{K}^2_1$. Next, $\bar{K}^2_2 \mid \bar{K}^2_1 \sim \mathrm{Bin}(\bar{K}^2_1, 1-\theta^{r_2}_{\tau-1})$ and in general

$$\bar{K}^2_t \mid \bar{K}^2_{t-1} \sim \mathrm{Bin}(\bar{K}^2_{t-1}, 1-\theta^{r_2}_{\tau-(t-1)}), \text{ i.e. } \bar{K}^2_t \sim \mathrm{Bin}(\bar{K}^2_0, \prod_{t'=\tau-t+1}^{\tau_N}(1-\theta^{r_2}_{t'})),$$

by the properties of binomially distributed random variables. This implies

$$\tilde{K}^2_t = n^{BB} - K^1_0 - \bar{K}^2_{\tau-t} \sim \mathrm{Bin}(n^{BB} - K^1_0, 1 - \prod_{t'=t+1}^{\tau_N}(1-\theta^{r_2}_{t'})).$$

□

So far we have considered the following three groups: individuals who account for D^0_0 (those who were not counted by neither of the processes K^1 or K^2), individuals who add to D^1_0 (counted by K^2) and those who are either part of D^3_0 or D^4_0. It remains to distinguish between the last two groups. This consists in deciding whether an individual who experienced a recombination between SL and N1 experiences any recombination between N1 and N2 during the sweep. Here, we have two possibilities to consider: either locus N2 did already recombine into the b-population before that individual leads to an increase of the process K^1. Or, the loci recombined together into the b-population and split within this background.
Conditional on K^1_0 we define the process which counts the individuals which are affected by either event. Let $0 \leqslant t \leqslant \tau_N$,

$$K^{34}_t := \#\{i \in [n^{BB}] : A^{t'}_{\tau_N}(i,1) \neq A^{t'}_{\tau_N}(i,2), B_{t'}(A^{t'}_{\tau_N}(i,2)) = 0 \text{ for some } t' \geqslant t, R(i,1) \geqslant 0\}. \tag{2.86}$$

If we assume that an individual's two neutral loci, of which one already found an ancestor of type b, will never coalesce again after separation and do not migrate back into the B-population (approximately true thanks to Propositions 2.11 and 2.12), K^{34}_t is exactly the number of those lineages out of K^1_0 where the two neutral loci are no longer connected within one individual and at least the second neutral locus resides in the b-population. In any case, K^{34}_t is exactly the number D^4_0 defined through $\tilde{\mathcal{A}}_4$. We will now prove the following Lemma.

Lemma 2.39. *Define for* $\tilde{\theta}^{34}_t = p^{r_2}_B(X_{t-1}, X_t) \cdot \mathbf{1}_{\{X_t = X_{t-1}+1\}}$ *the process*

$$\tilde{K}^{34} := (\tilde{K}^{34}_t)_{\tau_N \geqslant t \geqslant 0} \text{ with } \tilde{K}^{34}_{\tau_N} = 0 \text{ and}$$
$$\tilde{K}^{34}_{t-1} - \tilde{K}^{34}_t \mid X, K^1_0, (\tilde{K}^{34}_u)^{\tau_N}_{u=t} \sim \mathrm{Bin}(K^1_0 - \tilde{K}^{34}_t, \tilde{\theta}^{34}_t).$$

Then, for $J \in [N]$,

$$\mathbb{P}(K^{34}_t \neq \tilde{K}^{34}_t \text{ for some } t \geqslant \tau_J) \leqslant \frac{C}{\log N}.$$

Proof. The proof uses the same idea as the proof of Lemma 2.37 but is more involved in the calculation of the expectation of the difference $K_{t-1}^{34} - K_t^{34}$.

Analogously to (2.80) we define

$$\mathcal{G}_t^{12,k} := \{i \in [2N] : A_{\tau_N}^t(m,1) = A_{\tau_N}^t(m,2) = i \text{ for exactly } k \text{ different } m \in [n^{BB}]\},$$
$$k = 0, 1, \ldots, n^{BB},\ 0 \leqslant t \leqslant \tau_N.$$

We can proceed as follows by distinguishing between the different backgrounds of locus N2 and describing the event with the help of the variables from page 34.

$$\begin{aligned}
&\mathbb{E}[K_{t-1}^{34} - K_t^{34} \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}] = \sum_{k=0}^{n^{BB}} k \cdot \mathbb{P}\big(K_{t-1}^{34} - K_t^{34} = k \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big) \\
&= \sum_{k=0}^{n^{BB}} k \cdot \sum_{i=1}^{2N} \Big\{ \mathbb{P}\big(I_{t,1} = i, i \in \mathcal{G}_t^{12,k}, B_{t-1}(I_{t,3}) = 0, I_{t,6} = 1, B_t(i) = 0, B_t(I_{t,2}) = 0 \\
&\qquad \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big) + \mathbb{P}\big(I_{t,1} = i, i \in \mathcal{G}_t^{12,k}, B_{t-1}(I_{t,3}) = 0, I_{t,6} = 1, B_t(i) = 1, \\
&\qquad B_t(I_{t,2}) = 1, I_{t,5} = 0, \exists s < t : B_s(A_t^s(i,1)) = 0 \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big) \Big\}. \qquad (2.87)
\end{aligned}$$

We will calculate both probabilities separately.

$$\begin{aligned}
&\mathbb{P}\big(I_{t,1} = i, i \in \mathcal{G}_t^{12,k}, B_{t-1}(I_{t,3}) = 0, I_{t,6} = 1, B_t(i) = 0, B_t(I_{t,2}) = 0 \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big) \\
&= \mathbb{P}\big(I_{t,6} = 1\big) \cdot \mathbb{P}\big(B_{t-1}(I_{t,3}) = 0 \mid X\big) \cdot \mathbb{P}\big(i \in \mathcal{G}_t^{12,k}, B_t(i) = 0 \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big) \\
&\quad \cdot \mathbb{P}\big(B_t(I_{t,2}) = 0, I_{t,1} = i \mid i \in \mathcal{G}_t^{12,k}, B_t(i) = 0, X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big) \\
&= r_2 \cdot \frac{2N - X_{t-1}}{2N} \cdot \mathbb{P}\big(i \in \mathcal{G}_t^{12,k}, B_t(i) = 0 \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big) \\
&\quad \cdot \begin{cases} \frac{(2N - X_t)\cdot 1}{X_t^2 + (2N - X_t)^2 + s X_t (2N - X_t)} = \frac{\beta_{X_t}}{X_t}, & X_t = X_{t-1}, \\ \frac{1}{2N - X_t}, & X_t = X_{t-1} - 1. \end{cases} \qquad (2.88)
\end{aligned}$$

by similar arguments as for (2.82). The calculation of the other probability from (2.87) involves some more thoughts. Note that the condition $\{\exists s < t : B_s(A_t^s(i,1)) = 0\}$ does not provide any information on the type of $(i,1)$ at time $t-1$ if it is combined with the condition that $\{I_{t,5} = 0\}$.

$$\begin{aligned}
&\mathbb{P}\big(I_{t,1} = i, i \in \mathcal{G}_t^{12,k}, B_{t-1}(I_{t,3}) = 0, I_{t,6} = 1, B_t(i) = 1, B_t(I_{t,2}) = 1, \\
&\qquad \underbrace{I_{t,5} = 0, \exists s < t : B_s(A_t^s(i,1)) = 0}_{\text{no information on time } t-1} \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big) \\
&= \mathbb{P}\big(I_{t,6} = 1\big) \cdot \mathbb{P}\big(B_{t-1}(I_{t,3}) = 0 \mid X\big) \\
&\quad \cdot \mathbb{P}\big(B_t(I_{t,2}) = 1, I_{t,1} = i \mid i \in \mathcal{G}_t^{12,k}, B_t(i) = 1, X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big)
\end{aligned}$$

$$\cdot \underbrace{\mathbb{P}\big(I_{t,5}=0 \mid i\in\mathcal{G}_t^{12,k}, B_t(i)=1, \exists s<t: B_s(A_t^s(i,1))=0, X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big)}_{=1-\mathcal{O}(1/\log(N))}$$
$$\cdot \mathbb{P}\big(i\in\mathcal{G}_t^{12,k}, B_t(i)=1, \exists s<t: B_s(A_t^s(i,1))=0 \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big)$$
$$= r_2\cdot\frac{2N-X_{t-1}}{2N}\cdot\mathbb{P}\big(i\in\mathcal{G}_t^{12,k}, B_t(i)=1, \exists s<t: B_s(A_t^s(i,1))=0 \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big)$$
$$\cdot\begin{cases}\frac{\beta_{X_t}}{2N-X_t}, & X_t=X_{t-1},\\ \frac{1}{X_t}, & X_t=X_{t-1}+1,\end{cases}$$

where the last equality is true by (2.82). Using (2.88) and the above result in (2.87) gives us

$$\begin{aligned}&\mathbb{E}\big[K_{t-1}^{34}-K_t^{34}\mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big]\\ &= r_2\frac{2N-X_{t-1}}{2N}\sum_{k=0}^{n^{BB}}k\sum_{i=1}^{2N}\Big\{\mathbb{P}\big(i\in\mathcal{G}_t^{12,k}, B_t(i)=0\mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big)\big[\tfrac{\beta_{X_t}}{X_t}\mathbf{1}_{\{X_t=X_{t-1}\}}\\ &\quad+\tfrac{1}{2N-X_t}\mathbf{1}_{\{X_t=X_{t-1}-1\}}\big]+\Big(1-\mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)\Big)\big[\tfrac{\beta_{X_t}}{2N-X_t}\mathbf{1}_{\{X_t=X_{t-1}\}}+\tfrac{1}{X_t}\mathbf{1}_{\{X_t=X_{t-1}+1\}}\big]\\ &\quad\cdot\mathbb{P}\big(i\in\mathcal{G}_t^{12,k}, B_t(i)=1, \exists s<t: B_s(A_t^s(i,1))=0\mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big)\Big\}.\end{aligned}$$

Similar as in (2.83) we can now rewrite the sums over k and i. To this end, define the number

$$K_t^{b,34}:=\#\big\{m\in[n^{BB}]: B_t(A_{\tau_N}^t(m,1))=0, A_{\tau_N}^t(m,1)=A_{\tau_N}^t(m,2)\big\}.$$

Then,

$$\begin{aligned}&\sum_{k=0}^{n^{BB}}k\sum_{i=1}^{2N}\mathbb{P}\big(i\in\mathcal{G}_t^{12,k}, B_t(i)=0\mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big)\\ &=\mathbb{E}\big[\sum_{i=1}^{2N}\sum_{k=0}^{n^{BB}}k\cdot\mathbf{1}_{\{i\in\mathcal{G}_t^{12,k}, B_t(i)=0\}}\mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big]=\mathbb{E}\big[K_t^{b,34}\mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big],\end{aligned}$$

and, by definition of K_0^1 and K_t^{34} we further have

$$\begin{aligned}&\sum_{k=0}^{n^{BB}}k\sum_{i=1}^{2N}\mathbb{P}\big(i\in\mathcal{G}_t^{12,k}, B_t(i)=1, \exists s<t: B_s(A_t^s(i,1))=0\mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big)\\ &=\mathbb{E}\big[\sum_{i=1}^{2N}\sum_{k=0}^{n^{BB}}k\cdot\mathbf{1}_{\{i\in\mathcal{G}_t^{12,k}, B_t(i)=1, \exists s<t: B_s(A_t^s(i,1))=0\}}\mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big]\\ &=K_0^1-K_t^{34}-\mathbb{E}\big[K_t^{b,34}\mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}\big].\end{aligned}$$

Hence, letting

$$E_t^{b,34} := \mathbb{E}[K_t^{b,34} \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}],$$

and consulting the values of the recombination probabilities from Lemma 2.26 and Lemma 2.27 we get

$$\begin{aligned}
&\mathbb{E}[K_{t-1}^{34} - K_t^{34} \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}] \\
&= r_2 \frac{2N - X_{t-1}}{2N} \Big\{ \big[\tfrac{\beta_{X_t}}{X_t} \mathbf{1}_{\{X_t = X_{t-1}\}} + \tfrac{1}{2N - X_t} \mathbf{1}_{\{X_t = X_{t-1} - 1\}}\big] \mathbb{E}[K_t^{b,34} \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}] \\
&\quad + \Big[1 - \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)\Big] \big[\tfrac{\beta_{X_t}}{2N - X_t} \mathbf{1}_{\{X_t = X_{t-1}\}} + \tfrac{1}{X_t} \mathbf{1}_{\{X_t = X_{t-1}+1\}}\big] \\
&\quad \cdot \big[K_0^1 - K_t^{34} - \mathbb{E}[K_t^{b,34} \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}]\big] \Big\} \\
&= p_{bb}^{r_2}(X_{t-1}, X_t) \mathbf{1}_{\{X_t = X_{t-1}\}} E_t^{b,34} + p_B^{r_2}(X_{t-1}, X_t) \mathbf{1}_{\{X_t = X_{t-1}+1\}} \big[K_0^1 - K_t^{34} - E_t^{b,34}\big] \\
&\quad \cdot \Big[1 - \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)\Big] + r_2 \frac{2N - X_{t-1}}{2N} \Big[\tfrac{1}{2N - X_t} \mathbf{1}_{\{X_t = X_{t-1} - 1\}} E_t^{b,34} \\
&\quad + \tfrac{\beta_{X_t}}{2N - X_t} \mathbf{1}_{\{X_t = X_{t-1}\}} \big[K_0^1 - K_t^{34} - E_t^{b,34}\big] \Big[1 - \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)\Big]\Big] \\
&= p_B^{r_2}(X_{t-1}, X_t) \mathbf{1}_{\{X_t = X_{t-1}+1\}} \big[K_0^1 - K_t^{34}\big] \Big[1 - \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)\Big] \\
&\quad + \Big[p_{bb}^{r_2}(X_{t-1}, X_t) \mathbf{1}_{\{X_t = X_{t-1}\}} - p_B^{r_2}(X_{t-1}, X_t) \mathbf{1}_{\{X_t = X_{t-1}+1\}}\Big] \cdot E_t^{b,34} \\
&\quad + \mathcal{O}\Big(\tfrac{1}{N \cdot \log(N)}\Big) \big[\beta_{X_t} \mathbf{1}_{\{X_t = X_{t-1}\}} + \mathbf{1}_{\{X_t = X_{t-1} - 1\}}\big].
\end{aligned}$$

We will now show that we can define the process $\tilde{K}^{34}$ as stated in Lemma 2.39, that is, in particular independent of the unknown value of the expectation $E_t^{b,34}$ and still get sufficient correspondence in their expectations in order to be able to apply Lemma 2.35.
With $\tilde{\theta}_t^{34} = p_B^{r_2}(X_{t-1}, X_t) \cdot \mathbf{1}_{\{X_t = X_{t-1}+1\}}$ as defined in the statement of the Lemma the expectation of the auxiliary process is

$$\mathbb{E}[\tilde{K}_{t-1}^{34} - \tilde{K}_t^{34} \mid X, K_0^1, (\tilde{K}_u^{34})_{u=t}^{\tau_N}] = \tilde{\theta}_t^{34}[K_0^1 - \tilde{K}_t^{34}] = p_B^{r_2}(X_{t-1}, X_t) \mathbf{1}_{\{X_t = X_{t-1}+1\}} [K_0^1 - \tilde{K}_t^{34}],$$

and hence the difference in the expectations of both processes on the set $\{K_u^{34} = \tilde{K}_u^{34}, t \leqslant u \leqslant \tau_N\}$ is equal to the following:

$$\begin{aligned}
\varepsilon(N; X, K_0^1, K_t^{34}, E_t^{b,34}) &:= \mathcal{O}(\tfrac{1}{\log(N)}) p_B^{r_2}(X_{t-1}, X_t) \mathbf{1}_{\{X_t = X_{t-1}+1\}} \big[K_0^1 - K_t^{34}\big] \\
&\quad - \Big[p_{bb}^{r_2}(X_{t-1}, X_t) \mathbf{1}_{\{X_t = X_{t-1}\}} - p_B^{r_2}(X_{t-1}, X_t) \mathbf{1}_{\{X_t = X_{t-1}+1\}}\Big] \cdot E_t^{b,34} \\
&\quad - \mathcal{O}\Big(\tfrac{1}{N \cdot \log(N)}\Big) \big[\beta_{X_t} \mathbf{1}_{\{X_t = X_{t-1}\}} + \mathbf{1}_{\{X_t = X_{t-1} - 1\}}\big] \\
&= E_t^{b,34} \cdot \frac{r_2}{2N} \frac{2N - X_{t-1}}{X_t} \big[\mathbf{1}_{\{X_t = X_{t-1}+1\}} - \beta_{X_t} \mathbf{1}_{\{X_t = X_{t-1}\}}\big]
\end{aligned}$$

$$+ \mathcal{O}(\tfrac{1}{\log(N)}) p_B^{r_2}(X_{t-1}, X_t) \mathbf{1}_{\{X_t = X_{t-1}+1\}} \big[K_0^1 - K_t^{34}\big]$$
$$- \mathcal{O}\Big(\tfrac{1}{N \cdot \log(N)}\Big) \big[\beta_{X_t} \mathbf{1}_{\{X_t = X_{t-1}\}} + \mathbf{1}_{\{X_t = X_{t-1}-1\}}\big], \tag{2.89}$$

that is, on the set $\{K_u^{34} = \tilde{K}_u^{34}, t \leqslant u \leqslant \tau_N\}$ we have

$$\mathbb{E}(K_{t-1}^{34} - K_t^{34} \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}) = \mathbb{E}(\tilde{K}_{t-1}^{34} - \tilde{K}_t^{34} \mid X, K_0^1, (K_u^{34})_{u=t}^{\tau_N}) - \mathcal{O}(\varepsilon(N; X, K_0^1, K_t^{34}, E_t^{b,34})),$$

and can apply Lemma 2.35 in order to get a bound on the probability that K_t^{34} is different from $\tilde{K}_t^{34}$ for some t. By proceeding similarly as in (2.84) from the proof of Lemma 2.37 this bound reads as follows:

$$\mathbb{P}(K_t^{34} \neq \tilde{K}_t^{34} \text{ for some } \tau_N \geqslant t \geqslant \tau_J \mid X, K_0^1) \leqslant K_0^1 \cdot \sum_{t=\tau_J+1}^{\tau_N} \Big[\mathbb{P}(K_{t-1}^{34} - K_t^{34} \geqslant 2 \mid X, K_0^1) + \mathbb{P}(\tilde{K}_{t-1}^{34} - \tilde{K}_t^{34} \geqslant 2 \mid X, K_0^1) + \mathbb{E}\Big[\mathcal{O}(\varepsilon(N; X, K_0^1, K_t^{34}, E_t^{b,34})) \mid X, K_0^1\Big]\Big] + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big). \tag{2.90}$$

This inequality however will only be of use if we can show that in particular the error term given through $\varepsilon(N; X, K_0^1, K_t^{34}, E_t^{b,34})$ will only contribute with terms of order at most $1/\log(N)$ even we consider its sum over all t up to τ_N. This statement is given in the following Lemma.

Lemma 2.40. *We have the following bound for the function ε from* (2.89)*:*

$$\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N} \varepsilon(N; X, K_0^1, K_t^{34}, E_t^{b,34})\Big] \leqslant C/\log(N).$$

Proof. We first observe that we know from (2.63), that the expectation of the sum over all $\theta_t^{r_2}$ is bounded by a constant and hence the term in the second to last line in (2.89) only contributes with terms of order $1/\log(N)$ in the end.
Even more obvious is that the term in the last line of (2.89) will as well only give terms of order at most $1/\log(N)$ as we have only terms of order at most $1/(N\log(N)$ in each summand. It remains to bound the remaining term from ε. Consulting (2.37) and (2.38) from Lemma 2.23 , we can calculate the following expectation for $k \in [N]$:

$$\mathbb{E}\big[U_k - \beta_k H_k\big] = \mathbb{E}\Big[U_k - \frac{U_k + D_k}{2-s}\Big] = \mathbb{E}\Big[\frac{1-s}{2-s} \cdot U_k - \frac{U_{k-1} - 1}{2-s}\Big]$$
$$= \frac{[1-s][1-(1-s)^{k+1}][1-(1-s)^{2N-k}] - [1-(1-s)^k][1-(1-s)^{2N-k+1}] + s[1-(1-s)^{2N}]}{(2-s) \cdot s[1-(1-s)^{2N}]}$$

$$= \frac{-(1-s)^{k+2} + (1-s)^{2N+2} + (1-s)^k - (1-s)^{2N+1} - s(1-s)^{2N}}{(2-s)\cdot s(1-(1-s)^{2N})}$$
$$= \frac{(1-s)^k(1-(1-s)^2) - (1-s)^{2N}(s+(1-s)-(1-s)^2)}{(2-s)\cdot s(1-(1-s)^{2N})} = \frac{(1-s)^k - (1-s)^{2N}}{(1-(1-s)^{2N})}$$
$$= (1-s)^k \cdot \frac{(1-(1-s)^{2N-k})}{(1-(1-s)^{2N})} \geqslant 0. \tag{2.91}$$

With this we can continue using that the unknown expectation $E_t^{b,34}$ is bounded by the sample size, that is n^{BB} here, independent of N.

$$\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N} \varepsilon(N; X, K_0^1, K_t^{34}, E_t^{b,34})\Big]$$
$$= \mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N} E_t^{b,34} \cdot \frac{r_2}{2N}\frac{2N-X_{t-1}}{X_t}\big[\mathbf{1}_{\{X_t=X_{t-1}+1\}} - \beta_{X_t}\mathbf{1}_{\{X_t=X_{t-1}\}}\big]\Big] + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)$$
$$= \mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N} E_t^{b,34}\Big[\mathbf{1}_{\{X_t=X_{t-1}+1\}}\frac{r_2(2N-X_{t-1})}{2NX_{t-1}} - \mathbf{1}_{\{X_t=X_{t-1}\}}\beta_{X_t}\frac{r_2(2N-X_{t-1})}{2NX_{t-1}}\Big]\Big] + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)$$
$$\leqslant n^{BB}\cdot\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N}\frac{r_2(2N-X_{t-1})}{2NX_{t-1}}\Big\{\mathbf{1}_{\{X_t=X_{t-1}+1\}} - \mathbf{1}_{\{X_t=X_{t-1}\}}\beta_{X_t}\Big\}\Big] + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)$$
$$= n^{BB}\cdot\sum_{k=J}^{N}\frac{r_2(2N-k)}{2Nk}\mathbb{E}[U_k - \beta_k H_k] + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big)$$
$$= n^{BB}\cdot\frac{r_2}{2N}\cdot\sum_{k=J}^{N}\frac{2N-k}{k}(1-s)^k\cdot\underbrace{\frac{(1-(1-s)^{2N-k})}{(1-(1-s)^{2N})}}_{\leqslant 1} + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big), \quad \text{by (2.91)},$$
$$\leqslant n^{BB}\frac{r_2}{2N}\frac{2N-J}{J}\cdot\sum_{k=J}^{N}(1-s)^k + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big) \leqslant n^{BB}\cdot\frac{r_2}{J}\cdot\frac{1}{s} + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big) \leqslant \frac{C}{\log(N)}.$$

□

Returning to (2.90) it remains to bound the expectations of the other two terms. By the definition of $\tilde{\theta}_t^{34}$ and (2.61) we get

$$\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N}\mathbb{P}(\tilde{K}_{t-1}^{34} - \tilde{K}_t^{34} \geqslant 2 \mid X, K_0^1)\Big] \leqslant \binom{n^{BB}}{2}\cdot\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N}(\theta_t^{bb})^2 + (\theta_t^{r_2})^2\Big] \leqslant \frac{C}{(\log N)^2},$$

and by similar calculations as in (2.85) and using Propositions 2.12 and 2.13 we obtain

$$\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau_N}\mathbb{P}(K_{t-1}^{34} - K_t^{34} \geqslant 2 \mid X, K_0^1)\Big] \leqslant \frac{C}{\log(N)},$$

as K_t^{34} can only increase by more than one if we have a coalescence prior to the particular

split of the two neutral loci. Both bounds and Lemma 2.40 result in the final statement

$$\mathbb{P}\big(K_t^{34} \neq \tilde{K}_t^{34} \text{ for some } \tau_N \geqslant t \geqslant \tau_J\big) \leqslant \frac{C}{\log(N)}.$$

□

Lemma 2.41. *We have*

$$\tilde{K}_{\tau_J}^{34} \mid X, \tilde{K}_0^1 \sim \mathrm{Bin}(K_0^1, F_J^{34}),$$

with $F_J^{34} := 1 - \prod_{t=\tau_J+1}^{\tau}(1-\tilde{\theta}_t^{34})$ *and* $\tilde{\theta}_t^{34} = p_B^{r_2}(X_{t-1}, X_t) \cdot \mathbf{1}_{\{X_t = X_{t-1}+1\}}$.

The proof is exactly the same as for Lemma 2.38 and is thus omitted here.

Let us summarize the result for K_t^{34} at this point: we know that conditional on X and K_0^1 the number of individuals whose neutral loci find different b-ancestors is approximately binomially distributed with parameters K_0^1 and F_1^{34}. What is left to show is that we can approximately calculate the expectation of F_J^{34} and that it gives us $\mathbb{E}(F_J^{34}) = \mathbb{E}(F_J^2) + \mathcal{O}(1/\log(N)$ with F_J^2 as in Lemma 2.38. This can be done very easily by performing similar calculations as in (2.63) from the proof of Lemma 2.33.

Lemma 2.42. *With* $\tilde{\theta}_t^{34} = p_B^{r_2}(X_{t-1}, X_t) \cdot \mathbf{1}_{\{X_t = X_{t-1}+1\}}$ *as above and* $\tilde{\eta}_J^{34} := \sum_{t=\tau_J+1}^{\tau_N} \tilde{\theta}_t^{34}$ *we have*

$$\mathbb{E}\Big(\tilde{\eta}_J^{34}\Big) = \frac{r_2}{s}\sum_{k=J}^{N}\frac{1}{k} + \mathcal{O}\Big(\frac{1}{\log(N)}\Big).$$

Further, it holds that

$$\big|\mathbb{E}(F_J^{34}) - (1 - \exp(-\mathbb{E}(\tilde{\eta}_J^{34})))\big| \leqslant C/\log(N).$$

In particular, this Lemma implies that $\mathbb{E}(F_1^{34}) = q^{(2)} + \mathcal{O}(1/\log(N))$, with $q^{(2)}$ as defined in (2.18).

Proof. Recall the calculation of the expectation of a similar η in (2.63) and note that the term from the holds in $\theta_t^{r(2)} = p_B^{r(2)}(X_{t-1}, X_t)$ only contributed as term of order at most $1/\log(N)$ to the expectation of the there considered $\tilde{\eta}_k$. With this, we indeed get the same result with the same error bound using r_2 instead of $r(2)$:

$$\mathbb{E}[\tilde{\eta}_J^{34}] = \frac{r_2}{s}\sum_{k=J}^{N}\frac{1}{k} + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big).$$

The fact that we can indeed perform a Poisson approximation for F_J^{34} holds true as $\tilde{\theta}_t^{34} \leqslant p_B^{r_2}(X_{t-1}, X_t) \leqslant p_B^{r(2)}(X_{t-1}, X_t)$, and (2.61) from the proof of Lemma 2.33.
Applying the Mean value theorem and Jensen's Inequality ends the proof as described in (2.44):

$$\begin{aligned}
&\big|\mathbb{E}\big[F_J^{34}\big] - (1 - \exp(-\mathbb{E}[\tilde{\eta}_J^{34}]))\big| \\
&\leqslant \mathbb{E}\big|\exp(-\tilde{\eta}_J^{34}) - \exp\big(-\mathbb{E}[\tilde{\eta}_J^{34}]\big)\big| + \mathcal{O}\Big(\tfrac{1}{\log(N)}\Big) \leqslant \sqrt{\mathrm{Var}(\tilde{\eta}_J^{34})} + \frac{C}{\log(N)} \leqslant \frac{C}{\log(N)}.
\end{aligned}$$

□

Taking all previous results together, we can prove our claim due to the following observations: First, look at the joint distribution of the approximating processes $\tilde{K}_t := (\tilde{K}_t^1, \tilde{K}_t^2, \tilde{K}_t^{34})$ and note that we could have defined the processes K^2 and $\tilde{K}^2$ additionally conditional on K_0^{34} with the same result since K^{34} and K^2 operate on disjoint sets and the information on which individuals are not available for the process K^2 is already included in the condition K_0^1. With this argument in mind we get by the Lemmas 2.36, 2.38 and 2.41:

$$\begin{aligned}
&\mathbb{P}(\tilde{K}_0^1 = k_1, \tilde{K}_0^2 = k_2, \tilde{K}_0^{34} = k_{34} \mid X) = \mathbb{P}(\tilde{K}_0^2 = k_2, \tilde{K}_0^{34} = k_{34} \mid \tilde{K}_0^1 = k_1, X)\mathbb{P}(\tilde{K}_0^1 = k_1 \mid X) \\
&= \mathbb{P}(\tilde{K}_0^2 = k_2 \mid \tilde{K}_0^1 = k_1, X)\mathbb{P}(\tilde{K}_0^{34} = k_{34} \mid \tilde{K}_0^1 = k_1, X)\mathbb{P}(\tilde{K}_0^1 = k_1 \mid X) \\
&= \binom{n^{BB} - k_1}{k_2} (F_1^2)^{k_2}(1 - F_1^2)^{n^{BB} - k_1 - k_2} \binom{k_1}{k_{34}} (F_1^{34})^{k_{34}}(1 - F_1^{34})^{k_1 - k_{34}} \\
&\quad \cdot \binom{n^{BB}}{k_1} (F_1^1)^{k_1}(1 - F_1^1)^{n^{BB} - k_1} \\
&= \frac{n^{BB}!}{k_2! k_{34}! (k_1 - k_{34})! (n^{BB} - k_1 - k_2)!} [(1 - F_1^1)(1 - F_1^2)]^{n^{BB} - k_1 - k_2} [(1 - F_1^1)F_1^2]^{k_2} \\
&\quad \cdot [F_1^1(1 - F_1^{34})]^{k_1 - k_{34}} [F_1^1 F_1^{34}]^{k_{34}} \\
&= \mathbb{P}(\bar{K}_0^X = (n^{BB} - k_1 - k_2, k_2, k_1 - k_{34}, k_{34}) \mid X),
\end{aligned} \tag{2.92}$$

with $\bar{K}_0^X \mid X$ distributed multinomially,

$$\bar{K}_0^X \mid X \sim \mathrm{Mult}(n^{BB}; (1 - F_1^1)(1 - F_1^2), (1 - F_1^1)F_1^2, F_1^1(1 - F_1^{34}), F_1^1 F_1^{34}).$$

Second, let $K_t := (K_t^1, K_t^2, K_t^{34})$. Then, combining the statements from Lemma 2.36, 2.37 and 2.39 results in

$$\mathbb{P}(K_t \neq \tilde{K}_t \text{ for some } t \geqslant \tau_J) \leqslant \sum_{j \in \{1,2,34\}} \mathbb{P}(K_t^j \neq \tilde{K}_t^j \text{ for some } t \geqslant \tau_J) \leqslant \frac{C}{\log N}.$$

Last, consider the process D, defined in (2.25). From the construction of the processes K^1, K^2

and K^{34} we get the following for $d_0, d_1, d_2, d_3, d_4 \in [n^{BB}]$ with $\sum_{j=0}^4 d_j = n^{BB}$ and $d_2 = 0$:

$$\begin{aligned}&\mathbb{P}(D_0 = (d_0, d_1, d_2, d_3, d_4)) = \mathbb{P}(n^{BB} - K_0^1 - K_0^2 = d_0, K_0^2 = d_1, K_0^1 - K_0^{34} = d_3, K_0^{34} = d_4)\\ &= \mathbb{P}(K_0^1 = d_3 + d_4, K_0^2 = d_1, K_0^{34} = d_4),\end{aligned}$$

and therefore it follows by the above and (2.92) that

$$\begin{aligned}&\Big|\mathbb{P}(D_0 = (d_0, d_1, d_2, d_3, d_4)) - \mathbb{P}(\tilde{K}_0^1 = d_3 + d_4, \tilde{K}_0^2 = d_1, \tilde{K}_0^{34} = d_4)\Big|\\ &= \Big|\mathbb{P}(D_0 = (d_0, d_1, d_2, d_3, d_4)) - \mathbb{P}(\bar{K}_0^X = (\underbrace{n^{BB} - d_1 - d_3 - d_4}_{=d_0}, d_1, d_3, d_4))\Big| \leqslant \frac{C}{\log N}. \qquad (2.93)\end{aligned}$$

The only thing left to show is that $q^{(1)}, q^{(2)}$ from (2.18) have the correct form.

