

A mixing tree-valued process arising under neutral evolution with recombination

Andrej Depperschmidt, Étienne Pardoux, Peter Pfaffelhuber

▶ **To cite this version:**

Andrej Depperschmidt, Étienne Pardoux, Peter Pfaffelhuber. A mixing tree-valued process arising under neutral evolution with recombination. Electronic Journal of Probability, 2015, 20, pp.1-22. 10.1214/EJP.v20-4286 . hal-01237957

HAL Id: hal-01237957
https://amu.hal.science/hal-01237957

Submitted on 7 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Electron. J. Probab. **20** (2015), no. 94, 1–22.
ISSN: 1083-6489 DOI: 10.1214/EJP.v20-4286

A mixing tree-valued process arising under neutral evolution with recombination

Andrej Depperschmidt[1] Étienne Pardoux[2] Peter Pfaffelhuber[3]

Abstract

The genealogy at a single locus of a constant size N population in equilibrium is given by the well-known Kingman's coalescent. When considering multiple loci under recombination, the ancestral recombination graph encodes the genealogies at all loci in one graph. For a continuous genome $\mathbb{G}$, we study the tree-valued process $(\mathscr{T}_u^N)_{u\in\mathbb{G}}$ of genealogies along the genome in the limit $N\to\infty$. Encoding trees as metric measure spaces, we show convergence to a tree-valued process with càdlàg paths. In addition, we study mixing properties of the resulting process for loci which are far apart.

Keywords: Ancestral recombination graph; Kingman coalescent; tree-valued process; Gromov-Hausdorff metric.
AMS MSC 2010: Primary 92D15, Secondary 60G10; 60K35.
Submitted to EJP on May 5, 2015, final version accepted on September 9, 2015.
Supersedes arXiv:1505.01165.

1 Introduction

A large body of literature within the area of mathematical population genetics is dealing with models for populations of constant size. While finite models such as the Wright-Fisher or the Moran model all have their specificities, the limit of large populations – given some moments are bounded – leads to a unified framework with diffusions and genealogical trees as their main tools; see e.g. Ewens (2004). In finite population models of size N – frequently denoted Cannings models (Cannings, 1974) – the offspring distribution of all individuals in each generation is exchangeable and subject to the constraint of a constant population size.

Neutral evolution accounts for the fact that all individuals have the same chance to produce offspring in next generations. Recombination is the evolutionary force by which genetic material from more than one (i.e. two in all biologically relevant cases) parents

[1] Abteilung für Mathematische Stochastik, Albert-Ludwigs University of Freiburg, Eckerstr. 1, D - 79104 Freiburg, Germany. E-mail: depperschmidt@stochastik.uni-freiburg.de
[2] Aix Marseille Université, Institut de Mathématiques de Marseille (I2M/UMR 7373) 39, rue F. Joliot-Curie, F-13453 Marseille, France, E-mail: pardoux@cmi.univ-mrs.fr
[3] Abteilung für Mathematische Stochastik, Albert-Ludwigs University of Freiburg, Eckerstraße 1, D - 79104 Freiburg, Germany. E-mail: p.p@stochastik.uni-freiburg.de

is mixed in their offspring. Genealogies under neutral evolution without recombination are given through the famous Kingman coalescent (Kingman, 1982), a random binary tree where pairs of lines merge exchangeably in a Markovian fashion. Genealogies under recombination must deal with the fact that recombination events mix up genetical material from the parents. As a consequence, lines not only merge due to joint ancestry, but also split due to different ancestors for the genetic material along the genome. The resulting genealogy is encoded in the *Ancestral Recombination Graph* (ARG), which appeared already in Hudson (1983), but entered the mathematical literature only in Griffiths (1991); Griffiths and Marjoram (1997). This graph gives the genealogies of all genetic loci under stationarity at once; see also Figure 1.

The sequence of genealogies along the chromosome is most important for biological applications, and fast simulation and inference of such genealogies is a major research topic today (Rasmussen et al., 2014). While the ARG gives the sequence of genealogies from the present to the past, a construction of genealogies along the chromosome is possible as well (Wiuf and Hein, 1999; Leocard and Pardoux, 2010). The advantage of the latter approach is that it allows to approximate the full sequence by ignoring long-range dependencies, a fruitful research topic started by McVean and Cardin (2005).

The goal of the present paper is to study the sequence of genealogies along the genome, denoted $\mathbb{G}$, in the limit $N \to \infty$. Precisely, we will use the notion of (ultra-) metric measure spaces, introduced in the probabilistic community by Greven et al. (2009), in order to formalize genealogical trees, read off the sequence $(\mathscr{T}_u^N)_{u\in\mathbb{G}}$ from the ARG and let $N \to \infty$. As main results, we obtain convergence (Theorem 3.3) to an ergodic tree-valued process which has càdlàg paths and study its mixing properties (Theorem 3.5). We start by introducing our notation.

Remark 1.1 (Notation)**.** Let (E, r) be a metric space. We denote by $\mathcal{M}_1(E)$ the space of probability measures on E equipped with the Borel-σ-algebra $\mathcal{B}(E)$. The space $\mathcal{C}_b(E)$ consists of all continuous, bounded, real-valued functions defined on E. For a second metric space (F, r_F) and $\mu \in \mathcal{M}_1(E)$ and a measurable map $\varphi : E \to F$, the measure $\varphi_*\mu$ is the push-forward of μ under φ. We denote vectors $(x_1, x_2, \dots) \in E^{\mathbb{N}}$ by $\underline{x}$ and integrals will be frequently denoted by $\langle \nu, f\rangle := \int f d\nu$. Weak convergence of probability measures will be denoted by $\Rightarrow$. For $I \subseteq \mathbb{R}$, the space $\mathcal{D}_E(I)$ is the set of càdlàg functions $f : I \to E$.

2 Ancestry under recombination

For a set of loci $\mathbb{G}$, also called *genome* in the sequel, we aim to study the ancestry of individuals from a large population. The joint genealogy for all loci is given by the ancestral recombination graph (Section 2.1), from which we can read off genealogical trees at all loci $u \in \mathbb{G}$ (Section 2.2). In Section 2.3, we formalize random genealogies as metric measure spaces.

2.1 The ancestral recombination graph

In this section we give a formal definition of the *ancestral recombination graph (ARG)* which is a (slight) generalization of the definition from Griffiths and Marjoram (1997); see also the leftmost branching and coalescing graph in Figure 1.

Definition 2.1 (N-ancestral recombination graph)**.**

1. *For $a < b$, $\mathbb{G} := [a, b]$, $\rho > 0$ and a finite set $[N] := \{1, \dots, N\}$, the* N-ancestral recombination graph (ARG)*, denoted by $\mathcal{A} := \mathcal{A}^N := \mathcal{A}^N(\mathbb{G})$, starting with particles in the set $[N]$ is defined by the following Markovian dynamics:*

   (i) When there are $k \geq 2$ particles, two randomly chosen particles coalesce

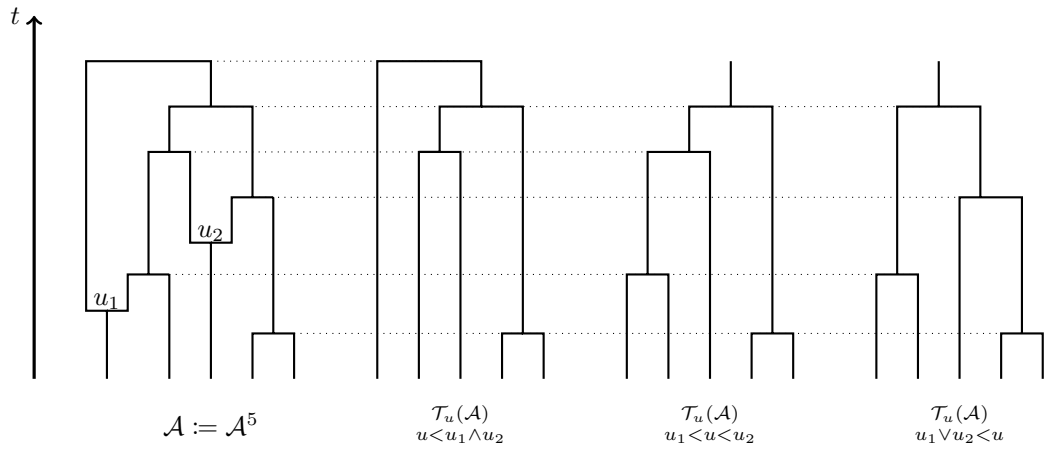


Figure 1: An example of an ARG $\mathcal{A}$ is given for $N = 5$ particles. From the resulting graph, where splitting events are marked by u_1 and u_2 (with $u_1 < u_2$), trees can be read off by using right and left particle at these splitting events. This results in three (realizations of) 5-Kingman coalescents.

(merge) at rate $\binom{k}{2}$ *and give rise to a single new particle.*

(ii) Each particle split*s in two at rate* $\rho(b-a)$*, resulting in a new* left *and a new* right *particle. Such a splitting event is marked by an independent, uniformly distributed random variable* $U \in \mathbb{G}$*.*

Denote by $\mathcal{A}_t$ *the set of particles at time* $t \geq 0$ *and stop when there is only one particle left.*

2. *The particle-counting process* $\mathcal{N} = (N_t)_{t\geq 0}$ *with* $N_t = \#\mathcal{A}_t$ *for* $\mathcal{A}$ *is a birth-and death chain. Precisely,* $\mathcal{N}$ *has birth rate* $b_k = \rho(b-a)k$ *and death rate* $d_k = \binom{k}{2}$*,* $k = 1, 2, \ldots$ *and is stopped at* $T = \inf\{t : N_t = 1\}$*.*

Since the birth rates are linear and the death rates are quadratic, the expectation of the stopping time T is finite; see Theorem 2.1 in Pardoux and Salamat (2009) for an explicit expression.

Remark 2.2 (Interpretation)**.** Clearly, within the above definition, some biological interpretation can be given.

1. The set $\mathbb{G}$ is the *genome*, i.e. the set of all loci (for any individual within the population). An element $u \in \mathbb{G}$ is called a *locus*.
2. The parameter ρ is the *recombination coefficient* per unit length.
3. The set $[N]$ represents N individuals sampled from a population and the particles within $\mathcal{A}_t$ form the ancestry of the individuals from $\mathcal{A}_0$ at time t in the past.
4. A coalescence event within $\mathcal{A}$ indicates joint ancestry.
5. Instead of talking about *left* and *right* particles to follow within the ARG, the biological language would rather suggest to talk about *upstream* and *downstream* genomic sequences.

Remark 2.3 (ARG as a limiting graph, single crossovers)**.**

1. The ARG arises as a limiting object within finite Moran models of population size $\widetilde{N}$ as $\widetilde{N} \to \infty$. In the model with recombination a finite population of $\widetilde{N}$ individuals, each carrying a genetic material indexed by $\mathbb{G}$, undergoes the following dynamics:

   (a) Every (unordered) pair of individuals *resamples* at rate 1, that is, one individuals dies and is replaced by an offspring of the other individual. The offspring carries the same genetic material as the parent.

   (b) At rate $\rho/\widetilde{N}$, every (unordered) pair $\{\ell, j\}$ of individuals *resamples with recombination*, that is, a resampling event occurs, individual j dies, say, a third individual r (chosen uniformly at random from all individuals) and a random U, distributed uniformly on $\mathbb{G}$, is chosen. Then, j is exchanged by an individual, which carries genetic material $[a, U)$ from ℓ and $[U, b]$ from r.

