



Ancestral Population Genomics

Julien Y. Dutheil and Asger Hobolth

Abstract

Borrowing both from population genetics and phylogenetics, the field of population genomics emerged as full genomes of several closely related species were available. Providing we can properly model sequence evolution within populations undergoing speciation events, this resource enables us to estimate key population genetics parameters such as ancestral population sizes and split times. Furthermore we can enhance our understanding of the recombination process and investigate various selective forces. With the advent of resequencing technologies, genome-wide patterns of diversity in extant populations have now come to complement this picture, offering an increasing power to study more recent genetic history.

We discuss the basic models of genomes in populations, including speciation models for closely related species. A major point in our discussion is that only a few complete genomes contain much information about the whole population. The reason being that recombination unlinks genomic regions, and therefore a few genomes contain many segments with distinct histories. The challenge of population genomics is to decode this mosaic of histories in order to infer scenarios of demography and selection. We survey modeling strategies for understanding genetic variation in ancestral populations and species. The underlying models build on the coalescent with recombination process and introduce further assumptions to scale the analyses to genomic data sets.

Key words Ancestral population, Coalescence, Demography, Divergence, Markov model, Migration, Recombination, Selection, Speciation

1 Introduction

We are in the population genomics era where data sets from the 1000 human genomes project [1], the great apes project [2], and the 1001 arabidopsis genomes project [3] are available. The underlying data sets contain genotypic information for thousands of individuals in one or several species, in the form of de novo sequenced genomes or variation compared to an available “reference” genome (a.k.a. *resequencing*). By comparing genomes from several individuals of the same species or closely related species, we can obtain information about split times, population sizes, recombination events, and selection in contemporary and ancestral

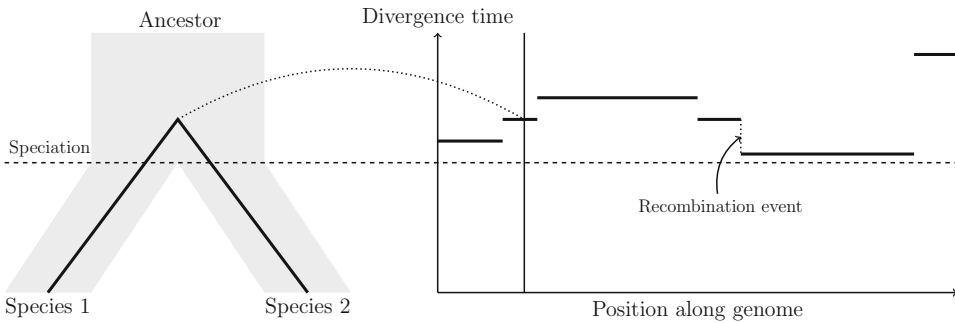


Fig. 1 Left: Isolation model of two species. Right: The coalescent process along the genomes of the two species. By comparing the two genomes we obtain information about the split time of the species and the ancestral population size. Furthermore the breakpoints along the genomes correspond to recombination events, so we also have information about the recombination process

species (*see* Fig. 1). In this chapter we discuss various models for obtaining this information.

Comparing homologous sequences available for a given locus to infer their degree of relatedness enables the discovery of the parental relationships of the sequences, depicted as a tree thereby named *genealogy*. When one sequence sampled from one individual of one species is compared with sequences from other species, the resulting genealogy contains information about the history of species, the so-called phylogeny. The phylogeny summarizes the relationship and the divergence times between the species.

Conversely, when sequences from several individuals within a species are sampled, we have access to the genetic variation in contemporary populations. The evolutionary forces that shape genetic variation within a species are genetic drift, mutation, recombination, and selection and are the subject of population genetics. The key modeling tool in population genetics is coalescent theory. Classical coalescent theory describes the genetic ancestry of a sample of homologous DNA sequences from the same species. This genealogical description includes times to common ancestry, which is measured back into the past.