Lemma 2.43. *For any $d_0, d_1, d_2, d_3, d_4 \in [n^{BB}]$ with $\sum_{j=0}^4 d_j = n^{BB}$ and $d_2 = 0$ it holds that*

$$\begin{aligned}&\Big|\mathbb{E}\Big[\big[(1-F_1^1)(1-F_1^2)\big]^{d_0}\big[(1-F_1^1)F_1^2\big]^{d_1}\big[F_1^1(1-F_1^{34})\big]^{d_3}\big[F_1^1F_1^{34}\big]^{d_4}\Big]\\ &\qquad - \big[(1-q^{(1)})(1-q^{(2)})\big]^{d_0}\big[(1-q^{(1)})q^{(2)}\big]^{d_1}\big[q^{(1)}(1-q^{(2)})\big]^{d_3}\big[q^{(1)}q^{(2)}\big]^{d_4}\Big| \leqslant \frac{C}{\log N}.\end{aligned}$$

Proof. Lemma 3.4.3 in [15] implies

$$\begin{aligned}&\Big|\mathbb{E}\Big[\big[(1-F_1^1)(1-F_1^2)\big]^{d_0}\big[(1-F_1^1)F_1^2\big]^{d_1}\big[F_1^1(1-F_1^{34})\big]^{d_2}\big[F_1^1F_1^{34}\big]^{d_3}\Big]\\ &\qquad - \big[(1-q^{(1)})(1-q^{(2)})\big]^{d_0}\big[(1-q^{(1)})q^{(2)}\big]^{d_1}\big[q^{(1)}(1-q^{(2)})\big]^{d_2}\big[q^{(1)}q^{(2)}\big]^{d_3}\Big|\\ &\leqslant \mathbb{E}\Big[d_0|(1-F_1^1)(1-F_1^2) - (1-q^{(1)})(1-q^{(2)})| + d_1|(1-F_1^1)F_1^2 - (1-q^{(1)})q^{(2)}|\\ &\qquad + d_2|F_1^1(1-F_1^{34}) - q^{(1)}(1-q^{(2)})| + d_3|F_1^1F_1^{34} - q^{(1)}q^{(2)}|\Big]\\ &\leqslant \mathbb{E}\Big[d_0\big(|(1-F_1^1) - (1-q^{(1)})| + |(1-F_1^2) - (1-q^{(2)})|\big) + d_1\big(|(1-F_1^1) - (1-q^{(1)})| + |F_1^2 - q^{(2)}|\big)\\ &\qquad + d_2\big(|F_1^1 - q^{(1)}| + |(1-F_1^{34}) - (1-q^{(2)})|\big) + d_3\big(|F_1^1 - q^{(1)}| + |F_1^{34} - q^{(2)}|\big)\Big]\\ &= \mathbb{E}\Big[n^{BB}|F_1^1 - q^{(1)}| + (d_0+d_1)|F_1^2 - q^{(2)}| + (d_2+d_3)|F_1^{34} - q^{(2)}|\Big]\end{aligned}$$

In Lemma 2.42 we have shown that $|F_1^{34} - F_1^2| \leqslant C/\log(N)$ for some constant C and thus

$$|F_1^{34} - q^{(2)}| \leqslant |F_1^{34} - F_1^2| + |F_1^2 - q^{(2)}| \leqslant |F_1^2 - q^{(2)}| + \mathcal{O}(1/\log(N)).$$

For $j = 1, 2$, terms of the type $\mathbb{E}|F_1^j - q_j|$ have already been treated in the proof of Proposition 2.15. Let $\eta_1^j := \sum_{j=1}^\tau \theta_t^{r_j}$ for $j = 1, 2$. Then, by (2.61), (2.63), (2.65), (2.66) and Lemma 3.10 in [36] we get similarly as in (2.44)

$$|F_1^j - q_j| \leqslant |F_1^j - (1 - \exp(-\eta_1^j))| + |\exp(-\eta_1^j) - \exp(-\mathbb{E}(\eta_1^j))|$$

$$+ \left|\exp(-\mathbb{E}(\eta_1^j)) - \exp\left(-\tfrac{r_j}{s}\log(2N)\right)\right|, \quad \text{and hence}$$
$$\mathbb{E}|F_1^j - q_j| \leqslant C/(\log N)^2 + C/\log(N) \leqslant C/\log(N).$$

This ends the proof as the error terms are of the correct order and will only be multiplied by finite values less or equal to n^{BB}. □

Applying the above Lemma to the previous results which gave Equation (2.93) then finishes the proof of Proposition 2.22. □

Proof of Proposition 2.19. The proof of this Proposition is almost identical to the proof of Lemma 5.2, and thus Proposition 2.5, from [36]. Any details can be taken from the above proof of Proposition 2.20 and thus we will be very brief here. Keep in mind that by Proposition 2.9 there are no further relevant events, such as migrations into the b-population, happening between times τ_N and τ.
For $K^{BB} := (K_t^{BB})_{\tau_N \leqslant t \leqslant \tau}$ with K_t^{BB} as defined in (2.22) we have $K_\tau^{BB} = 0$ and further we can verify that $K_{t-1}^{BB} - K_t^{BB} \in \{0, 1, \ldots, n\}$ for all $\tau_N + 1 \leqslant t \leqslant \tau$ and

$$\mathbb{E}(K_{t-1}^{BB} - K_t^{BB} \mid X, (K_u^{BB})_{t \leqslant u \leqslant \tau}) = (n - K_t^{BB})\theta_t^{BB},$$

with $\theta_t^{BB} = p_{BB}^{r_2}(X_{t-1}, X_t)$ from Lemma 2.28. Let us define another process $\tilde{K}^{BB} = (\tilde{K}_t^{BB})_{\tau_N \leqslant t \leqslant \tau}$ as follows:

$$\tilde{K}_\tau^{BB} = 0 \;\text{ and }\; \tilde{K}_{t-1}^{BB} - \tilde{K}_t^{BB} \mid X, (\tilde{K}_u^{BB})_{t \leqslant u \leqslant \tau} \sim \mathrm{Bin}(n - \tilde{K}_t^{BB}, \theta_t^{BB}). \tag{2.94}$$

Then, on the set where $\{\tilde{K}_t^{BB} = K_t^{BB}\}$ we have

$$\mathbb{E}(\tilde{K}_{t-1}^{BB} - \tilde{K}_t^{BB} \mid X, (\tilde{K}_u^{BB})_{t \leqslant u \leqslant \tau}) = (n - K_t^{BB})\theta_t^{BB} = \mathbb{E}(K_{t-1}^{BB} - K_t^{BB} \mid X, (K_u^{BB})_{t \leqslant u \leqslant \tau}),$$

and thus can apply Lemma 2.35 in the same way as it is done in the proof of Lemma 5.2 in [36] in order to attain for any $J \geqslant N$

$$\begin{aligned}&\mathbb{P}(K_t^{BB} \neq \tilde{K}_t^{BB} \text{ for some } t \geqslant \tau_J \mid X)\\ &\leqslant n \sum_{t=\tau_J+1}^{\tau} \Big\{\mathbb{P}(K_{t-1}^{BB} - K_t^{BB} \geqslant 2 \mid X) + \mathbb{P}(\tilde{K}_{t-1}^{BB} - \tilde{K}_t^{BB} \geqslant 2 \mid X)\Big\}\end{aligned}$$

The first part in this sum can be bounded thanks to Proposition 2.17:

$$\begin{aligned}\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau} \mathbb{P}(K_{t-1}^{BB} - K_t^{BB} \geqslant 2 \mid X)\Big] &\leqslant n \cdot \mathbb{P}(K_{t-1}^{BB} - K_t^{BB} \geqslant 2 \text{ for some } \tau \geqslant t \geqslant \tau_J + 1)\\ &\leqslant n \cdot \mathbb{P}(\exists i, m, j : G^j(i, m) \geqslant R_B^{\mathrm{rec}}(i) \geqslant \tau_J) \leqslant C/\log(N).\end{aligned}$$

The second part in the sum can be bounded by properties of binomially distributed random variables by using (2.46) from the proof of Proposition 2.14:

$$\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau}\mathbb{P}(\tilde{K}_{t-1}^{BB}-\tilde{K}_t^{BB}\geqslant 2\mid X)\Big]\leqslant\binom{n}{2}\mathbb{E}\Big[\sum_{t=\tau_J+1}^{\tau}\big(\theta_t^{BB}\big)^2\Big]\leqslant C/(\log(N))^2.$$

The definition (2.94) of $\tilde{K}^{BB}$ implies that

$$\tilde{K}_{\tau_J}^{BB}\mid X\sim\mathrm{Bin}(n,F_J^{BB}),$$

with $F_J^{BB}=1-\prod_{t=\tau_J+1}^{\tau}(1-\theta_t^{BB})$, $J\geqslant N$, and thus we have for any $d=0,1,\ldots,n$:

$$\big|\mathbb{P}(K_{\tau_J}^{BB}=d)-\binom{n}{d}\mathbb{E}\big[(F_J^{BB})^d(1-F_J^{BB})^{n-d}\big]\big|\leqslant C/\log(N),$$

We can now finish the proof with similar calculations as in (2.44) and in the proof of Lemma 2.43: applying Lemma 3.4.3 in [15] and consulting Proposition 2.14 and (2.47) from its proof results in

$$\begin{aligned}&\big|\mathbb{E}\big[(F_J^{BB})^d(1-F_J^{BB})^{n-d}\big]-(q_J^{BB})^d(1-q_J^{BB})^{n-d}\big|\\&\leqslant d\cdot\mathbb{E}|F_J^{BB}-q_J^{BB}|+(n-d)\cdot\mathbb{E}|1-F_J^{BB}-(1-q_J^{BB})|\leqslant c/\log(N).\end{aligned}$$

□

Proof of Proposition 2.21. Proposition 2.20 tells us that, approximately, we can decide independently for each individual what kind of relationship its neutral loci take at the beginning of the sweep. It also gives the success probabilities for the corresponding multinomial distribution. The proof of this Corollary now follows from straightforward calculations by the following thoughts:
First consider the case that only the first neutral locus of an individual sampled at time τ_N is relevant. It will be ancestral to the mutant if this individual accounts for an increment of D^0 or D^1, and it will be ancestral to some b-individual if it accounts for any other entry of the vector D. Hence, by the distribution from $\tilde{D}$ as given in (2.27) the probability that N1 originates from the mutant is

$$q_B^{\mathrm{N1}}=(1-q^{(1)})(1-q^{(2)})+(1-q^{(1)})q^{(2)}=1-q^{(1)},$$

and the probability for the complementary event is equal to

$$q_b^{\mathrm{N1}}=q^{(1)}(1-q^{(2)})+q^{(1)}q^{(2)}=q^{(1)}.$$

For the second neutral locus, these probabilities can be derived in the same way consulting (2.27): N2 is ancestral to the mutant only if the individual is counted by D^0, with probability $q_B^{\mathrm{N2}} = (1-q^{(1)})(1-q^{(2)})$ and it escapes the sweep in all other cases, that is, with probability

$$q_b^{\mathrm{N2}} = (1-q^{(1)})q^{(2)} + q^{(1)}(1-q^{(2)}) + q^{(1)}q^{(2)} = q^{(1)} + q^{(2)} - q^{(1)}q^{(2)}.$$

In order to calculate the probability that both neutral loci of an individual from time τ which reside in different B-individuals at time τ_N behave in a specific way, we just have to multiply the corresponding probabilities for the individual lines from time τ_N because of the multinomial and therefore independent character of D. As an example, the success probability corresponding to the first entry of $\bar{D}$ results as the product of q_B^{N1} and q_B^{N2}. All other success probabilities from the vector arise in the same way and finally Propositions 2.12 and 2.13 guarantee that $\bar{D}^3 \equiv 0$ with high probability. □

Proof of Proposition 2.22. The proof of this Proposition follows from straightforward calculations by the statements of the Propositions 2.19, 2.20 and Corollary 2.21.
From the definition of the processes $\bar{D}$ from (2.28) and D as in (2.25) we know that for sample size n at time τ and conditional on $K_{\tau_N}^{BB} = k$, the vector D^* is given through $D^* = \bar{D}(k) + D(n-k)$, that is,

$$\begin{aligned}
\mathbb{P}(D^* = d^*) &= \sum_{k=0}^{n} \mathbb{P}(D^* = d^* \mid K_{\tau_N}^{BB} = k) \cdot \mathbb{P}(K_{\tau_N}^{BB} = k) \\
&= \sum_{k=0}^{n} \mathbb{P}(\bar{D}(k) + D(n-k) = d^* \mid K_{\tau_N}^{BB} = k) \cdot \mathbb{P}(K_{\tau_N}^{BB} = k) \\
&= \sum_{k=0}^{n} \mathbb{P}(K_{\tau_N}^{BB} = k) \cdot \sum_{0 \leqslant d \leqslant d^*} \mathbb{P}(\bar{D}(k) = d^* - d, D(n-k) = d \mid K_{\tau_N}^{BB} = k) \qquad (2.95)
\end{aligned}$$

By Proposition 2.20 we know that a sample from time τ_N can be approximately constructed by a multinomial experiment: for each sampled individual we can independently draw the ancestral relationship for its two neutral loci. This in particular implies that, given the number k of individuals whose neutral loci split within the B-population, the vectors $\tilde{\bar{D}}(k)$ and $\tilde{D}(n-k)$ are independent.
We will show that the process $\tilde{D}^*$ defined in (2.32) has approximately, up to an error term of order at most $1/\log(N)$, the same distribution as the sum $\tilde{\bar{D}} + \tilde{D}$. As

$$\mathbb{P}((\bar{D}, D) \neq (\tilde{\bar{D}}, \tilde{D})) \leqslant \mathbb{P}(\bar{D} \neq \tilde{\bar{D}}) + \mathbb{P}(D \neq \tilde{D}) \leqslant C/\log(N)$$

by Proposition 2.20 and Corollary 2.21, this then shows that D^* and $\tilde{D}^*$ have approximately the same distribution.
Let us hence consider the sum $\tilde{\bar{D}} + \tilde{D}$ instead of the true processes in (2.95). Keep in mind

that $\tilde{\tilde{D}}^3 \equiv 0$, as well as $\tilde{D}^2 \equiv 0$, and that if $D^* = d^*$ we need $K_{\tau_N}^{BB} \geqslant d_2^*$. On the other hand, $K_{\tau_N}^{BB}$ can at most reach the value $n - d_3^*$. For $d^* = (d_0^*, d_1^*, d_2^*, d_3^*, d_4^*)$ with $\sum_{i=0}^4 d_i^* = n$ we get by (2.30) and (2.27):

$$\mathbb{P}(\tilde{\tilde{D}} + \tilde{D} = d^*) = \sum_{k=d_2^*}^{n-d_3^*} \mathbb{P}(\tilde{\tilde{D}} + \tilde{D} = d^* \mid K_0^{BB} = k) \cdot \mathbb{P}(K_0^{BB} = k)$$

$$= \sum_{k=d_2^*}^{n-d_3^*} \sum_{0 \leqslant d \leqslant d^*} \mathbb{P}(\tilde{D} = d, \tilde{\tilde{D}} = d^* - d \mid K_0^{BB} = k) \cdot \mathbb{P}(K_0^{BB} = k) \cdot \mathbf{1}_{\{d_0+d_1+d_4 = n-k-d_3^*,\, d_2=0,\, d_3=d_3^*\}}$$

$$= \sum_{k=d_2^*}^{n-d_3^*} \sum_{\substack{0 \leqslant d \leqslant d^* \\ d_2=0, d_3=d_3^*}} \mathbb{P}(\tilde{D} = d \mid K_0^{BB} = k) \mathbb{P}(\tilde{\tilde{D}} = d^* - d \mid K_0^{BB} = k) \mathbb{P}(K_0^{BB} = k) \mathbf{1}_{\{d_0+d_1+d_4 = n-k-d_3^*\}}$$

$$= \sum_{k=d_2^*}^{n-d_3^*} \sum_{\substack{0 \leqslant d \leqslant d^* \\ d_2=0, d_3=d_3^*}} \mathbf{1}_{\{d_0+d_1+d_4 = n-k-d_3^*\}} \frac{(n-k)!}{d_0! d_1! d_3^*! d_4!} [(1-q^{(1)})(1-q^{(2)})]^{d_0} [(1-q^{(1)})q^{(2)}]^{d_1}$$

$$\cdot [q^{(1)}(1-q^{(2)})]^{d_3^*} [q^{(1)}q^{(2)}]^{d_4} \frac{k!}{(d_0^*-d_0)!(d_1^*-d_1)! d_2^*! (d_4^*-d_4)!} [(1-q^{(1)})(1-q^{(1)})(1-q^{(2)})]^{d_0^*-d_0}$$

$$\cdot [(1-q^{(1)})[1-(1-q^{(1)})(1-q^{(2)})]]^{d_1^*-d_1} [q^{(1)}(1-q^{(1)})(1-q^{(2)})]^{d_2^*}$$

$$\cdot [q^{(1)}[1-(1-q^{(1)})(1-q^{(2)})]]^{d_4^*-d_4} \frac{n!}{k!(n-k)!} (q^{BB})^k (1-q^{BB})^{n-k}$$

$$= \binom{n}{d_0^*, d_1^*, d_2^*, d_3^*, d_4^*} [(1-q^{(1)})(1-q^{(2)})]^{d_0^*} [(1-q^{(1)})]^{d_1^*} [q^{BB}q^{(1)}(1-q^{(1)})(1-q^{(2)})]^{d_2^*} [q^{(1)}]^{d_4^*}$$

$$\cdot [q^{(1)}(1-q^{(2)})(1-q^{BB})]^{d_3^*} \sum_{k=d_2^*}^{n-d_3^*} \sum_{\substack{0 \leqslant d \leqslant d^* \\ d_2=0, d_3=d_3^*}} \binom{d_0^*}{d_0} \binom{d_1^*}{d_1} \binom{d_4^*}{d_4} \overbrace{(q^{BB})^{k-d_2^*}}^{=(q^{BB})^{d_0^*-d_0+d_1^*-d_1+d_4^*-d_4}} \cdot \underbrace{(1-q^{BB})^{n-k-d_3^*}}_{=(1-q^{BB})^{d_0+d_1+d_4}}$$

$$\cdot [(1-q^{(1)})]^{d_0^*-d_0} [1-(1-q^{(1)})(1-q^{(2)})]^{d_1^*+d_4^*-d_1-d_4} [q^{(2)}]^{d_1+d_4} \cdot \mathbf{1}_{\{d_0+d_1+d_4 = n-k-d_3^*\}}$$

$$= \binom{n}{d_0^*, d_1^*, d_2^*, d_3^*, d_4^*} [(1-q^{(1)})(1-q^{(2)})]^{d_0^*} [(1-q^{(1)})]^{d_1^*} [q^{BB}q^{(1)}(1-q^{(1)})(1-q^{(2)})]^{d_2^*}$$

$$\cdot [q^{(1)}(1-q^{(2)})(1-q^{BB})]^{d_3^*} [q^{(1)}]^{d_4^*} \sum_{k=d_2^*}^{n-d_3^*} \sum_{\substack{0 \leqslant d \leqslant d^* \\ d_2=0, d_3=d_3^*}} \binom{d_0^*}{d_0} \binom{d_1^*}{d_1} \binom{d_4^*}{d_4} [q^{BB}(1-q^{(1)})]^{d_0^*-d_0} [1-q^{BB}]^{d_0}$$

$$\cdot [(1-q^{BB})q^{(2)}]^{d_1+d_4} [q^{BB}(1-(1-q^{(1)})(1-q^{(2)}))]^{d_1^*+d_4^*-d_1-d_4} \cdot \mathbf{1}_{\{d_0+d_1+d_4+k-d_2^* = d_0^*+d_1^*+d_4^*\}}$$

$$= \binom{n}{d_0^*, d_1^*, d_2^*, d_3^*, d_4^*} [(1-q^{(1)})(1-q^{(2)})]^{d_0^*} [(1-q^{(1)})]^{d_1^*} [q^{BB}q^{(1)}(1-q^{(1)})(1-q^{(2)})]^{d_2^*}$$

$$\cdot [q^{(1)}(1-q^{(2)})(1-q^{BB})]^{d_3^*} [q^{(1)}]^{d_4^*} \cdot \sum_{k=0}^{n-d_2^*-d_3^*} \sum_{d_0=0}^{d_0^*} \binom{d_0^*}{d_0} [q^{BB}(1-q^{(1)})]^{d_0^*-d_0} [1-q^{BB}]^{d_0}$$

$$
\begin{aligned}
&\cdot \sum_{d_1=0}^{d_1^*} \binom{d_1^*}{d_1} [(1-q^{BB})q^{(2)}]^{d_1} [q^{BB}(1-(1-q^{(1)})(1-q^{(2)}))]^{d_1^*-d_1} \\
&\cdot \sum_{d_4=0}^{d_4^*} \binom{d_4^*}{d_4} [q^{BB}(1-(1-q^{(1)})(1-q^{(2)}))]^{d_4^*-d_4} [(1-q^{BB})q^{(2)}]^{d_4} \cdot \mathbf{1}_{\{d_0+d_1+d_4+k=d_0^*+d_1^*+d_4^*\}} \\
&\overset{(*)}{=} \binom{n}{d_0^*, d_1^*, d_2^*, d_3^*, d_4^*} [(1-q^{(1)})(1-q^{(2)})]^{d_0^*} [(1-q^{(1)})]^{d_1^*} [q^{BB}q^{(1)}(1-q^{(1)})(1-q^{(2)})]^{d_2^*} \\
&\cdot [q^{(1)}(1-q^{(2)})(1-q^{BB})]^{d_3^*} [q^{(1)}]^{d_4^*} \cdot \underbrace{[q^{BB}(1-q^{(1)})+1-q^{BB}]}_{=1-q^{BB}q^{(1)}}{}^{d_0^*} \cdot \\
&\cdot \underbrace{[(1-q^{BB})q^{(2)}+q^{BB}(q^{(1)}(1-q^{(2)})+q^{(2)})]}_{=q^{BB}q^{(1)}(1-q^{(2)})+q^{(2)}}{}^{d_1^*} [q^{BB}(q^{(1)}(1-q^{(2)})+q^{(2)})+(1-q^{BB})q^{(2)}]^{d_4^*} \\
&= \binom{n}{d_0^*, d_1^*, d_2^*, d_3^*, d_4^*} [(1-q^{(1)})(1-q^{(2)})(1-q^{(1)}q^{BB})]^{d_0^*} [(1-q^{(1)})[q^{(1)}q^{BB}(1-q^{(2)})+q^{(2)}]]^{d_1^*} \\
&\cdot [q^{BB}q^{(1)}(1-q^{(1)})(1-q^{(2)})]^{d_2^*} [q^{(1)}(1-q^{(2)})(1-q^{BB})]^{d_3^*} [q^{(1)}[q^{(1)}q^{BB}(1-q^{(2)})+q^{(2)}]]^{d_4^*}.
\end{aligned}
$$

The equality in $(*)$ can be seen by realizing that the restriction to those combinations of d_0, d_1, d_4 which fulfill $d_0 + d_1 + d_4 + k = d_0^* + d_1^* + d_4^*$ combined with the sum over all $k = 0, \ldots, d_0^* + d_1^* + d_4^*$ gives exactly the correct number of terms such that every summand from the first sum is multiplied with every possible product of terms from the second and third sum in order to apply the binomial theorem.
From equality (A.1) in the proof of Lemma 2.30 we deduce that

$$
|q^{(1)} - p_1| \leqslant \frac{C_1}{\log(N)}, \quad |q^{(2)} - p_2| \leqslant \frac{C_2}{\log(N)}, \quad |q^{BB} - p_2| \leqslant \frac{C_{BB}}{\log(N)}
$$

Applying Lemma 3.4.3, [15] again, leads to the following:

$$
\begin{aligned}
&\Big|\mathbb{P}(\tilde{\tilde{D}} + \tilde{D} = d^*) - \binom{n}{d_0^*, d_1^*, d_2^*, d_3^*, d_4^*} [(1-p_1)(1-p_2)(1-p_1p_2)]^{d_0^*} [(1-p_1)[p_1p_2(1-p_2)+p_2]]^{d_1^*} \\
&\cdot [p_2p_1(1-p_1)(1-p_2)]^{d_2^*} \cdot [p_1(1-p_2)(1-p_2)]^{d_3^*} \cdot [p_1[p_1p_2(1-p_2)+p_2]]^{d_4^*}\Big| \leqslant \frac{C}{\log(N)},
\end{aligned}
$$

as we have finitely many factors.
Comparing the above with the definition of $\bar{q}_{\mathrm{G1}} = (q_0, q_1, \ldots, q_4)$ from (2.8) shows that indeed

$$
\big|\mathbb{P}(\tilde{\tilde{D}} + \tilde{D} = d^*) - \mathbb{P}(\tilde{D}^* = d^*)\big| \leqslant C/\log(N),
$$

for some constant C. Applying the Triangle Inequality finishes the proof as noticed above. □

CHAPTER 3

Modeling a selective sweep with varying population size

In this chapter we consider the same biological question for two partially linked neutral loci: how does a selective sweep influence the neutral gene genealogy of a random sample taken at the time when the mutant allele has fixed in the population. However, we here study an essentially different model for the evolution of the population through time which in particular allows for varying population size. Precisely, we consider a version of the Darwinian model which was introduced in Section 1.4.

We will see that the ideas needed for the proofs here are very similar to those developed in Chapter 2. The critical point is to check that we can indeed transfer the methods to this setting of a Markovian birth and death process where the expectation of births, that is up-jumps, and deaths, corresponding to down-jumps, are not as easily derived as in the setting with constant population size. Further, it is not clear which events will still be negligible, in the same way as the concrete values of the event probabilities are not obvious, and need careful calculations.

In Section 1.4 we introduced a simple version of such an eco-evolutionary model and also stated some results concerning the dynamics of a population which evolves accordingly. A two-locus version (with one neutral locus) of this model was studied by Smadi in [38]. The focus there was to analyze different selective strengths of the mutant allele and also different recombination regimes, that is, different orders of the recombination probability, in order to obtain a result on the neutral allele proportions at the end of a selective sweep. In a joint work with Charline Smadi, we aimed at combining the results from the multi-locus model with the Darwinian approach in pursuance of a fine study of the so-called hitch-hiking effect under a biologically motivated model with varying population size. Precisely, we consider a three-locus model under the already introduced geometries (G1) and (G2) from Chapter 2 in order to gain insight into not only the proportions but also the neutral gene genealogies. While we will state the main results with respect to each geometry, we will only give the proofs for (G1) as it has the more difficult dependence structure and all results can be easily transfered to the second geometry (G2).

A crucial aspect of our model is that we suppose that the surrounding environment of the population has a certain carrying capacity, denoted by K, which was already introduced in

Definition 1.33. The equivalent of considering large population sizes as before is here then the examination of large capacities K. In contrast to the approximation order from the previous analysis from Chapter 2, that is, 1 over the logarithm of the population size, we will here in the model with varying size state results in the limit $K \to \infty$ without giving asymptotics.
All results in this chapter were obtained through the joint work with Charline Smadi. We here give a detailed overview on the model and present the main results in Section 3.1 and 3.2, respectively. In the subsequent discussion of the result we will in particular draw the connection to the results presented in Section 2.2 of the previous chapter. In the remaining parts of this chapter we will develop the main ideas and ingredients for the proof of the main theorems. We however abstain from stating proofs here which were mainly developed by Charline Smadi and refer to a preprint of our joint work instead. Nevertheless, we will describe the main ideas which are used and further detail one proof which was essentially the result of my previous work as described in Chapter 2.
Due to convenience, we will in most cases stick to the notation which was used in this joint work and do not adjust it such that it is consistent with the remainder of this thesis. This in particular considers the name of the selectively advantageous allele, here a, and the wild-type, A (in contrast to B and b, respectively, from Chapter 2).

3.1 An eco-evolutionary three-locus model with recombination

We consider a haploid, but sexually reproducing population which evolves according to a birth and death Markov process in continuous time. This can be understood as follows: there are always two individuals involved in a reproduction event and if no recombination occurs, a birth leads to a binary branching of the one parental line and both branches then carry the same alleles at all loci. In case of a recombination, the offspring can instead copy some genetic material from the other parent, which leads to a coalescence of the two parental lines forwards in time.
To be precise: we consider one locus under selection, SL, with alleles in $\mathcal{A} := \{A, a\}$ and two neighboring neutral loci N1 and N2 with alleles in the finite sets $\mathcal{B}$ and $\mathcal{C}$ respectively, and denote by $\mathcal{E} = \mathcal{A} \times \mathcal{B} \times \mathcal{C}$ the type space. As mentioned before we will study the two possible geometric alignments stated in (G1) and (G2). We will introduce the model and notations again only for (G1), their analogs for (G2) can be deduced straightforwardly.
The state of the population at any time $t \in \mathbb{R}_+$ is characterized by the numbers of individuals present at time t which carry a certain genotype. We denote by N^K this process, defined by

$$N^K = (N^K(t), t \geq 0) = \big((N^K_{\alpha\beta\gamma}(t))_{(\alpha,\beta,\gamma)\in\mathcal{E}}, t \geq 0\big), \tag{3.1}$$

where $N^K_{\alpha\beta\gamma}(t)$ denotes the number of $\alpha\beta\gamma$-individuals present at time t. Note that here the superscript ^{K} indicates the dependence of the process and not a scaling by K, opposed to the superscript $^{(K)}$ from Definition 1.34 and the subsequent statements in Section 1.4.
The recombination probabilities are defined in the same way as in Section 2.1 and depend on K in such a way that

$$\limsup_{K\to\infty} r_j \log K < \infty, \ j = 1, 2, \tag{3.2}$$

which is called the regime of *weak recombination* and motivated by Theorem 2 in [38] which states that in this scale we can observe a non-trivial signature by the recombinations on the neutral allele distribution. If the recombination probabilities are larger (which means that the neutral loci are more distant from the selected locus), there are many recombinations and the selective strength of the sweep is not strong enough, that is, the sweep is "too slow", in order to modify the neutral diversity at these sites.
Recombinations can lead to a mixing of the parental genetic material in the newborn, and hence, parents with types (α, β, γ) and $(\alpha', \beta', \gamma')$ in $\mathcal{E}$ can generate the following offspring:

possible genotype	event	probability
$\alpha\beta\gamma, \alpha'\beta'\gamma'$	no recombination	$(1-r_1)(1-r_2)$
$\alpha\beta'\gamma', \alpha'\beta\gamma$	one recombination between SL and N1	$r_1(1-r_2)$
$\alpha\beta\gamma', \alpha'\beta'\gamma$	one recombination between N1 and N2	$(1-r_1)r_2$
$\alpha\beta'\gamma, \alpha'\beta\gamma'$	two recombinations	$r_1 r_2$

(3.3)

We will see that the probability to witness a birth event with two simultaneous recombinations is very small.
Recall Definition 1.33 and Equation (1.23). We will now analogously introduce the biologically motivated eco-evolutionary parameters of this model. As we assume the alleles at the loci N1 and N2 to be neutral, the ecological parameters of an individual only depend on its allele α at the locus under selection. Let us denote the fertility of an individual as follows

$$f_\alpha \ = \ \text{the birth rate of an individual of genetic type } \alpha. \tag{3.4}$$

By addressing the complementary type of the allele α by $\bar{\alpha}$, we get the following result for the total birth rate of individuals of type $(\alpha, \beta, \gamma) \in \mathcal{E}$ in the process N^K:

$$b^K_{\alpha\beta\gamma}(n) = (1-r_1)(1-r_2) f_\alpha n_{\alpha\beta\gamma} + r_1(1-r_2) f_\alpha n_\alpha \frac{f_\alpha n_{\alpha\beta\gamma} + f_{\bar{\alpha}} n_{\bar{\alpha}\beta\gamma}}{f_a n_a + f_A n_A}$$

$$+ (1-r_1)r_2 f_\alpha \Big(\sum_{\gamma' \in \mathcal{C}} n_{\alpha\beta\gamma'} \Big) \frac{\sum_{\beta' \in \mathcal{B}} (f_\alpha n_{\alpha\beta'\gamma} + f_{\bar{\alpha}} n_{\bar{\alpha}\beta'\gamma})}{f_a n_a + f_A n_A}$$
$$+ r_1 r_2 f_\alpha \Big(\sum_{\beta' \in \mathcal{B}} n_{\alpha\beta'\gamma} \Big) \frac{\sum_{\gamma' \in \mathcal{C}} (f_\alpha n_{\alpha\beta\gamma'} + f_{\bar{\alpha}} n_{\bar{\alpha}\gamma'})}{f_a n_a + f_A n_A}, \tag{3.5}$$

where $n_{\alpha\beta\gamma}$ (resp. n_α) denotes the number of $\alpha\beta\gamma$-individuals (resp. α-individuals) present in the population and $n = (n_{\alpha\beta\gamma}, (\alpha,\beta,\gamma) \in \mathcal{E})$ is the current state of the population. This rate can be best understood by consulting the list from (3.3) and realizing that the total birth rate in the current population is equal to $f_a n_a + f_A n_A$.
An α-individual can die either a natural death (rate D_α), or it dies because of type-dependent competition. Recalling the definition of the competition kernel C in (1.21) then results in the total death rate of individuals carrying the alleles $(\alpha,\beta,\gamma) \in \mathcal{E}$ in the population process N^K:

$$d^K_{\alpha\beta\gamma}(n) = \left(D_\alpha + \frac{C_{\alpha,A}}{K} n_A + \frac{C_{\alpha,a}}{K} n_a \right) n_{\alpha\beta\gamma}. \tag{3.6}$$

Hence, N^K is a multitype birth and death process with rates given in (3.5) and (3.6). Often we are only interested in the total numbers $N^K_A(t)$ and $N^K_a(t)$ of A- and a-individuals, respectively, given by a birth and death process $\big(N^K_A(t), N^K_a(t)\big)_{t\geq 0}$ with the following rates:

$$\begin{aligned} b^K_\alpha(n) &= \sum_{(\beta,\gamma)\in\mathcal{B}\times\mathcal{C}} b^K_{\alpha\beta\gamma}(n) = f_\alpha n_\alpha, \\ d^K_\alpha(n) &= \sum_{(\beta,\gamma)\in\mathcal{B}\times\mathcal{C}} d^K_{\alpha\beta\gamma}(n) = \Big(D_\alpha + \frac{C_{\alpha,A}}{K} n_A + \frac{C_{\alpha,a}}{K} n_a \Big) n_\alpha. \end{aligned} \tag{3.7}$$

So far, we did not specify how the selective advantage of individuals with the a-allele is included into the model. However, as the birth and death rates are type-dependent it can obviously be integrated into the present framework. The advantage or disadvantage an individual with allele type α has in an $\bar{\alpha}$-population at equilibrium is summarized by the invasion fitness $S_{\alpha\bar{\alpha}}$ as already introduced in (1.30):

$$S_{\alpha\bar{\alpha}} := f_\alpha - D_\alpha - C_{\alpha,\bar{\alpha}} \bar{n}_{\bar{\alpha}}, \tag{3.8}$$

where the equilibrium density $\bar{n}_\alpha = (f_\alpha - D_\alpha)/C_{\alpha,\alpha}$ is defined as in (1.27). The role of the invasion fitness $S_{\alpha\bar{\alpha}}$ and the definition of the equilibrium density $\bar{n}_\alpha$ were already motivated in Section 1.4. We know from Proposition 1.35 that if $N^K_A(0)$ and $N^K_a(0)$ are of order K and K is large, the process $(N^K_A(0)/K, N^K_a(0)/K)$ is very close to the solution of the competitive Lotka-Volterra system from (1.29),

$$\dot{n}^{(z)}_\alpha = (f_\alpha - D_\alpha - C_{\alpha,A} n^{(z)}_A - C_{\alpha,a} n^{(z)}_a) n^{(z)}_\alpha, \quad n^{(z)}_\alpha(0) = z_\alpha, \quad \alpha \in \mathcal{A}, \tag{3.9}$$

during any finite time interval. In particular, (1.32) from Proposition 1.36 tells us that (3.9) has a unique stable equilibrium $(0, \bar{n}_a)$ and two unstable steady states $(\bar{n}_A, 0)$ and $(0, 0)$ if we assume that the following assumptions on the parameters hold true:

$$\bar{n}_A > 0, \quad \bar{n}_a > 0, \quad \text{and} \quad S_{Aa} < 0 < S_{aA}. \tag{3.10}$$

Moreover, if we define the conditional probability measure

$$\mathbb{P}(\cdot) := \mathbb{P}(\,\cdot \mid N_A^K(0) = \lfloor \bar{n}_A K \rfloor, N_a^K(0) = 1), \tag{3.11}$$

then Equation (39) from Lemma 3 in [13] states that in the limit of large carrying capacities K, the probability that the allele a fixes in the population is equal to the following:

$$\lim_{K\to\infty} \mathbb{P}(\mathrm{Fix}^K) = \frac{S_{aA}}{f_a} =: s, \tag{3.12}$$

where s is called rescaled invasion fitness. Here, the event of fixation of the a-allele and the extinction time of the A-population are defined as follows:

$$T_{\mathrm{ext}}^K := \inf\left\{t \geq 0 : N_A^K(t) = 0\right\}, \quad \text{and} \quad \mathrm{Fix}^K := \{T_{\mathrm{ext}}^K < \infty, N_a^K(T_{\mathrm{ext}}^K) > 0\}. \tag{3.13}$$

The condition in the measure from (3.12) is exactly the starting condition for a potential selective sweep, and the condition in (3.10) guarantees that we have a positive probability for the mutant allele to actually fix in the population such that the sweep is completed. We will now present our results under this evolutionary model.