   When considering the history of a sample of $N \ll \widetilde{N}$ individuals, two things can happen: First, if a resampling event of two individuals within the sample is hit, these individuals find a common ancestor, and their ancestral lines coalesce. Second, if a line hits a resampling event with recombination, the history of its genetic material is split at the corresponding U, and follows along two different lines. (Note that this happens at rate $(\widetilde{N}-1)\rho/\widetilde{N} \approx \rho$.) These two lines have a high chance to be outside the sample of N lines if $N \ll \widetilde{N}$. As we see, as $\widetilde{N} \to \infty$, the ancestry is properly described by the ARG as in Definition 2.1.

2. We assume here only *single crossovers*, i.e. the mix of the genetic material of ℓ and r is exactly as just described (rather than taking e.g. $[a, U_1] \cup (U_2, b]$ from ℓ and $(U_1, U_2]$ from r for some random variables $a \le U_1 < U_2 \le b$).

2.2 Trees derived from an ARG

In this section we describe sets of trees that can be read off from an ARG and discuss some of their properties as well as different constructions of the ARG. A construction of the ARG along the genome (see Remark 4.3) will be particularly useful in the sequel and will be explained in more detail in Section 4.1.

Definition 2.4 (Genealogical trees read off from $\mathcal{A}$)**.** *Let $\mathcal{A} = (\mathcal{A}_t)_{t\ge 0}$ be an ancestral recombination graph and $\mathcal{B} \subseteq [N]$ be a subset of the initial particles. For $u \in \mathbb{G}$ we read off the random tree $\mathcal{T}_u := \mathcal{T}_u^{\mathcal{B}} := \mathcal{T}_u^{\mathcal{B}}(\mathcal{A})$ (in the case $\mathcal{B} = [N]$, we also write $\mathcal{T}_u := \mathcal{T}_u^N := \mathcal{T}_u^N(\mathcal{A})$) as follows:*

(i) Start with particles in the set $\mathcal{B}$ and follow particles along $\mathcal{A}$.

(ii) Upon coalescence events within $\mathcal{A}$, followed particles are merged as well. If a coalescence event within $\mathcal{A}$ only involves a single followed particle, continue to follow the coalesced particle.

(iii) Upon a splitting event, consider its mark U. If $u \le U$, follow the left particle, and if $u > U$, follow the right particle.

We denote by $\mathcal{T}_{u,t} := \mathcal{T}_{u,t}^{\mathcal{B}} := \mathcal{T}_{u,t}^{\mathcal{B}}(\mathcal{A})$ the set of particles in $\mathcal{T}_u$ at time t. We denote by $\bullet_u$ the root of $\mathcal{T}_u$ which is the most recent common ancestor (MRCA) of all leaves at locus u.

Remark 2.5 ((N-)coalescent)**.** We will frequently use the notion of an N(-Kingman)-coalescent. This is a random tree arising by the following particle picture: Starting with N particles, each pair of particles coalesces exchangeably at rate 1. (Alternatively, we may say that the total coalescence rate when there are k particles is $\binom{k}{2}$ and upon

a coalescence event, a random pair is chosen to coalesce.) The tree is stopped when reaching a single particle which we denote by $\bullet$ in the sequel. It is well-known (see also Example 2.13) that this random tree converges as $N \to \infty$ to the Kingman's coalescent.

Remark 2.6 (Properties of $\mathcal{T}_u^{\mathcal{B}}(\mathcal{A})$)**.**

1. Since $\mathcal{A}$ is stopped upon reaching a single particle, and no splits occur within $\mathcal{T}_u$, the latter is certainly a tree. Moreover, its root may (see $\mathcal{T}_u$ for $u < u_1 \wedge u_2$ in Figure 1) or may not (see $\mathcal{T}_u$ for $u > u_1 \vee u_2$ in Figure 1) be identical to the node of the stopping particle within $\mathcal{A}$.

2. Note that for $M = \#\mathcal{B}$, each tree $\mathcal{T}_u^{\mathcal{B}}$ is an M-coalescent. Indeed, by exchangeability within Kingman's coalescent, any two particles within this tree coalesce at rate 1, independently of all others.

Remark 2.7 (Unused branches of ARG)**.** If R is the number of recombination events in an ARG $\mathcal{A}$, then we can bound the number of different trees in $(\mathcal{T}_u^N(\mathcal{A}))_{u\in\mathbb{G}}$. When following the left and right branches at each recombination point, we find 2^R different trees. However, since the R recombination points have marks $U_1, \dots, U_R$, there are at most $R+1$ different trees arising from $u < U_{(1)}, U_{(i)} \le u < U_{(i+1)}, i = 1, \dots, R-1$ and $u \ge U_{(R)}$ (where $U_{(i)}$ is the ith order statistic of $U_1, \dots, U_R$). However, it is possible that a branch within $\mathcal{A}$ which is only followed when considering $u \in (v, b]$ (for some $v \in \mathbb{G}$) carries a recombination event with mark $U < v$. In this case, $\mathcal{T}_{v-}^N = \mathcal{T}_{v+}^N$ which reduces the number of different trees. For a lower bound of the number of different trees with R recombination events in $\mathcal{A}$, we find a minimum of two different trees within $(\mathcal{T}_u^N(\mathcal{A}))_{u\in\mathbb{G}}$ if $R > 0$.

This somewhat inefficient procedure of generating recombination events which do not take effect on the level of trees has the advantage of mathematical clarity and has been used by Griffiths and Marjoram (1997). It is also possible to allow only recombination events which are used when reading off the trees $(\mathcal{T}_u^N(\mathcal{A}))_{u\in\mathbb{G}}$; see Hudson (1983). The latter procedure has the advantage of being more efficient in simulations.

Remark 2.8 (Construction of $(\mathcal{T}_u^N)_{u\in\mathbb{G}}$ along the genome)**.** Instead of constructing the process $(\mathcal{T}_u^N)_{u\in\mathbb{G}}$ from the present to the past along the ARG $\mathcal{A}$, Wiuf and Hein (1999) have shown that there is also a construction along the genome. We will recall this approach together with approximations of $(\mathcal{T}_u^N)_{u\in\mathbb{G}}$ related to this construction in Section 4.1.

Remark 2.9 (Outlook on Theorem 3.3)**.** Before we go on with introducing more objects needed to formulate our main results let us give an outlook on one of them.

1. Our goal is to study

$$\text{convergence of the process } (\mathcal{T}_u^N)_{u\in\mathbb{G}} \text{ as } N \to \infty. \tag{2.1}$$

   Since $\mathcal{T}_u^N$ is an N-coalescent for all $u \in \mathbb{G}$, and as the convergence of the N-coalescent to Kingman's coalescent as $N \to \infty$ is well-known, convergence of finite-dimensional distributions in (2.1) is not surprising. However, we will also show tightness of $\{(\mathcal{T}_u^N)_{u\in\mathbb{G}} : N \in \mathbb{N}\}$ in the space of càdlàg paths. This requires to define a proper topology on the space of trees, which we will do in the next section.

2. In our formulation of Theorem 3.3, two different sets of trees derived from an ARG will arise:

   (a) For $\mathcal{A}^N$, we will consider

$$\{\mathcal{T}_u^N(\mathcal{A}^N) : u \in \mathbb{G}\},$$

   which is the set of all trees with N leaves from an N-ARG.

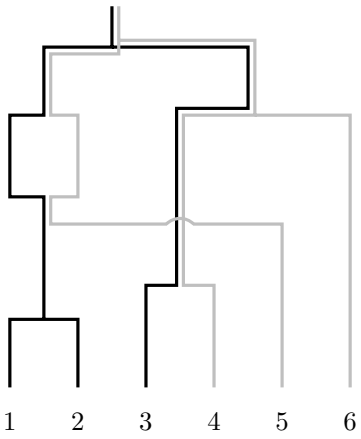


Figure 2: Two (interacting) trees with three leaves each read off from a joint ARG $\mathcal{A}^6$ starting with disjoint sets of leaves. The black tree (at locus $u = 0$) is $\mathcal{T}_0^{\{1,2,3\}}(\mathcal{A}^6)$, while the gray tree (at locus $u = v$) is $\mathcal{T}_v^{\{4,5,6\}}(\mathcal{A}^6)$.

(b) For $\mathcal{A}^n$ and $n_1, \dots, n_j \in \mathbb{N}$ with $n_1 + \cdots + n_j = n$, consider a partition $\{\mathcal{B}_1, \dots, \mathcal{B}_j\}$ of $[n]$ with $\#\mathcal{B}_1 = n_1, \dots, \#\mathcal{B}_j = n_j$ and $u_1, \dots, u_j \in \mathbb{G}$. Then, we consider the trees

$$\{\mathcal{T}_{u_i}^{\mathcal{B}_i}(\mathcal{A}^n) : i = 1, \dots, j\}.$$

These trees arise when considering an n-ARG and partition its initial particles into the sets $\mathcal{B}_1, \dots, \mathcal{B}_j$ and following their ancestry. See Figure 2 for an example of resulting trees and their interaction with each other.

2.3 Space of metric measure spaces

Any (finite or infinite) genealogical tree can be encoded by an (ultra-) metric space (X, r) where X is the set of leaves and r is the genealogical distance. If we equip the set of leaves with a sampling probability measure we obtain a *metric measure space* (X, r, μ). Spaces of metric measure spaces were introduced and their topological properties were studied in Gromov (2007)[1] and Greven et al. (2009). We now recall the space of (isometry classes of) metric measure spaces $\mathbb{M}$, the Gromov-weak topology on $\mathbb{M}$, and polynomials, which form a convergence determining algebra of functions

Definition 2.10 (mm-spaces)**.**

1. *A* metric measure space (mm-space) *is a triple (X, r, μ) where (X, r) is a complete and separable metric space and $\mu \in \mathcal{M}_1(X)$ with supp$(\mu) = X$. Two mm-spaces (X_1, r_1, μ_1) and (X_2, r_2, μ_2) are called* measure-preserving isometric *if there exists an isometry φ between X_1 and X_2 so that $\mu_2 = \varphi_* \mu_1$.*

2. *Being measure preserving isometric is an equivalence relation and we denote the equivalence class of (X, r, μ) by $\overline{(X, r, \mu)}$ and write*

$$\mathbb{M} := \left\{\chi = \overline{(X, r, \mu)} : (X, r, \mu) \text{ is a mm-space}\right\} \tag{2.2}$$

for the space of measure preserving isometry classes of mm-spaces.

In order to define a topology on $\mathbb{M}$, we use the notion of polynomials.

[1] The first edition of this book appeared in 1999. An even older french version from 1981 did not contain the chapter about metric measure spaces.