Molecular phylogenetics and population genetics have accumulated 50 years of methodological developments. The convergence of these two fields and their key mathematical and statistical tools is needed in order to fully understand genomic sequence alignments, because comparing genealogies and phylogenies is at the heart of the study of the speciation process [4].

We describe the interplay between population genetics and phylogenetics by reviewing the methods and models that have been developed to understand evolutionary history from genomic data (*see* Table 1 for a comparative summary of all methods).

Table 1
Methods comparison

Principle	ARG Approx.	Spec.	Parameters estimated	Rate variation/ sequencing errors	Data set	Reference
Infer genealogy from independent loci, use distribution of inferred divergence and topology counts to estimate parameters	Independent loci	I	T, N_A	—	Primates: 53 “random” autosomal intergenic non-repetitive DNA segments of 2–20 kb	[5]
Count alignment patterns, fit EM model to infer parameters	—	I	T_1, N_A	Correction with outgroup	Primates	[6]
Likelihood calculation under a demographic model, numerical integration over genealogies	Independent loci	I	T_1, T_2, N_A	Independent estimate of rate	Primates	[7]
	Independent loci	IM	$T_1, T_2, N_1, N_2, N_A, m_{1 \rightarrow 2}, m_{2 \rightarrow 1}$	RAS	Drosophila	[8]
	Independent loci	IM	$T_1, T_2, N_1, N_2, N_A, m_{1 \rightarrow 2}, m_{2 \rightarrow 1}$	Independent estimate of rate	Primates: same data as [9] restricted to human, chimpanzee, gorilla, and orangutan	[10]
Bayesian inference	Independent loci	I	$T_1, T_2, T_3, T_4, N_{A1}, N_{A2}, N_{A3}$	RAS + branch-specific departure from molecular clock	Primates: 15,000 neutral loci (7.4 Mb)	[9]
Integrating over a subset of candidate genealogies using a hidden Markov model	Markov process	I	T_1, T_2, N_{A1}, N_{A2}	—	Primates: 1 Mb alignment	CoalHMM [11]
	Markov process	I	$T_1, T_2, N_{A1}, N_{A2}, r$	RAS	Primates: 1 Mb alignment	CoalHMM [12]
Integrating over the discretized distribution of divergence for a pair of genomes	Markov process	I	T, N_A, r	—	Orangutans: two full genomes	CoalHMM [13]

(continued)

Table 1
(continued)

Principle	ARG Approx.	Spec.	Parameters estimated	Rate variation/ sequencing errors	Data set	Reference
Integrating over the discretized distribution of divergence for a pair of haploid genomes in a population	Markov process	—	$N_{t_{1...n}}, r$	—	Human diploid genomes	PSMC [14]
Integrating over the discretized distribution of divergence of the most recent coalescence event with multiple haploid genomes in a population	Markov process	—	$N_{t_{1...n}}, r$ fixed	—	Human diploid genomes	MSMC [15]
Use the conditional sampling distribution to approximate the integration over the discretized distribution of divergence coalescence events within multiple haploid genomes in one or more population(s)	Markov process + CSD	—	$N_{t_{1...n}}, r$ fixed	—	Human diploid genomes	diCal [16–18]
Extension of the pairwise sequentially Markov coalescent with site frequency spectrum based on many individuals	Markov process + Poisson random field	—	$N_{t_{1...n}}$	—	Unphased human genomes	SMC++ [19]
Bayesian sampling of ARG, using a discretized distribution of divergence time conditioned on multiple haploid genomes in a population	Markov process + “threading”	—	$N_{t_{1...n}}, r$	—	Human diploid genomes	ARGweaver [20]

This table summarizes and compares existing ancestral population genomics methods. Parameters correspond to the one in Fig. 4. RAS: Rate across site model, assuming an a priori distribution of evolutionary rate (usually a discretized gamma distribution) over alignment positions

2 Coalescent Theory and Speciation

We start by describing the standard coalescent model within one population. The coalescent model describes the shape of the genealogy of several sequences sampled from a single population. For more information on the coalescent, we refer to [21, 22] and [23]. This section describes the coalescent process as a chronological process. In the next section, we will see how it can be modeled as a spatial process along the genome. In subsequent sections we extend the standard model to include two or more populations. In the cases where multiple populations are present we describe both the isolation model and the isolation-with-migration model.