3.2 Results and discussion for a sample from a Darwinian population at the end of a selective sweep

In the following we will consider a random sample of a fixed number of $d \in \mathbb{N}$ individuals taken at the time of fixation T_{ext}^K of the mutant allele and again describe the ancestral relationships for all sampled neutral loci $[d; 1, 2] = \{(i, 1), (i, 2),\ i \in [d]\}$ by a marked partition $\Theta_d^K \in \mathcal{P}^*_{[d;1,2]}$. Recall the definition of the true partition Θ from Definition 2.4 as well as the Definition 2.5 of a marked $\bar{q}$-partition. We stress again that in such a partition there are no unmarked blocks with more than two elements and if there is an unmarked block of size two then both its elements correspond to loci from the same sampled individual. We will again formulate these concepts in the present setting as it turns out that also under this evolutionary model, the partition of a sample has the structure of a marked $\bar{p}$-partition for some vector $\bar{p}$.

Definition 3.1. *The partition $\Theta_d^K \in \mathcal{P}^*_{[d;1,2]}$ reflecting the true structure of the sample is defined as follows: each block of the partition Θ_d^K is composed of all those neutral alleles*

sampled at time T^K_{ext} which originate from the same given individual alive at the beginning of the sweep. The block containing the descendants of the mutant a (if such a block exists) is distinguished by the mark $*$.

Definition 3.2. *For $d \in \mathbb{N}$ let Δ_d denote the subset of $\mathcal{P}^*_{[d;1,2]}$ consisting of all marked $\bar{p}$-partitions*

$$\Delta_d = \{\pi : \pi \text{ is a marked } \bar{p}\text{-partition, } \bar{p} = (p_1, \dots, p_5) \in [0,1]^5, \ |\bar{p}| = \sum_{j=1}^{5} p_j = 1\}. \tag{3.14}$$

Further, for any $\pi \in \mathcal{P}^*_{[d;1,2]}$ we set:

$$\begin{aligned}
|\pi|_1 &= \#\{i \in [d] : (i,1) \text{ and } (i,2) \text{ belong to the marked block}\}\\
|\pi|_2 &= \#\{i \in [d] : (i,1) \text{ belongs to the marked block, } \{(i,2)\} \text{ is an unmarked block}\}\\
|\pi|_3 &= \#\{i \in [d] : (i,2) \text{ belongs to the marked block, } \{(i,1)\} \text{ is an unmarked block}\}\\
|\pi|_4 &= \#\{i \in [d] : \{(i,1),(i,2)\} \text{ is an unmarked block}\}\\
|\pi|_5 &= \#\{i \in [d] : \{(i,1)\} \text{ and } \{(i,2)\} \text{ are two distinct unmarked blocks}\}.
\end{aligned} \tag{3.15}$$

Then, $\sum_{k=1}^5 |\pi|_k \leqslant d$ with equality if $\pi \in \Delta_d$.

We will now define the probabilities for the vector $\bar{p}$ of the marked $\bar{p}$-partition which reflects the structure of the true partition in the limit of large carrying capacities. Let

$$\begin{aligned}
q_1 &:= \exp\Big(-\frac{r_1 \log K}{s}\Big), \qquad q_2 := \exp\Big(-\frac{r_2 \log K}{s}\Big),\\
\bar{q}_2 &:= \exp\Big(-\frac{f_a r_2 \log K}{|S_{Aa}|}\Big), \quad \text{and} \quad q_3 := \frac{f_a r_1}{f_a(r_1+r_2) - f_A r_2}\Big(q_2^{f_A/f_a} - q_1 q_2\Big),
\end{aligned} \tag{3.16}$$

where S_{aA} and S_{Aa} have been defined in (3.8), $s = S_{aA}/f_a$ as in (3.12), f_a, f_A as in (3.4) and r_1, r_2 fulfilling (3.2). We did not make any assumption on the sign of $f_a(r_1+r_2) - f_A r_2$, but q_3 can be written in the form $\delta(e^{-\mu} - e^{-\nu})/(\nu - \mu)$ for $\mu, \nu \in \mathbb{R}_+$, $\delta = r_1 \log(K)/s$, and hence it is well defined and non negative. Further, it is easy to check that $q_3 < 1$. We now define the entries of the appropriate $\bar{p} \in [0,1]^5$ which will quantify the law of Θ^K_d for large K in main result:

$$\begin{aligned}
&\text{Let } \ \bar{p} := (p_1, p_2, p_3, p_4, p_5) \ \text{ with}\\
&p_1 := q_1 q_2[1 - (1-q_1)(1-\bar{q}_2)], \quad p_2 := q_1[(1 - q_1 q_2) - q_2 \bar{q}_2 (1 - q_1)],\\
&p_3 := q_1 q_2 (1-\bar{q}_2)(1-q_1), \quad p_4 := \bar{q}_2 q_3, \quad p_5 := (1-q_1)(1 - q_1 q_2(1-\bar{q}_2)) - \bar{q}_2 q_3.
\end{aligned} \tag{3.17}$$

Note that $\sum_{1 \leqslant k \leqslant 5} p_k = 1$. Finally, we summarize all assumptions made on the parameters throughout this chapter as follows:

Assumption 3.3. *We assume that* $(N_A^K(0), N_a^K(0)) = (\lfloor \bar{n}_A K \rfloor, 1)$ *and that the conditions from* (3.2) *on the recombination probabilities and from* (3.10) *on the equilibrium densities and the fitness hold.*

With Definitions 3.1 and 3.2 in mind, we can now state our main results:

Theorem 3.4 (Genealogy of a sample, geometry (G1))**.** *Let* d *be in* $\mathbb{N}$*. Then under Assumption 3.3, we have for every* $\pi \in \mathcal{P}^*_{[d;1,2]}$

$$\lim_{K\to\infty} \left| \mathbb{P}(\Theta_d^K = \pi | \mathrm{Fix}^K) - \mathbf{1}_{\{\pi\in\Delta_d\}} {p_1}^{|\pi|_1} {p_2}^{|\pi|_2} {p_3}^{|\pi|_3} {p_4}^{|\pi|_4} {p_5}^{|\pi|_5} \right| = 0, \tag{3.18}$$

with p_k *as defined in* (3.17)*,* $k = 1, \ldots, 5$*.*

This statement can be interpreted in a similar way as Theorem 2.6. When K is large, Θ_d^K belongs to Δ_d with a probability close to one, and has marking probabilities as stated in (3.17). More precisely, in the limit $K \to \infty$ the true partition is a marked $\bar{p}$-partition which in particular implies that the d sampled individuals have asymptotically independent neutral genealogies.
With high probability, the neutral alleles of a given sampled individual i either originate from the first mutant a and belong to the marked block, or they escape the sweep and originate from an A individual. In the latter case they form a singleton or a double-singleton $\{(i,1),(i,2)\}$ as we could already witness in the main theorem of the previous chapter, Theorem 2.6. Hence, escaped neutral alleles of two distinct sampled individuals originate from distinct A-individuals in the limit $K \to \infty$.
However, the structure of $\bar{p}$ again implies that in general, the genealogies of the two neutral alleles of a given individual are not independent as the allele at the second neutral locus will migrate together with the first neutral locus in case of a recombination between SL and N1. This phenomenon is seen in the above probabilities as follows: consider for example the probability that $(i,1)$ and $(i,2)$ escape the sweep, which is $p_4 + p_5$. The probability that $(i,1)$ (resp. $(i,2)$) escapes the sweep however is $p_3 + p_4 + p_5$ (resp. $p_2 + p_4 + p_5$), and for every $K \in \mathbb{N}$ such that $r_1 \neq 0$ we have

$$(p_3 + p_4 + p_5)(p_2 + p_4 + p_5) = (1 - q_1)(1 - q_1 q_2) < (1 - q_1)(1 - q_1 q_2 + q_1 q_2 \bar{q}_2) = p_4 + p_5.$$

As the term $q_1 q_2 \bar{q}_2$ does not tend to 0 when K goes to infinity under the condition on the recombination on the recombination probabilities from (3.2), the only possibility to have an equality in the limit is the case where $r_1 \log K \ll 1$ or in other words when the probability to see a recombination between SL and N1 is negligible.

Let us now consider the second alignment N1-SL-N2:

Theorem 3.5 (Genealogy of a sample, geometry (G2))**.** *Let d be in $\mathbb{N}$. Then under Assumption 3.3, we have for every $\pi \in \mathcal{P}^*_{[d;1,2]}$*

$$\lim_{K\to\infty} \Big| \mathbb{P}(\Theta_d^K = \pi | \mathrm{Fix}^K) - \mathbf{1}_{\{\pi\in\Delta_d\}} [q_1 q_2]^{|\pi|_1} [q_1(1-q_2)]^{|\pi|_2} [(1-q_1)q_2]^{|\pi|_3} [(1-q_1)(1-q_2)]^{|\pi|_5} \Big| = 0, \tag{3.19}$$

with q_1, q_2 as defined in (3.16)

Once again this result implies that the neutral genealogies of the d sampled individuals are asymptotically independent. Furthermore, in the geometry (G2) we have independence between the neutral loci of one individual. Indeed, the result stated in Theorem 3.5 means that a neutral allele at locus Nj escapes the sweep with probability $1 - q_j$ independently of all other neutral alleles, including the allele at the other neutral locus of the same individual. This is due to the fact that in (G2) a recombination between SL and one neutral locus has no impact on the genetic background of the allele at the other neutral locus. Note in particular that there is no block of the form $\{(i,1),(i,2)\}$ in the limit partition, as the two neutral alleles will recombine at the same time only with a very small probability.

At first glance, the statements of both Theorems 3.4 and 3.5 look identical to the results from Theorems 2.6 and 2.7. However, there are two major differences: First, the above results make no statement concerning the rate of convergence, they only give a result in the limit of $K \to \infty$. The other difference lies in the structure of the marking probabilities summarized in the vector $\bar{p}$. We will be a bit more detailed on the latter issue: In [38], the author studied the two-locus version (SL-N1) of the here presented model. It was shown that the ancestral relationships in a sample taken at the end of the selective sweep correspond to the ones derived in [36] under the two-locus Moran model with recombination and selection when we replace the fitness by the rescaled invasion fitness S_{aA}/f_a. The first expectation is now, that we can find the same relation between the probabilities from (3.17) and the entries from $\bar{q}_{\mathrm{G1}}$ in (2.8) from the previous chapter. If we however make the analogous comparison and try to match the here obtained results for (G1) with the statement from Theorem 2.6 we observe an interesting phenomenon: the probabilities for the different types of ancestry only coincide if the birth rates of a- and A-individuals are the same, that is, if $f_a = f_A$ holds true. To give some heuristic on this observation we need to go back to the actual definition of the fitness, Definition 1.30, and recall Example 1.32: in biology, the fitness describes the ability to both survive and reproduce, and it is defined by the average contribution of an individual with a given genotype to the gene pool of the next generation. Hence a mutation which affects the fitness of an individual in a given environment can either act on the fertility (f_α in our model), or on the death rate, through the intrinsic rate (D_α) or by competition ($C_{\alpha,\alpha'}$), or on both. As we here get correspondence of both results only in the case of identical fertilities

for the A and the a individuals, we deduce that the selective strength of the mutation in the Moran model introduced in Section 2.1 only acts through the death rate. This answers the question which already came up in Example 1.32.

3.3 Dynamics of the sweep and proofs of the main results

A key ingredient of the underlying theory for this work is that we can divide the process of a selective sweep in three phases, as already mentioned in Section 1.4. This idea is similar to the separation of the time intervals $[0, \tau_N)$ and $[\tau_N, \tau]$ from the previous chapter. We will here as well consider each phase separately and show for each phase that some events have negligible probability. We will as well use the technique of Poisson approximations from Lemma A.2 throughout the calculations of the marking probabilities. However, the process which we consider in this chapter is more complex. Whereas one can use a symmetry argument and random walk theory in order to calculate the expected number of up-jumps and down-jumps during the sweep as given in Lemma 2.23, we here need to define rather complicated couplings for each phase in order to obtain approximate values for the needed expectations as well as bounds on the covariances needed for the Poisson approximation.

First of all however, we need to rigorously characterize the *three phases* of a selective sweep such that we can define appropriate and manageable probability measures for the respective time intervals. Here we strongly benefit from the ideas and results from Sections 3 and 4.2 from [38].

Keep in mind that also in this chapter we condition on the event of fixation of the advantageous allele by considering the measure $\mathbb{P}$ defined in (3.11). Throughout the remainder of this chapter, let $\varepsilon > 0$ be independent of K and as small as it needs to be such that all approximations hold.

In the following description it helps to visualize the process of the sweep as shown in the following graphic 3.3.1, obtained by Charline Smadi through a simulation with carrying capacity $K = 10000$ and invasion parameters as stated below the figure.

First phase. In the first phase, the size of the wild-type A-population stays close to its equilibrium value $\bar{n}_A K$ as long as the size of the mutant a-population has not hit $\lfloor \varepsilon K \rfloor$. Similar as in (3.10) of [38] we introduce an interval surrounding $\bar{n}_A K$

$$I_\varepsilon^K := \Big[K\Big(\bar{n}_A - 2\varepsilon \frac{C_{A,a}}{C_{A,A}}\Big), K\Big(\bar{n}_A + 2\varepsilon \frac{C_{A,a}}{C_{A,A}}\Big)\Big] \cap \mathbb{N}, \tag{3.20}$$

and the stopping times T_ε^K and $\tilde{T}_\varepsilon^K$, which denote the hitting time of $\lfloor \varepsilon K \rfloor$ by the mutant population and the exit time of I_ε^K by the resident population, respectively,

$$T_\varepsilon^K := \inf\{t \geqslant 0, N_a^K(t) = \lfloor \varepsilon K \rfloor\}, \quad \tilde{T}_\varepsilon^K := \inf\{t \geqslant 0, N_A^K(t) \notin I_\varepsilon^K\}, \tag{3.21}$$

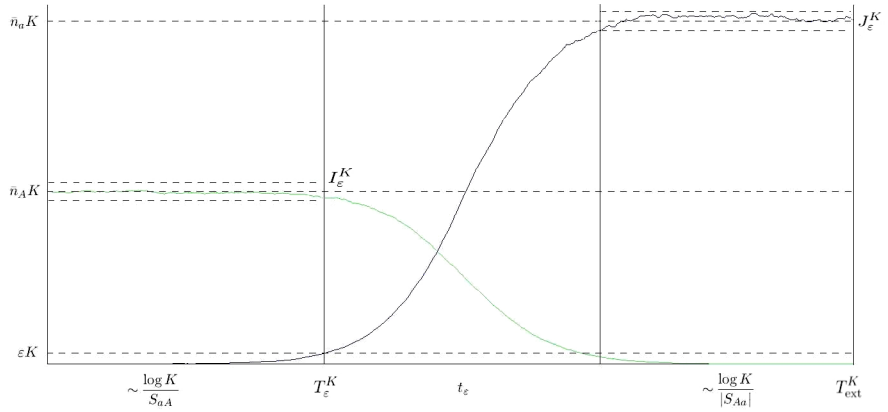


Figure 3.3.1: The three phases of a selective sweep; in this simulation $K = 10\,000$, $\bar{n}_a = 2\bar{n}_A = 2$ and $S_{aA} = |S_{Aa}| = 4$. Time is going from left to right.

then we can deduce from [13] (see Equation (A.6) in [38] for the details of the derivation) that the event of fixation of the mutant type and the event $\{T_\varepsilon^K \leqslant \tilde{T}_\varepsilon^K\}$ are very close, that is, their symmetric difference has small probability:

$$\limsup_{K\to\infty} \mathbb{P}\big(\big[\{T_\varepsilon^K \leqslant \tilde{T}_\varepsilon^K\} \cap (\mathrm{Fix}^K)^c\big] \cup \big[\mathrm{Fix}^K \cap \{T_\varepsilon^K > \tilde{T}_\varepsilon^K\}\big]\big) \leqslant c\varepsilon, \tag{3.22}$$

for a finite c and ε small enough, where we recall the definition of $\mathbb{P}(\cdot)$ from (3.11). The intuition is that once the invading type has reached a certain level while the wild-type did not increase too much, we have a high probability for the event of a selective sweep. From this point onwards, *first phase* will denote the time interval $[0, T_\varepsilon^K]$ when the a-population size is smaller than $\lfloor \varepsilon K \rfloor$.

Second phase. From Proposition 1.35 we know that when N_A^K and N_a^K are of order K, the rescaled population process $(N_A^K/K, N_a^K/K)$ is well approximated by the competitive Lotka-Volterra system (3.9) ((1.29), respectively). Moreover, Proposition 1.36 says that under condition (3.10) the dynamical system has a unique attracting equilibrium $(0, \bar{n}_a)$ for an initial condition z satisfying $z_a > 0$, with $\bar{n}_a$ the equilibrium density as defined in (1.27). In particular, if we introduce for $(n_A, n_a) \in \mathbb{N}^2$ the notation,

$$\mathbb{P}_{(n_A,n_a)}(\cdot) := \mathbb{P}(\cdot \mid N_A^K(0) = n_A, N_a^K(0) = n_a), \tag{3.23}$$

then Theorem 3 (b) in [13] offers a finer study on the solution of the system (1.29) in Proposition 1.35 and we get:

$$\lim_{K\to\infty} \sup_{z\in\Gamma} \mathbb{P}_{(\lfloor z_A K \rfloor, \lfloor z_a K \rfloor)}\Big(\sup_{0\leqslant t\leqslant t_\varepsilon, \alpha\in\mathcal{A}} \Big| \frac{N_\alpha^K(t)}{K} - n_\alpha^{(z)}(t) \Big| \geqslant \delta \Big) = 0, \tag{3.24}$$

for every $\delta > 0$ and $z_a > 0$, where

$$\begin{aligned}\Gamma &:= \Big\{z \in \mathbb{R}_+^{\mathcal{A}} : z_A K \in I_\varepsilon^K, z_a \in [\varepsilon/2, \varepsilon]\Big\},\\ t_\varepsilon(z) &:= \inf\big\{s \geqslant 0 : \forall t \geqslant s, n_A^{(z)}(t) \in [0, \varepsilon^2/2], n_a^{(z)}(t) \in [\bar{n}_a - \varepsilon/2, \bar{n}_a + \varepsilon/2]\big\},\\ t_\varepsilon &:= \sup\{t_\varepsilon(z) : z \in \Gamma\} < \infty,\end{aligned}$$

In the sequel, the expression *second phase* will be used synonymously for the time interval $[T_\varepsilon^K, T_\varepsilon^K + t_\varepsilon]$ when the population process is close to the system (3.9).

Third phase. Equation (3.24) also implies that

$$\lim_{K\to\infty} \mathbb{P}\Big(\tfrac{N_A^K(T_\varepsilon^K + t_\varepsilon)}{K} \in [\omega_1, \omega_2], \Big|\tfrac{N_a^K(T_\varepsilon^K + t_\varepsilon)}{K} - \bar{n}_a\Big| \leqslant \varepsilon,\ \Big|\Big(\tfrac{N_A^K(T_\varepsilon^K)}{K}, \tfrac{N_a^K(T_\varepsilon^K)}{K}\Big) \in \Gamma\Big) = 1, \quad (3.25)$$

where

$$2\omega_1 := \inf\{n_A^{(z)}(t_\varepsilon) : z \in \Gamma\} > 0, \quad \text{and} \quad \omega_2/2 := \sup\{n_A^{(z)}(t_\varepsilon) : z \in \Gamma\} \leqslant \varepsilon^2/2. \quad (3.26)$$

The *third phase*, which corresponds to the remaining time interval $[T_\varepsilon^K + t_\varepsilon, T_{\text{ext}}^K]$, can be seen as the symmetric counterpart of the first phase, where the roles of A and a are interchanged as the almost sure initial properties (given in (3.25)) are similar to the ones which define the end of the first phase. In particular, the a-population size stays close to its equilibrium $\bar{n}_a K$ while the A-population goes to extinction.

Approximations for the first and third phases

We will now define conditional probability measures for the first and third phase whose exact definition is motivated by the above observations and which approximate the measure defined in (3.11). The ideas in this section are similar to the ones which were developed in the proof of Lemma 4.3 in [38]. For $M := 3 + (f_a + C_{a,A})/C_{a,a}$, define the set

$$J_\varepsilon^K := \Big[K\Big(\bar{n}_a - M\varepsilon\Big), K\Big(\bar{n}_a + M\varepsilon\Big)\Big] \cap \mathbb{N}, \quad (3.27)$$

and the stopping times T_0^K, $T_\varepsilon^{(K,2)}$ and $\bar{T}_\varepsilon^K$, which denote respectively the hitting times of 0 and $\lfloor \varepsilon K \rfloor$ by the A-population, and the exit time of J_ε^K by the a-population during the third phase (i.e. after the time $T_\varepsilon^K + t_\varepsilon$):

$$\begin{aligned}T_0^K &:= \inf\{t \geqslant 0, N_A^K(T_\varepsilon^K + t_\varepsilon + t) = 0\},\\ T_\varepsilon^{(K,2)} &:= \inf\{t \geqslant 0, N_A^K(T_\varepsilon^K + t_\varepsilon + t) = \lfloor \varepsilon K \rfloor\},\\ \bar{T}_\varepsilon^K &:= \inf\{t \geqslant 0, N_a^K(T_\varepsilon^K + t_\varepsilon + t) \notin J_\varepsilon^K\}.\end{aligned} \quad (3.28)$$

Recall the stopping times T_ε^K and $\tilde{T}_\varepsilon^K$ as defined in (3.21) and define the event

$$\mathcal{N}_\varepsilon^K := \{T_\varepsilon^K \leqslant \tilde{T}_\varepsilon^K\} \cap \Big\{ \tfrac{N_A^K(T_\varepsilon^K + t_\varepsilon)}{K} \in [\omega_1, \omega_2], \Big| \tfrac{N_a^K(T_\varepsilon^K + t_\varepsilon)}{K} - \bar{n}_a \Big| \leqslant \varepsilon \Big\}, \tag{3.29}$$

which describes the state of the population at the end of the first phase and the beginning of the third phase. Then we have $\mathcal{N}_\varepsilon^K \cap \{T_0^K < T_\varepsilon^{(K,2)} \wedge \bar{T}_\varepsilon^K\} \subset \{T_\varepsilon^K < \tilde{T}_\varepsilon^K\}$ by definition and further get from the proof of Lemma 3 in [13] that

$$\lim_{K\to\infty} \mathbb{P}\big(\{T_\varepsilon^K \leqslant \tilde{T}_\varepsilon^K\} \cap (\mathcal{N}_\varepsilon^K \cap \{T_0^K < T_\varepsilon^{(K,2)} \wedge \bar{T}_\varepsilon^K\})^C\big) = 0. \tag{3.30}$$

Using these events we define two conditional probability measures associated with the first and third phases of the sweep:

$$\mathbb{P}^{(1)}(\cdot) := \mathbb{P}(\,\cdot \mid T_\varepsilon^K \leqslant \tilde{T}_\varepsilon^K), \quad \text{and} \quad \mathbb{P}^{(3)}(\cdot) := \mathbb{P}(\,\cdot \mid \mathcal{N}_\varepsilon^K \cap \{T_0^K < T_\varepsilon^{(K,2)} \wedge \bar{T}_\varepsilon^K\}). \tag{3.31}$$

Then, by (3.12), (3.22) and (3.30), there exist two finite constants c and ε_0 such that for $\varepsilon \leqslant \varepsilon_0$ and any measurable event $\mathcal{C}$

$$\sup_{i\in\{1,3\}} \limsup_{K\to\infty} \big|\mathbb{P}(\mathcal{C} \mid \mathrm{Fix}^K) - \mathbb{P}^{(i)}(\mathcal{C})\big| \leqslant c\varepsilon. \tag{3.32}$$

Instead of the conditional measure $\mathbb{P}$ we can therefore use the probabilities $(\mathbb{P}^{(i)}, i \in \{1,3\})$ which are easier to handle. Nevertheless, in the proof of Lemma 3.11, given in B, we need to define a different approximation of $\mathbb{P}^{(1)}$ in order to be able to justify a coupling with a supercritical birth and death process which is needed for the derivation of the expected number of up- and down-jumps. Keep in mind that in this sort of model an event only happens if the population size increases by one (birth) or decreases by one (death). There is no such thing as an event connected to a *hold* as opposed to the constant population model which we studied in Chapter 2.

3.3.1 Events impacting the neutral gene genealogies in each phase

Similar as in Section 2.1 we will now rigorously define the different types of reproduction events und will derive statements for the genealogies of neutral alleles with respect to each of the three phases of the sweep. We will later then give a fine study of the probabilities for the different kinds of reproduction events in Sections 3.5.2 and 3.5.4, very similar to the derivations from Section 2.4.3, and further combine the results in order to achieve the statements for the whole sweep.

First phase. As we actually work with the continuous time model in this chapter we need first need to define the times where there is a change in the population size. We introduce

the jump times of the population process as follows:

$$\tau_0^K = 0 \quad \text{and} \quad \tau_m^K = \inf\{t > \tau_{m-1}^K, N^K(t) \neq N^K(\tau_{m-1}^K)\}, \quad m \geqslant 1, \tag{3.33}$$

and the number of jumps during the first phase is denoted by $J^K(1)$,

$$J^K(1) = \inf\{m \in \mathbb{N}, N_a^K(\tau_m^K) = \lfloor \varepsilon K \rfloor\}. \tag{3.34}$$

Within the framework of this model we will now define the terms of coalescence and recombination for the neutral loci of two distinct individuals sampled at time T_ε^K with types denoted by $(\alpha, \beta, \gamma) \in \mathcal{E}$ and $(\alpha', \beta', \gamma') \in \mathcal{E}$, respectively. The definitions are essentially the same as for the model introduced in Section 2.1 of the previous chapter, but we use a different phrasing here.

We say that β and β' coalesce at time τ_m^K if they are carried by two different individuals at time τ_m^K and by the same individual at time τ_{m-1}^K. Seen forwards in time this corresponds to a birth and hence a copy of the neutral allele. Going backwards, it corresponds to the fusion of two neutral alleles into one, carried by one parent of the newborn. We define in the same way coalescent events at locus N2 (resp. both loci N1 and N2) for alleles γ and γ' (resp. allele pairings (β, γ) and (β', γ')).

We say that β (and/or γ) recombines at time τ_m^K from the α- to the α'-population if the individual carrying the allele β (and/or γ) at time τ_m^K is a newborn, carries the allele α inherited from it first parent, and has inherited its allele β (and/or γ) from a different individual carrying allele α'.

Note that we again are not interested in recombination events which do not change either the background of an allele or the current status of common ancestry of sampled neutral loci. This in particular means that in (G1) we do not keep track of recombinations of a pair (β, γ) within the α-population. Moreover, by a similar argument as in the proof of Theorem 2.7 in Section 2.3.3, we do not consider the recombination of β or γ within the α-population in (G2) as such an event has no impact on the other neutral locus.

Let us now describe the genealogical scenarios which determine the ancestral relationships between the neutral alleles of one individual during the each phase and which occur with positive probability when K is large. Recall the definitions in e1) and e2) on page 71 and define accordingly

$[2,1]_{A,i}^{rec}$: first $(i,2)$ recombines into the A-population, then $(i,1)$ recombines into the A-population and connects to a different individual than $(i,2)$,

$[12,2]_{A,i}^{rec}$: the tuple $(i,1),(i,2)$ recombines into the A-population, then a second recombination splits the two neutral loci inside the A-population.

Let us at first focus on the first phase and introduce for an individual i, sampled uniformly from the a-population at time T_ε^K, the following events in case of (G1):

$$\begin{aligned}
NR(i)^{(1)}: &\quad \text{there is no recombination into the } A\text{-population affecting } (i,1) \text{ or } (i,2) \text{ and}\\
&\quad \text{both neutral loci of the } i\text{-individual originate from the first mutant,}\\
R2(i)^{(1)}: &\quad \text{only the allele } (i,2) \text{ is affected by a recombination with the } A\text{-population,}\\
&\quad \text{hence } (i,1) \text{ originates from the first mutant and } (i,2) \text{ from an } A\text{-individual,}\\
R12(i)^{(1)}: &\quad \text{a recombination between SL and N1 from the } a\text{- into the } A\text{-population occurs}\\
&\quad \text{and both neutral alleles } (i,1) \text{ and } (i,2) \text{ originate from the same } A\text{-individual,}\\
R1|2(i)^{(1,\mathrm{G1})}: &\quad [2,1]_{A,i}^{rec,1} \cup [12,2]_{A,i}^{rec,1}.
\end{aligned} \tag{3.35}$$

For a better understanding consult again the drawing from Figure 2.2.1 in which for example individual 4 experienced the event $[2,1]_{A,i}^{rec}$ and individual 5 the event $[12,2]_{A,i}^{rec}$.
With the definition of the probabilities q_1, q_2, q_3 in (3.16) we get the following proposition concerning the neutral genealogies during the first phase under (G1).

Proposition 3.6. *Let i be an a-individual sampled uniformly at the end of the first phase. Under Assumption 3.3, there exist two finite constants c and ε_0 such that for every $\varepsilon \leqslant \varepsilon_0$,*

$$\begin{aligned}
\limsup_{K\to\infty} \Big\{ &\Big|\mathbb{P}^{(1)}(NR(i)^{(1)}) - q_1 q_2\Big| + \Big|\mathbb{P}^{(1)}(R2(i)^{(1)}) - q_1(1-q_2)\Big| \\
&+\Big|\mathbb{P}^{(1)}(R12(i)^{(1)}) - q_3\Big| + \Big|\mathbb{P}^{(1)}(R1|2(i)^{(1,\mathrm{G1})}) - (1-q_1-q_3)\Big|\Big\} \leqslant c\varepsilon.
\end{aligned}$$

As the sum of all q_j-combinations in Proposition 3.6 equals one, this statement implies that for large K, the sum of probabilities of the four events is as well close to one. In particular, in the limit $K \to \infty$ we exclusively observe the events described in (3.35).
Let us now analogous introduce the possible genealogical trajectories for geometry (G2) during the first phase:

$$\begin{aligned}
NR(i)^{(1)}, R2(i)^{(1)}: &\quad \text{are defined as for geometry (G1)}\\
R1(i)^{(1)}: &\quad \text{only } (i,1) \text{ is affected by a recombination into the } A\text{-population;}\\
&\quad (i,2) \text{ originates from the first mutant and } (i,1) \text{ from an } A\text{-individual}\\
R1|2(i)^{(1,\mathrm{G2})}: &\quad (i,1) \text{ and } (i,2) \text{ are affected by a recombination with the}\\
&\quad A\text{-population; they originate from two distinct } A\text{-individuals}
\end{aligned} \tag{3.36}$$

In Section 3.5.2 we will prove Proposition 3.6, the below stated asymptotics for the neutral

genealogies under (G2) follow similarly.

Proposition 3.7. *Let i be an a-individual sampled uniformly at the end of the first phase. Under Assumption 3.3, there exist two finite constants c and ε_0 such that for every $\varepsilon \leqslant \varepsilon_0$,*

$$\limsup_{K\to\infty} \Big\{ \Big|\mathbb{P}^{(1)}(NR(i)^{(1)}) - q_1q_2\Big| + \Big|\mathbb{P}^{(1)}(R2(i)^{(1)}) - q_1(1-q_2)\Big| \\ + \Big|\mathbb{P}^{(1)}(R1(i)^{(1)}) - (1-q_1)q_2\Big| + \Big|\mathbb{P}^{(1)}(R1|2(i)^{(1,\mathrm{G2})}) - (1-q_1)(1-q_2)\Big| \Big\} \leqslant c\varepsilon.$$

Both statements are in correspondence with the results for the time interval $[0, \tau_N)$ from the previous chapter (if again the parameters of this model are chosen accordingly, as already mentioned in Section 3.2).

Second phase. As already mentioned, the duration of the second phase is of order $\mathcal{O}(1)$. Now, by the assumption on the recombination probabilities made in (3.2), they are negligible with respect to one and consequently, we expect that no recombination event impacts the genealogies of the neutral loci occurs during the second phase. More precisely, let us sample uniformly two distinct a-individuals i and j at the end of the second phase and introduce the events:

$$\begin{aligned} NR(i)^{(2)}: &\quad \text{there is no recombination affecting } (i,1) \text{ or } (i,2), \\ NC(i,j)^{(2)}: &\quad \text{there is no coalescence between the neutral genealogies of} \\ &\quad \text{individuals } i \text{ and } j. \end{aligned} \tag{3.37}$$

Then we have the following result on the neutral genealogies during the second phase, for both geometries (G1) and (G2). We will briefly mention the idea of a proof it in Section 3.5.4.

Proposition 3.8. *Let i and j be two distinct a-individuals sampled uniformly at the end of the second phase. Then under Assumption 3.3,*

$$\lim_{K\to\infty} \mathbb{P}^{(1)}(NR(i)^{(2)} \cap NC(i,j)^{(2)}) = 1.$$

Third phase. Finally, when K is large, there is only one event occurring with positive probability during the third phase which might modify the ancestry of the neutral alleles of an individual i sampled at the end of the sweep in geometry (G1):

$$\begin{aligned} R2(i)^{(3,\mathrm{G1})}: &\quad \text{a recombination between loci N1 and N2 occurs and separates} \\ &\quad (i,1) \text{ and } (i,2) \text{ within the } a\text{-population}, \end{aligned} \tag{3.38}$$

Hence, we get the analogous statement as in Section 2.3.1 where we claimed that after time τ_N the only relevant event was a separation within the mutant B-population, and thus there

is no recombination into the wild-type population during the third phase. If we here define the events

$NR(i)^{(3)}$: there is no recombination affecting $(i,1)$ or $(i,2)$ and they both originate from the same a-individual at the end of the second phase

$NC(i,j)^{(3)}$: defined as $NC(i,j)^{(2)}$ for two distinct individuals sampled uniformly at the end of the sweep.

$NRA(i)^{(3)}$: no neutral allele of individual i recombines from the a to the A population.

we indeed get the following proposition on the neutral genealogies during the third phase under (G1).

Proposition 3.9. *Let i and j be two distinct a-individuals sampled uniformly at the end of the sweep. Under Assumption 3.3, there exist two finite constants c and ε_0 such that for every $\varepsilon \leqslant \varepsilon_0$,*

$$\limsup_{K\to\infty}\Big\{\Big|\mathbb{P}^{(3)}(R2(i)^{(3,\mathrm{G1})}) - (1-\bar{q}_2)\Big| + \Big|\mathbb{P}^{(3)}(NR(i)^{(3)}) - \bar{q}_2\Big| + \Big|\mathbb{P}^{(3)}(NC(i,j)^{(3)}) - 1\Big| + \Big|\mathbb{P}^{(3)}(NRA(i)^{(3)}) - 1\Big|\Big\} \leqslant c\sqrt{\varepsilon}.$$

Note that we do not need an analogous statement for the separation of the two neutral loci under geometry (G2), as mentioned before.