Definition 2.11 (Distance matrix distribution, polynomials and the Gromov-weak topology)**.** *Let (X,r,μ) be a mm-space and $\mathcal{X}=\overline{(X,r,\mu)}$.*

1. *We define the function $R:=R^{(X,r)}:X^{\mathbb{N}}\to\mathbb{R}_+^{\binom{\mathbb{N}}{2}}$ by $R(\underline{x})=(r(x_i,x_j))_{1\le i<j}$ and define the* distance matrix distribution

$$\nu^{\mathcal{X}}=R_*\mu^{\otimes\mathbb{N}}\in\mathcal{M}_1\Big(\mathbb{R}_+^{\binom{\mathbb{N}}{2}}\Big).$$

2. *A function $\Phi:\mathbb{M}\to\mathbb{R}$ is called* polynomial *if there is a bounded measurable function $\phi:\mathbb{R}_+^{\binom{\mathbb{N}}{2}}\to\mathbb{R}$, depending only on finitely many coordinates, such that for all $\mathcal{X}=\overline{(X,r,\mu)}\in\mathbb{M}$ we have*

$$\Phi(\mathcal{X})=\langle\mu^{\otimes\mathbb{N}},\phi\circ R\rangle=\langle\nu^{\mathcal{X}},\phi\rangle. \tag{2.3}$$

   The smallest n for which there is a function ϕ which depends only on coordinates $(r_{ij})_{1\le i<j\le n}$ so that (2.3) *holds is called the* degree of the polynomial Φ*. We then write $\Phi^{n,\phi}$ instead of Φ to stress the dependence on n and ϕ. The space of bounded continuous polynomials is denoted*

$$\Pi^0:=\Big\{\Phi^{n,\phi}:n\in\mathbb{N},\phi\in\mathcal{C}_b(\mathbb{R}_+^{\binom{\mathbb{N}}{2}})\Big\}. \tag{2.4}$$

   (Sometimes we will abuse notation and write $\phi((r_{ij})_{1\le i<j\le n}):=\phi((r_{ij})_{1\le i<j})$.)

3. *The smallest topology on $\mathbb{M}$ for which all functions Π^0 are continuous is called the* Gromov-weak topology*. For this topology, $\mathcal{X}_n\xrightarrow{n\to\infty}\mathcal{X}$ if and only if $\nu^{\mathcal{X}_n}\overset{n\to\infty}{\Longrightarrow}\nu^{\mathcal{X}}$.*

Remark 2.12 (Some properties of polynomials)**.**

1. We stress that for $\mathcal{X}=\overline{(X,r\mu)}$, the measure $R_*\mu^{\otimes\mathbb{N}}$ does not depend on the representative and hence $\nu^{\mathcal{X}}$ is well-defined.

2. Given two polynomials $\Phi^{n,\phi}$ and $\Psi^{m,\psi}$ one can show that the product $\Phi^{n,\phi}\cdot\Psi^{m,\psi}$ is a polynomial of degree $n+m$; see Remark 2.8(i) in Greven et al. (2013). The space Π^0 is an algebra which separates points; see Section $3\frac{1}{2}.5.$ in Gromov (2007) and Proposition 2.6 in Greven et al. (2009).

3. The space $\mathbb{M}$ equipped with the Gromov-weak topology is Polish. Later in Section 2.4 we will give the Gromov-Prohorov metric on $\mathbb{M}$, which is complete and metrizes the Gromov-weak topology (see Theorem 5 and Proposition 5.6 in Greven et al. (2009)). We will also give the Gromov-Hausdorff metric and other metrics on $\mathbb{M}$, which we will need in the formulation and proof of Theorem 3.3.

Example 2.13 (Kingman coalescent tree as a metric measure space)**.** Consider the N-coalescent from Remark 2.5. Let $K^N=[N]=\{1,\dots,N\}$ be the set of leaves and let the metric r^N be the usual tree distance, i.e. $r^N(i,j)$ is twice the time to the MRCA of i and j, $1\le i,j\le N$. Finally, let μ^N be the uniform measure on K^N. Then $\mathcal{K}^N:=\overline{(K^N,r^N,\mu^N)}$ is an equivalence class of a metric measure space. Furthermore, by Theorem 4 in Greven et al. (2009) there exist an $\mathbb{M}$-valued random variable $\mathcal{K}^\infty$ such that

$$\mathcal{K}^N\overset{N\to\infty}{\Longrightarrow}\mathcal{K}^\infty. \tag{2.5}$$

The limiting object $\mathcal{K}^\infty$ is called the *Kingman measure tree*.

2.4 Metrics on metric (measure) spaces

We now recall the definitions of several distances on $\mathbb{M}$ that we will use in the sequel. While the Hausdorff distance is a metric on closed subsets of a metric space, the Prohorov and total variation distances are metrics on probability measures on a metric space. All three distances can be turned into distances on $\mathbb{M}$.

Definition 2.14 (Hausdorff, Prohorov and total variation distances)**.** *Let (Z,d) be a metric space and $\mu_1,\mu_2 \in \mathcal{M}_1(Z)$.*

1. *The* Hausdorff distance *between two non-empty subsets $A,B \subseteq Z$ is defined by*

$$d_{\mathrm{H}}(A,B) := \inf\{\varepsilon > 0 : A \subseteq B^\varepsilon, B \subseteq A^\varepsilon\}. \tag{2.6}$$

   where for $A \subseteq Z$

$$A^\varepsilon := \{x \in Z : d(x,A) < \varepsilon\} = \{z \in Z : \exists y \in A, d(y,z) < \varepsilon\}.$$

2. *The* Prohorov distance *of μ_1,μ_2 is defined by*

$$d_{\mathrm{P}}(\mu_1,\mu_2) := \inf\{\varepsilon > 0 : \mu_1(F) \le \mu_2(F^\varepsilon) + \varepsilon, \forall F \subseteq Z,\ \textit{closed}\}. \tag{2.7}$$

3. *The* total variation distance *of μ_1,μ_2 is defined by*

$$d_{\mathrm{TV}}(\mu_1,\mu_2) := \sup_{A \in \mathcal{B}(Z)} |\mu_1(A) - \mu_2(A)|. \tag{2.8}$$

Remark 2.15 (Total variation distance)**.**

1. If Z is finite, the total variation distance of probability measures μ_1 and μ_2 on Z is given by

$$d_{\mathrm{TV}}(\mu_1,\mu_2) = \frac{1}{2}\sum_{z \in Z} |\mu_1(\{z\}) - \mu_2(\{z\})|. \tag{2.9}$$

2. Recall that the Prohorov distance of two probability measures is bounded by their total variation distance.

For all three notions just defined, we now recall the corresponding "Gromov-versions" which are (semi-)metrics on $\mathbb{M}$. The idea is always to find an optimal isometric embedding into a third metric space and compute there the usual distance of the images of the spaces and measures.

Definition 2.16 (Gromov distances)**.** *Let $\mathcal{X}_1 = \overline{(X_1,r_1,\mu_1)}$ and $\mathcal{X}_2 = \overline{(X_2,r_2,\mu_2)}$ be mm-spaces. Moreover, let $\varphi_1 : X_1 \to Z$ and $\varphi_2 : X_2 \to Z$ be isometric embeddings into a common complete and separable metric space (Z,d).*

1. *The* Gromov-Hausdorff distance *of $\mathcal{X}_1$ and $\mathcal{X}_2$ is defined by*

$$d_{\mathrm{GH}}(\mathcal{X}_1,\mathcal{X}_2) := \inf_{\varphi_1,\varphi_2,Z} d_{\mathrm{H}}(\varphi_1(X_1),\varphi_2(X_2)).$$

2. *The* Gromov-Prohorov metric *of $\mathcal{X}_1$ and $\mathcal{X}_2$ is defined by*

$$d_{\mathrm{GP}}(\mathcal{X}_1,\mathcal{X}_2) := \inf_{\varphi_1,\varphi_2,Z} d_{\mathrm{P}}((\varphi_1)_*\mu_1,(\varphi_2)_*\mu_2).$$

3. *The* Gromov total variation distance *of $\mathcal{X}_1$ and $\mathcal{X}_2$ is defined by*

$$d_{\mathrm{GTV}}(\mathcal{X}_1,\mathcal{X}_2) := \inf_{\varphi_1,\varphi_2,Z} d_{\mathrm{TV}}((\varphi_1)_*\mu_1,(\varphi_2)_*\mu_2).$$

Remark 2.17 (Properties)**.** Let us recall or rather state some known and some obvious properties of the distances just introduced.

1. The distances $d_{\mathrm{GH}}, d_{\mathrm{GP}}, d_{\mathrm{GTV}}$ are well-defined. As can be seen by considering isometries between elements of one isometry class the distances do not depend on the representative.

2. Since the Gromov-Hausdorff distance only uses the metric spaces (X_1, r_1) and (X_2, r_2) but not the measures μ_1, μ_2, it is only a pseudo-metric on $\mathbb{M}$.

3. According to Lemma 5.4 and Proposition 5.6 in Greven et al. (2009) the Gromov-Prohorov metric d_{GP} is indeed a metric on $\mathbb{M}$ and the metric space $(\mathbb{M}, d_{\mathrm{GP}})$ is complete and separable. Moreover, it metrizes the Gromov-weak topology by Theorem 5 of Greven et al. (2009).

4. Since the Prohorov distance is bounded by the total variation distance, we also find that

$$d_{\mathrm{GP}}(\mathcal{X}_1, \mathcal{X}_2) \leq d_{\mathrm{GTV}}(\mathcal{X}_1, \mathcal{X}_2). \tag{2.10}$$

   This will be useful later since the total variation distance is usually easier to compute or estimate than the Prohorov distance.

3 Main results

We now formalize the trees $\{\mathcal{T}_u^N(\mathcal{A}^N) : u \in \mathbb{G}\}$ as mm-spaces. Our results, then, are dealing with these $\mathbb{M}$-valued random processes. In particular, Theorem 3.3 studies convergence as $N \to \infty$. Since $\mathcal{T}_u^N(\mathcal{A}^N)$ is an N-coalescent for all $u \in \mathbb{G}$, the resulting process is stationary. It is even mixing, as Theorem 3.5 shows.

Definition 3.1 ($\mathcal{T}_u$ as an mm-space)**.** *Let $\mathcal{A} = \mathcal{A}^N(\mathbb{G})$ for $N \in \mathbb{N}$ and $\mathbb{G} = [a, b]$ with $a < b$ be an N-ARG. For $u \in \mathbb{G}$ and $\mathcal{B} \subseteq [N]$, let $\mathcal{T}_u^{\mathcal{B}}(\mathcal{A})$ be as in Definition 2.4. As in Example 2.13, let r be the usual tree-distance and μ the uniform measure on $\mathcal{B}$. Then, $(\mathcal{B}, r, \mu)$ is a metric measure space. Its isometry class will be denoted $\mathcal{T}_u := \mathcal{T}_u^{\mathcal{B}} := \mathcal{T}_u^{\mathcal{B}}(\mathcal{A})$ in the sequel. If $\mathcal{B} = [N]$ we write $\mathcal{T}_u := \mathcal{T}_u^N := \mathcal{T}_u^N(\mathcal{A})$.*

In Theorem 3.3, we need the notion of the variation of a function which we briefly recall.

Remark 3.2 (Variation)**.** Let (E, r) be a metric space and $f : I \to E$ for $I \subset \mathbb{R}$. The *variation of f with respect to r* on subintervals $[a, b] \subset I$ is defined by

$$\mathrm{V}_a^b(f) := \sup\Big\{\sum_{i=1}^{k} r(f(t_i), f(t_{i-1})) : \ k \in \mathbb{N},\ a = t_0 < t_1 < \cdots < t_k = b\Big\}. \tag{3.1}$$

Theorem 3.3 (Convergence of N-ARGs)**.** *Let $\mathcal{T}^N := (\mathcal{T}_u^N(\mathcal{A}))_{u\in\mathbb{G}}$ be as in Definition 3.1. Then, $\mathcal{T}^N \xRightarrow{N\to\infty} \mathcal{T}$ on $\mathcal{D}_{\mathbb{M}}(\mathbb{G})$ for some process $\mathcal{T}$. The (law of the) process $\mathcal{T} = (\mathcal{T}_u)_{u\in\mathbb{G}}$ is uniquely given as follows:*
For each $j \in \mathbb{N}$, $u_1, \ldots, u_j \in \mathbb{G}$, $n_1, \ldots, n_j \in \mathbb{N}$, let $\mathcal{T}_{u_i}^{\mathcal{B}_i}$ be as in Remark 2.9 and $\underline{\underline{R}}_i$ be the distances of leaves $\mathcal{B}_i$ within $\mathcal{T}_{u_i}^{\mathcal{B}_i}$. Then, for $\Phi_i = \Phi^{n_i,\phi_i} \in \Pi^0, i = 1, \ldots, j$,

$$\mathbb{E}[\Phi_1(\mathcal{T}_{u_1}) \cdots \Phi_j(\mathcal{T}_{u_j})] = \mathbb{E}[\phi_1(\underline{\underline{R}}_1) \cdots \phi_j(\underline{\underline{R}}_j)]. \tag{3.2}$$

The paths of $\mathcal{T}$ are almost surely of finite variation with respect to Gromov-Prohorov, Gromov total variation and Gromov-Hausdorff metrics.

Remark 3.4 (Path-properties of $\mathcal{T}$)**.** We can ask about path-properties of the limiting process $\mathcal{T}$. Let us give an example: Let N_u^ε be the number of ε-balls that are needed to cover $\mathcal{T}_u$. In other words $N_{u,}^\varepsilon$ is the number of families in $\mathcal{T}_u$ at some distance ε from the leaves. For fixed $u \in \mathbb{G}$, we know that $\mathcal{T}_u$ is a Kingman coalescent and hence we have $\varepsilon N_u^\varepsilon \xrightarrow{\varepsilon\to 0} 2$ almost surely. (See e.g. (35) in Aldous (1999).) Is it also true that

$$\mathbb{P}(\varepsilon N_u^\varepsilon \xrightarrow{\varepsilon\to 0} 2 \text{ for all } u \in \mathbb{G}) = 1?$$

Similar questions arise for well-known almost sure properties of Kingman's coalescent regarding the family-sizes near the leaves; see Aldous (1999), Chapter 4.2.

For our next result, we need to extend the ARG to $\mathbb{G} = (-\infty, \infty)$. This can be done using some projective property (see also Lemma 4.11): Let $\mathbb{G}_n \uparrow (-\infty, \infty)$. Clearly, (3.2) gives the finite-dimensional distributions of $(\mathcal{T}_u)_{u\in\mathbb{G}_n}$ for every n. In particular, for $m < n$, we see that the projection of $(\mathcal{T}_u)_{u\in\mathbb{G}_n}$ to $\{u \in \mathbb{G}_m\}$ is the same as $(\mathcal{T}_u)_{u\in\mathbb{G}_m}$ and therefore (3.2) defines a projective family of probability measures which can by Kolmogorov's extension Theorem be extended to the law of $(\mathcal{T}_u)_{u\in\mathbb{R}}$.

Theorem 3.5 (Mixing properties)**.** *For $n \in \mathbb{N}$ let $\Psi = \Psi^{n,\psi}$ and $\Phi = \Phi^{n,\phi}$ be polynomials of (at most) degree n and $(\mathcal{T}_u)_{u\in\mathbb{G}}$ be the extension of the limit process $\mathcal{T}$ from Theorem 3.3 to $\mathbb{G} = \mathbb{R}$. Then, there exists a positive finite constant C only depending on n such that*

$$|\mathbb{E}[\Psi(\mathcal{T}_0)\Phi(\mathcal{T}_u)] - \mathbb{E}[\Psi(\mathcal{T}_0)]\mathbb{E}[\Phi(\mathcal{T}_u)]| \le \frac{C}{\rho^2 u^2}\|\psi\|_\infty\|\phi\|_\infty, \quad u > 0. \tag{3.3}$$

Remark 3.6 (Dependency on n)**.** In our proof we will show that (3.3) holds with C replaced by $\frac{2n^4}{9+7\rho u+\rho^2u^2} \le 2n^4/9$. Therefore, the bound is only useful in the limit $u \to \infty$ for fixed n. If $n \to \infty$ and u is fixed, the trivial bound $|\mathbb{E}[\Psi(\mathcal{T}_0)\Phi(\mathcal{T}_u)] - \mathbb{E}[\Psi(\mathcal{T}_0)]\mathbb{E}[\Phi(\mathcal{T}_u)]| \le 2\|\psi\|_\infty\|\phi\|_\infty$ holds as well. It would be desirable to obtain a bound similar to (3.3) uniformly in n and u, but this seems to be out of reach with the techniques we develop here.

4 Preliminaries

4.1 Construction of $(\mathcal{T}_u^N)_{u\in\mathbb{G}}$ along the genome

In Wiuf and Hein (1999), a construction of an N-ARG was given, which results in the same trees $(\mathcal{T}_u^N)_{u\in\mathbb{G}}$ (or $(\mathcal{T}_u^N)_{u\in\mathbb{G}}$) in distribution as described in Definition 2.4 (and Definition 3.1), but constructs this tree-valued process *along the genome*, i.e. by starting at $u \in \mathbb{G}$ and then letting the trees evolve when moving along $\mathbb{G}$. This construction which we recall below will be helpful in all further proofs. See also Leocard and Pardoux (2010).

Definition 4.1 (N-ARG' along the genome)**.**

1. *For $a < b$, $\mathbb{G} = [a, b]$ and $N \in \mathbb{N}$, construct an evolving pair $(\mathcal{G}, \mathcal{T}) = (\mathcal{G}_n, \mathcal{T}_n)_{n=0,1,2,\dots}$, where $\mathcal{G}_n$ is a graph and $\mathcal{T}_n$ is a tree as follows:*
2. *Start with an N-coalescent $\mathcal{G}_0 = \mathcal{T}_0$ with set of leaves $[N]$ (which is continued indefinitely, i.e. not stopped upon hitting a single line) and in each step do the following (with $U_0 = a$):*
   (a) *Measure the length of the (vertical branches of the) graph $L_n = L(\mathcal{G}_n)$, choose a uniformly distributed point $X_n \in \mathcal{G}_n$ and an independent exponential random variable ξ_n.*
   (b) *From X_n, change the graph as follows: Create a split event at X_n, and mark it by $U_n = U_{n-1} + \xi_n/(L_n\rho)$. The line in $\mathcal{G}_n$ which starts at X_n is called the* left

branch *and a new* right branch *is created. Starting from this recombination event, if we follow the new branch into the past (i.e. away from the leaves), it coalesces with one of the other branches of* $\mathcal{G}_n$ *as in Kingman's coalescent tree, i.e. this branch coalesces with any of the other branches at rate 1, the various possible coalescence times being mutually independent (and of course only the shortest one is effective). The resulting graph is* $\mathcal{G}_{n+1}$ *and* $\mathcal{T}_{n+1}$ *is given by following* $\mathcal{T}_n$ *until the right branch at* X_n *is hit and waiting until this branch coalesces back with* $\mathcal{T}_n$*. (There is a chance that* $X_n \notin \mathcal{T}_n$*; in this case we have* $\mathcal{T}_{n+1} = \mathcal{T}_n$*.)*

Stop when $U_n > b$ *and set* $\mathcal{G} = \mathcal{G}_n$*.*

3. *Let us now consider the final graph* $\mathcal{G}$*. This is a coalescing-splitting random graph (similar to* $\mathcal{A}$ *in Definition 2.1) and therefore can also be considered as an evolving set of particles which coalesce and split, where splitting events are marked by some element of* $\mathbb{G}$ *and are continued by a* left *and* right *branch. Hence, we can define for a subset* $\mathcal{B} \subseteq [N]$ *and* $u \in \mathbb{G}$ *the random tree* $\mathcal{S}_u := \mathcal{S}_u^{\mathcal{B}} := \mathcal{S}_u^{\mathcal{B}}(\mathcal{G})$ *as in Definition 2.4. Again, we set* $\mathcal{S}_u^N := \mathcal{S}_u^{[N]}$*.*

Remark 4.2 (Properties of $(\mathcal{S}_u^{\mathcal{A}_0})_{u\in\mathbb{G}}$)**.**

1. From Wiuf and Hein (1999), the graphs $\mathcal{A}$ and $\mathcal{G}$ have the same distribution, hence the same is true for the processes $(\mathcal{T}_u^N)_{u\in\mathbb{G}}$ and $(\mathcal{S}_u^N)_{u\in\mathbb{G}}$.

2. Let $\mathcal{L} \subseteq \mathcal{S}_v^N$ be a line of length ℓ starting in some leaf $x \in [N]$ and reaching in the direction of the root. Then, $\mathcal{L} \subseteq \mathcal{S}_w^N$ for some $w \in \mathbb{G}$ if and only if no recombination marks $\mathcal{L}$ before reaching w. By construction, this happens with probability $e^{-\rho|w-v|\ell}$. Moreover, let $u < v < w$. Then, (given $\mathcal{S}_v^N$), $\mathcal{L} \subseteq \mathcal{S}_u^N$ with probability $e^{-\rho|v-u|\ell}$, independent of the event $\mathcal{L} \subseteq \mathcal{S}_w^N$.

3. By construction, $\mathcal{G}$ from above is an ARG and therefore, its length has finite expectation. Hence, the construction above terminates almost surely.

Remark 4.3 (Approximating the ARG using a Markov process along the genome)**.** One striking feature of the construction of Wiuf and Hein (1999) is that it allows to approximate $(\mathcal{T}_u^N)_{u\in\mathbb{G}}$ by a Markov process by changing the dynamics of the construction $(\mathcal{G}_n, \mathcal{T}_n)$ in order to obtain a Markovian dynamics $\mathcal{T} = (\mathcal{T}_n)$:

1. The following approximation was used by McVean and Cardin (2005): Instead of (a) in Definition 4.1, measure the length of the (vertical branches of the) tree $L_n = L(\mathcal{T}_n)$, choose a uniformly distributed point $X_n \in \mathcal{T}_n$ and an independent exponential random variable ξ_n. Then, instead of (b), change the graph as follows: Create a split event at X_n, and mark it by $U_n = U_{n-1} + \xi_n/(L_n\rho)$. Delete the branch which connects X_n to its ancestral node from $\mathcal{T}_n$, all other lines are available for coalescence. Then, start a new branch in X_n which coalesces with all other available branches rate 1. The resulting tree is $\mathcal{T}_{n+1}$. In particular, by only allowing coalescences with $\mathcal{T}_u^N$, this approximation becomes a Markov process, also called the Sequentially Markov Coalescent (SMC). (In SMC', Chen et al. (2009) use almost the same construction but without deleting the branch connecting X_n to its ancestral node for the set of lines available for coalescence, leading to a better approximation.)

2. Another simulation software based on the Markovian Coalescent Simulator (MaCS) from Chen et al. (2009), has all lines from the last k genealogical trees as *available* for coalescence after a recombination event.