Multinomial statement. For proving the Theorems 3.4 and 3.5 we finally need a statement which makes it possible to deduce a joint result for the whole sample from the above Propositions. In [38] it was argued that one can directly transfer the binomial statement from Proposition 2.6 in [36] under the Moran model to the setting of a Darwinian model. Hence, here we will essentially use the results from the Propositions 2.20 and 2.19 without transferring the calculations into the framework of the here considered model. Nevertheless, we will restate the Propositions with the here used notation.
Let $\Theta_d^{(K,1)} \in \mathcal{P}^*_{[d;1,2]}$ denote the analog of Θ_d^K where the d individuals are sampled at the end of the first phase and not at the end of the sweep. Further, denote by $|R2^{(3,\mathrm{G1})}|_d$ (resp. $|NR^{(3)}|_d$) the number of a-individuals in a d-sample taken at the end of the sweep whose neutral alleles originate from two distinct a-individuals (resp. from the same a-individual) at the beginning of the third phase. Then we have the following result:

Proposition 3.10 (cf. Propositions 2.19, 2.20)**.** *Let $d \in \mathbb{N}$. Then, under Assumption 3.3 and geometry* (G1)*, there exist two finite constants c and ε_0 such that for every $\varepsilon \leqslant \varepsilon_0$, the ancestral relationships of a d-sample taken at the end of the first phase satisfy for every*

$(m_1,\dots,m_4)\in\mathbb{Z}_+^4$:

$$\limsup_{K\to\infty}\Big|\mathbb{P}^{(1)}(|\Theta_d^{(K,1)}|_k=m_k,1\leqslant k\leqslant 4) - \mathbf{1}_{\{m_1+m_2+m_3+m_4=d\}}\frac{d!}{m_1!m_2!m_3!m_4!}(q_1q_2)^{m_1}(q_1(1-q_2))^{m_2}q_3^{m_3}(1-q_1-q_3)^{m_4}\Big|\leqslant c\sqrt{\varepsilon},$$

with q_1,q_2,q_3 *from* (3.16) *and* $|\Theta_d^{(K,1)}|_1=\#\{i:NR(i)^{(1)}\}$, $|\Theta_d^{(K,1)}|_2=\#\{i:R2(i)^{(1)}\}$, $|\Theta_d^{(K,1)}|_3=\#\{i:R12(i)^{(1)}\}$ *and* $|\Theta_d^{(K,1)}|_4=\#\{i:R1|2(i)^{(1,\mathrm{G1})}\}$.
In the same way, the neutral genealogy of a d-sample taken at the end of the sweep satisfies for every $(m_1,m_2)\in\mathbb{Z}_+^2$:

$$\limsup_{K\to\infty}\Big|\mathbb{P}^{(3)}((|R2^{(3,\mathrm{G1})}|_d,|NR^{(3)}|_d)=(m_1,m_2))-\mathbf{1}_{\{m_1+m_2=d\}}\frac{d!}{m_1!m_2!}(1-\bar{q}_2)^{m_1}\bar{q}_2^{m_2}\Big|\leqslant c\sqrt{\varepsilon}.$$

Again, this proposition is a key result in the sense that we now only need to focus on individual neutral genealogies to get results on the joint genealogy of a d-sample.

3.3.2 Proof of Theorem 3.4

We will only briefly state the idea (in case of the geometric alignment (G1)) and refrain from detailed calculations as all the ingredients were already stated in the previous section, Propositions 3.6, 3.8, 3.9 and 3.10. Theorem 3.5 follows in the same way.
The procedure for proving Theorem 3.4 is as follows: we will consider one uniformly sampled individual from the end of the selective sweep and list for each claimed ancestral relationship from (3.15) some specific events which lead to such a configuration for the one individual i. Then we use the above stated propositions and show that the sum of the probabilities of all listed events equals one up to terms of order $\sqrt{\varepsilon}$ for an arbitrarily small, fixed ε, which implies that in the limit $K\to\infty$ we indeed have only these possibilities for the neutral genealogy of an individual i. Applying Proposition 3.10 then finishes the proof as it leads to the joint statement for the whole random sample of $d\in\mathbb{N}$ individuals.
Recall for that purpose the ancestral relationships $\mathcal{A}_0,\dots,\mathcal{A}_4$ as defined in (2.20) where individuals fulfilling $\mathcal{A}_j$ will account to $|\pi|_{j+1}$, $j=0,\dots,4$. Further, have again a look at the events which were defined in (3.35), (3.37) and on page 128 in Section 3.3.1.
1. Events leading to $\mathcal{A}_0$. We will distinguish between whether the two neutral loci of individual i separated within the a-population during the third phase of the sweep or not.

$$\begin{aligned} ev_{11} &:= R2(i)^{(3,\mathrm{G1})}\cap NR(i1)^{(2)}\cap NR(i2)^{(2)}\cap NC(i1,i2)^{(2)}\\ &\qquad\cap[NR(i1)^{(1)}\cup R2(i1)^{(1)}]\cap NR(i2)^{(1)} \qquad (3.39)\\ ev_{12} &:= NR(i)^{(3)}\cap NR(i)^{(2)}\cap NR(i)^{(1)}, \end{aligned}$$

where we denote by $i1$ and $i2$ the labels of the parents of the first and second neutral loci of i, respectively, at the end of the second phase (the way we label the a-individuals has no importance as they are exchangeable).

2. Events leading to $\mathcal{A}_1$. Again we list two events where the first considers a split in the a-population towards the end of the sweep and the second the complementary event.

$$\begin{aligned} ev_{21} &:= R2(i)^{(3,\mathrm{G1})} \cap NR(i1)^{(2)} \cap NR(i2)^{(2)} \cap NC(i1,i2)^{(2)} \cap [NR(i1)^{(1)} \cup R2(i1)^{(1)}] \\ &\qquad \cap [R12(i2)^{(1)} \cup R1|2(i2)^{(1)} \cup R2(i2)^{(1)}] \\ ev_{22} &:= NR(i)^{(3)} \cap NR(i)^{(2)} \cap R2(i)^{(1)}. \end{aligned} \tag{3.40}$$

For understanding the first event we recall that the second locus will also migrate in case of a recombination between SL and N1.

3. Events leading to $\mathcal{A}_2$. Here, a separation of the neutral loci is necessary for the occurrence of the relationship $\mathcal{A}_2$ as we showed that the events aAa, $2R$ and $R2a$ as defined on page 136 are negligible. We hence consider only one trajectory for $\mathcal{A}_2$:

$$\begin{aligned} ev_{31} := R2(i)^{(3,\mathrm{G1})} &\cap NR(i1)^{(2)} \cap NR(i2)^{(2)} \cap NC(i1,i2)^{(2)} \\ &\cap [R12(i1)^{(1)} \cup R1|2(i1)^{(1,\mathrm{G1})}] \cap NR(i2)^{(1)}. \end{aligned} \tag{3.41}$$

4. Events leading to $\mathcal{A}_3$. By a similar reasoning as above we only consider the following trajectory for $\mathcal{A}_3$:

$$ev_{41} := NR(i)^{(3)} \cap NR(i)^{(2)} \cap R12(i)^{(1)}. \tag{3.42}$$

5. Events leading to $\mathcal{A}_4$. Finally, we list two possibilities for the occurrence of two singletons $\{(i,1)\}$ and $\{(i,2)\}$ in the partition of the sample.

$$\begin{aligned} ev_{51} &:= R2(i)^{(3,\mathrm{G1})} \cap NR(i1)^{(2)} \cap NR(i2)^{(2)} \cap NC(i1,i2)^{(2)} \\ &\qquad \cap [R12(i1)^{(1)} \cup R1|2(i1)^{(1)}] \cap [R12(i2)^{(1)} \cup R1|2(i2)^{(1)} \cup R2(i2)^{(1)}] \\ ev_{52} &:= NR(i)^{(3)} \cap NR(i)^{(2)} \cap R1|2(i)^{(1,\mathrm{G1})}. \end{aligned} \tag{3.43}$$

In (3.31) we defined the probability measures $\mathbb{P}^{(1)}$ and in such a way that the following identity holds for all measurable events $\mathcal{C}^{(1)}$, $\mathcal{C}^{(2)}$ and $\mathcal{C}^{(3)}$ occurring during the first, second and third phase, respectively,

$$\mathbb{P}(\mathcal{C}^{(1)} \cap \mathcal{C}^{(2)} \cap \mathcal{C}^{(3)} \mid \mathrm{Fix}^K) = \mathbb{P}^{(3)}(\mathcal{C}^{(3)} \mid \mathcal{C}^{(1)} \cap \mathcal{C}^{(2)}) \cdot \mathbb{P}^{(1)}(\mathcal{C}^{(2)} \mid \mathcal{C}^{(1)}) \cdot \mathbb{P}^{(1)}(\mathcal{C}^{(1)}) + O_K(\varepsilon), \tag{3.44}$$

with $\mathbb{P}$ as defined in (3.11) and $O_K(\varepsilon)$ a function of K and ε which satisfies

$$\limsup_{K\to\infty} |O_K(\varepsilon)| \leqslant c\varepsilon, \tag{3.45}$$

for $\varepsilon \leqslant \varepsilon_0$ where ε_0 and c are finite.
This enables us to calculate the probabilities of the above chain of events by applying the results from the Propositions 3.6, 3.8 and 3.9 where the corresponding probabilities were already determined for the limit of large K. We will not give all the details of the calculations here but we indeed in the end get the entries of the vector $\bar{p}$ as defined in (3.17). As an example consider the probability of $\mathcal{A}_0$. Note that we use the exchangeability of lineages within one background and, implicitly, a resampling argument as in the proof of Theorem 2.6 for each phase. In addition we deduce from Proposition 3.10 that we can treat distinct individuals independently such that the events for the (possibly) two ancestors $i1$ and $i2$ of i are independent. In particular, this independence holds also true for the second phase, as there are no events impacting the neutral genealogies of the sampled individuals.

$$\begin{aligned}
\mathbb{P}(ev_{11} \uplus ev_{12}) &= \mathbb{P}^{(3)}(R2(i)^{(3,\mathrm{G1})}) \cdot 1 \cdot \mathbb{P}^{(1)}([NR(i1)^{(1)} \cup R2(i1)^{(1)}] \cap NR(i2)^{(1)}) \\
&\quad + \mathbb{P}^{(3)}(NR(i)^{(3)}) \cdot 1 \cdot \mathbb{P}^{(1)}(NR(i)^{(1)}) + O_K(\varepsilon) \\
&= (1-\bar{q}_2) \cdot q_1 q_2 \cdot (q_1 q_2 + q_1(1-q_2)) + \bar{q}_2 q_1 q_2 + O_K(\varepsilon) \\
&= (1-\bar{q}_2) q_1^2 q_2 + \bar{q}_2 q_1 q_2 + O_K(\sqrt{\varepsilon}) = p_1 + O_K(\sqrt{\varepsilon}).
\end{aligned}$$

□

3.4 Number of births and deaths during the selective sweep

In this section we will only state the results on the expected number of up-jumps and down-jumps during the different phases of the sweep, and refrain from giving any proofs as the derivations in this part of the joint project were essentially done by Charline Smadi. We will however give an intuition of the basic idea and state the proof of the first Lemma 3.11 of this section in the Appendix B in order to give some insight into the calculations.

Births and deaths during the first phase

In the calculations for the probabilities of the events described in (3.35) we need the following approximations for the number of up-jumps of the a-population and further a bound on the covariance for the Poisson approximations. Recall the definition of the jump times of the a-population process during the first phase from (3.33) and consider some $k < \lfloor \varepsilon K \rfloor$. Then

we define the number of up-jumps from k to $k+1$ during the first phase by

$$U_k^K(1) := \#\{m, \tau_m^K \leqslant T_\varepsilon^K, (N_a^K(\tau_m^K), N_a^K(\tau_{m+1}^K)) = (k, k+1)\}. \tag{3.46}$$

Further, for $\varepsilon < S_{aA}/(2C_{a,A}C_{A,a}/C_{A,A} + C_{a,a})$ we define two constants which depend on the ecological parameters of the model:

$$s - \frac{2C_{a,A}C_{A,a} + C_{a,a}C_{A,A}}{f_a C_{A,A}}\varepsilon =: s_-(\varepsilon) \leqslant s \leqslant s_+(\varepsilon) := s + 2\frac{C_{a,A}C_{A,a}}{f_a C_{A,A}}\varepsilon. \tag{3.47}$$

Those values are motivated by the coupling with a supercritical birth and death process (which will be described in the appendix). Recall (3.20) and introduce a real number λ_ε

$$\lambda_\varepsilon := (1 - s_-(\varepsilon))^3(1 - s_+(\varepsilon))^{-2}, \tag{3.48}$$

which belongs to $(0, 1)$ for ε small enough.

Lemma 3.11. *There exist three positive finite constants c, K_0 and ε_0 such that for $K \geqslant K_0$ and $\varepsilon \leqslant \varepsilon_0$:*

If $j \leqslant k < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K$,

$$\left|\mathbb{E}^{(1)}_{(n_A,j)}[U_k^K(1)] - \frac{1 - (1-s)^{\lfloor \varepsilon K \rfloor - k} - (1-s)^{k+1}}{s}\right| \leqslant c\varepsilon. \tag{3.49}$$

If $k < j < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K$,

$$\mathbb{E}^{(1)}_{(n_A,j)}[U_k^K(1)] \leqslant \frac{(1 - s_-(\varepsilon))^{j-k}}{s_+(\varepsilon)s_-^2(\varepsilon)}. \tag{3.50}$$

If $k' \leqslant k < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K$,

$$\left|\mathrm{Cov}^{(1)}_{(n_A,j)}(U_k^K(1), U_{k'}^K(1))\right| \leqslant c\Big(\lambda_\varepsilon^{(k-k')/2} + \varepsilon\Big). \tag{3.51}$$

As mentioned before, a proof of this lemma is given in the Appendix B.
From Lemma 3.11 we can deduce an approximate value for the expected number of down-jumps by the same argument as given in (2.38).

Corollary 3.12. *Let $\varepsilon \leqslant \varepsilon_0$ and $K \geqslant K_0$ where ε_0 and K_0 are as in Lemma 3.11 and define the total number of downcrossings from k to $k-1$,*

$$D_k^K(1) := \#\{m, \tau_m^K \leqslant T_\varepsilon^K, (N_a^K(\tau_m^K), N_a^K(\tau_{m+1}^K)) = (k, k-1)\}, \tag{3.52}$$

then by definition of the probability $\mathbb{P}^{(1)}$, for every $2 \leqslant k < \lfloor \varepsilon K \rfloor$

$$D_k^K(1) = U_{k-1}^K(1) - 1, \quad \mathbb{P}^{(1)} - a.s.,$$

and for $j < k < \lfloor \varepsilon K \rfloor$ *and* $n_A \in I_\varepsilon^K$,

$$\left| \mathbb{E}^{(1)}_{(n_A,j)}[D_k^K(1)] - \frac{1-s}{s}(1-(1-s)^{\lfloor \varepsilon K \rfloor - k} - (1-s)^{k-1}) \right| \leqslant c\varepsilon, \tag{3.53}$$

where c *is a finite constant.*

Note that the expressions for the expectations both in Lemma 3.11 and Corollary 3.12 are actually similar looking to the results from Lemma 2.23.
We will further need the number of up-jumps from k to $k+1$ during an excursion above or below l. Denote by $\sigma_l^K(1)$ the number of the jump when the a-population first hits a level l before the end of the first phase, $l < \lfloor \varepsilon K \rfloor$,

$$\sigma_l^K(1) := \inf\{m, \tau_m^K \leqslant T_\varepsilon^K, N_a^K(\tau_m^K) = l\}, \tag{3.54}$$

and for $1 \leqslant k, l < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K$,

$$U^K_{n_A,l,k}(1) := \#\Big\{ m < \sigma_l^K(1), (N_a^K(\tau_m^K), N_a^K(\tau_{m+1}^K)) = (k, k+1) \Big\}. \tag{3.55}$$

Then, if we denote by μ_ε the real number

$$\mu_\varepsilon := (1 - s_-(\varepsilon))^2 (1 - s_+(\varepsilon))^{-1}, \tag{3.56}$$

which belongs to $(0,1)$ for ε small enough, we can derive the following bounds:

Lemma 3.13. *There exist three positive finite constants* c, K_0 *and* ε_0 *such that for* $K \geqslant K_0$, $\varepsilon \leqslant \varepsilon_0$, $1 \leqslant k < l < \lfloor \varepsilon K \rfloor$ *and* $n_A \in I_\varepsilon^K$,

$$\mathbb{E}^{(1)}_{(n_A,k+1)}[U^K_{n_A,k,l}(1) \mid \sigma_k^K(1) < \infty] \vee \mathbb{E}^{(1)}_{(n_A,l-1)}[U^K_{n_A,l,k}(1)] \leqslant c\mu_\varepsilon^{l-k}.$$

For the first phase, we only need one more expression which for once focuses on the A-population rather than the mutant population. We analogously define for $k < \lfloor \varepsilon K \rfloor$ the number of up-jumps of the A-population when the a-population is of size k:

$$\mathcal{U}_k^K(1) := \#\{m, \tau_m^K \leqslant T_\varepsilon^K, N_A^K(\tau_{m+1}^K) - N_A^K(\tau_m^K) = 1, N_a^K(\tau_m^K) = k\}, \tag{3.57}$$

The derivation of the bounds for the expectations and covariances of these quantities is more evolved than the results for the a-population and is again omitted here as the techniques were developed by Charline Smadi. See [11] for more details.

Lemma 3.14. *There exist three finite constants* c, K_0 *and* ε_0 *such that for* $K \geqslant K_0$, $\varepsilon \leqslant \varepsilon_0$

and $k < \lfloor \varepsilon K \rfloor$,

$$\begin{aligned} \Big| \mathbb{E}^{(1)}\Big[\sum_{i=1}^{k} \mathcal{U}_i^K(1) \Big] - \frac{f_A \bar{n}_A K \log k}{s f_a} \Big| &\leqslant cK(1+\sqrt{\varepsilon}\log k), \\ \mathrm{Var}^{(1)}\Big(\sum_{i=1}^{k} \mathcal{U}_i^K(1) \Big) &\leqslant cK^2(1+\sqrt{\varepsilon}\log^2 k). \end{aligned} \tag{3.58}$$

Births and deaths during the second and third phase

Recall the definition of the jump numbers τ_m^K in (3.33) and denote by $U^K(2)$ the total number of up-jumps of the a-population during the second phase which corresponded to the time interval $[T_\varepsilon^K, T_\varepsilon^K + t_\varepsilon]$:

$$U^K(2) := \#\{m, T_\varepsilon^K < \tau_m^K \leqslant T_\varepsilon^K + t_\varepsilon, N_a^K(\tau_{m+1}^K) - N_a^K(\tau_m^K) = 1\}. \tag{3.59}$$

Further introduce the event

$$C_\varepsilon^K := \{N_a^K(t) \geqslant \varepsilon^2 K/4, \forall\ T_\varepsilon^K \leqslant t \leqslant T_\varepsilon^K + t_\varepsilon\}. \tag{3.60}$$

Then, we get the following almost sure statement in the limit of large K:

Lemma 3.15.

$$\lim_{K\to\infty} \mathbb{P}^{(1)}(U^K(2) > K \log\log K \mid C_\varepsilon^K) = 0.$$

Finally, define the number of up-jumps of the a-population during the third phase when N_A^K equals $k \leqslant \lfloor \varepsilon K \rfloor$

$$U_k^K(3) := \#\{m, T_\varepsilon^K + t_\varepsilon < \tau_m^K \leqslant T_{\mathrm{ext}}^K, N_a^K(\tau_{m+1}^K) - N_a^K(\tau_m^K) = 1, N_A^K(\tau_m^K) = k\}. \tag{3.61}$$

We now state an approximation for the expectation of $U_k^K(3)$ which can be proven in a similar way as Lemma 3.14.

Lemma 3.16. *There exist three finite constants c, ε_0 and K_0 such that for $\varepsilon \leqslant \varepsilon_0$ and $K \geqslant K_0$,*

$$\begin{aligned} \Big| \mathbb{E}^{(3)}\Big[\sum_{i=1}^{k} U_i^K(3) \Big] - \frac{f_a \bar{n}_a K \log k}{\bar{s} f_A} \Big| &\leqslant cK(1+\sqrt{\varepsilon}\log k), \\ \mathrm{Var}^{(3)}\Big(\sum_{i=1}^{k} U_i^K(3) \Big) &\leqslant cK^2(1+\varepsilon\log^2 K). \end{aligned}$$

3.5 The genealogy of the two neutral loci of one individual

With the results from the previous section at hand, we can now proceed similar as in Section 2.4.3 and calculate the marking probabilities for one individual. We will consider the three phases separately, starting with the first phase. All results in this part strongly base on the methods which were developed in [36] and refined in Section 2.4.3 of the previous chapter.

3.5.1 Coalescence and recombination

We start with giving the probabilities for coalescence and recombination under this model with varying population size. Recall the definition in (2.15) and let

$$r_j^* := r_1 + \mathbf{1}_{\{j=2\}}(r_2 - 2r_1r_2), \quad j \in \{1,2\} \tag{3.62}$$

denote the probability to have exactly one recombination in front of locus Nj. Further let $p^{(c_j)}_{\alpha\alpha'}(n)$, $j \in \{1,2\}$, be the probability that, conditionally on the current state of the population $(N_A^K(\tau_{m-1}^K), N_a^K(\tau_{m-1}^K)) = n \in \mathbb{N}^2$ and on the birth of an individual carrying allele α at time τ_m^K (i.e. $N_\alpha^K(\tau_m^K) - N_\alpha^K(\tau_{m-1}^K) = 1$) the genealogies of two randomly chosen neutral alleles, located at locus Nj and associated respectively with alleles α and α' at time τ_m^K, coalesce at τ_m^K. Then the following Lemma follows from Lemma 7.1 in [38] by replacing the recombination probabilities appropriately.

Lemma 3.17. *For every $n = (n_A, n_a) \in \mathbb{N}^2$, $\alpha \in \mathcal{A}$ and $j \in \{1,2\}$,*

$$p^{(c_j)}_{\alpha\alpha}(n) = \frac{2}{n_\alpha(n_\alpha+1)}\Big(1 - \frac{r_j^* f_{\bar\alpha} n_{\bar\alpha}}{f_A n_A + f_a n_a}\Big) \quad \textit{and} \quad p^{(c_j)}_{\alpha\bar\alpha}(n) = \frac{r_j^* f_{\bar\alpha}}{(n_\alpha+1)(f_A n_A + f_a n_a)}.$$

In the same way let $p^{(r_j)}_{\alpha_1\alpha_2}(n)$ denote the probability to have a recombination from the α_1- into the α_2-population precisely (and exclusively) before locus Nj, again conditionally on $(N_A^K(\tau_{m-1}^K), N_a^K(\tau_{m-1}^K)) = n \in \mathbb{N}^2$ and on the birth of an individual carrying allele α_1 at time τ_m^K. Further, we denote by $p^{(r_{12})}_{\alpha_1\alpha_2}(n)$ the probability to have a double recombination under the same conditions. Letting $r_{12}^* := r_1 r_2$ we obtain the following result:

Lemma 3.18. *For $n \in \mathbb{N}^2$, $\alpha \in \{A,a\}$ and $j \in \{1,2,12\}$, we have*

$$p^{(r_j)}_{\alpha\alpha}(n) = \frac{r_j^* f_\alpha (n_\alpha - 1)}{(n_\alpha+1)(f_A n_A + f_a n_a)} \quad \textit{and} \quad p^{(r_j)}_{\alpha\bar\alpha}(n) = \frac{r_j^* f_{\bar\alpha} n_{\bar\alpha}}{(n_\alpha+1)(f_A n_A + f_a n_a)}.$$

Proof. The second equality is stated in (7.2) in Remark 4 in [38]. Conditional on the birth of an α-individual and the state of the process at the $(m-1)$-th jump, the probability of picking the newborn when choosing an individual at random amongst the α-individuals is equal to $1/(n_\alpha+1)$. A recombination before the locus Nj (or before locus $N1$ and locus $N2$ if $j = 12$)

happens with probability r_j^*, independent of all other events. Finally, the probability that the second parent belongs to the same background as the newborn but is different from the first parent is equal to $f_\alpha(n_\alpha - 1)/(f_A n_A + f_a n_a)$. □

We will often use approximate values of the above defined recombination events, given in the following remark.

Remark 3.19. *Let us recall the definition of I_ε^K in (3.20). Then there exist three finite constants c, ε_0 and K_0 such that for $\varepsilon \leq \varepsilon_0$, $K \geq K_0$, $j \in \{1, 2, 12\}$ and $(n_A, k) \in I_\varepsilon^K \times \{1, ..., \lfloor \varepsilon K \rfloor\}$,*

$$(1 - c\varepsilon)\frac{r_j^*}{k+1} \leqslant \inf_{n_A \in I_\varepsilon^K} p_{aA}^{(r_j^*)}(n_A, k) \leqslant \sup_{n_A \in I_\varepsilon^K} p_{aA}^{(r_j^*)}(n_A, k) \leqslant \frac{r_j^*}{k+1}. \tag{3.63}$$

Similar,

$$(1 - c\varepsilon)\frac{r_2}{n_A} \leqslant p_{AA}^{(r_2)}(n_A, k) \leqslant \frac{r_2}{n_A}, \quad \text{and} \quad p_{aa}^{(r_2)}(n_A, k) \leqslant \frac{f_a}{f_A}\frac{r_2}{n_A}. \tag{3.64}$$

3.5.2 The genealogy of the two neutral loci in the first phase

We will here first bound the probabilities of events which would lead to a different structure of the partition of a sample and subsequently calculate the probabilities of events which have non-negligible probabilities. For a rigorous formulation we need to define an equivalent notation as on page 35 for the previously considered model. With the definition of the m-th jump time from (3.33) in mind, we define for $k \in \{1, 2, (1, 2)\}$ and $m \in \mathbb{N}$,

$$\begin{aligned}(\alpha i k)_m := \{&m \leqslant J^K(1) \text{ and the } k\text{-th locus/loci of the } i\text{-th individual is/are} \\ &\text{associated to an allele } \alpha \text{ at the } m\text{-th jump time}\}.\end{aligned} \tag{3.65}$$

The notation $(\alpha i 1)_m, (\alpha' i 2)_m$ here implies that the two neutral loci of individual i are associated to two distinct individuals at the m-th jump time, for any $\alpha, \alpha' \in \mathcal{A}$.
We will prove that we can neglect the following event combinations: sample $d \in \mathbb{N}$ distinct individuals uniformly at the end of the sweep and define

aAa : a neutral allele recombines from the a-population to the A-population, and then (backwards in time) back into the a-population

CR : two neutral alleles coalesce in the a-population, and then (backwards in time) recombine into the A-population

CA : two neutral alleles coalesce and at least one of them carries the allele A at the time of coalescence

$2R$: a neutral allele takes part in a double recombination (i.e. a recombination before N1 and a recombination before N2 at the same birth event)

$R2a$: a recombination separates the two neutral loci of an individual within the a-population

Lemma 3.20. *There exist three positive finite constants c, K_0 and ε_0 such that for $\varepsilon \leqslant \varepsilon_0$ and $K \geqslant K_0$*

$$\mathbb{P}^{(1)}(aAa) + \mathbb{P}^{(1)}(CR) + \mathbb{P}^{(1)}(2R) + \mathbb{P}^{(1)}(R2a) \leqslant \frac{c}{\log K}, \quad \textit{and} \quad \mathbb{P}^{(1)}(CA) \leqslant \frac{c \log K}{K}.$$

Proof. The probabilities of events aAa, CR and CA were already bounded in Lemma 7.3 and (7.12) in [38]. Hence we just have to bound the probability of $2R$ and $R2a$. If a neutral allele experiences a double recombination, it happens either when it is associated with an allele a or with an allele A. From Lemma 3.18 and the fact that r_1 and r_2 are of order $1/\log K$ we get for $k < \lfloor \varepsilon K \rfloor$:

$$\sup_{n_A \in I_\varepsilon^K} \Big(p_{aa}^{(r_{12})}(n_A, k) + p_{aA}^{(r_{12})}(n_A, k) \Big) \leqslant \frac{c}{(k+1)\log^2 K},$$
$$\sup_{n_A \in I_\varepsilon^K} \Big(p_{Aa}^{(r_{12})}(n_A, k) + p_{AA}^{(r_{12})}(n_A, k) \Big) \leqslant \frac{c}{K \log^2 K}.$$

Recall the definitions of $U_k^K(1)$ and $\mathcal{U}_k^K(1)$ in (3.46) and (3.57), respectively. As a birth of an α-individual is needed to have a recombination from the α- to the α'-population, we can bound the probability to have a double recombination by

$$\mathbb{P}^{(1)}(2R) \leqslant \frac{c}{\log^2 K} \mathbb{E}^{(1)} \Big[\sum_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \Big(\frac{U_k^K(1)}{k+1} + \frac{\mathcal{U}_k^K(1)}{K} \Big) \Big].$$

By applying inequality (3.49) and Lemma 3.58 we succeed in bounding $\mathbb{P}^{(1)}(2R)$ by a constant over $\log K$. It remains to consider the event $R2a$ of a recombination within the a-population. This is done similarly as in the proof of Proposition 2.10. Define as in (2.14) the first time (with respect to the backwards-in-time process) that this event happens:

$$\begin{aligned} R_{aa}^{(1)}(i) := \sup\{m,\, m \leqslant J^K(1) \text{ and both neutral loci of the } i\text{-th individual are} \\ \text{associated to different } a\text{-individuals at the } m-1\text{-th jump}\}, \end{aligned} \tag{3.66}$$

where again $\sup \varnothing = -\infty$. Then,

$$\mathbb{P}^{(1)}(R_{aa}^{(1)}(i) \geqslant 0) = \sum_{l=1}^{\lfloor \varepsilon K \rfloor - 1} \mathbb{P}^{(1)}(R_{aa}^{(1)}(i) \geqslant 0, N_a^K(\tau^K_{R_{aa}^{(1)}(i)}) = l)$$

$$
\begin{aligned}
&= \sum_{l=1}^{\lfloor \varepsilon K \rfloor -1} \sum_{m<\infty} \mathbb{P}^{(1)}(m \leqslant J^K(1), N_a^K(\tau_{m-1}^K) = l, N_a^K(\tau_m^K) = l+1, \\
&\qquad\qquad (ai1)_m, (ai2)_m, \forall m' > m : (ai12)_{m'}) \\
&\leqslant \sum_{l=1}^{\lfloor \varepsilon K \rfloor -1} \sum_{m<\infty} \sup_{n_A \in I_\varepsilon^K} \Big(p_{aa}^{(r_2)}(n_A, l) \mathbb{P}^{(1)}_{(n_A, l+1)}(\forall m \geqslant 0 : (ai12)_m) \Big) \\
&\qquad\qquad \cdot \mathbb{P}^{(1)}(m \leqslant J^K(1), N_a^K(\tau_{m-1}^K) = l, N_a^K(\tau_m^K) = l+1) \\
&\leqslant \sum_{l=1}^{\lfloor \varepsilon K \rfloor -1} \frac{c}{K \log K} \mathbb{E}^{(1)}[U_l^K(1)] \leqslant \frac{c}{\log K},
\end{aligned}
$$

by (3.49) and (3.64). □

To simplify the notations we will denote the union of all negligible events by

$$NE := aAa \cup CR \cup CA \cup 2R \cup R2a. \tag{3.67}$$

We will now come to the calculation of the marking probabilities. As they are very similar to the once performed in Section 2.4.3 we will only detail one of them, namely the calculation of the event $[2,1]_{A,i}^{rec}$.

The two loci of one individual separate within the A-population - the event $[2,1]_{A,i}^{rec}$

Let us first state an equivalent of Lemma 2.30 for this context. As the proof follows the same idea as the proof of the statement from the previous chapter, we will not rewrite it within this framework.

Lemma 3.21. *Let $(c_N, N \in \mathbb{N})$ be a bounded sequence of $\mathbb{R}^*$. Then there exists a finite constant c such that*

$$\limsup_{N\to\infty} \Big| \sum_{l=1}^{N} \frac{e^{\frac{c_N}{\log N} \log l}}{l+1} - \frac{\log N}{c_N}(e^{c_N} - 1) \Big| \leqslant c.$$

Proposition 3.22. *Let i be an a-individual sampled uniformly at the end of the first phase. There exist two finite constants c and ε_0 such that for $\varepsilon \leqslant \varepsilon_0$,*

$$\limsup_{K\to\infty} \Big| \mathbb{P}^{(1)}([2,1]_{A,i}^{rec}) - \Big[\frac{r_2}{r_1+r_2} - e^{-\frac{r_1}{s}\log\lfloor \varepsilon K \rfloor} + \frac{r_1}{r_1+r_2} e^{-\frac{r_1+r_2}{s}\log\lfloor \varepsilon K \rfloor} \Big] \Big| \leqslant c\sqrt{\varepsilon}.$$

Recall (3.33) and define similar to the definition in (2.10) for $k \in \{1, 2, \{1,2\}\}$ and $m \in \mathbb{N}$,

$$
\begin{aligned}
R(i,k) := \sup\{ m : m \leqslant J^K(1) \text{ and the } k\text{-th locus/loci of the } i\text{-th individual} \\
\text{is/are associated to an allele } A \text{ at the } (m-1)\text{-th jump time}\},
\end{aligned} \tag{3.68}
$$

the last jump (forwards in time) when the k-th locus/loci of the i-th individual belongs to the A-population (with $\sup \varnothing = -\infty$). The idea for proving Proposition 3.22 is the same as in the proof of Lemma 2.34: we decompose the event $[2,1]^{rec}_{A,i}$ according to the different possible a-population sizes when the first (backwards in time) recombination between N1 and N2 occurs.

$$\begin{aligned}\mathbb{P}^{(1)}([2,1]^{rec}_{A,i}) &= \mathbb{P}^{(1)}(R(i,2) > R(i,1) \geqslant 0) \\ &= \sum_{l=1}^{\lfloor \varepsilon K \rfloor - 1} \mathbb{P}^{(1)}(R(i,2) > R(i,1), N_a^K(\tau^K_{R(i,2)}) = l) \\ &\qquad \cdot \mathbb{P}^{(1)}(R(i,1) \geqslant 0 \mid R(i,2) > R(i,1), N_a^K(\tau^K_{R(i,2)}) = l).\end{aligned} \tag{3.69}$$

In the following Lemma, which then gives results in the proof of Proposition 3.22, we consider separately the two probabilities of the above product:

Lemma 3.23. *There exist three finite constants c, K_0 and ε_0 such that for $K \geq K_0$ and $\varepsilon \leq \varepsilon_0$,*

$$\begin{aligned}\Big| \mathbb{P}^{(1)}(R(i,2) > R(i,1), N_a(\tau^K_{R(i,2)}) = l) \\ - r_2 \frac{1-(1-s)^{\lfloor \varepsilon K \rfloor - l} - (1-s)^{l+1}}{s(l+1)} e^{-\frac{r_1+r_2}{s} \log \frac{\lfloor \varepsilon K \rfloor}{l}} \Big| \leqslant \frac{c\sqrt{\varepsilon}}{l \log K}\end{aligned} \tag{3.70}$$

and

$$\Big| \mathbb{P}^{(1)}(R(i,1) \geq 0 \mid R(i,2) > R(i,1), N_a(\tau^K_{R(i,2)}) = l) - \sum_{k=1}^{l-1} \frac{r_1}{s(k+1)} e^{-\frac{r_1}{s} \log \frac{l-1}{k}} \Big| \leqslant c\sqrt{\varepsilon}. \tag{3.71}$$

Proof of Proposition 3.22. From Lemma 3.23 and Equation (3.69) we get the existence of a finite c such that for K large enough and ε small enough,

$$\begin{aligned}\mathbb{P}^{(1)}([2,1]^{rec}_{A,i}) &\leqslant \sum_{l=1}^{\lfloor \varepsilon K \rfloor} \Big(r_2 \frac{1-(1-s)^{\lfloor \varepsilon K \rfloor - l} - (1-s)^{l+1}}{s(l+1)} \exp\Big(-\frac{r_1+r_2}{s} \log\Big(\frac{\lfloor \varepsilon K \rfloor}{l} \Big) \Big) + \frac{c\sqrt{\varepsilon}}{l \log K} \Big) \\ &\qquad \cdot \Big(\sum_{k=1}^{l-1} \frac{r_1}{s(k+1)} \exp\Big(-\frac{r_1}{s} \log\Big(\frac{l-1}{k} \Big) \Big) + c\sqrt{\varepsilon} \Big) \\ &\leqslant \sum_{l=1}^{\lfloor \varepsilon K \rfloor} \frac{r_2}{s(l+1)} \exp\Big(-\frac{r_1+r_2}{s} \log\Big(\frac{\lfloor \varepsilon K \rfloor}{l} \Big) \Big) \Big(1 - \exp\Big(-\frac{r_1}{s} \log l \Big) \Big) \Big) + c\sqrt{\varepsilon} \\ &\leqslant \Big(\frac{r_2}{r_1+r_2} - e^{-r_1 \log K/s} + \frac{r_1}{r_1+r_2} e^{-(r_1+r_2)\log K/s} \Big) + c\sqrt{\varepsilon},\end{aligned}$$

where we used Lemma 3.21 twice. To get the lower bound we use similar approximations and Lemma A.1. This ends the proof. □

The remainder of this section is devoted to the proof of Lemma 3.23.