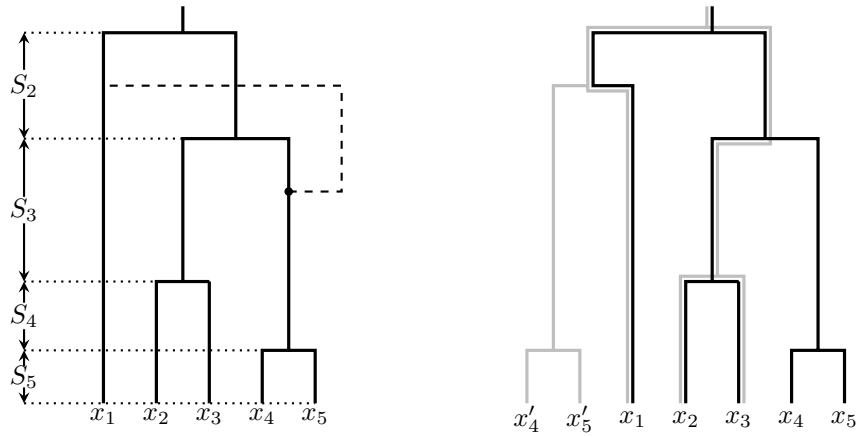


Figure 3: A conditional embedding of the trees $\mathcal{S}_u^5$ and $\mathcal{S}_v^5$ into a common tree.

While approximations such as SMC and MaCS make genome-wide computer simulations under recombination feasible, their construction differs from the ARG at least for loci which are far apart. In particular, Theorem 3.5 would not be true for these approximations.

4.2 Conditional distances of trees and first upper bounds

To obtain useful bounds on expected tree distances introduced in the previous section we will condition on one of the trees and use the construction of the tree-valued process along the genome from Section 4.1. Let us start with an illustrative example.

Example 4.4 (A conditional embedding of two trees and upper bounds; Figure 3)**.** Given the tree $\mathcal{S}_u^5$ drawn in solid lines on the left of Figure 3, we can read off the times $S_2, \dots, S_5$ which the tree spends with exactly $2, \dots, 5$ lines. These random variables are independent and each S_k is exponentially distributed with mean $1/\binom{k}{2}$. The black bullet indicates a recombination event with a mark between u and v, and $\mathcal{S}_v^5$ can be read off by following the dashed line.

On the right hand side of Figure 3 the trees $\mathcal{S}_u^5$ (black) and $\mathcal{S}_v^5$ (gray) are embedded into a common tree. The embedding of $\mathcal{S}_v^5$ into the common tree is given by $x_i \mapsto x_i, i = 1, 2, 3$ and $x_i \mapsto x'_i, i = 4, 5$. Using (2.10) and (4.1) below it is easily seen that if the trees $\mathcal{S}_u^5$ and $\mathcal{S}_v^5$ are equipped with uniform probability measure, we have

$$d_{\mathrm{GP}}(\mathcal{S}_u^5, \mathcal{S}_v^5) \le d_{\mathrm{GTV}}(\mathcal{S}_u^5, \mathcal{S}_v^5) \le 2/5.$$

From the same embedding we see also that the Gromov-Hausdorff distance of $\mathcal{S}_u^5$ and $\mathcal{S}_v^5$, conditioned on $\mathcal{S}_u^5$ is bounded by the tree distance of $\{x'_4, x'_5\}$ and x_1 which is twice the time of back coalescence of the dashed recombination line. Thus, in this particular case we have $d_{\mathrm{GH}}(\mathcal{S}_u^5, \mathcal{S}_v^5) \le 2(S_2 + \cdots + S_5)$. It could of course happen that the dashed line coalesces back after the time of the MRCA of $\mathcal{S}_u^5$. In this case the upper bound would be $2(S_1 + \cdots + S_5)$, where S_1 is another independent exponentially distributed random variable with mean 1.

First, we explain a way to compute the Gromov total variation distance explicitly.

Remark 4.5 (How to compute d_{GTV})**.** If the spaces X_1, X_2 are finite with $\#X_1 = \#X_2 = N$ and μ_i is the uniform distribution on $X_i, i = 1, 2$, then one can compute the Gromov total variation distance explicitly. There are isometric embeddings φ_i of X_i into a common finite metric space (Z, d), $i = 1, 2$, so that Z can be decomposed in three disjoint

sets $Z = Z_1 \uplus Z_2 \uplus Z_{\text{joint}}$ with the property $\varphi_i(X_i) = Z_i \cup Z_{\text{joint}}$, $i = 1, 2$. The optimal embeddings are the ones for which $\#Z_{\text{joint}}$ is maximal. Using such optimal embeddings we have

$$d_{\text{GTV}}(\mathcal{X}_1, \mathcal{X}_2) = \frac{1}{2}\mu_1(\varphi_1^{-1}(Z_1)) + \frac{1}{2}\mu_2(\varphi_2^{-1}(Z_2)) + \frac{1}{2}\sum_{z \in Z_{\text{joint}}} |\mu_1(\varphi_1^{-1}(\{z\})) - \mu_2(\varphi_2^{-1}(\{z\}))| = \frac{\#Z_1}{N} = \frac{\#Z_2}{N}. \quad (4.1)$$

As it turns out, to obtain upper bounds on Gromov total variation and therefore also the Gromov-Prohorov distances it is helpful to introduce yet another distance which is only well-defined on trees from the processes $(\mathcal{T}_u^{\mathcal{A}_0})_{u \in \mathbb{G}}$.

Definition 4.6 (Auxiliary distance)**.** *Let $(\mathcal{T}_u^N)_{u \in \mathbb{G}}$ and $\mathcal{A}$ be as in Definition 2.1. Recall the MRCA within $\mathcal{T}_u^N$, denoted by $\bullet_u$, from Definition 2.4. For $u, v \in \mathbb{G}$, we decompose $[N]$ into two subsets:*

1. *We denote by $[N]_{\bullet_u, v}$ the set of all $x \in [N]$ such that the path within $\mathcal{T}_u^N$ from x to $\bullet_u$ is* not *hit by a splitting event marked with $U \in [u \wedge v, u \vee v]$.*
2. *Then $[N]^c_{\bullet_u, v}$ is the set of all $x \in [N]$ with such a splitting event on the path from x to $\bullet_u$ in $\mathcal{T}_u^N$.*

We define

$$d_{\text{aux}}^{\bullet_u, v}(\mathcal{T}_u^N, \mathcal{T}_v^N) := \frac{\#[N]^c_{\bullet_u, v}}{N}. \quad (4.2)$$

Proposition 4.7 (Properties of the auxuliary distance)**.** *Let d_{GTV} be as in Definition 2.16, d_{aux}, as in Definition 4.6, $(\mathcal{T}_u^N)_{u \in \mathbb{G}}$ as in Definition 2.4 and $\mathcal{T}_u^N = \overline{\mathcal{T}_u^N}$ as in Definition 3.1.*

1. *The distance d_{aux} is an upper bound for d_{GTV} in the sense that, for all $u, v \in \mathbb{G}$*
$$d_{\text{GTV}}(\mathcal{T}_u^N, \mathcal{T}_v^N) \le d_{\text{aux}}^{\bullet_u, v}(\mathcal{T}_u^N, \mathcal{T}_v^N)$$
2. *For $u, v, w \in \mathbb{G}$ with $u < v < w$, the distances $d_{\text{aux}}^{\bullet_v, u}(\mathcal{T}_u^N, \mathcal{T}_v^N)$ and $d_{\text{aux}}^{\bullet_v, w}(\mathcal{T}_v^N, \mathcal{T}_w^N)$ are independent given $\mathcal{T}_v^N$.*

Proof. 1. Since the Gromov-Total-Variation distance is defined as the infimum over all embeddings into a common mm-space, it suffices to bound the Total-Variation distance on a concrete embedding. Therefore, we use Remark 4.5 and write $Z = Z_1 \uplus Z_2 \uplus Z_{\text{joint}}$ with $Z_{\text{joint}} = [N]_{\bullet_u, v}$ and Z_1, Z_2 are two copies of $[N]^c_{\bullet_u, v}$. Distances on Z_{joint} within $\mathcal{T}_u^N$ and $\mathcal{T}_v^N$ are identical by construction and we see from (4.1) that

$$d_{\text{GTV}}(\mathcal{T}_u^N, \mathcal{T}_v^N) \le d_{\text{TV}}(\mathcal{T}_u^N, \mathcal{T}_v^N) \le \frac{\#Z_1}{N} = \frac{\#[N]^c_{\bullet_u, v}}{N} = d_{\text{aux}}^{\bullet_u}(\mathcal{T}_u^N, \mathcal{T}_v^N).$$

2. From Definition 4.1 and Remark 4.2, we see that the triple $(\mathcal{T}_u^N, \mathcal{T}_v^N, \mathcal{T}_w^N)$ can be constructed starting with $\mathcal{T}_v^N$ (with $a = v, b = w$). The resulting graph $\mathcal{G}$ is then used to again use the same procedure with initial state $(\mathcal{G}, \mathcal{T}_v^N)$ and use the same procedure (with $a = v$ and $b = u$) this time moving to the left of v. In total, this results in marking $\mathcal{T}_v^N$ at rates $\rho(w-v)$ and $\rho(v-u)$ independently. Leaves which are marked at rate $\rho(w-v)$ ($\rho(v-u)$) until $\bullet_v$ is hit, are elements of $[N]^c_{\bullet_v, u}$ ($[N]_{\bullet_v, w}$). Importantly, since the marking on $[u, v]$ and $[v, w]$ are independent (given $\mathcal{T}_v^N$) – see Remark 4.2 – the claim follows. □

Remark 4.8 (Bound of d_{GTV} by d_{aux} not sharp)**.** It can be easily seen that the strict inequality $d_{\text{GTV}}(\mathcal{T}_u^N, \mathcal{T}_v^N) < d_{\text{aux}}^{\bullet_u, v}(\mathcal{T}_u^N, \mathcal{T}_v^N)$ is possible. This happens for instance if the dashed line in Figure 3 coalesces immediately with the same line.

Proposition 4.9 (Bounds on d_{aux})**.** *Let $v, w \in \mathbb{G}$ and $(\mathcal{T}_u^N)_{u\in\mathbb{G}}$ be as in Definition 4.1. For $\mathcal{T}_v^N$, let $S_2, S_3, \dots$ be the duration for which $2, 3, \dots$ lines are present in the tree. We have*

$$\mathbb{E}\big[d_{\mathrm{aux}}^{\bullet_v,w}(\mathcal{T}_v^N, \mathcal{T}_w^N)|\mathcal{T}_v^N\big] \le \rho|w-v| \sum_{k=2}^{N} S_k. \tag{4.3}$$

Proof. For $x \in \mathcal{A}_0$, let $\mathcal{L}_x \subseteq \mathcal{T}_v^N$ be the path from x to $\bullet_v$. Its length is given by the tree height $L \coloneqq S_2 + \cdots + S_N$.

Given $\mathcal{T}_v^N$, $x \in [N]_{\bullet_v,w}$ with probability $1 - e^{-\rho|w-v|L}$ by Remark 4.2.2. Hence, by exchangeability and the definition of d_{aux},

$$\mathbb{E}\big[d_{\mathrm{aux}}^{\bullet_v,w}(\mathcal{T}_v^N, \mathcal{T}_w^N)|\mathcal{T}_v^N\big] = \frac{1}{N}\sum_{x\in[N]} \mathbb{P}(x \in [N]^c_{\bullet_v,w}) = 1 - e^{-\rho|w-v|L} \le \rho|w-v|L.$$

□

4.3 Projective properties of the ARG

It is well-known that Kingman's coalescent is projective in the sense that the tree spanned by a sample of n leaves from an N-coalescent has the same distribution as an n-coalescent. The same holds for the ARG as we will show next.