Proof of Equation (3.70)*, Lemma 3.23.*
We can decompose the event $\{R(i,2) > R(i,1), N_a(\tau^K_{R(i,2)}) = l\}$ according to the jump number of the (backwards in time) first recombination. Recall the definition of $NR(i)^{(1)}$ in (3.35). We will again use this event but with a slightly different meaning, as the initial condition of (N^K_A, N^K_a) will not necessarily be $(\bar{n}_A K, 1)$. It will however still correspond to the absence of any recombination before the end of the first phase. We recall that whenever we do not indicate the initial conditions for probabilities, expectations and covariances it means that $(N^K_A(0), N^K_a(0)) = (\bar{n}_A K, 1)$.
With the definition of $(\alpha ik)_m$ in (3.65) we get

$$\begin{aligned}
&\mathbb{P}^{(1)}(R(i,2) > R(i,1), N_a(\tau^K_{R(i,2)}) = l)\\
&= \sum_{m>1} \mathbb{P}^{(1)}(m < J^K(1), N^K_a(\tau^K_m) = l, N^K_a(\tau^K_{m+1}) = l+1, (ai1)_m, (Ai2)_m, \forall m' > m : (ai12)_{m'})\\
&\leqslant \sum_{m>1} \sup_{n_A \in I^K_\varepsilon} \Big(p^{(r_1)}_{aA}(n_A, l) \mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)} \mid N^K_a(\tau^K_1) = l+1) \Big)\\
&\qquad \cdot \mathbb{P}^{(1)}(m < J^K(1), N_a(\tau^K_{m-1}) = l, N^K_a(\tau^K_m) = l+1), \qquad (3.72)
\end{aligned}$$

and the same expression with the infimum on $n_A \in I^K_\varepsilon$ for a lower bound. We first focus on the second probability in the sum and aim at proving the existence of two finite constants c and ε_0 such that for every $\varepsilon \leqslant \varepsilon_0$, $n_A \in I^K_\varepsilon$ and $l < \lfloor \varepsilon K \rfloor$,

$$\limsup_{K\to\infty} \left| \mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)} \mid N^K_a(\tau^K_1) = l+1) - e^{-\frac{r_1+r_2}{s} \log(\lfloor \varepsilon K \rfloor / l)} \right| \leqslant c\sqrt{\varepsilon}. \qquad (3.73)$$

We will use the same idea as described on page 65 in the previous chapter. Let

$$\mathcal{N}^{(1)} := \sigma\big(N^K(\tau^K_m), m \leqslant J^K(1)\big) \qquad (3.74)$$

be the σ-algebra generated by the process $N^K = (N^K_A, N^K_a)$ during the first phase. The first step consists in working conditionally on the trajectory of the process N^K, describing this probability as a product of conditional probabilities close to one, and derive a Poisson approximation. To this aim, we define for $m \in \mathbb{Z}_+$:

$$\theta^{(r_1+r_2)}(m) := \mathbf{1}_{\{m < J^K(1)\}} \mathbf{1}_{\{N^K_a(\tau^K_{m+1}) - N^K_a(\tau^K_m) = 1\}} (p^{(r_1)}_{aA} + p^{(r_2)}_{aA})(N^K(\tau^K_m))$$

Then, similarly as for example in (2.45), we have for $n_A \in I^K_\varepsilon$ and $l < \lfloor \varepsilon K \rfloor$

$$\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)} \mid N^K_a(\tau^K_1) = l+1, \mathcal{N}^{(1)}) = [1 - (p^{(r_1)}_{aA} + p^{(r_2)}_{aA})(n_A, l)] \prod_{m=1}^{\infty} (1 - \theta^{(r_1+r_2)}(m)).$$

For sake of simplicity we will approximate the probability

$$\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)} \mid \mathcal{N}^{(1)}) = \prod_{m=0}^{\infty}(1-\theta^{(r_1+r_2)}(m)),$$

which is easier to handle, and satisfies

$$\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)} \mid N_a^K(\tau_1^K) = l+1, \mathcal{N}^{(1)}) - \mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)} \mid \mathcal{N}^{(1)}) \leqslant r_1 + r_2.$$

If we introduce the variable,

$$\eta_l^{(12)} := \mathbf{1}_{\{N_a^K(0)=l\}} \sum_{m=1}^{\infty} \theta^{(r_1+r_2)}(m),$$

we get by following the path of the proof of Lemma 2.32:

$$\begin{aligned}\mathbb{E}^{(1)}_{(n_A,l)}\Big| \prod_{m=1}^{\infty}(1-\theta^{(r_1+r_2)}(m)) - \exp(-\eta_l^{(12)})\Big| &\leqslant \mathbb{E}^{(1)}_{(n_A,l)}\Big[\sum_{m=1}^{\infty}(\theta^{(r_1+r_2)}(m))^2\Big] \\ &\leqslant (r_1+r_2)^2\Big(\sum_{k=1}^{\lfloor \varepsilon K\rfloor}\frac{\mathbb{E}^{(1)}_{(n_A,l)}[U_k^K(1)]}{(k+1)^2}\Big) \leqslant \frac{c}{\log^2 K},\end{aligned} \tag{3.75}$$

for any $n_A \in I_\varepsilon^K, l < \lfloor \varepsilon K\rfloor$ and a finite c, where we used the value for the expectation from (3.49) in Lemma 3.11 and the bound on the recombination probabilities in (3.2). Define an approximation of the random variable $\eta_l^{(12)}$ by

$$\tilde{\eta}_l^{(12)} := \mathbf{1}_{\{N_a^K(0)=l\}} \sum_{m=1}^{\infty} \theta^{(r_1+r_2)}(m)\mathbf{1}_{\{N_a^K(\tau_m^K)\geqslant l\}}, \tag{3.76}$$

and aim at proving that the difference $\mathbb{E}^{(1)}_{(n_A,l)}[\eta_l^{(12)} - \tilde{\eta}_l^{(12)}]$ is negligible for every $n_A \in I_\varepsilon^K$. This latter can be bounded as follows:

$$\begin{aligned}0 &\leqslant \mathbb{E}^{(1)}_{(n_A,l)}[\eta_l^{(12)} - \tilde{\eta}_l^{(12)}] = \mathbb{E}^{(1)}_{(n_A,l)}\Big[\sum_{m=1}^{\infty}\theta^{(r_1+r_2)}(m)\mathbf{1}_{\{N_a^K(\tau_m^K)<l\}}\Big] \\ &\leqslant \sum_{k=1}^{l-1}\mathbb{E}^{(1)}_{(n_A,l)}[U_k^K(1)] \sup_{n_A\in I_\varepsilon^K} p_{aA}^{(r_1+r_2)}(k,n_A) \leqslant \sum_{k=1}^{l-1}\frac{(1-s_-(\varepsilon))^{l-k}}{s_+(\varepsilon)s_-^2(\varepsilon)}\frac{(r_1+r_2)}{k+1} \leqslant c\frac{(r_1+r_2)}{l},\end{aligned} \tag{3.77}$$

for a finite c and ε small enough. We used (3.50) to bound the number of up-jumps, (3.63) to bound the recombination probability, and received the last inequality with the help of Lemma A.1. The expected value of $\tilde{\eta}_l^{(12)}$ can be calculated approximately by using the results from

Section 3.4 and Remark 3.19. For an upper bound we get by using (3.50) and again (3.63)

$$\mathbb{E}^{(1)}_{(n_A,l)}[\tilde{\eta}^{(12)}_l] \leqslant \sum_{k=l}^{\lfloor \varepsilon K\rfloor -1} \frac{r_1+r_2}{k}\mathbb{E}^{(1)}_{(n_A,l)}[U^K_k(1)] \leqslant (1+c\varepsilon)\frac{r_1+r_2}{s}\log\Big(\frac{\lfloor \varepsilon K\rfloor}{l}\Big) + \frac{c}{\log K}, \quad (3.78)$$

for a finite c and ε small enough. For the lower bound we get similarly by using (3.50), (3.63) and Lemma A.1,

$$\mathbb{E}^{(1)}_{(n_A,l)}[\tilde{\eta}^{(12)}_l] \geqslant (1-c\varepsilon)\frac{r_1+r_2}{s}\log\Big(\frac{\lfloor \varepsilon K\rfloor}{l}\Big) - \frac{c}{\log K}, \quad (3.79)$$

for a finite c and ε small enough. The last step consists in bounding the variance of $\tilde{\eta}^{(12)}_l$. As the calculation of this variance is quite involved, we introduce an approximation of $\tilde{\eta}^{(12)}_l$, namely

$$\begin{aligned}\tilde{\tilde{\eta}}^{(12)}_l &:= \mathbf{1}_{\{N^K_a(0)=l\}}\sum_{m=1}^{\infty}\mathbf{1}_{\{N_a(\tau^K_m)\geqslant l\}}\mathbf{1}_{\{N_a(\tau^K_{m+1})-N_a(\tau^K_m)=1\}}\frac{r_1+r_2}{N_a(\tau^K_m)+1}\\ &= \mathbf{1}_{\{N^K_a(0)=l\}}\sum_{k=l}^{\lfloor \varepsilon K\rfloor -1}\frac{r_1+r_2}{k+1}U^K_k(1).\end{aligned}$$

Thanks to Equation (3.63) we get $(1-c\varepsilon)\tilde{\tilde{\eta}}^{(12)}_l \leqslant \tilde{\eta}^{(12)}_l \leqslant \tilde{\tilde{\eta}}^{(12)}_l$ for a finite c and ε small enough, which implies

$$\begin{aligned}&\Big|\hat{\mathrm{Var}}_{(n_A,l)}\tilde{\eta}^{(12)}_l - \hat{\mathrm{Var}}_{(n_A,l)}\tilde{\tilde{\eta}}^{(12)}_l\Big| \leqslant c\varepsilon\mathbb{E}^{(1)}_{(n_A,l)}\Big[\Big(\tilde{\tilde{\eta}}^{(12)}_l\Big)^2\Big]\\ &\leqslant c\varepsilon(r_1+r_2)^2\sum_{k,k'=l}^{\lfloor \varepsilon K\rfloor -1}\frac{\mathbb{E}^{(1)}[(U^K_k(1))^2]+\mathbb{E}^{(1)}[(U^K_{k'}(1))^2]}{(k+1)(l+1)} \leqslant c\varepsilon.\end{aligned} \quad (3.80)$$

Here we used that $U^K_k(1)$ and $U^K_{k'}(1)$ are smaller than geometric random variables with parameter $s_-(\varepsilon)$. Hence, it is sufficient to bound $\hat{\mathrm{Var}}_{(n_A,l)}\tilde{\tilde{\eta}}^{(12)}_l$. By (3.51) and (3.2) we get:

$$\begin{aligned}\hat{\mathrm{Var}}_{(n_A,l)}\tilde{\tilde{\eta}}^{(12)}_l &= (r_1+r_2)^2\sum_{k,k'=l}^{\lfloor \varepsilon K\rfloor -1}\frac{\mathrm{Cov}^{(1)}_{(n_A,l)}(U^K_k,U^K_{k'})}{(k+1)(k'+1)}\\ &\leqslant 2(r_1+r_2)^2\sum_{k\leqslant k'=l}^{\lfloor \varepsilon K\rfloor -1}\frac{\lambda_\varepsilon^{(k'-k)/2}+\varepsilon}{(k+1)(k'+1)} \leqslant c\frac{\log\lfloor \varepsilon K\rfloor}{\log^2 K}(c+\varepsilon\log\lfloor \varepsilon K\rfloor).\end{aligned} \quad (3.81)$$

Recalling (3.80) and again (3.2), we finally obtain

$$\hat{\mathrm{Var}}_{(n_A,l)}\tilde{\eta}^{(12)}_l \leqslant c\varepsilon, \quad (3.82)$$

for a finite c and ε small enough. Following the instructions described on page 65 by using

the inequalities (3.75), (3.77), (3.82), (3.78) and (3.79) ends the proof of equation (3.73). To conclude this proof, we have to control the two other probabilities in the decomposition (3.72). As the first probability in the sum describes the event of a recombination between the two neutral loci from the a- into the A-population, we get from (3.63):

$$\left|\mathbb{P}^{(1)}_{(n_A,l)}((ai1)_0,(Ai2)_0 \mid N_a^K(\tau_1^K)=l+1,(ai12)_1)-\frac{r_2}{l+1}\right| \leqslant \frac{c\varepsilon}{l\log K},$$

for a finite c and ε small enough. Together with (3.73) and (3.46) this leads to,

$$\begin{aligned}
&\mathbb{P}^{(1)}(R(i,2)>R(i,1),N_a(\tau^K_{R(i,2)})=l)\\
&\leqslant (1+c\sqrt{\varepsilon})\frac{r_2}{l+1}e^{-\frac{r_1+r_2}{s}\log\left(\frac{\lfloor\varepsilon K\rfloor}{l}\right)}\sum_{m<\infty}\mathbb{P}^{(1)}(m<J^K(1),N_a^K(\tau_m^K)=l,N_a^K(\tau_{m+1}^K)=l+1)\\
&=(1+c\sqrt{\varepsilon})\frac{r_2}{l+1}e^{-\frac{r_1+r_2}{s}\log\left(\frac{\lfloor\varepsilon K\rfloor}{l}\right)}\mathbb{E}^{(1)}[U_l^K(1)],
\end{aligned}$$

for a finite c, ε small enough and K large enough. We similarly get a lower bound and end up the proof of Equation (3.70) by applying (3.49). □

Proof of Equation (3.71)*, Lemma 3.23.*
We will decompose the event considered here according to the value of N_a^K when the second (backwards in time) recombination occurs. Recall (3.54) and define similarly to τ_k^* from (2.4) the last hitting of k by N_a^K during the first phase,

$$\zeta_k^K(1):=\sup\{m\leqslant T_\varepsilon^K, N_a^K(\tau_m^K)=k\},\quad 1\leqslant k\leqslant\lfloor\varepsilon K\rfloor. \tag{3.83}$$

Further define the events

$$\begin{aligned}
NR(l,\sigma,i):=\{&\text{the first locus of individual } i \text{ sampled at jump time } \sigma_l^K(1)\\
&\text{does not recombine from the } a\text{- to the } A\text{-population between 0 and } \tau^K_{\sigma_l^K(1)}\}
\end{aligned}$$

and $NR(l,\zeta,i)$ which is defined similarly but with $\zeta^K(1)$ instead of $\sigma^K(1)$. Then Bayes' rule leads to:

$$\begin{aligned}
&\mathbb{P}^{(1)}(R(i,1)\geqslant 0 \mid R(i,2)>R(i,1),N_a(\tau^K_{R(i,2)})=l)\\
&\quad=\sum_{k=1}^{\lfloor\varepsilon K\rfloor}\mathbb{P}^{(1)}(R(i,1)\geqslant 0,N_a(\tau^K_{R(i,1)})=k \mid R(i,2)>R(i,1),N_a(\tau^K_{R(i,2)})=l),\\
&\quad\leqslant\sum_{k=1}^{\lfloor\varepsilon K\rfloor}\sum_{m<\infty}\Big(\sup_{n_A\in I_\varepsilon^K}p_{aA}^{(r_1)}(n_A,k)\mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\zeta,i)\mid N_a^K(\tau_1^K)=k+1)\Big)\\
&\qquad\qquad\cdot\mathbb{P}^{(1)}(m<R(i,2),N_a^K(\tau_m^K)=k,N_a^K(\tau_{m+1}^K)=k+1\mid N_a(\tau^K_{R(i,2)})=l).
\end{aligned} \tag{3.84}$$

The lower bound is obtained by taking the infimum for $n_A \in I_\varepsilon^K$ and replacing $NR(l,\zeta,i)$ by $NR(l,\sigma,i)$. Note that the first probability is a known recombination probability which can be evaluated thanks to (3.63). The ideas used to approximate the second probability are the same as those developed to approximate $\mathbb{P}^{(1)}_{n_A,l}(NR(i)^{(1)})$. In particular, we use a decomposition similar to the one described on page 65. However, similar as in the proof of Lemma 2.34 we have a supplementary difficulty due to the randomness of $N_a(\tau^K_{R(i,2)})$. In the previous case we were interested in an event before the first hitting of $\lfloor \varepsilon K \rfloor$, while in the current case, the conditioning on the value of $N_a^K(\tau^K_{R(i,2)})$ does not tell us how many times N_a^K has hit this value before. This is why we needed to introduce $NR(l,\sigma,i)$ and $NR(l,\zeta,i)$. First we prove that with high probability the a-population size is bigger when the (backwards in time) first recombination occurs than when the second, of locus $(i,1)$, occurs. Note that by (3.49) and Lemma 3.13 there exists a finite c such that for every $l \leqslant k < \lfloor \varepsilon K \rfloor$:

$$\begin{aligned}&\sum_{m<\infty} \mathbb{P}^{(1)}(m < R(i,2), N_a^K(\tau_m^K) = k, N_a^K(\tau_{m+1}^K) = k+1 \mid R(i,2) > R(i,1), N_a(\tau^K_{R(i,2)}) = l)\\ &\leqslant \mathbb{E}^{(1)}[U_l^K(1)] \sup_{n_A \in I_\varepsilon^K} \mathbb{E}^{(1)}_{(n_A,l)}[U_{n_A,l,k} | N_a^K(\tau_1^K) = l+1] \leqslant c\mu_\varepsilon^{k-l},\end{aligned}$$

where according to (3.56), $\mu_\varepsilon \in (0,1)$ for ε small enough. Hence, recalling (3.84) and Remark 3.19, we obtain for $k \geqslant l$

$$\mathbb{P}^{(1)}(R(i,1) \geqslant 0, N_a(R(i,1)) \geqslant l \mid R(i,2) > R(i,1), N_a(\tau^K_{R(i,2)}) = l) \leqslant cr_1 \sum_{k=l}^{\lfloor \varepsilon K \rfloor} \frac{\mu_\varepsilon^{k-l}}{k+1} \leqslant \frac{c}{\log K},$$

for a finite c and ε small enough. We therefore can ignore all $k \geqslant l$ in the sum in (3.84) and continue with the case $k < l$. Here, we can bound the sum as follows:

$$\begin{aligned}&\mathbb{E}^{(1)}[U_k^K(1)] - \sup_{n_A \in I_\varepsilon^K} \mathbb{E}^{(1)}_{(n_A,l)}[U^K_{n_A,l,k}(1) | N_a^K(\tau_1^K) = l-1] \mathbb{E}^{(1)}[U_l^K(1)]\\ &\leqslant \sum_{m<\infty} \mathbb{P}^{(1)}(m < R(i,2), N_a^K(\tau_m^K) = k, N_a^K(\tau_{m+1}^K) = k+1\\ &\qquad\qquad \mid R(i,2) > R(i,1), N_a(\tau^K_{R(i,2)}) = l) \;\leqslant\; \mathbb{E}^{(1)}[U_k^K(1)].\end{aligned}$$

Considering the difference between the two bounds within equation (3.84) yields

$$\sum_{k=1}^{l-1} \frac{r_1}{k+1} \sup_{n_A \in I_\varepsilon^K} \mathbb{E}^{(1)}_{(n_A,l)}[U^K_{n_A,l,k}(1) | N_a^K(\tau_1^K) = l-1] \mathbb{E}^{(1)}[U_l^K(1)] \leqslant cr_1 \sum_{k=1}^{l-1} \frac{\mu_\varepsilon^{l-k}}{k+1} \leqslant \frac{c}{\log K},$$

for a finite c by (3.49) and Lemma 3.13 and thus we can work with $\mathbb{E}^{(1)}(U_k^K(1))$ as an approximation for the sum.

Define for $m \geqslant 1$,

$$\theta^{(r_1)}(m) := \mathbf{1}_{m<J^K(1)}\mathbf{1}_{\{N_a^K(\tau^K_{m+1})-N_a^K(\tau^K_m)=1\}}p^{(r_1)}_{aA}(N^K(\tau^K_m))$$

We again focus on the study of $\mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\zeta,i))$ which equals $\mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\zeta,i) \mid N_a^K(\tau_1^K) = k+1)$ up to r_1. By conditioning on the jump process we get for $n_A \in I_\varepsilon^K$ and $l < \lfloor \varepsilon K \rfloor$,

$$\prod_{m=0}^{\zeta_l^K(1)} (1-\theta^{(r_1)}(m)) \leqslant \mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\zeta,i) \mid \mathcal{N}^{(1)}), \tag{3.85}$$

and the same expression with σ replacing ζ. We define the corresponding parameters for the Poisson approximation as follows:

$$\eta^{(1),-}_{k,l} := \mathbf{1}_{\{N_a^K(0)=k\}} \sum_{m=1}^{\sigma_l^K(1)} \theta^{(r_1)}(m), \quad \text{and} \quad \eta^{(1),+}_{k,l} := \mathbf{1}_{\{N_a^K(0)=k\}} \sum_{m=1}^{\zeta_l^K(1)-1} \theta^{(r_1)}(m).$$

We will show that both can actually be approximated by:

$$\tilde{\eta}^{(1)}_{k,l} := \mathbf{1}_{\{N_a^K(0)=k\}} \sum_{m=1}^{\zeta_l^K(1)-1} \theta^{(r_1)}(m)\mathbf{1}_{\{k\leqslant N_a^K(\tau_m^K)\leqslant l\}}. \tag{3.86}$$

Recall Definitions (3.46), (3.52) and (3.55). On the one hand, for $n_A \in I_\varepsilon^K$ and $k < \lfloor \varepsilon K \rfloor$,

$$\begin{aligned}
\mathbb{E}^{(1)}_{(n_A,k)}[\eta^{(1),+}_{k,l} - \tilde{\eta}^{(1)}_{k,l}] &= \mathbb{E}^{(1)}\Big[\sum_{m=1}^{\zeta_l^K(1)-1} \theta^{(r_1)}(m)(\mathbf{1}_{\{N_a^K(\tau_m^K)<k\}} + \mathbf{1}_{\{N_a^K(\tau_m^K)>l\}})\Big] \\
&\leqslant \sum_{j=1}^{k-1} \sup_{n_A\in I_\varepsilon^K} p^{(r_1)}_{aA}(n_A,j)\mathbb{E}^{(1)}[D_k^K(1)] \sup_{n_A\in I_\varepsilon^K} \mathbb{E}^{(1)}_{(n_A,k)}[U^K_{n_A,k,j}|N_a^K(\tau_1^K)=k-1] \\
&\qquad + \sum_{j=l+1}^{\lfloor \varepsilon K\rfloor} \sup_{n_A\in I_\varepsilon^K} p^{(r_1)}_{aA}(n_A,j)\mathbb{E}^{(1)}[U_l^K(1)] \sup_{n_A\in I_\varepsilon^K} \mathbb{E}^{(1)}_{(n_A,l)}[U^K_{n_A,l,j}|N_a^K(\tau_1^K)=l+1],
\end{aligned}$$

where we used that in the first phase, under $\mathbb{P}^{(1)}$, the number of excursions below k (resp. above l) is equal to $D_k^K(1)$ (resp. $U_l^K(1)-1$). Applying the inequalities (3.49), (3.50), Lemma 3.13, and Equation (3.63), we get the existence of a finite c such that for ε small enough:

$$\mathbb{E}^{(1)}_{(n_A,k)}\left(\eta^{(1),+}_{k,l} - \tilde{\eta}^{(1)}_{k,l}\right) \leqslant cr_1 \sum_{j=1}^{\lfloor \varepsilon K\rfloor} \frac{\mu_\varepsilon^{|j-l|}}{j+1} \leqslant \frac{c}{\log K},$$

as $\mu_\varepsilon \in (0,1)$ for ε small enough and by (3.2). On the other hand,

$$\begin{aligned}\mathbb{E}^{(1)}_{(n_A,k)}[|\eta^{(1),-}_{k,l} - \tilde{\eta}^{(1)}_{k,l}|] &\leqslant \mathbb{E}^{(1)}\Big[\sum_{m=1}^{\sigma^K_l(1)} \theta^{(r_1)}(m)\mathbf{1}_{\{N^K_a(\tau^K_m)<k\}} + \sum_{m=\sigma^K_l(1)+1}^{\zeta^K_l(1)-1} \theta^{(r_1)}(m)\mathbf{1}_{\{k\leqslant N^K_a(\tau^K_m)\leqslant l\}}\Big] \\ &\leqslant r_1\Big[\mathbb{E}^{(1)}[D^K_k(1)]\sum_{j=1}^{k-1}\frac{\sup_{n_A\in I^K_\varepsilon}\ \mathbb{E}^{(1)}_{(n_A,k)}[U^K_{n_A,k,j}|N^K_a(\tau^K_1)=k-1]}{j} \\ &\qquad + \mathbb{E}^{(1)}[D^K_l(1)]\sum_{j=k}^{l-1}\frac{\sup_{n_A\in I^K_\varepsilon}\ \mathbb{E}^{(1)}_{(n_A,l)}[U^K_{n_A,l,j}|N^K_a(\tau^K_1)=l-1]}{j}\Big] \\ &\leqslant cr_1\Big(\sum_{j=1}^{k-1}\frac{\mu_\varepsilon^{k-j}}{j+1} + \sum_{j=k}^{l-1}\frac{\mu_\varepsilon^{l-j}}{j+1}\Big) \leqslant \frac{c}{\log K}.\end{aligned}$$

This shows that both bounds in (3.85) coincide up to terms of small order and thus we only need to calculate the approximation of one product by using $\tilde{\eta}^{(1)}_{k,l}$ for the Poisson approximation. From (3.75) we deduce that this approximation holds true up to terms of order $1/\log^2 K$. Recalling once again the recipe described on page 65, we see that it only remains to calculate the expected value of $\tilde{\eta}^{(1)}_{k,l}$ and to bound its variance. The expectation can be bounded from below and above in the same way as the expected value of $\tilde{\eta}^{(12)}_l$ from the previous part in (3.78) and (3.79):

$$(1-c\varepsilon)\frac{r_1}{s}\log\Big(\frac{l-1}{k}\Big) - \frac{c}{\log K} \leqslant \mathbb{E}^{(1)}_{(n_A,k)}\Big[\tilde{\eta}^{(1)}_{k,l}\Big] \leqslant (1+c\varepsilon)\frac{r_1}{s}\log\Big(\frac{l-1}{k}\Big) + \frac{c}{\log K}. \tag{3.87}$$

A comparison of the definitions of $\tilde{\eta}^{(1)}_{k,l}$ in (3.86) and $\tilde{\eta}^{(12)}_k$ in (3.76) shows that the variance of $\tilde{\eta}^{(1)}_{k,l}$ can be bounded by the same expression, that is, a constant times ε. All in all, this gives us for K large enough, ε small enough, $k<l<\lfloor \varepsilon K\rfloor$ and $n_A \in I^K_\varepsilon$:

$$\Big|\mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\zeta,i) - \exp\Big(-\frac{r_1}{s}\log\Big(\frac{l-1}{k}\Big)\Big)\Big| \leqslant c\sqrt{\varepsilon}, \tag{3.88}$$

and the same inequality with σ instead of ζ. In conclusion, recalling (3.84), and using (3.63) and (3.88) we get the existence of a finite c such that for K large enough and ε small enough,

$$\begin{aligned}\Big|\mathbb{P}^{(1)}(R(i,1)\geqslant 0 \mid R(i,2)>R(i,1), N_a(\tau^K_{R(i,2)})=l) \\ - \sum_{k=1}^{l-1}\frac{r_1}{k+1}\exp\Big(-\frac{r_1}{s}\log\Big(\frac{l-1}{k}\Big)\Big)\mathbb{E}^{(1)}[U^K_k(1)]\Big| \leqslant c\sqrt{\varepsilon}.\end{aligned}$$

Applying (3.49) and Lemma A.1 yields Equation (3.71) and hence ends the proof of Lemma 3.23. □

3.5.3 Proof of Proposition 3.6

With similar ideas we can calculate the probability of the event that the two neutral loci recombine together into the A-population and then separate within the wild-type population.

Proposition 3.24. *Let i be an individual sampled uniformly at the end of the first phase. There exist two finite constants c and ε_0 such that for $\varepsilon \leqslant \varepsilon_0$,*

$$\limsup_{K\to\infty} \Big| \mathbb{P}^{(1)}([12,2]^{rec}_{A,i}) - r_1 \Big[\frac{1 - e^{-\frac{r_1+r_2}{s}\log\lfloor \varepsilon K\rfloor}}{r_1+r_2} + \frac{e^{-\frac{r_1+r_2}{s}\log\lfloor \varepsilon K\rfloor} - e^{-\frac{f_A r_2}{f_a s}\log\lfloor \varepsilon K\rfloor}}{r_1+r_2-f_A r_2/f_a} \Big] \Big| \leqslant c\sqrt{\varepsilon}.$$

This result together with Proposition 3.22 gives us the asymptotic probability of the relation $R1|2(i)^{(1,\mathrm{G1})} = [2,1]^{rec,1}_{A,i} \cup [12,2]^{rec,1}_{A,i}$ as defined in (3.35). The remaining statements of Proposition 3.6 which concern the events $R2(i)^{(1)}$ and $R12(i)^{(1)}$ follow with very similar calculations, realizing that one the one hand

$$\mathbb{P}^{(1)}(R2(i)^{(1)}) = \mathbb{P}^{(1)}(R(i,2) \geqslant 0, R(i,1) = -\infty) + \mathbb{P}^{(1)}(aAa),$$

with aAa as defined on page 136, and on the other hand

$$\begin{aligned}
&\mathbb{P}^{(1)}(R12(i)^{(1)}) = \mathbb{P}^{(1)}(R(i,1) = R(i,2) \geqslant 0, (i,12)^{(1)}_A) + \mathbb{P}^{(1)}(aAa \cup CA) \\
&= \sum_{l=1}^{\lfloor \varepsilon K\rfloor} \Big(1 - \mathbb{P}^{(1)}(R1|2(i)^{(1,\mathrm{G1})} \mid R(i,1) = R(i,2) \geqslant 0, N_a(\tau^K_{R(i,2)}) = l)\Big) \\
&\qquad \cdot \mathbb{P}^{(1)}(R(i,1) = R(i,2) \geqslant 0, N_a(\tau^K_{R(i,2)}) = l) + \mathbb{P}^{(1)}(NE(3)),
\end{aligned}$$

and all needed probabilities can be calculated as in the previous section, hence we give no further details here.

3.5.4 The genealogy of the two neutral loci in the second and third phase

We will here briefly describe the ideas of the proofs concerning the neutral genealogy of an individual during the second and third phase. The following proof of Proposition 3.8 was performed by Charline Smadi. For the sake of completeness and to emphasize the different nature of the second phase in contrast to the remainder of the sweep we will recite it at this point.

Proof of Proposition 3.8

Recall the event C^K_ε from (3.60) and note that for any jump time during the second phase, that is, for any τ^K_m with $T^K_\varepsilon \leqslant \tau^K_m \leqslant T^K_\varepsilon + t_\varepsilon$, and any $j \in \{1,2\}$ we can deduce from Lemmas

3.18 and 3.17 that

$$p_{aA}^{(r_j)}(N^K(\tau_m^K)) \leqslant \frac{8r_j}{\varepsilon^2 K} \quad \text{and} \quad p_{aa}^{(c_j)}(N^K(\tau_m^K)) \leqslant \frac{32}{\varepsilon^4 K^2}. \tag{3.89}$$

We will detail the derivation for the event $NR(i)^{(2)}$ as defined on page 127. The statement concerning the absence of any coalescent event during the second phase follows in the same way, using the above bound on the coalescence probability. For $m \in \mathbb{N}$ we get from (3.89)

$$\mathbb{P}^{(1)}(NR(i)^{(2)} \mid U^K(2) = m, C_\varepsilon^K) \geqslant \Big(1 - \frac{8(r_1+r_2)}{\varepsilon^2 K}\Big)^m. \tag{3.90}$$

If K is large enough we have $\log(1 - 8(r_1+r_2)/(\varepsilon^2 K)) \geqslant -10(r_1+r_2)/(\varepsilon^2 K)$ and hence

$$\begin{aligned}\mathbb{P}^{(1)}(NR(i)^{(2)} \mid C_\varepsilon^K) &\geqslant \Big(1 - \mathbb{P}^{(1)}(U^K(2) > K\log\log K \mid C_\varepsilon^K)\Big) e^{K\log\log K \log(1-\frac{8(r_1+r_2)}{\varepsilon^2 K})} \\ &\geqslant \Big(1 - \mathbb{P}^{(1)}(U^K(2) > K\log\log K \mid C_\varepsilon^K)\Big) e^{-\frac{10(r_1+r_2)K\log\log K}{\varepsilon^2 K}},\end{aligned}$$

where we use the bound on the recombination probabilities given in (3.2). According to (3.24), N_a^K is smaller than $2\bar{n}_a K$ on the time interval $[T_\varepsilon^K, T_\varepsilon^K + t_\varepsilon]$ with probability close to 1. In this case, we can bound the total number of births $U^K(2)$, as defined in (3.59), by the sum of $2\bar{n}_a K$ independent and identically distributed Poisson random variables with parameter $f_a t_\varepsilon$. The strong law of large numbers then yields

$$\lim_{K\to\infty} \mathbb{P}^{(1)}(U^K(2) > K\log\log K \mid C_\varepsilon^K) = 0.$$

We apply again (3.24) in order to get

$$\lim_{K\to\infty} \mathbb{P}^{(1)}(C_\varepsilon^K) = 1, \quad \text{and hence we have} \quad \lim_{K\to\infty} \mathbb{P}^{(1)}(NR(i)^{(2)}) = 1.$$

Proof of Proposition 3.9

The proof of the asymptotic probability of $R2(i)^{(3,\mathrm{G1})}$ is similar to the previous calculations in the proof of Proposition 2.14, hence we refrain from giving any details. Note however that it extensively uses Lemma 3.16 for the approximate value of the number of up-jumps during the third phase and the bound on the variance when performing the Poisson approximation. Similar as in the proof of Proposition 2.14, the derivations leading to a sufficient bound on the variance are more evolved than the calculations in Section 3.5.2. Further, all calculations are performed with respect to the measure $\mathbb{P}^{(3)}$, defined in (3.31).
The proofs for showing that in the limit of $K \to \infty$ we indeed have no recombination into the A-population and no coalescence of any two sampled individuals during the third phase follow a similar idea as the above proof of Proposition 3.8 and are hence omitted here.

CHAPTER 4

Some properties of Λ-coalescents in population genetics

In this chapter we present a recursion formula for the expected height of the coalescent tree corresponding to some measure Λ, and will further state a spatial algorithm for the ancestral recombination graph for the more general Λ-coalescent, introduced in Definition 1.11 in Section 1.2. In both sections, we will consider the case of a very simple measure $\Lambda = \delta_a$, with δ_a being the Dirac measure at some $a \in [0, 1]$, as an example.
Before that, we will investigate some properties of the Λ-coalescent which are relevant in this context and in particular lead to a way of constructing the coalescent. We introduce this so-called Poissonian construction by the same approach which was chosen in the notes of Berestycki, [4]. Afterwards we offer an algorithm in R which generates a Λ-coalescent for a given number of individuals.

4.1 Construction of the Λ-coalescent

We start with analyzing the rates of a Λ-coalescent process and define the event of a p-merger to this end.

Definition 4.1 (cf. Definition 3.3 in [4])**.** *Let $\mathcal{P}$ be $\mathcal{P}_{\mathbb{N}}$ or $\mathcal{P}_n$ for some finite $n \in \mathbb{N}$. Define the operation $\star$ as follows: for two partitions $\pi, \pi' \in \mathcal{P}$ the partition $\pi^* = \pi \star \pi'$ is obtained by merging (coalescing) all those blocks of π whose labels are together in one block of π'.*
Further let $\kappa_p \in \mathcal{P}$ denote a random partition with only one block containing more than one element, obtained by tossing a coin with success probability p for each element: in case of success, the element is in the distinguished block, otherwise it is a singleton. Then, for $p \in (0, 1]$ and $\pi \in \mathcal{P}$ we call the operation $\pi \mapsto \pi \star \kappa_p$ a p-merger.

This corresponds to performing a Bernoulli(p)-experiment for each block of the partition π and coagulating all those whose trial had a successful outcome. The nature of κ_p can be understood in the context of Λ-coalescents when recalling that we only allow for one merger at a time. Hence, there cannot be more blocks with more than one element in κ_p, otherwise we could produce simultaneous merging events.

Recall the rates for a Λ-coalescent as stated in (1.5), Definition 1.11, and write

$$\lambda_{b,k} = \int_0^1 p^k (1-p)^{b-k} p^{-2} \Lambda(dp). \tag{4.1}$$

In the above formulation, we can detect two parts hidden in the rates: first, there is some binomial choosing included in $\lambda_{b,k}$: with success probability p we mark k blocks, and the remaining $b-k$ blocks are not marked with probability $1-p$. The missing binomial coefficient addresses the fact that any k-tuple of blocks merges with the above rate which means that the total rate of events where k out of b blocks merge is equal to

$$\lambda_{b,k}^{\#} = \binom{b}{k} \lambda_{b,k}. \tag{4.2}$$

Thus, we recognize the structure of a p-merger.
The second part in the integral from (4.1), the factor $p^{-2}\Lambda(dp)$, then plays the role of the rate at which a merger related to the success probability p happens.
Indeed, if we consider a process where a p-merger happens with rate $\mu(dp)$ then the rates for some k-merger when there are b blocks left equals $\int_0^1 p^k(1-p)^{b-k}\mu(dp)$, due to the above definition of a p-merger. As in [4] we can now argue why $\mu(dp)$ may be written as $p^{-2}\Lambda(dp)$ for some finite measure Λ: the process is only well-defined if the rate at which at least two out of b blocks coalesce is finite. Recalling the definition in (4.2) then leads to the following formulation of this constraint:

$$\int_0^1 \binom{b}{2} p^2 \mu(dp) < \infty.$$

This condition implies that we can define $\mu(dp) = p^{-2}\Lambda(dp)$ for a finite measure Λ.
The above described way of interpreting the rates of a Λ-coalescence leads exactly to the so-called Poissonian construction of the process. It is convenient to assume the following decomposition:

$$\Lambda = \tilde{\Lambda} + a\delta_0, \quad \text{where} \quad \Lambda(\{0\}) = a \quad (\text{or equivalently} \quad \tilde{\Lambda}(\{0\}) = 0). \tag{4.3}$$

Algorithm 4.2 (Corollary 3.1, [4])**.** Let $\mathcal{N}$ be a Poisson point process on $S := (0,1] \times \mathbb{R}_+$ with intensity $p^{-2}\tilde{\Lambda}(dp) \otimes dt$, with $\tilde{\Lambda}$ as in (4.3). Let $(p_i, t_i)_{i \geqslant 1}$ be the points of the corresponding Poisson process. Then the process $\Pi^\Lambda := (\Pi_t^\Lambda)_{t \geqslant 0}$ can be constructed by performing a p_i-merger at time t_i for each point (p_i, t_i) of the Poisson process. In addition, every pair of blocks merges at rate a.

Note that the additional pair coalescence rate only appears if the measure Λ has an atom at zero. We will use this kind of construction when describing the spatial algorithm for an ARG

in Section 4.3. Let us now describe an algorithm with which one can generate a genealogy of a sample whose ancestral lines evolved according to a Λ-coalescent for (three) different types of measures Λ. The actual R-code is stated in the Appendix C.

Algorithm 4.3. The function `lambdacoal <- function(n,Lambdatype,a,alpha)` generates a Λ-coalescent tree for `n` individuals, where Λ is either a Dirac measure at some point `a` (if `Lambdatype = 0`), the uniform measure (if `Lambdatype = 1`), or a Beta-coalescent (if `Lambdatype = 2`) with parameter `alpha`, meaning that $\Lambda(dx)$ is equal to the Beta$(2-\alpha, \alpha)$ distribution.
The output is a list composed of an array `genealogy` where we can read the times of coalescent events, the number of lineages which merge at an event and the labels (numbers in $\{1, \ldots, n\}$) of the lines which merge. We first draw the time of the next merger from an Exponential random variable with rate equal to the total rate of events given the number of lines which are still present in the partition. Then we choose the type of merger according to the rates of the Λ-coalescent and subsequently sample uniformly at random the labels of the lines which will merge at that event. After each coalescent event we relabel the lines which took part in the merger and name them by the smallest index of all lineages which coalesced. The array `genealogy` then fully determines the shape of the coalescent tree. Further, the total branch length is calculated straightforwardly and returned as the list entry `branch.length`. See C.1 for further details.
Two exemplary outputs are shown in Figure 4.1.1.

4.2 The expected height in a Λ-coalescent

As already pointed out in Section 1.3.2, the number of neutral mutations on a coalescent tree depends only on the mutation rate and the branch length. In the infinite sites model, a comparison between individuals concerning the number and locations of mutations on their ancestral lines results in the number of so-called segregating sites: locations in the genome at which the studied individuals differ. This quantity is of interest in particular as the newer genotyping methods provide whole sequences. After an alignment, we can extract exactly the number of segregating sites from the given data. Having this quantity at hand, one can statistically extract probable underlying mutation rates and also the type of the genealogical tree of the population, provided that one has the knowledge of which tree is likely to give which results. To this end, the study of the total branch length and also the height, corresponding to the time to the most recent common ancestor, is essential in order to draw conclusions on the evolutionary behavior of the sequenced population.
In the case of the Kingman coalescent, we can easily derive the expected height and length of the coalescent tree and thus induce the expected number of mutations occurring during

```
> lambdacoal(10,2,0,0.8)
$genealogy
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]     [,11] [,12]
[1,]    1    2    3    4    5    6    7    8    9    10 0.0000000     0
[2,]    1    2    3    4    5    6    7    8    9     1 0.1688833     2
[3,]    1    2    3    4    5    6    7    8    1     1 0.2044172     2
[4,]    1    2    3    4    5    6    7    6    1     1 0.2240150     2
[5,]    1    1    1    1    1    1    1    1    1     1 0.4383743     7

$branch.length
[1] 3.665935
```

(a) Output for $n = 10$, Lambdatype $= 2$ and parameter $\alpha = 0.8$, that is, Λ is distributed $\sim$Beta$(1.2, 0.8)$.

```
> lambdacoal(10,0,0.2,0)
$genealogy
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]      [,11] [,12]
[1,]    1    2    3    4    5    6    7    8    9    10 0.00000000     0
[2,]    1    2    2    4    2    2    7    8    9    10 0.01239973     4
[3,]    1    1    1    4    1    1    7    8    1    10 0.11985310     3
[4,]    1    1    1    4    1    1    7    8    1     8 0.13840286     2
[5,]    1    1    1    4    1    1    7    7    1     7 0.13956449     2
[6,]    1    1    1    1    1    1    7    7    1     7 0.62548314     2
[7,]    1    1    1    1    1    1    1    1    1     1 0.75609175     2

$branch.length
[1] 2.692539
```

(b) Output for $n = 10$, Lambdatype $= 0$ and parameter $a = 0.2$, that is, $\Lambda = \delta_{0.2}$.

Figure 4.1.1: Some examples obtained by the algorithm `lambdacoal`.

the time until the population (or sample) has found its most recent common ancestor. As mentioned in Remark 1.12, the Kingman coalescent is a Λ-coalescent with $\Lambda = \delta_0$. The rates are then easily calculated and, more important, as there are only two-mergers, straightforward calculations lead to an expression of the expectation of the tree height:

$$\mathbb{E}(\mathcal{H}^{\Pi^{K,n}}) = \sum_{j=2}^{n} \frac{2}{j(j-1)} = \sum_{j=2}^{n} \left[\frac{-2}{j} + \frac{2}{j-1} \right] = 2\left(1 - \frac{1}{n}\right).$$

For more general coalescent trees, it is difficult to arrive at analytical statements on both quantities, the height and the length. Here, one often uses the connections of Λ-coalescents to other processes which are better understood: Bertoin and Le Gall [6] and later Birkner et all [9] showed that Beta-coalescents are in fact duals of generalized Fleming-Viot processes (also called Λ-Fleming-Viot processes). Further, Berestycki, Berestycki and Schweinsberg [3] proved an embedding of the Beta-coalescent for $1 < \alpha < 2$ in α-stable continuous-state

branching processes via the corresponding height process. Based on this theory, they derived asymptotic statements on the distribution of the number and clustering of mutations under the infinite sites model.

In this Section here, we will not review or pursue the techniques given through duality. Instead, we go back to the definition of the rates of the Λ-coalescent and derive a recursive formula for the expected height of the coalescent tree. We further give an algorithm in MuPAD in the case of a simple δ_a-measure (where a can be chosen from $(0,1]$ or left unspecified) in order to show the usefulness of such a formula. The code itself is given in the Appendix, Algorithm C.2.

The statement is obtained by expressing the time to the most recent common ancestor through different random variables, where we use ideas from Schweinsberg [35] given in a different context.

Recall the definition of the total rate $\lambda_{b,k}^{\#}$ with which k out of b blocks merge in a Λ-coalescent and define the total rate of events when there are b blocks present through

$$\lambda_b = \sum_{k=2}^{b} \lambda_{b,k}^{\#} \tag{4.4}$$

and the rate with which more than $b-k$ lines merge as

$$\bar{\lambda}_b^k = \sum_{j=b-k+1}^{b} \lambda_{b,j}^{\#}. \tag{4.5}$$

Finally, denote the rate at which the number of blocks decreases when b blocks are present by γ_b. Keep in mind that a merger of k lines into one decreases the number of blocks by $k-1$, hence

$$\gamma_b = \sum_{k=2}^{b} (k-1)\lambda_{b,k}^{\#}. \tag{4.6}$$

The recursive formula then reads as follows:

Theorem 4.4. *Let Λ be a finite measure on $[0,1]$ and consider the Λ-coalescent $\Pi^{[n]}$ for some $n \in \mathbb{N}$ defined through rates as stated in (4.1). Define for $z \in \{2,3,\ldots,n\}, x \in [z-1], y \in [z]$ the functions*

$$g^{(x,y)}(z) = \begin{cases} \binom{z}{z-x+1} \frac{\lambda_{z,z-x+1}}{\bar{\lambda}_z^x} & \text{if } x = y, x \neq z \\ 1 - \binom{z}{z-x+1} \frac{\lambda_{z,z-x+1}}{\bar{\lambda}_z^x} & \text{if } y = z \\ 0 & \text{else,} \end{cases} \tag{4.7}$$

and

$$f^{(x,y)}(z) = \frac{g^{(x,y)}(z)}{\gamma_y}. \tag{4.8}$$

Then the following expression for the expected height of the coalescent $\Pi^{[n]}$ *holds true for* $n \geqslant 3$*:*

$$\begin{aligned}\mathbb{E}(\mathcal{H}^{\Pi^{[n]}}) = \frac{1}{\gamma_n} + \sum_{j=2}^{n-1} \sum_{k_1=j+1}^{n} [f^{(j,j)}(k_1) + f^{(j,k_1)}(k_1)] \Big[\prod_{l_1=j+1}^{k_1-1} g^{(l_1,k_1)}(k_1) \Big] \\ \cdot \Big\{ \sum_{k_2=k_1+1}^{n} g^{(k_1,k_1)}(k_2) \Big[\prod_{l_2=k_1+1}^{k_2-1} g^{(l_2,k_2)}(k_2) \Big] \\ \cdot \Big\{ \sum_{k_3=k_2+1}^{n} g^{(k_2,k_2)}(k_3) \Big[\prod_{l_3=k_2+1}^{k_3-1} g^{(l_3,k_3)}(k_3) \Big] \cdot \Big\{ \dots \\ \cdot \Big\{ \sum_{k_{n-j}=k_{n-j-1}+1}^{n} g^{(k_{n-j-1},k_{n-j-1})}(k_{n-j}) \Big[\prod_{l_{n-j}=k_{n-j-1}+1}^{k_{n-j}-1} g^{(l_{n-j},k_{n-j})}(k_{n-j}) \Big] \Big\} \dots \Big\} \Big\} \Big\},\end{aligned} \tag{4.9}$$

where we use the convention that empty products and empty sums are equal to 1 (which implies that in particular the innermost product equals 1, as $k_{n-j} \leqslant n$ *and* $k_{n-j-1} \geqslant n-1$*, hence* $k_{n-j-1}+1 > k_{n-j}-1$*). For* $n = 2$*, it holds that*

$$\mathbb{E}(\mathcal{H}^{\Pi^{[2]}}) = \frac{1}{\gamma_2} = \frac{1}{\lambda_{2,2}}.$$

Proof. We will first derive the shape of the functions g and f from above and then show how the above expression for the expectation arises.
In addition to the above definitions from (4.1) and (4.4) to (4.6), we use the notation from Definition 1.13 and further let

$$J_k = J_k^{\Pi^{[n]}} = N_{k-1}^{\Pi^{[n]}} - N_k^{\Pi^{[n]}}$$

be the height of the k-th jump. As already mentioned in Section 1.2, a Λ-n-coalescent process does not necessarily take all possible states of block numbers in between n and 1. It is therefore convenient to define the states which are actually visited by the corresponding jump chain as follows (cf. [35], p.6):

$$L_n(j) = \min\{m \geqslant j : |\Pi_t^{[n]}| = m \text{ for some } t \geqslant 0\}.$$

Then $L_n(j) \equiv N_{k-1}$ for all j with $N_k < j \leqslant N_{k-1}$, or, in other words, $L_n(j)$ takes (N_{k-1} −

$N_k) = J_k$-times the value N_{k-1}. In particular,

$$L_n(j) \text{ is either equal to } j \text{ or we have } L_n(j) = L_n(j+1). \tag{4.10}$$

Since we start with the partition Δ_n we have $L_n(n) \equiv n$. Recall the precise Definition 1.11 of a Λ-coalescent and note that

$$\begin{aligned}
&\mathbb{E}(T_k \mid N_{k-1}) = 1/\lambda_{N_{k-1}} \mathbf{1}_{\{N_{k-1}>1\}}, \quad \text{by definition of the total jump rate, and} \\
&\mathbb{E}(J_k \mid N_{k-1}) = \sum_{j=2}^{N_{k-1}} (j-1) \cdot \mathbb{P}(J_k = j-1 \mid N_{k-1}) = \gamma_{N_{k-1}}/\lambda_{N_{k-1}} \mathbf{1}_{\{N_{k-1}>1\}},
\end{aligned}$$

as on $\{N_{k-1} > 1\}$ it holds that

$$\mathbb{P}(J_k = j-1 | N_{k-1}) = \binom{N_{k-1}}{j} \frac{\lambda_{N_{k-1},j}}{\lambda_{N_{k-1}}} \mathbf{1}_{\{N_{k-1}>1\}},$$

by definition of J_k. With this we can rewrite the expected height of the coalescence process as follows (cf. proofs of Lemmas 6 and 7, [35]):

$$\begin{aligned}
&\mathbb{E}(\mathcal{H}^{\Pi^{[n]}}) = \sum_{k=1}^{n-1} \mathbb{E}(T_k) = \sum_{k=1}^{n-1} \mathbb{E}\big[\mathbb{E}(T_k \mid N_{k-1})\big] = \sum_{k=1}^{n-1} \mathbb{E}\Big[\frac{1}{\lambda_{N_{k-1}}} \mathbf{1}_{\{N_{k-1}>1\}}\Big] \\
&= \sum_{k=1}^{n-1} \mathbb{E}\Big[\frac{1}{\gamma_{N_{k-1}}} \mathbb{E}(J_k \mid N_{k-1}) \mathbf{1}_{\{N_{k-1}>1\}}\Big] = \sum_{k=1}^{n-1} \mathbb{E}\Big[\mathbb{E}\Big(\frac{\mathbf{1}_{\{N_{k-1}>1\}}}{\gamma_{N_{k-1}}} \cdot J_k \mid N_{k-1}\Big)\Big] \\
&= \sum_{k=1}^{n-1} \mathbb{E}\Big(\frac{\mathbf{1}_{\{N_{k-1}>1\}}}{\gamma_{N_{k-1}}} J_k\Big) = \sum_{j=2}^{n} \mathbb{E}\Big(\frac{1}{\gamma_{L_n(j)}}\Big), \text{ by definition of } L_n(j) \\
&= \frac{1}{\gamma_n} + \sum_{j=2}^{n-1} \mathbb{E}\Big[\mathbb{E}\Big(\frac{1}{\gamma_{L_n(j)}} \mid L_n(j+1)\Big)\Big] \\
&= \frac{1}{\gamma_n} + \sum_{j=2}^{n-1} \mathbb{E}\Big[\sum_{k=j}^{n} \frac{1}{\gamma_k} \mathbb{P}\Big(L_n(j) = k \mid L_n(j+1)\Big)\Big].
\end{aligned} \tag{4.11}$$

We claim that the following expression is equal to the conditional probability in line (4.11):

$$\begin{aligned}
\mathbb{E}\left(\mathbf{1}_{\{L_n(j)=k\}} \mid L_n(j+1)\right) = \mathbf{1}_{\{k=j\}} &\binom{L_n(j+1)}{L_n(j+1)-j+1} \frac{\lambda_{L_n(j+1),L_n(j+1)-j+1}}{\bar{\lambda}^j_{L_n(j+1)}} \\
&+ \mathbf{1}_{\{L_n(j+1)=k\}} \left[1 - \binom{k}{k-j+1} \frac{\lambda_{k,k-j+1}}{\bar{\lambda}^j_k}\right].
\end{aligned} \tag{4.12}$$

Note that at most one of the above indicator functions can be nonzero, as $L_n(j+1) = k$ implies that $k \geqslant j+1 > j$.

The above equality (4.12) holds since the right-hand side is $\sigma(L_n(j+1))$-measurable and further, we have

$$\mathbb{P}(L_n(j)=j \mid L_n(j+1)=m) = \frac{\binom{m}{m-j+1}\lambda_{m,m-j+1}}{\bar{\lambda}_m^j} = 1-\mathbb{P}(L_n(j)=m \mid L_n(j+1)=m),$$

by the following argument: if $m-j$ of m present blocks merged to one, then the number of blocks would go from m to $m-(m-j)+1=j+1$ and consequently $L_n(j+1)=j+1$. Hence, the event $\{L_n(j+1)=m\}$ tells us that more than $m-j$ blocks will merge when we are in state m. The total rate for such a merger is $\bar{\lambda}_m^j$, as defined in (4.5). A specific jump to j blocks then occurs if we mark any $m-j+1$ of the m present blocks with the measure specific rate over the total rate $\bar{\lambda}_m^j$. By (4.10), $\mathbb{P}(L_n(j)=k \mid L_n(j+1)=m)$ can only be nonzero for $k \in \{j,m\}$.
This implies that for all events $\{L_n(j+1)=m\} \in \sigma(L_n(j+1))$ with $m \geqslant j+1$ it holds that

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\{L_n(j)=k\}}\mathbf{1}_{\{L_n(j+1)=m\}}\right] = \mathbb{P}(L_n(j)=k \text{ and } L_n(j+1)=m)\\
&= \begin{cases} \mathbb{P}(L_n(j)=j \mid L_n(j+1)=m)\mathbb{P}(L_n(j+1)=m) & \text{if } k=j\\ \mathbb{P}(L_n(j)=m \mid L_n(j+1)=m)\mathbb{P}(L_n(j+1)=m) & \text{if } k=m\\ 0 & \text{else}\end{cases}\\
&= \begin{cases} \binom{m}{m-j+1}\frac{\lambda_{m,m-j+1}}{\bar{\lambda}_m^j}\mathbb{P}(L_n(j+1)=m) & \text{if } k=j\\ \left(1-\binom{m}{m-j+1}\frac{\lambda_{m,m-j+1}}{\bar{\lambda}_m^j}\right)\mathbb{P}(L_n(j+1)=m) & \text{if } k=m\\ 0 & \text{else}\end{cases}
\end{aligned}$$

which is exactly the expectation of the product of $\mathbf{1}_{\{L_n(j+1)=m\}}$ with the right-hand side of (4.12):

$$\begin{aligned}
\mathbb{E}\Bigg[\mathbf{1}_{\{L_n(j+1)=m\}} \cdot \Bigg\{ &\mathbf{1}_{\{k=j\}}\binom{L_n(j+1)}{L_n(j+1)-j+1}\frac{\lambda_{L_n(j+1),L_n(j+1)-j+1}}{\bar{\lambda}^j_{L_n(j+1)}}\\
&+\mathbf{1}_{\{L_n(j+1)=k\}}\left[1-\binom{k}{k-j+1}\frac{\lambda_{k,k-j+1}}{\bar{\lambda}_k^j}\right]\Bigg\}\Bigg] = \mathbb{E}\left[\mathbf{1}_{\{L_n(j)=k\}}\mathbf{1}_{\{L_n(j+1)=m\}}\right].
\end{aligned}$$

Together with the abbreviating functions $g^{(x,y)}(z)$ and $f^{(x,y)}(z)$ from (4.7) and (4.8), this can be plugged into the expression for the expectation of $\mathcal{H}^{\Pi^{[n]}}$ from (4.11):

$$\mathbb{E}(\mathcal{H}^{\Pi^{[n]}}) - \frac{1}{\gamma_n} = \sum_{j=2}^{n-1}\mathbb{E}\Big[\sum_{k=j}^{n}\frac{1}{\gamma_k}\mathbf{1}_{\{k=j\}}\binom{L_n(j+1)}{L_n(j+1)-j+1}\frac{\lambda_{L_n(j+1),L_n(j+1)-j+1}}{\bar{\lambda}^j_{L_n(j+1)}}$$

$$
\begin{aligned}
&\quad + \sum_{k=j}^{n} \frac{1}{\gamma_k} \mathbf{1}_{\{L_n(j+1)=k\}} \Big[1 - \binom{k}{k-j+1} \frac{\lambda_{k,k-j+1}}{\bar{\lambda}_k^j} \Big] \Big] \\
&= \sum_{j=2}^{n-1} \mathbb{E}\Big[\frac{1}{\gamma_j} g^{(j,j)}(L_n(j+1)) + \sum_{k=j+1}^{n} \frac{1}{\gamma_k} g^{(j,k)}(L_n(j+1)) \Big] \\
&= \sum_{j=2}^{n-1} \sum_{l_1=j+1}^{n} \Big[f^{(j,j)}(l_1) + \sum_{k=j+1}^{n} f^{(j,k)}(l_1) \Big] \mathbb{P}(L_n(j+1) = l_1) \\
&= \sum_{j=2}^{n-1} \sum_{l_1=j+1}^{n} \Big[f^{(j,j)}(l_1) + f^{(j,l_1)}(l_1) \Big] \mathbb{E}[\mathbf{1}_{\{L_n(j+1)=l_1\}}].
\end{aligned}
$$

The idea is now, to recursively take the term for the highest value of j out of the sum and condition on the value of the next $L_n(j)$. In the above sum, this means that we consider the term for $j = n-1$ separately in order to be able to condition on $L_n(j+2)$ in the expectation. Here, we use that $L_n(n) = 1$ almost surely. This procedure will be continued successively until the highest value of j equals the starting point of the sum, namely 2. We write down the first and an intermediate state (where terms for $j = n-4$ are taken out of the sum) of this method and deduce from that the formula in (4.9).

$$
\begin{aligned}
&\mathbb{E}(\mathcal{H}^{\Pi^{[n]}}) - \frac{1}{\gamma_n} = \sum_{j=2}^{n-1} \sum_{l_1=j+1}^{n} [f^{(j,j)}(l_1) + f^{(j,l_1)}(l_1)] \mathbb{E}[\mathbf{1}_{\{L_n(j+1)=l_1\}}] \\
&= f^{(n-1,n-1)}(n) + f^{(n-1,n)}(n) + \sum_{j=2}^{n-2} \sum_{l_1=j+1}^{n} [f^{(j,j)}(l_1) + f^{(j,l_1)}(l_1)] \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{L_n(j+1)=l_1\}} \,|\, L_n(j+2)]] \\
&= f^{(n-1,n-1)}(n) + f^{(n-1,n)}(n) + \sum_{j=2}^{n-2} \sum_{l_1=j+1}^{n} [f^{(j,j)}(l_1) + f^{(j,l_1)}(l_1)] \\
&\quad \cdot \mathbb{E}[\mathbf{1}_{\{l_1=j+1\}} g^{j+1,j+1}(L_n(j+2)) + \mathbf{1}_{\{L_n(j+2)=l_1\}} g^{j+1,L_n(j+2)}(L_n(j+2))] \\
&= f^{(n-1,n-1)}(n) + f^{(n-1,n)}(n) + \sum_{j=2}^{n-2} \sum_{l_1=j+1}^{n} [f^{(j,j)}(l_1) + f^{(j,l_1)}(l_1)] \\
&\quad \cdot \sum_{l_2=j+2}^{n} \mathbb{P}(L_n(j+2) = l_2) [\mathbf{1}_{\{l_1=j+1\}} g^{(j+1,j+1)}(l_2) + \mathbf{1}_{\{l_2=l_1\}} g^{(j+1,l_2)}(l_2)] \\
&= f^{(n-1,n-1)}(n) + f^{(n-1,n)}(n) \\
&\quad + \sum_{j=2}^{n-2} [f^{(j,j)}(j+1) + f^{(j,j+1)}(j+1)] \sum_{l_2=j+2}^{n} g^{(j+1,j+1)}(l_2) \mathbb{P}(L_n(j+2) = l_2) \\
&\quad + \sum_{j=2}^{n-2} \sum_{l_1=j+2}^{n} [f^{(j,j)}(l_1) + f^{(j,l_1)}(l_1)] g^{(j+1,l_1)}(l_1) \mathbb{P}(L_n(j+2) = l_1) \\
&= \dots \\
&= f^{(n-1,n-1)}(n) + f^{(n-1,n)}(n)
\end{aligned}
$$

$$
\begin{aligned}
&+ \big[f^{(n-2,n-2)}(n-1) + f^{(n-2,n-1)}(n-1)\big] g^{(n-1,n-1)}(n) \\
&+ \big[f^{(n-2,n-2)}(n) + f^{(n-2,n)}(n)\big] g^{(n-1,n)}(n) \\
&+ \big[f^{(n-3,n-3)}(n-2) + f^{(n-3,n-2)}(n-2)\big] \{g^{(n-2,n-2)}(n-1) g^{(n-1,n-1)}(n) \\
&\quad + g^{(n-2,n-2)}(n) g^{(n-1,n)}(n)\} \\
&+ \big[f^{(n-3,n-3)}(n-1) + f^{(n-3,n-1)}(n-1)\big] g^{(n-2,n-1)}(n-1) g^{(n-1,n-1)}(n) \\
&+ \big[f^{(n-3,n-3)}(n) + f^{(n-3,n)}(n)\big] g^{(n-2,n)}(n) g^{(n-1,n)}(n) \\
&+ \big[f^{(n-4,n-4)}(n-3) + f^{(n-4,n-3)}(n-3)\big] \Big\{g^{(n-3,n-3)}(n-2) \{g^{(n-2,n-2)}(n-1) g^{(n-1,n-1)}(n) \\
&\quad + g^{(n-2,n-2)}(n) g^{(n-1,n)}(n)\} + g^{(n-3,n-3)}(n-1) g^{(n-2,n-1)}(n-1) g^{(n-1,n-1)}(n) \\
&\quad + g^{(n-3,n-3)}(n) g^{(n-2,n)}(n) g^{(n-1,n)}(n)\Big\} \\
&+ \big[f^{(n-4,n-4)}(n-2) + f^{(n-4,n-2)}(n-2)\big] g^{(n-3,n-2)}(n-2) \{g^{(n-2,n-2)}(n-1) g^{(n-1,n-1)}(n) \\
&\quad + g^{(n-2,n-2)}(n) g^{(n-1,n)}(n)\} \\
&+ \big[f^{(n-4,n-4)}(n-1) + f^{(n-4,n-1)}(n-1)\big] g^{(n-3,n-1)}(n-1) g^{(n-2,n-1)}(n-1) g^{(n-1,n-1)}(n) \\
&+ \big[f^{(n-4,n-4)}(n) + f^{(n-4,n)}(n)\big] g^{(n-3,n)}(n) g^{(n-2,n)}(n) g^{(n-1,n)}(n) \\
&+ \sum_{j=2}^{n-4} \big[f^{(j,j)}(j+1) + f^{(j,j+1)}(j+1)\big] \\
&\quad \cdot \Big\{g^{(j+1,j+1)}(j+2) g^{(j+2,j+2)}(j+3) \sum_{l_4=j+4}^{n} g^{(j+3,j+3)}(l_4) \mathbb{P}(L_n(j+4) = l_4) \\
&\quad + g^{(j+1,j+1)}(j+2) \sum_{l_3=j+4}^{n} g^{(j+2,j+2)}(l_3) g^{(j+3,l_3)}(l_3) \mathbb{P}(L_n(j+4) = l_3) \\
&\quad + g^{(j+1,j+1)}(j+3) g^{(j+2,j+3)}(j+3) \sum_{l_4=j+4}^{n} g^{(j+3,j+3)}(l_4) \mathbb{P}(L_n(j+4) = l_4) \\
&\quad + \sum_{l_2=j+4}^{n} g^{(j+1,j+1)}(l_2) g^{(j+2,l_2)}(l_2) g^{(j+3,l_2)}(l_2) \mathbb{P}(L_n(j+4) = l_2)\Big\} \\
&+ \sum_{j=2}^{n-4} \big[f^{(j,j)}(j+2) + f^{(j,j+2)}(j+2)\big] g^{(j+1,j+2)}(j+2) \\
&\quad \cdot \Big\{g^{(j+2,j+2)}(j+3) \sum_{l_4=j+4}^{n} g^{(j+3,j+3)}(l_4) \mathbb{P}(L_n(j+4) = l_4) \\
&\quad + \sum_{l_3=j+4}^{n} g^{(j+2,j+2)}(l_3) g^{(j+3,l_3)}(l_3) \mathbb{P}(L_n(j+4) = l_3)\Big\} \\
&+ \sum_{j=2}^{n-4} \big[f^{(j,j)}(j+3) + f^{(j,j+3)}(j+3)\big] g^{(j+1,j+3)}(j+3) g^{(j+2,j+3)}(j+3) \\
&\quad \cdot \sum_{l_4=j+4}^{n} g^{(j+3,j+3)}(l_4) \mathbb{P}(L_n(j+4) = l_4)
\end{aligned}
$$

$$+\sum_{j=2}^{n-4}\sum_{l_1=j+4}^{n}[f^{(j,j)}(l_1)+f^{(j,l_1)}(l_1)]g^{(j+1,l_1)}(l_1)g^{(j+2,l_1)}(l_1)\mathbb{P}(L_n(j+4)=l_1).$$

The statement of the theorem given in (4.9) follows inductively by careful observation. □

The MuPAD Code, which implements the above formula in the case of $\Lambda = \delta_a$, is stated only in C.2 but we will briefly explain the idea behind it at this point. The crucial factor is that we have nested sums whose entries and also starting points depend on the previous value. It is not clear how one can implement this in a straightforward way. Here, we have chosen a procedure which first generates the nested sums and products with the help of placeholders in the form of unspecified arrays: one for the index k_i, one for the values of the function g and one for the product over the respective g. Starting from the sum with index k_{n-j-1} in the j-th step, we substitute one after the other the placeholders for the running variables k_i. In the end, we need to substitute the remaining placeholder with the corresponding expressions of the functions g and f. For details, see the code stated in C.2. Note that the program should also work for more complicated measures if we defined the rates and accordingly the functions g and f differently in the beginning of the program.
With the program we calculated some particular values for $n = 5, 10, 15, 20$, displayed in Figure 4.2.1 below. Further, exemplary results for the expected height $\mathcal{H}^n$ for an unspecified $a \in (0,1)$ and small n obtained through the algorithm have the form

$$\mathbb{E}(\mathcal{H}^4) = \frac{10a^3-40a^2+56a-27}{6a^3-25a^2+36a-18},$$
$$\mathbb{E}(\mathcal{H}^5) = \frac{40a^6-315a^5+1055a^4-1926a^3+2025a^2-1166a+288}{24a^6-190a^5+639a^4-1172a^3+1240a^2-780a+180}.$$

n	$a=0.1$	$a=0.5$	$a=0.8$
5	1.5913	1.5227	1.3573
10	1.8151	1.8353	1.6744
15	1.9024	2.0011	1.8478
20	1.9526	2.1144	1.9640

Figure 4.2.1: Values for the expected height for some a in $\Lambda = \delta_a$.