Lemma 4.10 (Projectivity in N for the N-ARG)**.** *Let $\mathbb{G} = [a, b]$, $\rho > 0$ and $\mathcal{A}$ be an N-ARG (with $\mathcal{A}_0 = [N] = \{1, \dots, N\}$). Let $\mathcal{B} \subseteq [N]$ with $\#\mathcal{B} = n$ and let $\pi_{\mathcal{B}}\mathcal{A}^N$ be the random graph which arises from the particle system starting with particles $\mathcal{B}$ and following them along $\mathcal{A}$. (Upon coalescence events within $\mathcal{A}^N$, followed particles merge as well. If a coalescence event within $\mathcal{A}$ only involves a single followed particle, continue to follow the coalesced particle. Splitting events hitting a followed particle are followed as well.) Then, $\pi_{\mathcal{B}}\mathcal{A}^N$ is an n-ARG.*

Proof. It suffices to consider the particle-counting process of $\pi_{\mathcal{B}}\mathcal{A}^N$ since the fine-structure of $\mathcal{A}^N$ is exchangeable. Clearly, pairs of particles coalesce at rate 1 and every particle splits at the same rate $\rho(b-a)$. Hence, it coincides with the particle-counting process of $\mathcal{A}^n$ and we are done. □

The projectivity of the N-ARG along the genome is stated next:

Lemma 4.11 (Projectivity in $\mathbb{G}$ of the N-ARG)**.** *Let $\mathbb{G} = [a, b]$, $\rho > 0$, $\mathcal{A}^N(\mathbb{G})$ be an N-ARG and $\mathbb{H} = [c, d] \subseteq \mathbb{G}$. Let $\pi_{\mathbb{H}}\mathcal{A}^N$ be the random graph which arises from the particle system starting with particles $[N]$ and following them along $\mathcal{A}$. (Upon coalescence events within $\mathcal{A}^N$, followed particles merge as well. If a coalescence event within $\mathcal{A}$ only involves a single followed particle, continue to follow the coalesced particle. Splitting events hitting a followed particle are followed as well if the mark falls in $\mathbb{H}$.) Then, $\pi_{\mathbb{H}}\mathcal{A}^N$ equals $\mathcal{A}^N(\mathbb{H})$ in distribution.*

Proof. It suffices to see that (i) the recombination events at loci in $[a, b] \setminus [c, d]$ split off ancestral material not in $[c, d]$ and therefore don't change the genealogical trees in $\pi_{\mathbb{H}}\mathcal{A}^N$ and (ii) coalescences with such lines don't appear in genealogical trees in $\mathbb{H}$. Leaving out these recombination events hence leads to $\mathcal{A}^N(\mathbb{H})$, so the claimed equality follows. □

5 Proof of Theorem 3.3

The proof of Theorem 3.3 requires three steps. First, in order to obtain existence of limiting processes along subsequences of $(\mathcal{T}^N)_{N=1,2,\dots}$, we have to prove (see Section 5.1)

$$\text{The family } (\mathcal{T}^N)_{N=1,2,\dots} \text{ is tight in } \mathcal{D}_{\mathbb{M}}(\mathbb{G}). \tag{5.1}$$

Second, we show in Section 5.2

$$\text{For any limiting process } \mathcal{T} \text{ along a subsequence of } (\mathcal{T}^N)_{N=1,2,\dots}, \text{ (3.2) holds.} \tag{5.2}$$

This equation determines uniquely the finite-dimensional distributions of $\mathcal{T}$ since polynomials are separating; in particular, since the right hand side of (3.2) does not depend on the subsequence, uniqueness of the limiting process follows. Third, bounds on the variation process of $\mathcal{T}$ are given in Section 5.3 such that

$$\text{The paths of } \mathcal{T} \text{ are a.s. of finite variation with respect to Gromov total variation, Gromov-Prohorov and Gromov-Hausdorff metrics.} \tag{5.3}$$

5.1 Tightness in $\mathcal{D}_{\mathbb{M}}(\mathbb{G})$

In order to prove tightness in the sense (5.1), we rely on Theorem 13.6 in Billingsley (1999) (see also Theorem 3.8.8 in Ethier and Kurtz (1986)), i.e. we will show that there is $C > 0$ such that

$$\limsup_{N\to\infty} \mathbb{E}[d_{\mathrm{GTV}}(\mathcal{T}^N_{-h}, \mathcal{T}^N_0) \cdot d_{\mathrm{GTV}}(\mathcal{T}^N_0, \mathcal{T}^N_h)] \le Ch^2. \tag{5.4}$$

Since $d_{\mathrm{GP}} \le d_{\mathrm{GTV}}$, this implies tightness with respect to the Gromov-weak topology.

We assume without loss of generality that the interval $[-h, h]$ is contained in $\mathbb{G}$. Note also that $(\mathcal{T}^N_u)_{u\in\mathbb{G}}$ is stationary. We combine Proposition 4.7 and Proposition 4.9 and write, for $S_k \sim \mathrm{Exp}\binom{k}{2}$,

$$\begin{aligned}
\mathbb{E}[d_{\mathrm{GTV}}(\mathcal{T}^N_{-h}, \mathcal{T}^N_0) \cdot d_{\mathrm{GTV}}(\mathcal{T}^N_0, \mathcal{T}^N_h)] &\le \mathbb{E}[d^{\bullet_0,-h}_{\mathrm{aux}}(\mathcal{T}^N_{-h}, \mathcal{T}^N_0) \cdot d^{\bullet_0,h}_{\mathrm{aux}}(\mathcal{T}^N_0, \mathcal{T}^N_h)] \\
&= \mathbb{E}\big[\mathbb{E}[d^{\bullet_0,-h}_{\mathrm{aux}}(\mathcal{T}^N_{-h}, \mathcal{T}^N_0)|\mathcal{T}^N_0] \cdot \mathbb{E}[d^{\bullet_0,h}_{\mathrm{aux}}(\mathcal{T}^N_0, \mathcal{T}^N_h)|\mathcal{T}^N_0]\big] \\
&\le \rho^2 h^2 \cdot \mathbb{E}\Big[\Big(\sum_{k=2}^N S_k\Big)^2\Big] \le 7\rho^2 h^2.
\end{aligned}$$

The last estimate follows from the following elementary computation:

$$\begin{aligned}
\mathbb{E}\Big[\Big(\sum_{k=2}^N S_k)\Big)^2\Big] &= \sum_{k=2}^N \mathbb{E}[S_k^2] + 2 \sum_{2\le k<\ell\le N} \mathbb{E}[S_k]\mathbb{E}[S_\ell] \\
&\le \sum_{k=2}^\infty \mathbb{E}[S_k^2] + \Big(\sum_{k=2}^\infty \mathbb{E}[S_k]\Big)^2 \\
&= \sum_{k=2}^\infty \frac{8}{k^2(k-1)^2} + \Big(\sum_{\ell=2}^\infty \frac{2}{\ell(\ell-1)}\Big)^2 = 8\Big(\frac{\pi^2}{3} - 3\Big) + 4 < 7.
\end{aligned}$$

Therefore, we have proved (5.4).

5.2 Finite-dimensional distributions

Let $\Phi_i = \Phi^{n_i,\phi_i}, i = 1, \dots, j$ be (bounded) polynomials and $(\mathcal{T}^N_u)_{u\in\mathbb{G}}$ be as in Definition 3.1 with $\mathcal{T}^N_u = \overline{\mathcal{T}^N_u}$, where $\mathcal{T}^N_u = ([N], r^N_u, \mu)$ is as in Definition 2.4. Then, μ is the uniform distribution on $[N]$ and r^N_u gives distances between elements of $[N]$ in $\mathcal{T}^N_u$. Let

$$A_{n,N} = \bigcup_{1\le i<j\le n} \{\underline{x} \in [N]^n : x_i = x_j\} \subseteq [N]^n$$

be the event that some entry in $\underline{x} \in [N]^n$ appears twice. Then, we find that $\mu^{\otimes n}(A_{n,N}) = \mathcal{O}(1/N)$ (for fixed n) as $N \to \infty$. Therefore, by construction, setting $\underline{x}_{kl} = (x_k, \dots, x_l)$ and $r(\underline{x}_{kl}, \underline{x}_{kl}) = (r(x_i, x_j))_{k \le i,j \le l}$, $\overline{n}_0 = 0$, $\overline{n}_i = n_1 + \cdots + n_i$,

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{E}[\Phi_1(\mathcal{T}^N_{u_1}) \cdots \Phi_j(\mathcal{T}^N_{u_j})] &= \lim_{N\to\infty} \mathbb{E}\Big[\prod_{i=1}^{j} \int \mu^{n_i}(d\underline{x}_{1n_i}) \phi_i(r^N_{u_i}(\underline{x}_{1n_i}, \underline{x}_{1n_i}))\Big] \\
&= \lim_{N\to\infty} \mathbb{E}\Big[\int \mu^n(d\underline{x}_{1n}) \prod_{i=1}^{j} \phi_i(r^N_{u_i}(\underline{x}_{\overline{n}_{i-1},\overline{n}_i}, \underline{x}_{\overline{n}_{i-1},\overline{n}_i}))\Big] \\
&= \lim_{N\to\infty} \mathbb{E}\Big[\int \mu^n(d\underline{x}_{1n}) \mathbb{1}_{A^c_{n,N}} \prod_{i=1}^{j} \phi_i(r^N_{u_i}(\underline{x}_{\overline{n}_{i-1},\overline{n}_i}, \underline{x}_{\overline{n}_{i-1},\overline{n}_i}))/\mu^n(A^c_{n,N})\Big] \\
&= \mathbb{E}[\phi_1(\underline{\underline{R}}_1) \cdots \phi_j(\underline{\underline{R}}_j)].
\end{aligned}$$

5.3 Finite variation

In this subsection we prove the last part of Theorem 3.3, namely that the paths of the process $\mathcal{T} = (\mathcal{T}_u)_{u\in\mathbb{G}}$ are of finite variation with respect to Gromov total variation, Gromov-Prohorov metric and Gromov-Hausdorff (semi-)metric on $\mathbb{M}$. Recall the definition of variation with respect to a metric in (3.1). First we show that for $d \in \{d_{\mathrm{GTV}}, d_{\mathrm{GP}}, d_{\mathrm{GH}}\}$ that for $u, v \in \mathbb{G}$, $u < v$ there is a positive finite constant $C = C(\rho)$ such that

$$\mathbb{E}[d(\mathcal{T}^N_u, \mathcal{T}^N_v)] \le C(v - u). \tag{5.5}$$

Once this is proven for a metric d, for any interval $[a, b] \subset \mathbb{G}$ and a partition $a = u_0 < u_1 < \cdots < u_k = b$ of that interval we have by the first part of Theorem 3.3

$$\mathbb{E}[d(\mathcal{T}_{u_i}, \mathcal{T}_{u_{i-1}})] \le \limsup_{N\to\infty} \mathbb{E}[d(\mathcal{T}^N_{u_i}, \mathcal{T}^N_{u_{i-1}})] \le C(u_i - u_{i-1}). \tag{5.6}$$

Then it follows easily

$$\mathbb{E}\Big[\sum_{i=1}^{k} d(\mathcal{T}_{u_i}, \mathcal{T}_{u_{i-1}})\Big] \le C(b - a).$$

Since the right hand side does not depend on particular partition of $[a, b]$, this shows that the variation of $\mathcal{T}$ with respect to d has finite expectation on finite intervals. Thus, the paths of $\mathcal{T}$ are almost surely of finite variation with respect d.