4.3 A spatial algorithm for the ancestral recombination graph for a measure Λ

The spatial algorithm that we propose here in order to generate an ARG for a Λ-coalescent is very similar to the one by Wiuf and Hein, as stated in Algorithm 1.27. Naturally it differs in

Step 1, as we will generate the coalescent tree for position 0 according to the rates determined by the measure Λ. This will in general also result in non-binary mergers, as seen in Section 4.1. On the contrary, the recombination mechanism works exactly in the same way here, since the probability to simultaneously witnessing a coalescent and a recombination event tends to zero in the scaling limit of large population sizes. This was already reasoned by Hein, Schierup and Wiuf in [23] and shown again within the frame of a diploid model with simultaneous mergers by Birkner, Blath and Eldon, [8].
The essential difference between our algorithm and Algorithm 1.27 therefore lies in the reattachment mechanism of the newly created line. Depending on the concrete nature of the measure Λ, a new line may merge together with any other line present in the graph and thus add a binary merger to it. In addition, it can join any already existing coalescent event according to the rates of that particular Λ-coalescent. We will state the exact details below and then give (and visualize) an explicit example.

Definition 4.5. *We denote by $\mathcal{G}(p)$ the graph which explains the ancestry of all points $x \leqslant p$ on the sampled sequences. For any graph $\mathcal{G}$ let $t^{\mathcal{G}} = (t_1^{\mathcal{G}}, t_2^{\mathcal{G}}, \ldots)$ be the vector of the times of coalescent events in $\mathcal{G}$, with $t_1^{\mathcal{G}} < t_2^{\mathcal{G}} < \ldots$. Further, let $m^{\mathcal{G}} = (m_1^{\mathcal{G}}, m_2^{\mathcal{G}}, \ldots)$, where $m_i^{\mathcal{G}}$ is the number of lines (blocks) which merge at the i-th merger, that is, at time $t_i^{\mathcal{G}}$.*
We denote the number of different lines (blocks) present at time $t-$ in the graph $\mathcal{G}$ by $A_t^{\mathcal{G}}$. By $t-$ we mean that in case of a merger at time t, $A_t^{\mathcal{G}}$ gives the number of blocks right before the merger happens.

Algorithm 4.6. Suppose that the population evolves in the limit according to a Λ-coalescent for some finite measure $\Lambda = \tilde{\Lambda} + a\delta_0$ with $\tilde{\Lambda}(\{0\})$ (cf. Equation (4.3)) and denote the rates corresponding to $\tilde{\Lambda}$ by

$$\tilde{\lambda}(b,k) := \int_0^1 p^k (1-p)^{b-k} p^{-2} \tilde{\Lambda}(dp).$$

Sample finitely many sequences from the population and identify each with the interval $[0, \rho/2]$ with $\rho/2$ the scaled recombination rate (see (1.15)). Then we can perform the following steps in order to generate the ancestral recombination graph:

1. Generate the local coalescent tree $\mathcal{T}(0) = \mathcal{G}(0)$ for position 0 according to the rates defined in (4.1) (using for example the Poissonian construction from Algorithm 4.2 or the program `lambdacoal` as described in Algorithms 4.3, C.1).

Perform the steps 2. to 6. as stated in Algorithm 1.27 in order to choose the next break point on the sequence, $P_i = P_{i-1} + p_i$ with $p_i \sim \mathrm{Exp}(B_{i-1})$, and its position on the current graph G_{i-1} via $t_i \sim \mathcal{U}(0, B_{i-1})$. Then this position t_i corresponds to a time T_i with $t_j^{G_{i-1}} < T_i < t_{j+1}^{G_{i-1}}$ a.s. for some j and $t_j^{G_{i-1}}$ as defined in Definition 4.5.

7. Coalesce the new recombined edge e_i to the graph G_{i-1} according to the following rule:

   draw $T_{\text{bin}} \sim \text{Exp}\Big(A_{T_i}^{G_{i-1}}\Big\{a + \tilde{\lambda}(A_{T_i}^{G_{i-1}} + 1, 2)\Big\}\Big)$, (4.13)

   and merge the new edge e_i with a uniformly chosen line from G_{i-1}

   at time $T_i + T_{\text{bin}}$ if $T_{\text{bin}} < t_{j+1}^{G_{i-1}}$,

   else

   draw $p_{\text{merge}} \sim \text{Bernoulli}\left(\dfrac{\tilde{\lambda}(A_{T_i}^{G_{i-1}} + 1, m_{j+1}^{G_{i-1}} + 1)}{\tilde{\lambda}(A_{T_i}^{G_{i-1}}, m_{j+1}^{G_{i-1}})}\right)$, (4.14)

   and let the new edge e_i join the merger at $t_{j+1}^{G_{i-1}}$ if $p_{\text{merge}} = 1$,

   else

   repeat the above steps with rates adjusted to the number of present lineages

   in the graph G_{i-1}: replace the T_i in $A_{T_i}^{G_{i-1}}$ in (4.13) and (4.14) by $t_{j+1}^{G_{i-1}}+, t_{j+2}^{G_{i-1}}+, ...$

   and consider, if necessary, the next mergers in the Bernoulli experiment.

   As defined in Definition 4.5, $A_{T_i}^{G_{i-1}}$ is the number of lineages present in the graph G_{i-1} right before we add the new edge e_i.

Proof. As reasoned above, we only need to show that the attachment mechanism as described in Step 7 matches the dynamics of the considered Λ-coalescent. We consider therefore separately the rate for an attachment of the new edge e_i to exactly one other lineage in the graph and the probability for joining an existing coalescent event.

1. The attachment of e_i results in a new binary merger in the new graph G_i.

The first part of the according rate results from the factor in front of the δ_0 measure, a. This Kingman component implies that any two lines merge at rate a and thus the edge e_i is attached in a 2-merger to the graph G_{i-1} with rate a times the number of lines present at the time. Recall that in the Kingman coalescent this resulted in a coalescence rate of $\binom{k}{2}$ with k lines present in the tree. Here however, one line, e_i is already fixed as one out of two lineages of the coalescent event. This leads to the rate $a \cdot A_{T_i}^{G_{i-1}}$ instead of $a \cdot \binom{A_{T_i}^{G_{i-1}}+1}{2}$.

As a second part we have to think of those lines which where the only line marked to be part of a p-merger for some p by the Poisson process for the construction of the coalescent process due to a point (p, t). The new line present in the graph can also get marked with that probability p leading to a coalescent event of e_i and the other lineage at time t. To determine the probability for this event, we have to bear in mind that such marking of one line can not be seen when looking at the graph as such an event does not result in a visible merger. Knowing the total graph up to some position x, we thus can not condition on such a "1-merger". Instead, we need to use the rate of a 2-merger when there are $A_{T_i}^{G_{i-1}} + 1$ lineages present: $\tilde{\lambda}(A_{T_i}^{G_{i-1}} + 1, 2)$. As this rate refers to a specific merger, we have to multiply it with

$\binom{A_{T_i}^{G_{i-1}}}{1} = A_{T_i}^{G_{i-1}}$ as well (again, the new line is already chosen as one participating line).
The total rate for such a binary coalescent event of e_i as long as there are $A_{T_i}^{G_{i-1}}$ lines present in the graph therefore equals

$$a \cdot A_{T_i}^{G_{i-1}} + A_{T_i}^{G_{i-1}} \cdot \tilde{\lambda}(A_{T_i}^{G_{i-1}} + 1, 2) = A_{T_i}^{G_{i-1}} \cdot \Big\{ a + \tilde{\lambda}(A_{T_i}^{G_{i-1}} + 1, 2) \Big\},$$

as claimed in (4.13). Clearly, the number of lineages has to be adjusted to the lesser number whenever we encounter a coalescent event prior to the random time T_{bin}. Then we need to draw a new time T_{bin} according to the adjusted exponential distribution and repeat the procedure.

Additionally, we need to consider the multiple merger property of the considered coalescent:
2. Joining a k-merger for $k \geqslant 2$, starting with $k = m_{j+1}^{G_{i-1}}$.
Given that we observe an $m_{j+1}^{G_{i-1}}$-merger out of $A_{t_{j+1}^{G_{i-1}}}^{G_{i-1}}$ individuals at time $t_{j+1}^{G_{i-1}}$, we have to derive the probability with which the new lineage e_i also gets marked and thus joins the other lineages in the coalescent event. Let $\mathfrak{p}$ denote the random variable from the Poisson point process whose outcome is the success probability for the marking procedure. Further, define

$$\mathfrak{p}^{A,k} \sim \mathfrak{p} \mid \text{ a specific } (A,k)\text{-merger happened } .$$

We are now interested in the law of $\mathfrak{p}^{A,k}$.

$$\begin{aligned}
\mathbb{P}(\mathfrak{p}^{A,k} \leqslant c) &= \frac{\mathbb{P}(\mathfrak{p} \leqslant c, \text{ a specific } (A(t),k)\text{-merger happened })}{\mathbb{P}(\text{ a specific } (A(t),k)\text{-merger happened })} \\
&= \frac{\mathbb{P}(\mathfrak{p} \leqslant c, \text{ a specific } (A(t),k)\text{-merger happened, a merger happened })}{\mathbb{P}(\text{ a specific } (A(t),k)\text{-merger happened, a merger happened })} \\
&= \frac{\mathbb{P}(\mathfrak{p} \leqslant c, \text{ a specific } (A(t),k)\text{-merger happened } \mid \text{ a merger happened })}{\mathbb{P}(\text{ a specific } (A(t),k)\text{-merger happened } \mid \text{ a merger happened })} \\
&= \frac{\int_0^c p^k (1-p)^{A(t)-k} p^{-2} \Lambda(\mathrm{d}p)}{\int_0^1 p^k (1-p)^{A(t)-k} p^{-2} \Lambda(\mathrm{d}p)} =: \nu([0,c]),
\end{aligned}$$

where ν denotes the probability measure of $\mathfrak{p}^{A,k}$.
We emphasize here that ν is a measure which is defined through an integral, i.e. $\nu(B) = \int_B f \,\mathrm{d}\Lambda$, $B \subset [0,1]$, where $f(p) = \frac{p^k(1-p)^{A-k}p^{-2}}{\int_0^1 p^k(1-p)^{A-k}p^{-2}\Lambda(\mathrm{d}p)}$ is called the density of ν with respect to Λ (as f is nonnegative on the interval $[0,1]$, the definition as a density makes sense).
Then we know that for any nonnegative h we have the following identity (cf. [7], Thm. 16.11):

$$\int_B h \,\mathrm{d}\nu = \int_B hf \,\mathrm{d}\Lambda \tag{4.15}$$

for any measurable set $B \subset [0,1]$.
Our aim is to calculate the probability that the $(A+1)$-st lineage also gets marked if we have a merger (A,k). For better readability, we introduce the following notation:

$$\mathbb{P}^{A,k}(\cdot) := \mathbb{P}(\cdot \mid \text{a specific } (A,k)\text{-merger happened })$$

The corresponding conditional expectation is denoted analogously.

$$\begin{aligned}
&\mathbb{P}^{A,k}((A+1)\text{-st lineage gets marked}) = \mathbb{E}^{A,k}(\mathbf{1}_{(A+1)\text{-st lineage gets marked}}) \\
&= \mathbb{E}^{A,k}\big[\mathbb{E}^{A,k}(\mathbf{1}_{(A+1)\text{-st lineage gets marked}} \mid \mathfrak{p})\big] \\
&= \mathbb{E}^{A,k}(\mathfrak{p}), \text{ as } \mathfrak{p} \text{ is the success probability} \\
&= \mathbb{E}(\mathfrak{p} \mid \text{ a specific } (A,k)\text{-merger happened }) = \mathbb{E}(\mathfrak{p}^{A,k}).
\end{aligned}$$

As we have already determined the law of $\mathfrak{p}^{A,k}$, we can now easily calculate the expectation with the help of (4.15):

$$\begin{aligned}
\mathbb{E}(\mathfrak{p}^{A,k}) &= \int_{[0,1]} p \,\mathrm{d}\nu(p) = \int_{[0,1]} p \cdot \frac{p^k(1-p)^{A-k}p^{-2}}{\int_0^1 p^k(1-p)^{A-k}p^{-2}\Lambda(\,\mathrm{d}p)} \,\mathrm{d}\Lambda(p), \qquad \text{by (4.15)} \\
&= \frac{\int_0^1 p^{k-1}(1-p)^{A-k}\Lambda(\,\mathrm{d}p)}{\int_0^1 p^{k-2}(1-p)^{A-k}\Lambda(\,\mathrm{d}p)} = \frac{\lambda_{A+1,k+1}}{\lambda_{A,k}},
\end{aligned}$$

which actually corresponds to our intuition. Inserting the current number of lines $(A^{G_{i-1}}_{t^{G_{i-1}}_{j+1}})$ present in the graph for A and replacing k by the number of lines which merged $(m^{G_{i-1}}_{j+1})$, the claim from (4.14) follows.
When following the ancestral lines backwards in time, we have to draw a new time T_{bin} each time the number of lines changes. In addition, as long as the new edge e_i has not coalesced with the graph, we have to draw a new p_{merge} for each merger we encounter. This explains the algorithm. □

We will now give an example including pictures of all intermediate graphs.

Example 4.7. We consider the Bolthausen-Sznitman coalescent for $\Lambda(dx) = dx$, that is, the uniform measure on $[0,1]$, and take a sample of $n = 5$ individuals, each represented by a sequence of length $[0,1]$, and length is again measured in expected number of recombinations.
Step 1 and 2. With `lambdacoal` we obtain a local tree for position 0 and its total branch length, see Figure 4.3.1. The times of coalescent events are then given in the last column of the output of `lambdacoal`, $t^{\mathcal{T}(0)} = (0.591, 0.783, 1.003)$, with corresponding merger types $m^{\mathcal{T}(0)} = (3,2,2)$.
Step 4. Draw (here executed with R) from $\sim \mathrm{Exp}(3.971)$ in order to get the first recombination

```
> lambdacoal(5,1,0,0)
$genealogy
     [,1] [,2] [,3] [,4] [,5]      [,6] [,7]
[1,]    1    2    3    4    5 0.0000000    0
[2,]    1    2    3    2    2 0.5910379    3
[3,]    1    2    2    2    2 0.7825443    2
[4,]    1    1    1    1    1 1.0031783    2

$branch.length
[1] 3.970977
```

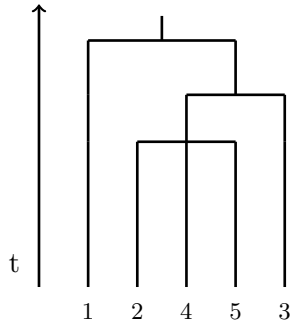


Figure 4.3.1: Step 1: local tree for position 0.

point on the sequence (if it exists): $p_1 = P_1 = 0.220$.

Step 5. Draw the location of the recombination event on the tree from $\sim \mathcal{U}(0, 3.971)$: $t_1 = 2.776$. If we walk through the tree from the bottom-left to the right, $2.776 = 1.003 + 0.591 + 0.783 + 0.399$ then corresponds to the line describing the fifth individual, at the time $T_1 = 0.399$ (note that we assume the individuals to be exchangeable here so it does not matter in which way the individuals are displayed in the graph; further, as we draw from a uniform distribution the pattern with which we measure the branch length is not important).

Step 6. Split the corresponding line and thus create a new (floating) edge e_1 (see Figure 4.3.2).

Step 7. Determine the time and type of the reattachment of e_1: first, note that Λ has no atom at zero, hence $a = 0$ in (4.13). At time $T_1 < 0.591$, there are $A_{T_1}^{\mathcal{T}(0)} = 5$ lineages present in the graph and the needed rate $\lambda(6, 2)$ is easily calculated: $\lambda(6, 2) = 1/5$. Consequently, we draw from $\sim \text{Exp}(1)$: $T_{bin} = 0.634$, which means $T_1 + T_{bin} > t_1^{\mathcal{T}(0)} = 0.591$ and therefore, it is not used. Next, we need to check if e_1 joins the merger at time $t_1^{\mathcal{T}(0)}$. We have $\lambda(6, 4) = 1/30, \lambda(5, 3) = 1/12$ and hence need to draw from $\sim$ Bernoulli(6/15): $p_{merge} = 1$ and thus e_1 takes part in this merging event. See Figure 4.3.2 for a visualization of the new graph G_1. The new branch length is $B_1 = B_0 + (0.591 - 0.399) = 4.163$.

We repeat the steps 4 to 8 of the algorithm.

Searching for the next recombination point on the sequences: $p_2 \sim \text{Exp}(4.163)$ gives us $p_2 = 0.145$ and hence $P_2 = 0.220 + 0.145 = 0.365 < 1$. Sampling the position from $\sim \mathcal{U}(0, 4.163)$ results in $t_2 = 0.729$. This corresponds to the lineage of the first individual and $T_2 = t_2$. We create the second new line e_2 by splitting the line at T_2, when there are $A_{T_2}^{G_1} = 3$ lineages present in the graph. As $\lambda(4, 2) = 1/3$, we draw again from $\sim \text{Exp}(1)$: $T_{bin} = 1.024$ and again $T_2 + T_{bin} > t_2^{G_1}$, hence, we dismiss this time. Check now if e_2 joins the merger of $m_2^{G_1} = 2$ lines at time $t_2^{G_1}$: we calculate $\lambda(4, 3) = 1/6$ and $\lambda(3, 2) = 1/2$ and thus draw from $\sim$ Bernoulli(1/3): $p_{merge} = 0$, meaning that e_2 does not join this merger. The new line

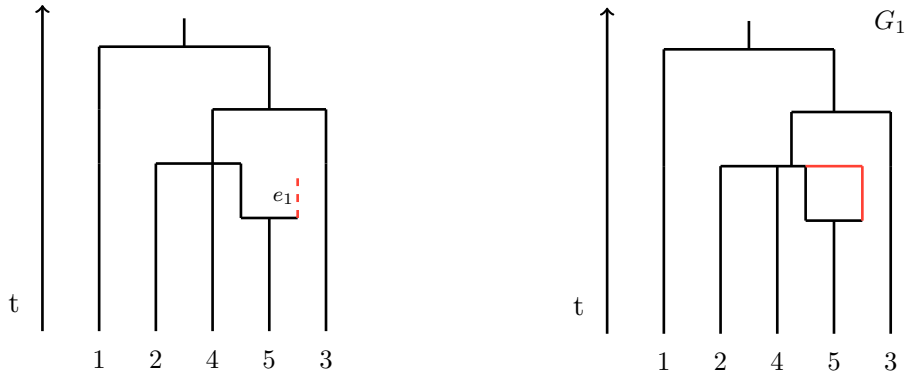


Figure 4.3.2: Step 6 and 7: generate the graph G_1.

will not join the graph before time $t_2^{G_1}$. We draw the next time for a binary merger while there are two other lineages present in the graph: $T_{bin} \sim \text{Exp}(1)$, with outcome $T_{bin} = 0.439$. Again, $T_2 + T_{bin} > t_3^{G_1}$ and T_{bin} is not used. Check if e_2 joins the 2-merger at time $t_3^{G_1}$: $\lambda(3,3) = 1/2$, $\lambda(2,2) = 1$, thus, draw from $\sim$ Bernoulli(1/2). Here, $p_{merge} = 0$ and hence we need to continue with drawing the time it takes the new line to coalesce with the root: $T_{bin} \sim \text{Exp}(1)$, with outcome $T_{bin} = 0.206$. See Figure 4.3.3 for the shape of the new graph G_2 with branch length $B_2 = B_1 + 0.274 + 2 \cdot 0.206 = 4.849$.

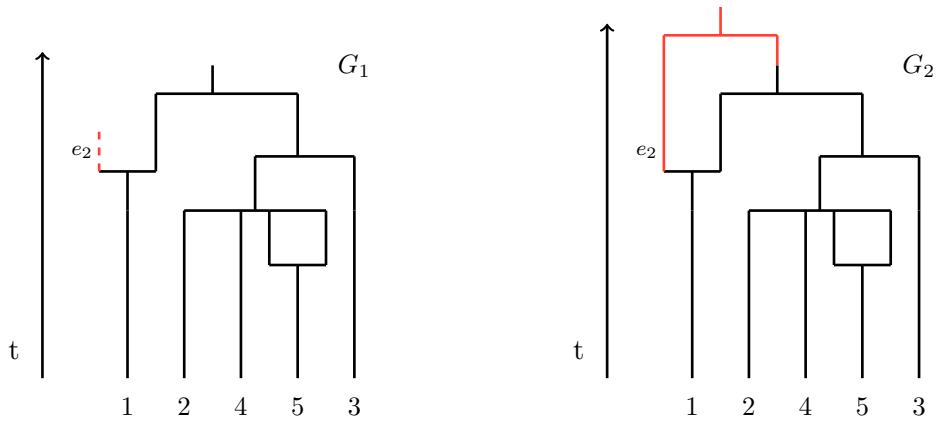


Figure 4.3.3: Step 6 and 7: generate the graph G_2.

Again, repeat the steps of the algorithm by searching for the next recombination point.
$p_3 \sim \text{Exp}(4.849)$ gives us $p_3 = 0.413$ and hence $P_3 = 0.365 + 0.413 = 0.778 < 1$. Sampling the position from $\sim \mathcal{U}(0, 4.849)$ results in $t_3 = 3.970$, which corresponds to a position on the rightmost line at time $T_3 = 0.879$ with $t_3^{G_2} < T_3 < t_4^{G_2}$. Create the third new line e_3 by splitting the line at T_3, when there are $A_{T_3}^{G_1} = 3$ lineages present in the graph. Sampling from $\sim \text{Exp}(1)$ gives us $T_{bin} = 0.103$ and in this case, $T_{bin} + T_3 = 0.982 < t_4^{G_2}$, hence we coalesce e_3

with one of the three present lines at time 0.982 chosen uniformly at random: we sample line 1 to be the partner in the binary coalescent event. See Figure 4.3.4 for the resulting graph G_3 of total branch length $B_3 = B_2 + 0.103 = 4.952$.

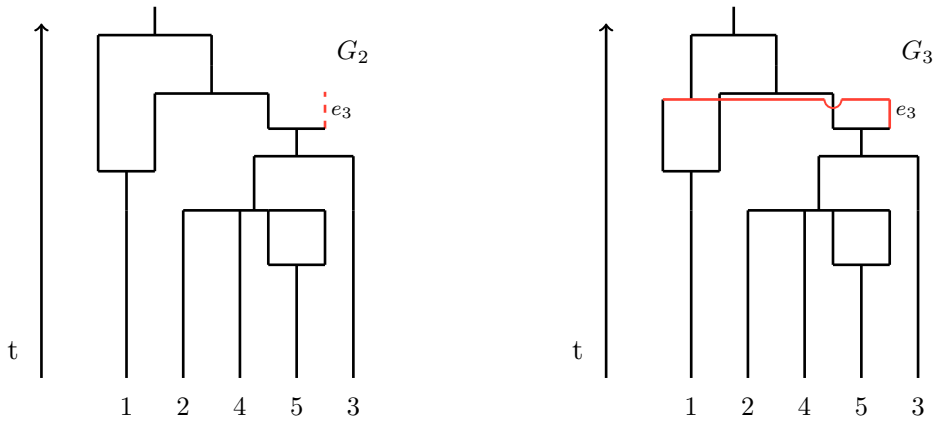


Figure 4.3.4: Step 6 and 7: generate the graph G_3.

A random draw from $\sim \text{Exp}(4.849)$ yields $p_4 = 0.388$ and hence $P_4 = 0.778 + 0.388 = 1.166 > 1$ and the algorithm stops as the alleged recombination exceeds the sequence length. The graph G_3 in Figure 4.3.5 therefore is the resulting ancestral recombination graph G^{ARG} which explains the ancestry of the five sequences $[0, 1]$. The vectors of coalescent times and mergers are $t^{G^{ARG}} = (0.591, 0.783, 0.982, 1.003, 1.209)$, and $m^{G^{ARG}} = (4, 2, 2, 2, 2)$, respectively.

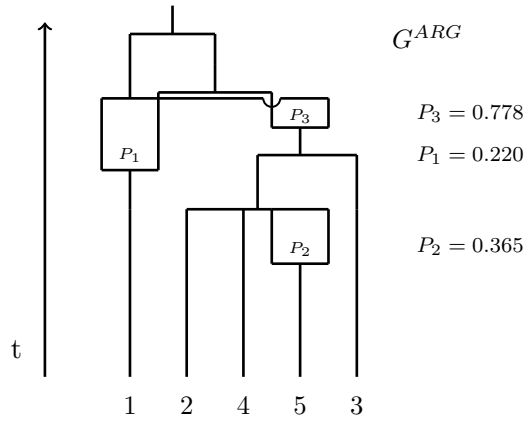


Figure 4.3.5: The ancestral recombination graph.

♢

APPENDIX A

Appendix of Chapter 2

We will first restate Lemma 3.5 from [36] as it is used quite often during the evaluation of error terms.

Lemma A.1 (Lemma 3.5, [36]). *If $a > 1$, there is a constant C depending on a but not on N so that*

$$\sum_{j=1}^{N} \frac{a^j}{j} \leqslant C \cdot \frac{a^N}{N}.$$

We refer to an idea of the proof to [36] and continue here with proving Lemma 2.30.

Proof of Lemma 2.30. The proof consists of two similar steps where the sums in the expression on the left-hand side are approximated via Riemann integrals. We start with the sum within the exponential. Note for both parts of the proof that the function $1/x^\alpha$ is monotonically decreasing for positive x, α.
1. Show that

$$\sum_{k=l+1}^{M} \frac{1}{k} = \int_{l+1}^{M} \frac{1}{x} dx + \mathcal{O}\Big(\tfrac{1}{l+1}\Big).$$

From the properties of the lower and upper sum we get the following inequality:

$$\begin{aligned} & \sum_{k=l+2}^{M} \frac{1}{k} \leqslant \int_{l+1}^{M} \frac{1}{x} dx \leqslant \sum_{k=l+1}^{M-1} \frac{1}{k} \\ \Leftrightarrow \quad & -\frac{1}{l+1} \leqslant \int_{l+1}^{M} \frac{1}{x} dx - \sum_{k=l+1}^{M} \frac{1}{k} \leqslant -\frac{1}{M} \\ \Rightarrow \quad & \Big| \int_{l+1}^{M} \frac{1}{x} dx - \sum_{k=l+1}^{M} \frac{1}{k} \Big| \leqslant \frac{1}{l+1}. \end{aligned}$$

The value of the integral can be easily calculated and we get

$$\begin{aligned}
&\sum_{k=l+1}^{M} \frac{1}{k} = \log(M) - \log(l+1) + \mathcal{O}\Big(\tfrac{1}{l+1}\Big) = \log\Big(\frac{M}{l+1}\Big) + \mathcal{O}\Big(\tfrac{1}{l+1}\Big) \\
\Rightarrow \quad &\exp\Big(-\frac{r}{s}\sum_{k=l+1}^{M}\frac{1}{k}\Big) = \exp\Big(-\frac{r}{s}\log\Big(\frac{M}{l+1}\Big)\Big)\cdot\exp\Big(-\frac{r}{s}\mathcal{O}\Big(\tfrac{1}{l+1}\Big)\Big) \\
&\qquad = \exp\Big(-\frac{r}{s}\log\Big(\frac{M}{l+1}\Big)\Big) + \mathcal{O}\Big(\tfrac{1}{(l+1)\log(N)}\Big) \\
&\qquad = \Big(\frac{M}{l+1}\Big)^{-\frac{r}{s}} + \mathcal{O}\Big(\tfrac{1}{(l+1)\log(N)}\Big). \qquad \text{(A.1)}
\end{aligned}$$

2. Show that

$$\sum_{l=1}^{M}(l+1)^{\frac{r}{s}-1} = \int_2^M x^{\frac{r}{s}-1}dx + \mathcal{O}(1).$$

Similar as above we can bound the integral by lower and upper sum as follows:

$$\begin{aligned}
&\sum_{l=3}^{M} l^{\frac{r}{s}-1} \leqslant \int_2^M x^{\frac{r}{s}-1}dx \leqslant \sum_{l=2}^{M-1} l^{\frac{r}{s}-1} \\
\Leftrightarrow \quad &-2^{\frac{r}{s}-1} - (M+1)^{\frac{r}{s}-1} \leqslant \int_2^M x^{\frac{r}{s}-1}dx - \sum_{l=2}^{M+1} l^{\frac{r}{s}-1} \leqslant -M^{\frac{r}{s}-1} - (M+1)^{\frac{r}{s}-1} \\
\Rightarrow \quad &\Big|\int_2^M x^{\frac{r}{s}-1}dx - \sum_{l=1}^{M}(l+1)^{\frac{r}{s}-1}\Big| \leqslant C,
\end{aligned}$$

for some constant C, as

$$k^{\frac{r}{s}-1} = \frac{1}{k}\exp\Big(\frac{r}{s}\log(k)\Big) \leqslant \frac{c}{k} \leqslant c, \quad \text{for all } k = 1, 2, \ldots, 2N, c \text{ a constant.}$$

The integral is easy to calculate,

$$\int_2^M x^{\frac{r}{s}-1}dx = \Big[\frac{s}{r}x^{\frac{r}{s}}\Big]_2^M = \frac{s}{r}\Big[M^{\frac{r}{s}} - 2^{\frac{r}{s}}\Big],$$

and putting (A.1) from step 1 and the above together results in the statement of the Lemma:

$$\begin{aligned}
&\frac{r}{s}\sum_{l=1}^{M}\frac{1}{l+1}\exp\Big(-\frac{r}{s}\sum_{k=l+1}^{M}\frac{1}{k}\Big) = \frac{r}{s}\sum_{l=1}^{M}\frac{1}{l+1}\Big[\Big(\frac{l+1}{M}\Big)^{\frac{r}{s}} + \mathcal{O}\Big(\tfrac{1}{(l+1)\log(N)}\Big)\Big] \\
&= \frac{r}{s}\frac{1}{M^{\frac{r}{s}}}\sum_{l=1}^{M}(l+1)^{\frac{r}{s}-1} + \mathcal{O}\Big(\tfrac{1}{(\log N)^2}\Big) \\
&= \frac{r}{s}\frac{1}{M^{\frac{r}{s}}}\Big\{\frac{s}{r}\Big[M^{\frac{r}{s}} - 2^{\frac{r}{s}}\Big] + \mathcal{O}(1)\Big\} + \mathcal{O}\Big(\tfrac{1}{(\log N)^2}\Big)
\end{aligned}$$

$$= 1 - \Big(\frac{2}{M}\Big)^{\frac{r}{s}} + \mathcal{O}\Big(\frac{1}{M^{\frac{r}{s}}(\log N)} + \frac{1}{(\log N)^2}\Big) = 1 - \exp\Big(-\frac{r}{s}\log\Big(\frac{M}{2}\Big)\Big) + \mathcal{O}\Big(\frac{1}{\log N}\Big)$$
$$= 1 - \exp\Big(-\frac{r}{s}\log(M)\Big) + \mathcal{O}\Big(\frac{1}{\log N}\Big),$$

as $M^{\frac{r}{s}} = \exp\big(\frac{r}{s}\log(M)\big)$ and

$$1 \leqslant \exp\Big(\frac{r}{s}\log(M)\Big) \leqslant \exp\Big(\frac{\tilde{r}}{s}\frac{\log(M)}{\log(2N)}\Big) \leqslant \exp\Big(\frac{\tilde{r}}{s}\Big) \leqslant c,$$

for a constant c, for all $M = 1, \ldots, 2N$. Applying (A.1) again ends the proof.

□

We will here adapt Theorem 3.6.1 from [15] to our purposes and give a short proof which follows the *second proof* of the theorem in [15]. We will use the notation $\mathcal{R}_i$ for the considered recombination of lineage i (hence $\mathcal{R}_i$ stands for $R(i,1), R_B^{\mathrm{rec}}(i), \ldots$) and denote by θ the corresponding recombination probability given X.

Lemma A.2 (cf. Theorem 3.6.1, [15])**.** *Let $m \in \{1, \ldots, 2N\}$. For $\eta_m := \sum_{t=\tau_m+1}^{\tau} \theta_t$,*

$$|\mathbb{P}(\mathcal{R}_i \geqslant \tau_m \mid X) - (1 - \mathrm{e}^{-\eta_m})| \leqslant \sum_{t=\tau_m+1}^{\tau} \theta_t^2.$$

Proof of Lemma A.2. Conditional on X define for m and t with $\tau_m \leqslant t \leqslant \tau$ the random variable which says whether or not the particular recombination happened at time t:

$$Y_{m,t} \text{ with } \mathbb{P}(Y_{m,t} = 1 \mid X) = \theta_t,\ \mathbb{P}(Y_{m,t} = 0 \mid X) = 1 - \theta_t.$$

Let $S_m := Y_{m,\tau_m+1} + \ldots + Y_{m,\tau}$, thus if $S_m > 0$, there was at least one recombination in the time interval $[\tau_m + 1, \tau]$. We show that the distribution of S_m can be approximated by a Poisson distribution with mean η_m.
Denote with $\mu_{m,t}$ the distribution of $Y_{m,t}$ and with μ_m the distribution of S_m, i.e. $\mu_m = \mu_{m,\tau_m+1} * \ldots * \mu_{m,\tau}$.
Let $\nu_{m,t}$ be a Poisson distribution with mean θ_t, ν_m be a Poisson distribution with mean η_m. By properties of Poisson distributed random variables, it holds that $\nu_m = \nu_{m,\tau_m+1} * \ldots * \nu_{m,\tau}$.
We will use the following general results from Lemmas 3.6.2, 3.6.3 and 3.6.4 from [15] without giving the proofs: let $\mu_1, \mu_2, \nu_1, \nu_2$ be probability measures.

1. $\|\mu_1 * \mu_2 - \nu_1 * \nu_2\| \leqslant \|\mu_1 \cdot \mu_2 - \nu_1 \cdot \nu_2\|$.
2. $\|\mu_1 \cdot \mu_2 - \nu_1 \cdot \nu_2\| \leqslant \|\mu_1 - \nu_1\| + \|\mu_2 - \nu_2\|$.
3. Let μ be the measure with $\mu(1) = p$ and $\mu(0) = 1 - p$. Let ν be a Poisson distribution with mean p. Then $\|\mu - \nu\| \leqslant 2p^2$.

With this,

$$\begin{aligned}
&\|\mu_m - \nu_m\| = \|\mu_{m,\tau_m+1} * \ldots * \mu_{m,\tau} - \nu_{m,\tau_m+1} * \ldots * \nu_{m,\tau}\| \\
&\overset{1.}{\leqslant} \|\mu_{m,\tau_m+1} \cdot (\mu_{m,\tau_m+2} * \ldots * \mu_{m,\tau}) - \nu_{m,\tau_m+1} \cdot (\nu_{m,\tau_m+2} * \ldots * \nu_{m,\tau})\| \\
&\overset{2.}{\leqslant} \|\mu_{m,\tau_m+1} - \nu_{m,\tau_m+1}\| + \|\mu_{m,\tau_m+2} * \ldots * \mu_{m,\tau} - \nu_{m,\tau_m+2} * \ldots * \nu_{m,\tau}\| \\
&\leqslant \ldots \text{ iterate } \ldots \\
&\leqslant \sum_{t=\tau_m+1}^{\tau} \|\mu_{m,t} - \nu_{m,t}\| \overset{3.}{\leqslant} \sum_{t=\tau_m+1}^{\tau} 2 \cdot \theta_t^2.
\end{aligned}$$

With the definition of the total variation distance we now have

$$\sup_A |\mu_m(A) - \nu_n(A)| \leqslant \sum_{t=\tau_m+1}^{\tau} \theta_t^2 \tag{A.2}$$

Considering the specific set $A = \{0\}$, we get

$$\begin{aligned}
|\mathbb{P}(\mathcal{R}_i \geqslant \tau_m \mid X) - (1 - \mathrm{e}^{-\eta_m})| &= |1 - \mathbb{P}(S_m = 0) - (1 - \nu_m(0))| \\
&= |1 - \mu_m(0) - (1 - \nu_m(0))| \\
&= |\mu_m(0) - \nu_n(0)| \\
&\overset{(A.2)}{\leqslant} \sum_{t=\tau_m+1}^{\tau} \theta_t^2.
\end{aligned}$$

□

APPENDIX B

Appendix of Chapter 3

We will here give an intuition as to how the statement of Lemma 3.11 can be derived. All derivations and ideas used in this addendum were provided by Charline Smadi.