For Gromov total variation and Gromov-Prohorov metrics (5.5) follows from $d_{\mathrm{GP}} \le d_{\mathrm{GTV}} \le d_{\mathrm{aux}}$ and (4.3). For the Gromov-Hausdorff metric (5.5) is shown in the following lemma.

Lemma 5.1. *There is a positive finite constant C independent of N so that for any $u, v \in \mathbb{G}$, $u < v$ we have*

$$\mathbb{E}[d_{\mathrm{GH}}(\mathcal{T}^N_u, \mathcal{T}^N_v)] \le C\rho(v - u). \tag{5.7}$$

Proof. Given the tree $\mathcal{T}^N_u$ as before we denote by $S_2, \dots, S_N$ the time for which exactly $2, \dots, N$ lines are present in the tree (cf. Figure 3). The random variables $S_2, \dots, S_N$ are independent and S_k is exponentially distributed with mean $1/\binom{k}{2}$.

Along the branches of $\mathcal{T}^N_u$ recombination events occur at rate $\rho(v - u)$. When a recombination event occurs at level k, that is during the period of time with exactly k lines in the tree $\mathcal{T}^N_u$, then the resulting extra line coalesces back into the tree $\mathcal{T}^N_u$ at some time after the recombination, that is at level ℓ for some $1 \le \ell \le k$. We also need to

consider the level $\ell = 1$ because it might be the case that the extra line coalesces back into the tree $\mathcal{T}_u^N$ after all lines of $\mathcal{T}_u^N$ have coalesced with each other.

Let S_1 be exponentially distributed with mean 1. Furthermore, for $k \geq 2$ and $1 \leq \ell \leq k$ let $A_{k,\ell}$ be the event that along the k branches of $\mathcal{T}_u^N$ during time S_k, at least one recombination event occurs that separates the trees $\mathcal{T}_u^N$ and $\mathcal{T}_v^N$, and the resulting extra line coalesces back into the tree $\mathcal{T}_u^N$ during time S_ℓ. Figure 3 shows an example of the event $A_{3,2}$.

Then, ignoring the probability of no coalescence during S_k (hence bounding this probability from above by 1),

$$\begin{aligned}\mathbb{P}\big[A_{k,\ell}|\mathcal{T}_u^N\big] &\leq (1 - e^{-\rho(v-u)kS_k}) \prod_{m=\ell+1}^{k-1} e^{-mS_m}(1 - e^{-\ell S_\ell}) \\ &\leq \rho(v-u)kS_k\ell S_\ell \prod_{m=\ell+1}^{k-1} e^{-mS_m}.\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}\Big[\prod_{m=\ell+1}^{k-1} e^{-mS_m}\Big] &= \prod_{m=\ell+1}^{k-1} \frac{1}{1+2/(m-1)} = \prod_{m=\ell}^{k-2} \frac{1}{1+2/m} \\ &= \exp\Big(-\sum_{m=\ell}^{k-2} \log(1+2/m)\Big) \leq \exp\Big(-\sum_{m=\ell}^{k-2} 1/m\Big) \\ &\leq \exp\Big(-\int_\ell^{k-2} \tfrac{1}{x} dx\Big) = \frac{\ell}{k-2}.\end{aligned}$$

Furthermore, given $\mathcal{T}_u^N$, on the event $A_{k,\ell}$ we have

$$d_{\mathrm{GH}}(\mathcal{T}_u^N, \mathcal{T}_v^N) \leq 2\sum_{j=\ell}^{N} S_j.$$

It follows that for some $C_0, C_1, C_2, C_3 > 0$, which don't depend on N, ρ, u and v,

$$\begin{aligned}\limsup_{N\to\infty} \mathbb{E}[d_{\mathrm{GH}}(\mathcal{T}_u^N, \mathcal{T}_v^N)] &\leq \limsup_{N\to\infty} \sum_{\ell\leq k\leq N} \mathbb{E}\Big[2\sum_{j=\ell}^{N} S_j; A_{k,\ell}\Big] \\ &\leq \limsup_{N\to\infty} C_0 \sum_{\ell\leq k\leq N} \sum_{j=\ell}^{N} \mathbb{E}[S_j]\mathbb{P}\big[A_{k,\ell}\big] \\ &\leq \limsup_{N\to\infty} C_1\rho(v-u) \sum_{\ell\leq k\leq N} \sum_{j=\ell}^{N} k\ell \frac{\ell/k}{\binom{j}{2}\binom{k}{2}\binom{\ell}{2}} \\ &\leq \limsup_{N\to\infty} C_2\rho(v-u) \sum_{\ell\leq k\leq N} \frac{1}{\ell k^2} = C_3\rho(v-u),\end{aligned}$$

which shows the assertion. □

6 Proof of Theorem 3.5

Theorem 3.5 claims that correlations between trees $\mathcal{T}_0$ and $\mathcal{T}_v$ decrease with $\mathcal{O}(1/v^2)$. Such correlations come with coalescence times present in $\mathcal{T}_0$ and $\mathcal{T}_v$. Before we come to the proof of Theorem 3.5, we study such joint coalescences.

6.1 Covariances of coalescence times

Lemma 6.1 (Covariance of distances at $u = 0$ and $u = v$)**.** *Let $\Phi = \Phi^{2,\phi}$. Then,*

$$\mathbb{P}(\Phi^{2,\phi}(\mathcal{T}_0) = \Phi^{2,\phi}(\mathcal{T}_v) \text{ for all } \phi) = \frac{2}{9 + 13\rho v + 2\rho^2 v^2}. \tag{6.1}$$

Proof. From Theorem 3.3, we see that the left hand side of (6.1) is given as follows: Let $\mathcal{A}^4$ be an ARG starting with four lines, $R_{12,0}$ be the distance of the pair $1, 2$ at $u = 0$ and $R_{34,v}$ the distance of $3, 4$ at $u = v$. Then, using (3.2), $\Phi^{2,\phi}(\mathcal{T}_0) = \Phi^{2,\phi}(\mathcal{T}_v)$ for all ϕ has the same probability as $R_{12,0} = R_{34,v}$ (recall that ϕ is a function in the definition of polynomials is a function of pairwise distances; see (2.3)).

Hence, the LHS of (6.1) equals x, where

$$\begin{aligned} x &= \mathbb{P}(R_{12,0} = R_{34,v}) \\ y &= \mathbb{P}(R_{12,0} = R_{23,v}) \\ z &= \mathbb{P}(R_{12,0} = R_{12,v}) \end{aligned}$$

The first event in the ARG starting in four lines, can be:

(i) coalescence of one of the pairs (1,3), (2,3), (1,4), (2,4)

(ii) coalescence of one of the pairs (1,2), (3,4)

(iii) Some recombination event.

In case (iii), the probability x is not changed after the recombination event, in case (ii), there is no way that the event $R_{12,0} = R_{34,v}$ (hence has probability 0). In case (i), however, the probability is the same as in an ARG with three lines, that lines 1 and 2 at locus 0 coalesce at the same time as lines 2 and 3 at locus v. This probability is defined to be y. Similar arguments for a first-event-decomposition in the probabilities y and z lead to

$$\begin{aligned} x &= \frac{2}{3}y + \frac{1}{3} \cdot 0, \\ y &= \frac{\rho v}{\rho v + 3}x + \frac{1}{\rho v + 3}z + \frac{2}{\rho v + 3} \cdot 0, \\ z &= \frac{2\rho v}{2\rho v + 1}y + \frac{1}{2\rho v + 1} \cdot 1. \end{aligned} \tag{6.2}$$

Solving this linear system gives the result. □

Using the last lemma, we immediately obtain another useful result.

Corollary 6.2 (Samples of size $n \geq 2$)**.** *Let $n \geq 2$, $\mathcal{A}^{2n}$ be a $2n$-ARG, $R_{ij,u}$ be the distance of the pair i, j at position u for $u \in \{0, v\}$. Then,*

$$\begin{aligned} \mathbb{P}(R_{ij,0} = R_{k\ell,v} \text{ for some } 1 \leq i < j \leq n;\, n+1 \leq k < \ell \leq 2n) \\ \leq \binom{n}{2}^2 \frac{2}{9 + 13\rho v + 2\rho^2 v^2}. \end{aligned}$$

6.2 An auxiliary random graph

For the proof of Theorem 3.5, we recall the $2n$-ARG $\mathcal{A}^{2n}$ for loci $v \in \{0, u\}$. We let $R_{ij,v}$ be the distance between i, j at locus v for $v \in \{0, u\}$. We set $\mathcal{T}_0 = \mathcal{T}_0^{\{1,\ldots,n\}}$ and $\mathcal{T}_v = \mathcal{T}_v^{\{n+1,\ldots,2n\}}$ (recall the notation from Definition 2.4). The following events can happen:

1. "Intra-tree" coalescence events: If in $\mathcal{A}^{2n}$ two particles coalesce and both particles belong to $\mathcal{T}_v$, then the total number of particles *and* the number of particles in $\mathcal{T}_v$ decreases, $v \in \{0, u\}$.

2. "Inter-tree" coalescence events: If two particles (in $\mathcal{A}^{2n}$) coalesce and one of the particles is present in $\mathcal{T}_u \setminus \mathcal{T}_0$, and the other is present in $\mathcal{T}_0 \setminus \mathcal{T}_u$, then the total number of particles decreases but the numbers of particles within $\mathcal{T}_0$ and $\mathcal{T}_u$ are preserved.

3. "Splitting recombination" events: If a particle, present in the overlap of the trees $\mathcal{T}_u \cap \mathcal{T}_0$, recombines with mark $U \in [0, u]$, then the particle splits in two new particles, one present in $\mathcal{T}_u \setminus \mathcal{T}_0$, the other one present in $\mathcal{T}_0 \setminus \mathcal{T}_u$.

We call a branch in $\mathcal{A}^{2n}$ a *single line* if it belongs to $(\mathcal{T}_0 \setminus \mathcal{T}_u) \cup (\mathcal{T}_u \setminus \mathcal{T}_0)$, whereas branches in $\mathcal{T}_0 \cap \mathcal{T}_u$ are called *double lines*. Intra-coalescence occur simultaneously within $\mathcal{T}_0$ and $\mathcal{T}_u$ if two double lines coalesce. These are the events that make $\mathcal{T}_0$ and $\mathcal{T}_v$ dependent. We will call such an event *joint coalescence*.

We now define a random graph $\widehat{\mathcal{A}}^{2n}$ based on $\mathcal{A}^{2n}$ such that we can couple $\mathcal{T}_0$ and $\mathcal{T}_u$ with two independent trees $\widehat{\mathcal{T}}_0$ and $\widehat{\mathcal{T}}_u$, both having the distribution of a Kingman's n-coalescent; see Lemma 6.5.

Definition 6.3 (An auxiliary random graph)**.** *Define a random graph $\widehat{\mathcal{A}}^{2n}$, from which two trees $\widehat{\mathcal{T}}_0$ and $\widehat{\mathcal{T}}_u$ can be read off as in Definition 2.4, as follows: Starting with $2n$ single lines, where $1, \dots, n \in \widehat{\mathcal{T}}_0 \setminus \widehat{\mathcal{T}}_u$ and $n+1, \dots, 2n \in \widehat{\mathcal{T}}_u \setminus \widehat{\mathcal{T}}_0$, the dynamics of the lines in $\widehat{\mathcal{A}}^{2n}$ are as follows (see Figure 4):*

(i) Each pair of single lines coalesces at rate 1. The result can be a single line (if both lines belong to $\widehat{\mathcal{T}}_0$ or both to $\widehat{\mathcal{T}}_u$) or a double line.