Coupling with supercritical birth and death processes during the first phase

The overall idea is to couple the process N_a^K during the first phase with two supercritical birth and death processes in such a way that the number of a-individuals is almost surely always in between the values of the two processes. That enables us to use known results from the theory on birth and death processes in order to draw conclusions on the dynamics of our process N_a^K.
Another important ingredient for the proof of the results on the number of jumps during the first phase, is to express the condition $\{T_\varepsilon^K \leqslant \tilde{T}_\varepsilon^K\}$ in the definition (3.31) of the probability measure $\mathbb{P}^{(1)}$ in a different and particularly more convenient way.
Recall the definition of the values $s_-(\varepsilon)$ and $s_+(\varepsilon)$ from (3.47) and keep in mind that by definition, the first phase of the selective sweep lasts until the random time T_ε^K from (3.21). By easy calculations we can check that for $t < T_\varepsilon^K \wedge \tilde{T}_\varepsilon^K$ the death rate $d_a^K(N^K(t))$ as defined in (3.7) satisfies

$$1 - s_+(\varepsilon) \leqslant \frac{d_a^K(N^K(t))}{f_a N_a^K(t)} = 1 - s + \frac{C_{a,A}}{f_a K}(N_A^K(t) - \bar{n}_A K) + \frac{C_{a,a}}{f_a K} N_a^K(t) \leqslant 1 - s_-(\varepsilon). \quad \text{(B.1)}$$

If we now apply the results from the coupling Theorem 2 in [13], we can construct the processes Z_ε^-, (N_A^K, N_a^K) and Z_ε^+ on the same probability space such that almost surely:

$$Z_\varepsilon^-(t) \leqslant N_a^K(t) \leqslant Z_\varepsilon^+(t), \quad \text{for all } t < \tilde{T}_\varepsilon^K \wedge T_\varepsilon^K. \quad \text{(B.2)}$$

Here, for $* \in \{-, +\}$, Z_ε^* is a birth and death process with initial state 1, and individual birth rates f_a and death rates $f_a(1 - s_*(\varepsilon))$.

Inspired by (B.1) we will now redefine the probability measure $\mathbb{P}^{(1)}$ from (3.31): we define

the events

$$\begin{aligned}\mathcal{L}_\varepsilon^K &:= \Big\{1 - s_+(\varepsilon) \leqslant \frac{d_a^K(N^K(t))}{f_a N_a^K(t)} \leqslant 1 - s_-(\varepsilon), \forall t < T_\varepsilon^K\Big\}, \text{ and} \\ \mathcal{H}_\varepsilon^K &:= \{N_A^K(t) \in I_\varepsilon^K, \forall t < T_\varepsilon^K\},\end{aligned} \tag{B.3}$$

and can verify that

$$\{T_\varepsilon^K \leqslant \tilde{T}_\varepsilon^K\} = \{T_\varepsilon^K < \infty\} \cap \mathcal{L}_\varepsilon^K \cap \mathcal{H}_\varepsilon^K.$$

Hence the probability $\mathbb{P}^{(1)}$ can be also defined as follows

$$\mathbb{P}^{(1)}(\cdot) = \mathbb{P}(\,\cdot \mid T_\varepsilon^K < \infty, \mathcal{L}_\varepsilon^K, \mathcal{H}_\varepsilon^K). \tag{B.4}$$

Before giving the proof of Lemma 3.11 we cite a general result from [1] on birth and death processes:

Proposition B.1 ([1])**.** *Let $Z = (Z_t)_{t\geqslant 0}$ be a birth and death process with birth and death rates b and d. For $i \in \mathbb{Z}^+$, $T_i = \inf\{t \geqslant 0, Z_t = i\}$ and $\mathbb{P}_i$ is the law of Z when $Z_0 = i$. Then for $i, j, k \in \mathbb{Z}_+^3$ such that $i < j < k$,*

$$\mathbb{P}_j(T_k < T_i) = \frac{1 - (d/b)^{j-i}}{1 - (d/b)^{k-i}}. \tag{B.5}$$

Further, we define the stopping time

$$\sigma_u^K := \inf\{t \geqslant 0, N_a^K(t) = \lfloor u \rfloor\}, \quad u \in \mathbb{R}_+, \tag{B.6}$$

which denotes the time of the first hitting of $\lfloor u \rfloor$ by the process N_a^K (in contrast to the jump number $\sigma_l^K(1)$ from (3.54)). In the following we denote by $\tilde{Z}^{(s)}$, $0 < s < 1$, a random walk with jumps ± 1 where up-jumps occur with probability $1/(2-s)$ and down-jumps with probability $(1-s)/(2-s)$. Further, the law of $\tilde{Z}^{(s)}$ when the initial state is $i \in \mathbb{N}$ is denoted by $\mathcal{P}_i^{(s)}$. Finally we define for every $\rho \in \mathbb{R}_+$ the stopping time

$$\tau_\rho := \inf\{n \in \mathbb{Z}_+, \tilde{Z}_n^{(s)} = \lfloor \rho \rfloor\}. \tag{B.7}$$

Proof of Lemma 3.11

The general idea is to construct two geometrically distributed random variables such that the number $U_k^K(1)$ of jumps from k to $k+1$ can be bounded from below and above by these variables. The approach is in fact similar but more evolved than in the proof of Lemma 3.2 (here Lemma 2.23) from [36] which was given in Section 2.4.1.

We will start with approximating the value of $\mathbb{E}^{(1)}_{(n_A,j)}[U^K_k(1)]$ for $j \leqslant k < \lfloor \varepsilon K \rfloor$. As we always consider the process conditioned on the event of a selective sweep, which means $\{T^K_\varepsilon < \infty\}$ for the first phase, the number of a-individuals will necessarily jump from k to $k+1$ at some point during the first phase. Once N^K_a has reached the state $k+1$ it can either go directly to $\lfloor \varepsilon K \rfloor$ without ever returning to k, or it jumps back to k. In the latter case, we are in the same situation again where the condition $\{T^K_\varepsilon < \infty\}$ implies that we necessarily hit the value $k+1$ another time.

We first approximate the probability that there is exactly one jump from k to $k+1$. In the case of the varying population size model, we do not know the value of N^K_A when N^K_a hits k for the first time, and hence we bound the probability by the extreme values of $N^K_A \in I^K_\varepsilon$. Recalling the stopping times defined in (B.6) and (B.7) and the new definition of the measure $\mathbb{P}^{(1)}$ given in (B.4) we can derive the upper bound as follows:

$$\begin{aligned} \mathbb{P}^{(1)}_{(n_A,j)}(U^K_k(1) = 1) &\leqslant \sup_{n_A \in I^K_\varepsilon} \mathbb{P}^{(1)}_{(n_A,k+1)}(T^K_\varepsilon < \sigma^K_k) \\ &= \sup_{n_A \in I^K_\varepsilon} \frac{\mathbb{P}_{(n_A,k+1)}(T^K_\varepsilon < \sigma^K_k \mid \mathcal{L}^K_\varepsilon, \mathcal{H}^K_\varepsilon)}{\mathbb{P}_{(n_A,k+1)}(T^K_\varepsilon < \infty \mid \mathcal{L}^K_\varepsilon, \mathcal{H}^K_\varepsilon)} \leqslant q_k^{(s_+(\varepsilon),s_-(\varepsilon))}, \end{aligned} \tag{B.8}$$

where for $(s_1, s_2) \in (0,1)^2$

$$q_k^{(s_1,s_2)} := \frac{\mathcal{P}^{(s_1)}_{k+1}(\tau_{\varepsilon K} < \tau_k)}{\mathcal{P}^{(s_2)}_{k+1}(\tau_{\varepsilon K} < \tau_0)},$$

with $\mathcal{P}^{(s)}_i$ as described above. Similarly, we can show that $\mathbb{P}^{(1)}_{(n_A,j)}(U^K_k(1) = 1) \geqslant q_k^{(s_-(\varepsilon),s_+(\varepsilon))}$. Repeating this idea, we can approximate the probability that there are only two jumps from k to $k+1$, given that there are at least two jumps, and so on. From this we infer that we can construct two geometric random variables G_1 and G_2 on a possibly enlarged probability space, with respective parameters $q_k^{(s_+(\varepsilon),s_-(\varepsilon))} \wedge 1$ and $q_k^{(s_-(\varepsilon),s_+(\varepsilon))}$, such that almost surely

$$G_1 \leqslant U^K_k(1) \leqslant G_2. \tag{B.9}$$

In particular, these inequalities hold true for the expected values of the above variables and hence we get by applying (B.5)

$$\begin{aligned} &\frac{(1-(1-s_+(\varepsilon))^{\lfloor \varepsilon K \rfloor - k})(1-(1-s_-(\varepsilon))^{k+1})}{s_+(\varepsilon)(1-(1-s_-(\varepsilon))^{\lfloor \varepsilon K \rfloor})} \\ \leqslant \mathbb{E}^{(1)}_{(n_A,j)}[U^K_k(1)] \;&\leqslant\; \frac{(1-(1-s_-(\varepsilon))^{\lfloor \varepsilon K \rfloor - k})(1-(1-s_+(\varepsilon))^{k+1})}{s_-(\varepsilon)(1-(1-s_+(\varepsilon))^{\lfloor \varepsilon K \rfloor})}. \end{aligned} \tag{B.10}$$

Recalling the definitions of $s_-(\varepsilon)$ and $s_+(\varepsilon)$ in (3.47) the definition of s in (3.12), we can perform straightforward calculations in order to obtain (3.49).

Let us now assume that $k < j$. Then by a similar logic we get

$$\begin{aligned}\mathbb{P}^{(1)}_{(n_A,j)}(U^K_k(1) \geqslant 1) &\leqslant \sup_{n_A \in I^K_\varepsilon} \mathbb{P}_{(n_A,j)}(\sigma^K_k < T^K_\varepsilon | T^K_\varepsilon < \infty, \mathcal{L}^K_\varepsilon, \mathcal{H}^K_\varepsilon) \\ &= \sup_{n_A \in I^K_\varepsilon} \frac{\mathbb{P}_{(n_A,j)}(T^K_\varepsilon < \infty | \sigma^K_k < T^K_\varepsilon, \mathcal{L}^K_\varepsilon, \mathcal{H}^K_\varepsilon)\mathbb{P}_{(n_A,j)}(\sigma^K_k < T^K_\varepsilon | \mathcal{L}^K_\varepsilon, \mathcal{H}^K_\varepsilon)}{\mathbb{P}_{(n_A,j)}(T^K_\varepsilon < \infty | \mathcal{L}^K_\varepsilon, \mathcal{H}^K_\varepsilon)} \\ &\leqslant \frac{\mathcal{P}^{(s_+(\varepsilon))}_k(\tau_{\varepsilon K} < \tau_0)\mathcal{P}^{(s_-(\varepsilon))}_j(\tau_k < \tau_{\varepsilon K})}{\mathcal{P}^{(s_-(\varepsilon))}_j(\tau_{\varepsilon K} < \tau_0)} \leqslant \frac{(1 - s_-(\varepsilon))^{j-k}}{s_+(\varepsilon)s_-(\varepsilon)}.\end{aligned}$$

In this case, the same proof as for (B.9) leads to:

$$\mathbb{E}^{(1)}_{(n_A,j)}[U^K_k(1) | U^K_k(1) \geqslant 1] \leqslant \left(q_k^{(s_-(\varepsilon), s_+(\varepsilon))}\right)^{-1} \leqslant s_-^{-1}(\varepsilon),$$

where we used (A.14) from Lemma A.2 in [38] for the last inequality. This ends the proof of (3.50).

The proof of (3.51) is nearly the same as the proof of the second part of (7.10) in Lemma 7.2, [38]. The only difference is that we consider the process after the time τ^K_m in the current case, and after 0 in the previous work. Hence we only give a sketch of the proof and refer to [38] for the details. Recall (3.83) and define the number of the up-jumps of N^K_a before and after $\zeta^K_{k'}(1)$ for $k', k < \lfloor \varepsilon K \rfloor$,

$$\begin{aligned}U^{(K,1)}_{k',k}(1) &:= \#\{m < \zeta^K_{k'}(1), (N^K_a(\tau^K_m), N^K_a(\tau^K_{m+1})) = (k, k+1)\}, \\ U^{(K,2)}_{k',k}(1) &:= \#\{m \geqslant \zeta^K_{k'}(1), \tau^K_m \leqslant T^K_\varepsilon, (N^K_a(\tau^K_m), N^K_a(\tau^K_{m+1})) = (k, k+1)\}.\end{aligned}$$

Then

$$U^K_k(1) = U^{(K,1)}_{k',k}(1) + U^{(K,2)}_{k',k}(1), \quad \mathbb{P}^{(1)} - a.s.,$$

for $k' \leqslant k < \lfloor \varepsilon K \rfloor$, and

$$\begin{aligned}\left|\mathrm{Cov}^{(1)}_{(n_A,j)}(U^K_{k'}(1), U^K_k(1))\right| \leqslant &\left(\mathbb{E}^{(1)}_{(n_A,j)}[(U^K_{k'}(1))^2]\mathbb{E}^{(1)}_{(n_A,j)}[(U^{(K,1)}_{k',k}(1))^2]\right)^{1/2} \\ &+ \left|\mathrm{Cov}^{(1)}_{(n_A,j)}(U^K_{k'}(1), U^{(K,2)}_{k',k}(1))\right|.\end{aligned}$$

We can bound the terms on the right-hand side by the following observations: by the coupling from (B.9) and (A.14) from [38], $\mathbb{E}^{(1)}_{(n_A,j)}[(U^K_{k'}(1))^2]$ can be bounded by $2/s^2_-(\varepsilon)$. The second

expectation, $\mathbb{E}^{(1)}_{(n_A,j)}[(U^{(K,1)}_{k',k}(1))^2]$, is of order $\lambda_\varepsilon^{k-k'} < 1$, with λ_ε as in (3.48). The intuition for this is that in case of the measure conditioned on fixation we do not expect to see larger excursions above some k' as this would imply that we needed to jump down many times which is unlikely under $\mathbb{P}^{(1)}$. Hence we expect that the number of upcrossings from k to $k+1$ during an excursion above k' decreases geometrically with k.

The last step consists in bounding the covariance which comes down to proving that the path of N^K_a after $\zeta^K_{k'}(1)$, the last hitting of k', only weakly depends on the trajectory before $\zeta^K_{k'}(1)$. As $\mathbb{P}^{(1)}$ implies that $N^K_A \in I^K_\varepsilon$, with I^K_ε as defined in (3.20), the size of the A-population experiences only small changes which implies that the value of N^K_A at time $\zeta^K_{k'}(1)$ only weakly depends on the past.

□

APPENDIX C

Appendix of Chapter 4

C.1 R code to generate a Λ-coalescent

Algorithm C.1. The algorithm as describe in words in Algorithm 4.3 can be realized by the following R-code:

```
lambdacoal <- function(n,Lambdatype,a,alpha){
 #specify measure Lambda through density function flambda
 #uniform measure
 if(Lambdatype==1){
 flambda <- function(x) {1}
 }
 #Beta with parameter alpha
 if(Lambdatype==2){
   flambda <- function(x) {dbeta(x,2-alpha,alpha)}
 }

 #define the array of rates
 if(Lambdatype > 0){
 lambdaarray <- array(0,c(n,n))
 for(b in 2:n){
   for(k in 2:b){
     integrand <- function(x) {x^(k-2)*(1-x)^(b-k)*flambda(x)}
     lambdaarray[b,k] <- integrate(integrand, lower = 0, upper = 1)$value
   }}}

 # for atom at a:
 if(Lambdatype == 0){
 lambdaarray <- array(0,c(n,n))
 for(b in 2:n){
   for(k in 2:b){
     lambdaarray[b,k] <- a^(k-2)*(1-a)^(b-k)
   }}}
```

```
  #array of binomial coefficients
  binomialcoeff <- array(0,c(n,n))
  for(b in 2:n){
    for(k in 2:b){
      binomialcoeff[b,k] <- choose(b,k)
    }}

  # reset parameters
  t <- 0
  tsum <- 0
  i <- 2
  m <- n+1
  blength <- 0

  #initialize array ancestry
  ancestry <- array(0,c(n,n+2))
  for(j in 1:n){ancestry[,j] <- j}

  while(m>2){
    totalrate <- as.numeric(lambdaarray[m-1,]%*%binomialcoeff[m-1,])
    t <- rexp(1,totalrate)
    tsum <- tsum + t
    multiprob <- (lambdaarray[m-1,]*binomialcoeff[m-1,])/totalrate
    mergertype <- rmultinom(1,1,multiprob)
    merger <- which(mergertype == 1)
      ancestry[i,n+2] <- merger
      lin <- sample(unique(ancestry[i-1,c(1:n)]),merger)
      name <- min(lin)
      for(j in 1:length(lin)){
        ancestry[c(i:n),lin[j]] <- name
        ancestry[c(i:n),c(1:n)][ancestry[c(i:n),c(1:n)]==lin[j]] <- name
      }
      ancestry[i,n+1] <- tsum
      blength <- blength + (m-1)*(tsum-ancestry[i-1,n+1])
      m <- length(unique(ancestry[i,c(1:n)])) + 1
      i <- i+1
    }

  result <- list(genealogy=ancestry[c(1:i-1),],branch.length=blength)
  return(result)
}
```

C.2 MuPAD code realizing the result of Theorem 4.4

Algorithm C.2. First, we define the rates λ, $\tilde{\lambda}$ and γ and with this define the functions g and f from (4.7) and (4.8). We will directly define the $f(a, j, k_1)$ of the program as the sum $f^{(j,j)}(k_1) + f^{(j,k_1)}(k_1)$.

```
assume(a, Type::Interval([0],[1])):
lamb := (b,k,a) -> a^(k-2)*(1-a)^(b-k):
lambtild := (b,k,a) -> sum(binomial(b,i)*lamb(b,i,a), i = b-k+1..b):
gamm := (b,a) -> sum((k-1)*binomial(b,k)*lamb(b,k,a),k=2..b):

g := (a,x,y,z) ->
(if x=y and z=y then 0
elif x=y then binomial(z,z-x+1)*lamb(z,z-x+1,a)/lambtild(z,x,a)
elif z=y then 1-binomial(z,z-x+1)*lamb(z,z-x+1,a)/lambtild(z,x,a)
else 0
end_if):

f := (a,x,z) -> 1/(gamm(x,a))*g(a,x,x,z) + 1/(gamm(z,a))*g(a,x,z,z):
```

As already described in Section 4.2, the idea is to first construct the correct number of nested sums and then replace the placeholders with the real g's. The placeholders will be 3-dim array `GG` representing the last 3 arguments of g and a 2-dim array `PP` for the products which appear in the expression of the expected height. Further, we will use another placeholder-array `vv` which stands for the index of the sum which is considered in the corresponding step. We make extensive use of the command `subs` which substitutes some object by another in a given expression (here the object will be the placeholder and the expression some value of g or an index of a sum). Note that the command for commenting lines is `//` in MuPAD.

```
expheightDelta := proc(n,a)
  // initialize local and global variables
  local j,i,k,l,m,summ0,summ1,summ2,r,rr,k1,prodg,RR;
  save expheight,vv,GG,PP;

begin

// we will exclude n=2 and n=3 as we do not need the formula in these case
if n=2 then expheight := 1/gamm(n,a);
elif n=3 then expheight := 1/gamm(n,a) + f(a,2,3);
else
```

```
// the entries of PP define the start-1 and end+1 of the product
PP := array(1..n,1..n);
vv := array(1..1,1..n);
GG := array(1..n,1..n,1..n);

expheight := 0;

for j from 2 to n-1 do
 if j=n-1 then
 else

// array with last column ones
// used to construct the sums
RR := array(1..n-j,1..n+1);
for k from 1 to n-j do
  RR[k,n+1] := 1;
end_for;
RR[1,n] := GG[vv[1,n-j-1],vv[1,n-j-1],n];

// the i-loop deals with the sums with index k_2 to k_{n-j-1} (k_{n-j} is
// 'calculated' in R[1,n]), hence n-j-i goes from n-j-1 to 2

for i from 1 to n-j-2 do
  // the r-loop constructs for a value i all corresponding nested sums,
  // starting with the highest possible value rr
  for r from n-i to n do
    rr := n+n-i-r;
    summ0 := subs(RR[i,rr+1], vv[1,n-j-i]=rr);
    RR[i+1,rr] := GG[vv[1,n-j-1-i], vv[1,n-j-1-i],rr]*PP[vv[1,n-j-1-i],rr]*summ0;
    if rr < n then
      RR[i+1,rr] := RR[i+1,rr] + RR[i+1,rr+1];
    end_if;
  end_for;
end_for;

end_if;
```

```
summ1 := 0;

// the value of i from the above for-loop is exactly what we need to creat
// this loop over the possible k_1
for k1 from j+1 to n do
  summ2:=subs(RR[i,k1+1],vv[1,1]=k1);
  summ1 := summ1 + f(a,j,k1)*PP[j,k1]*summ2;
end_for;

expheight := expheight + summ1;
end_for;

// we need to replace the placeholder arrays with the corresponding functions
for k from 2 to n do
  for m from 2 to n do
    for l from 2 to n do
      expheight := subs(expheight, GG[k,l,m]=g(a,k,l,m));
    end_for;
  end_for;
end_for;

for k from 2 to n do
  for m from k+1 to n do
    prodg := 1;
    for l from k+1 to m-1 do
      prodg := prodg*g(a,l,m,m);
    end_for;
    expheight := subs(expheight, PP[k,m]=prodg);
  end_for;
end_for;

expheight:=simplify(expheight+1/gamm(n,a));
end_if;

return(expheight);
end_proc;
```

Notation

$(\alpha ik)_m$: the k-th locus of individual i is associated to an allele α at the m-th jump time, page 136

$\tilde{\mathcal{A}}_0^t(i), \ldots, \tilde{\mathcal{A}}_4^t(i)$: ancestral relationships of an individual i, page 47

$A_t^u(i,j)$: label of the ancestor living at time u of (i,j) living at time $t > u$, page 35

$B_t(i)$: indicates the allelic type of i at time t, page 35

β_k: $= k(2N-k)/[k^2 + (2N-k)^2 + sk(2N-k)]$, page 57

$b(\alpha)$: birth rate of an individual of type α, page 27

c_N: time scale to get to the coalescent, page 10

$C(\alpha, \alpha')$: pressure felt by an individual of type α through an individual of type α', page 27

$D(\alpha)$: intrinsic death rate of an individual of type α, page 27

$d(\alpha)$: total death rate of an individual of type α, page 28

$\Delta_{\mathbb{N}}, \Delta_n$: partition of $\mathbb{N}$, n, resp., in singletons, page 7

f_α: birth rate of an individual of type α, page 115

(G1): first geometric alignment, SL−N1−N2, page 32

(G2): second geometric alignment, N1−SL−N2, page 32

$G^{j_i,j_m}(i,m), G^j(i,m)$: first time the (i,j_i) and (m,j_m) find a common ancestor, page 41

$\mathcal{H}, \mathcal{H}^{\Pi^{[n]}}$: height of the coalescent tree corresponding to $\Pi^{[n]}$, page 9

I_ε^K: interval containing the equilibrium size $\bar{n}_A K$ of the A-population, page 121

$I_{t,1}, \ldots, I_{t,6}$: evolution describing random variables in the Moran model, page 35

K: carrying capacity, scaling parameter, page 28

$\mathcal{L}, \mathcal{L}^{\Pi^{[n]}}$: length of the coalescent tree corresponding to $\Pi^{[n]}$, page 9

$\subset, \overset{k}{\prec}$: see equation 1.2, page 6

$\mathrm{Mult}(n; p_1, \ldots, p_k)$: multinomial distribution with n trials and k categories with success probabilities $p_1, \ldots, p_k$, page 2

$[N]$: $= \{1, 2, \ldots, N\}$, page 2

$[n; 1, 2]$: $= \{(i, 1), (i, 2),\ i \in [n] := \{1, \ldots, n\}\}$, page 37

$\bar{n}_\alpha$: equilibrium density of a monomorphic type α-population, page 29

$N^K = ((N^K_{\alpha\beta\gamma}(t))_{(\alpha,\beta,\gamma)\in\mathcal{E}}, t \geq 0)$: population process in the Darwinian model, page 114

ν, ν^t: offspring distribution in a Cannings model, page 2

$\mathbb{P}$: probability measure conditioned on fixation of the mutant allele, page 34

$\mathbb{P}^{(1)}$: probability measure for the first phase of the selective sweep, page 124

$\mathbb{P}^{(3)}$: probability measure for the third phase of the selective sweep, page 124

$p^{c,j}_{\bullet_1,\bullet_2}(X_{t-1}, X_t)$: coalescence probability for types $\bullet_1, \bullet_2$, page 59

$\Pi^K, \Pi^{K,n}$: Kingman coalescent process on $\mathcal{P}_\mathbb{N}$,$\mathcal{P}_n$, resp., page 7

$\Pi^\Lambda, \Pi^{\Lambda,n}$: Λ-coalescent process on $\mathcal{P}_\mathbb{N}$,$\mathcal{P}_n$, resp., page 8

$\Pi^\mathbb{N}, \Pi^{[n]}$: coalescent process on $\mathcal{P}_\mathbb{N}$,$\mathcal{P}_n$, resp., page 6

$\mathcal{P}_\mathbb{N}$: space of all partitions of $\mathbb{N}$, page 6

$\mathcal{P}_n$: space of all partitions of $[n]$, page 6

$p^r_\bullet(X_{t-1}, X_t)$: recombination probability into another background, page 59

$p^{r_2}_{bb}(X_{t-1}, X_t), p^{r_2}_{BB}(X_{t-1}, X_t)$: recombination within the b-population, resp. B-population, page 60

$Q_{\bar{q}}$: distribution of a marked $\bar{q}$-partition of $[n; 1, 2]$, page 38

r_j: probability for a recombination before between SL and the neutral locus Nj, page 33

$\rho/2$: scaled recombination rate, page 16

$R(i, j)$: first time the (i, j) migrates into the b-population, page 40

$R^{\mathrm{rec}}_B(i)$: first time the $(i, 1)$ and $(1, 2)$ find different B-ancestors, page 41

$R^{\mathrm{rec}}_b(i)$: first time the $(i, 1)$ and $(1, 2)$ separate within the b-population, page 41

s^G, s: selection coefficient (of the genotype G), page 25

$S(\alpha', \alpha)$: invasion fitness of a mutant of type α' in an α-population, page 30

$\mathcal{S}_i$: event $\{R_B^{\text{rec}}(i) \geqslant \tau_N\}$, page 46

$\tau, \tau_m \tau_m^*$: fixation time of the mutant B; first and last hitting of the state m of X_t, page 34

$\tau_0^K, \tau_1^K, \ldots$: jump times of N^K, page 125

T_ε^K: first time that $N_a^K(t)$ hits $\varepsilon K\rfloor$, page 121

T_{ext}^K: time of extinction of the A population, page 117

Θ: the true partition of a sample at the end of a selective sweep, page 37

$\theta/2$: scaled mutation rate, page 22

θ_t^{BB}: $= p_{BB}^{r_2}(X_{t-1}, X_t)$, page 66

θ_t^{bb}: $= p_{bb}^{r_2}(X_t, X_{t+1})$, page 73

$\theta_t^{r_j}$: $= p_B^{r_j}(X_t, X_{t+1})$, $j = 1, 2$, page 82

$T_{mrca}, T_{mrca}^{\Pi^{[n]}}$: time to the most recent common ancestor (for the coalescent process $\Pi^{[n]}$), page 7

$\mathcal{T}(p)$: local tree for position p, page 20

$\tilde{T}_\varepsilon^K$: first time $N_A^K(t)$ exits I_ε^K, page 121

$\mathcal{U}(A)$: uniform distribution on the set A, page 3

$U_k^K(j)$: number of up-jumps of N_a^K from k to $k+1$ in phase $j = 1, 3$, page 132

$U_{k,j}, D_{k,j}, H_{k,j}$: up- and down-jumps, holds starting in state k after time τ_j, page 56

$\mathcal{U}_k^K(1)$: number of up-jumps of N_A^K from k to $k+1$ in phase 1 when $N_a^K(\tau_m^K) = k$, page 133

$U_{n_A,l,k}^K(1)$: number of up-jumps of N_a^K from k to $k+1$ in phase 1 during an excursion above or below l, page 133

$X = (X)_{t \geqslant 0}$: Markov chain describing the path of the advantageous allele B, page 34

Bibliography

[1] K.B. Athreya and P. Ney. *Branching Processes*. Dover Books on Mathematics Series. Dover Publications, 2004.

[2] Nicholas H Barton. The effect of hitch-hiking on neutral genealogies. *Genetical Research*, 72(02):123–133, 1998.

[3] Julien Berestycki, Nathanaël Berestycki, and Jason Schweinsberg. Beta-coalescents and continuous stable random trees. *The Annals of Probability*, pages 1835–1887, 2007.

[4] N. Berestycki. Recent progress in coalescent theory. *Ensaios Matematicos*, 16:1–193, 2009.

[5] Jean Bertoin. *Random fragmentation and coagulation processes*. Cambridge University Press Cambridge, 2006.

[6] Jean Bertoin and Jean-François Le Gall. The bolthausen–sznitman coalescent and the genealogy of continuous-state branching processes. *Probability theory and related fields*, 117(2):249–266, 2000.

[7] P. Billingsley. *Probability and measure*. Wiley-India, 2008.

[8] M. Birkner, J. Blath, and B. Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Arxiv preprint arXiv:1203.4950*, 2012.

[9] Matthias Birkner, Jochen Blath, Marcella Capaldo, Alison Etheridge, Martin Möhle, Jason Schweinsberg, and Anton Wakolbinger. *Alpha-stable branching and beta-coalescents*. WIAS, 2004.

[10] Erwin Bolthausen and A-S Sznitman. On ruelle's probability cascades and an abstract cavity method. *Communications in mathematical physics*, 197(2):247–276, 1998.

[11] Rebekka Brink-Spalink and Charline Smadi. Genealogies of two linked neutral loci after a selective sweep in a large population of varying size. *In preparation*, 2014.

[12] C Cannings. The latent roots of certain markov chains arising in genetics: a new approach, i. haploid models. *Advances in Applied Probability*, pages 260–290, 1974.

[13] Nicolas Champagnat. A microscopic interpretation for adaptive dynamics trait substitution sequence models. *Stochastic processes and their applications*, 116(8):1127–1160, 2006.

[14] R. Durrett. *Probability models for DNA sequence evolution*. Springer Verlag, 2008.

[15] Rick Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

[16] B. Eldon and J. Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172(4):2621, 2006.

[17] A. Etheridge, P. Pfaffelhuber, and A. Wakolbinger. An approximate sampling formula under genetic hitchhiking. *The Annals of Applied Probability*, 16(2):685–729, 2006.

[18] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[19] Nicolas Fournier and Sylvie Méléard. A microscopic probabilistic description of a locally regulated population and macroscopic approximations. *Annals of applied probability*, pages 1880–1919, 2004.

[20] Robert C Griffiths. The two-locus ancestral graph. *Lecture Notes-Monograph Series*, pages 100–117, 1991.

[21] Robert C Griffiths and Paul Marjoram. An ancestral recombination graph. *Institute for Mathematics and its Applications*, 87:257, 1997.

[22] Matthew Hamilton. *Population genetics*. John Wiley & Sons, 2011.

[23] J. Hein, M.H. Schierup, and C. Wiuf. *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, USA, 2005.

[24] Motoo Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893, 1969.

[25] Motoo Kimura and James F Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49(4):725, 1964.

[26] John FC Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982.

[27] John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[28] Martin Möhle. Weak convergence to the coalescent in neutral population models. *Journal of applied probability*, pages 446–460, 1999.

[29] C. Neuhauser and SM Krone. Ancestral processes with selection. *Theoretical population biology*, 51:210–237, 1997.

[30] Claudia Neuhauser and Stephen M Krone. The genealogy of samples in models with selection. *Genetics*, 145(2):519–534, 1997.

[31] P. Pfaffelhuber and A. Studeny. Approximating genealogies for partially linked neutral loci under a selective sweep. *Journal of mathematical biology*, 55(3):299–330, 2007.

[32] Jim Pitman. Coalescents with multiple collisions. *Annals of Probability*, pages 1870–1902, 1999.

[33] Jim Pitman. *Combinatorial stochastic processes*, volume 32. Springer, 2006.

[34] Serik Sagitov et al. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, 36(4):1116–1125, 1999.

[35] J. Schweinsberg. A necessary and sufficient condition for the λ-coalescent to come down from infinity. 2000.

[36] J. Schweinsberg and R. Durrett. Random partitions approximating the coalescence of lineages during a selective sweep. *The Annals of Applied Probability*, 15(3):1591–1651, 2005.

[37] Jason Schweinsberg. Coalescent processes obtained from supercritical galton–watson processes. *Stochastic processes and their applications*, 106(1):107–139, 2003.

[38] Charline Smadi. An eco-evolutionary approach of adaptation and recombination in a large population of varying size. *arXiv preprint arXiv:1402.4104*, 2014.

[39] Leocard Stephanie et al. Selective sweep and the size of the hitchhiking set. *Advances in Applied Probability*, 41(3):731–764, 2009.

[40] John Wakeley. *Coalescent theory: an introduction*, volume 1. Roberts & Company Publishers Greenwood Village, Colorado, 2009.

[41] C. Wiuf and J. Hein. Recombination as a point process along sequences. *Theoretical population biology*, 55(3):248–259, 1999.

Curriculum Vitae

Rebekka Brink-Spalink

June 1st, 1986	born in Buxtehude
June 2005	Abitur,
	Halepaghen-Schule Buxtehude
October 2005	Study of Mathematics with minor in Theoretical Physics,
- October 2010	Georg-August-University of Göttingen
September 2008	Erasmus semester in France,
- December 2008	Université de Rennes I
October 2010	Diplom, advisor: Prof. Dr. Robert Schaback
since November 2010	Ph.D. Student and member of the Research Training Group 1644,
	Scaling Problems in Statistics,
November 2010	research assistant at the Department of Animal Breeding
- May 2011	and Genetics,
May 2011	research assistant at the Institute of Mathematical Stochastics,
- December 2014	Georg-August-University of Göttingen,
	Ph.D. project supervised by Prof. Dr. Anja Sturm