(ii) Each pair of lines where one is a single line and the other a double line coalesces at rate 1. The resulting line is a double line.

(iii) Each double line splits at rate ρu into two single lines.

(iv) Between each pair of double lines there is a coalescence/splitting event *at rate 2. This event produces a double line and a single line. With probability $1/2$ the resulting single line is in $\widehat{\mathcal{T}}_0$ or in $\widehat{\mathcal{T}}_u$, respectively.*

Remark 6.4 (Properties of $\widehat{\mathcal{T}}_0$ and $\widehat{\mathcal{T}}_u$)**.** Note that the above dynamics in $\widehat{\mathcal{A}}^{2n}$ are the same as in $\mathcal{A}^{2n}$ except for the coalescence/splitting event described in (iv). The corresponding event in $\mathcal{A}^{2n}$ was called joint coalescence above. In particular, we remark that we can perfectly couple $\mathcal{A}^{2n}$ with $\widehat{\mathcal{A}}^{2n}$ until the first coalescence/splitting event occurs.

Lemma 6.5 (Properties of $\widehat{\mathcal{A}}^{2n}$)**.** *We note the following properties of $\widehat{\mathcal{A}}^{2n}$:*

1. *$\widehat{\mathcal{T}}_0$ and $\widehat{\mathcal{T}}_u$ are independent and distributed as n-coalescents.*

2. *If we couple $\mathcal{A}^{2n}$ and $\widehat{\mathcal{A}}^{2n}$ until the first event (iv) happens, and let them evolve independently otherwise, then*

$$\{\text{no event (iv) happens}\} \subseteq \{\mathcal{T}_0 = \widehat{\mathcal{T}}_0\} \cap \{\mathcal{T}_u = \widehat{\mathcal{T}}_u\}.$$

3. *For $n \geq 2$, there is $C = C(n) > 0$ such that*

$$\mathbb{P}(\text{no event (iv) happens}) \leq C/(\rho^2 u^2).$$

Proof. 1. Obviously in $\widehat{\mathcal{A}}^{2n}$ each pair of lines in $\widehat{\mathcal{T}}_0$ coalesces at rate 1 and also each pair of lines in $\widehat{\mathcal{T}}_u$ coalesces at rate 1, so that both trees are Kingman's coalescents, and the trees are independent by construction.
2. Denoting by A the event that at a coalescence/splitting event in $\widehat{\mathcal{A}}^{2n}$ occurs, we have $\mathcal{A}^{2n} = \widehat{\mathcal{A}}^{2n}$ on A^c by construction.
3. By construction, A occurs at rate 2 for every pair of lines within $\mathcal{Z}_{\mathrm{joint}}$. For non-negative integers a, b and c we indicate by $\mathbb{P}_{abc}$ computations of probabilities within $\widehat{\mathcal{A}}^{2n}$ with start in

- a single lines within $\widehat{\mathcal{T}}_0$,
- b double lines,
- c single lines within $\widehat{\mathcal{T}}_u$.

Then, for

$$x = \mathbb{P}_{202}(A), \qquad y = \mathbb{P}_{111}(A), \qquad z = \mathbb{P}_{020}(A),$$

we obtain the same set of equations as in (6.2) with the last one replaced by

$$z = \frac{2\rho u}{2\rho u + 2} y + \frac{2}{2\rho u + 2} \cdot 1.$$

Solving this system gives

$$x = \frac{2}{9 + 7\rho u + \rho^2 u^2},$$

which shows the assertion for $n = 2$. As in Corollary 6.2, we obtain for all $n = 2, 3, \ldots$

$$\mathbb{P}(\text{some event (iv) happens}) \leq \binom{n}{2}^2 \frac{2}{9 + 7\rho u + \rho^2 u^2}$$

which concludes the proof. □

6.3 Proof of Theorem 3.5

Let $\Psi = \Psi^{n,\psi}$ and $\Phi = \Phi^{n,\phi}$. According to Theorem 3.3, we need to consider a $2n$-ARG $\mathcal{A}^{2n}$ and let $R_{ij,v}$ be the distance between i, j at locus v for $v \in \{0, u\}$. Writing $\underline{\underline{R}}_0 := (R_{ij,0})_{1 \leq i,j \leq n}$, $\underline{\underline{R}}_v := (R_{ij,v})_{n+1 \leq i,j \leq 2n}$, Theorem 3.3 gives

$$\mathbb{COV}[\Psi(\mathcal{T}_0), \Phi(\mathcal{T}_u)] = \mathbb{COV}[\psi(\underline{\underline{R}}_0), \phi(\underline{\underline{R}}_u)]. \tag{6.3}$$

Let $\widehat{\mathcal{T}}_0, \widehat{\mathcal{T}}_u$ be as in Lemma 6.5, which are coupled with $\mathcal{T}_0, \mathcal{T}_u$ before the first coalescence/splitting event happens. Let $\underline{\underline{\widehat{R}}}_0$ and $\underline{\underline{\widehat{R}}}_u$ be the (finite) distance matrices that correspond to $\widehat{\mathcal{T}}_0$ and $\widehat{\mathcal{T}}_u$. Slightly abusing the notation we write

$$\psi_0 = \psi(\underline{\underline{R}}_0), \quad \phi_u = \phi(\underline{\underline{R}}_u), \quad \widehat{\psi}_0 = \psi(\underline{\underline{\widehat{R}}}_0) \quad \text{and} \quad \widehat{\phi}_u = \phi(\underline{\underline{\widehat{R}}}_u).$$

Denoting by A the event that a coalescence/splitting event in $\widehat{\mathcal{A}}^{2n}$ occurs, we have using Lemma 6.5

$$\begin{aligned}
\mathbb{E}[\psi_0 \phi_u] &= \mathbb{E}[\psi_0 \phi_u \mathbb{1}_{A^c}] + \mathbb{E}[\psi_0 \phi_u \mathbb{1}_A] \\
&= \mathbb{E}[\widehat{\psi}_0 \widehat{\phi}_u \mathbb{1}_{A^c}] + \mathbb{E}[\psi_0 \phi_u \mathbb{1}_A] \\
&= \mathbb{E}[\widehat{\psi}_0 \widehat{\phi}_u] - \mathbb{E}[\widehat{\psi}_0 \widehat{\phi}_u \mathbb{1}_A] + \mathbb{E}[\psi_0 \phi_u \mathbb{1}_A] \\
&= \mathbb{E}[\widehat{\psi}_0] \mathbb{E}[\widehat{\phi}_u] - \mathbb{E}[\widehat{\psi}_0 \widehat{\phi}_u \mathbb{1}_A] + \mathbb{E}[\psi_0 \phi_u \mathbb{1}_A] \\
&= \mathbb{E}[\psi_0] \mathbb{E}[\phi_u] - \mathbb{E}[\widehat{\psi}_0 \widehat{\phi}_u \mathbb{1}_A] + \mathbb{E}[\psi_0 \phi_u \mathbb{1}_A].
\end{aligned}$$

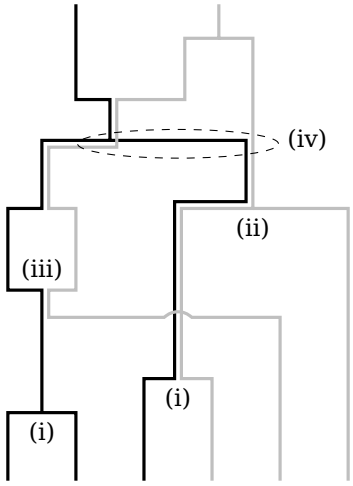


Figure 4: Reading off trees at different loci starting with disjoint sets of leaves from modified ARG $\widehat{A}^{2n}$ with $n = 3$. Some of the events are annotated according to the description. The dashed ellipsis encloses the event which is not possible in the original ARG, cf. Figure 2.

It follows now that for $C = 2\mathbb{P}(A)$, we have

$$|\mathbb{E}[\psi_0\phi_u] - \mathbb{E}[\psi_0]\mathbb{E}[\phi_u]| \leq C\|\psi\|_\infty\|\phi\|_\infty,$$

which, in view of Lemma 6.5.3. shows the assertion of Theorem 3.5 .

Acknowledgments This research was supported by the DFG through the priority program 1590, and in particular through grant Pf-672/6-1 to PP.

References

Aldous, D. J. (1999). Deterministic and stochastic models for coalescence (aggregation and coagulation): a review of the mean-field theory for probabilists. *Bernoulli 5*(1), 3–48. MR-1673235

Billingsley, P. (1999). *Convergence of probability measures* (Second ed.). Wiley Series in Probability and Statistics: Probability and Statistics. New York: John Wiley & Sons Inc. A Wiley-Interscience Publication. MR-1700749

Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Advances in Appl. Probability 6*, 260–290. MR-0343949

Chen, G. K., P. Marjoram, and J. D. Wall (2009). Fast and flexible simulation of dna sequence data. *Genome research 19*(1), 136–142.

Ethier, S. N. and T. G. Kurtz (1986). *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc. Characterization and convergence. MR-0838085

Ewens, W. J. (2004). *Mathematical population genetics. I* (Second ed.), Volume 27 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York. Theoretical introduction. MR-2026891

Greven, A., P. Pfaffelhuber, and A. Winter (2009). Convergence in distribution of random metric measure spaces (Λ-coalescent measure trees). *Probab. Theory Related Fields 145*(1-2), 285–322. MR-2520129

Greven, A., P. Pfaffelhuber, and A. Winter (2013). Tree-valued resampling dynamics

Martingale problems and applications. *Probab. Theory Related Fields 155*(3-4), 789–838. MR-3034793

Griffiths, R. (1991). The two-locus ancestral graph. In *Selected Proceedings of the Sheffield Symposium on Applied Probability (Sheffield, 1989)*, Volume 18 of *IMS Lecture Notes Monogr. Ser.*, pp. 100–117. Hayward, CA: Inst. Math. Statist. MR-1193063

Griffiths, R. C. and P. Marjoram (1997). An ancestral recombination graph. In *Progress in population genetics and human evolution (Minneapolis, MN, 1994)*, Volume 87 of *IMA Vol. Math. Appl.*, pp. 257–270. Springer, New York. MR-1493031

Gromov, M. (2007). *Metric structures for Riemannian and non-Riemannian spaces* (English ed.). Modern Birkhäuser Classics. Boston, MA: Birkhäuser Boston Inc. Based on the 1981 French original, With appendices by M. Katz, P. Pansu and S. Semmes, Translated from the French by Sean Michael Bates. MR-2307192

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology 23*(2), 183 – 201.

Kingman, J. (1982). The coalescent. *Stochastic Process. Appl. 13*(3), 235–248. MR-0671034

Leocard, S. and E. Pardoux (2010). Evolution of the ancestral recombination graph along the genome in case of selective sweep. *J. Math. Biol. 61*(6), 819–841. MR-2726452

McVean, G. A. T. and N. J. Cardin (2005). Approximating the coalescent with recombination. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences 360*(1459), 1387–1393.

Pardoux, E. and M. Salamat (2009). On the height and length of the ancestral recombination graph. *J. Appl. Probab. 46*(3), 669–689. MR-2560895

Rasmussen, M. D., M. J. Hubisz, I. Gronau, and A. Siepel (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genet 10*(5), e1004342.

Wiuf, C. and J. Hein (1999). Recombination as a point process along sequences. *Theoretical population biology 55*(3), 248–259.