2.1 The Standard Coalescent Model

The standard coalescent model is a continuous-time approximation of the neutral Wright–Fisher model. In the Wright–Fisher model the number of chromosomes $2N$ (we consider diploid organisms) is fixed in each non-overlapping generation. Each chromosome in a new generation chooses its ancestor uniformly at random from the previous generation.

Consider two chromosomes. The probability of the two chromosomes choosing the same ancestor is $1/(2N)$ and the probability of the two chromosomes not finding a common ancestor is $1 - 1/(2N)$. Let R_2 denote the number of generations back in time when the two individuals find a most recent common ancestor (MRCA). By repeating the argument above, the probability of the two chromosomes not finding a common ancestor r generations back in time is

$$P(R_2 > r) = \left(1 - \frac{1}{2N}\right)^r.$$

If we scale time t in units of $2N$, i.e., set $r = 2Nt$, we get

$$P(R_2 > r) = \left(1 - \frac{1}{2N}\right)^r = \left(1 - \frac{1}{2N}\right)^{2Nt} \approx e^{-t},$$

where the approximation is valid for large N . In coalescent time units the waiting time $T_2 = R_2/(2N)$ before coalescence of two individuals is therefore exponentially distributed with mean one.

These considerations can be extended to multiple individuals. In general the time T_n before two of n individuals coalesce is exponentially distributed with rate $\binom{n}{2}$.

The waiting time W_n for a sample of n individuals to find the most recent common ancestor (MRCA) is given by

$$W_n = T_n + T_{n-1} + \cdots + T_2,$$

where T_k are independent exponential random variables with parameter $\binom{k}{2}$; see Fig. 2 for an illustration. It follows that the mean of W_n is

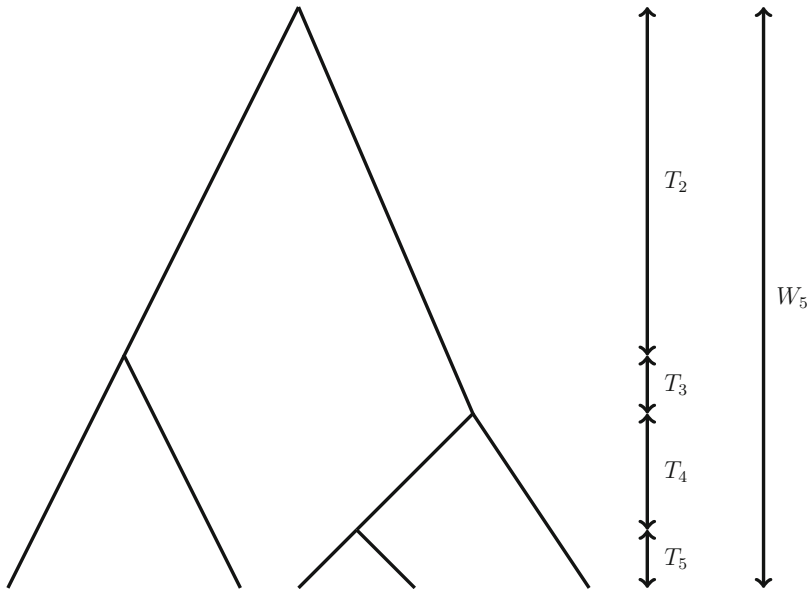


Fig. 2 Illustration of the coalescent process. The waiting time before two out of n individuals coalesce is T_n and the time before a sample of n individuals find common ancestry is W_n

$$\begin{aligned}
 E[W_n] &= \sum_{k=2}^n E[T_k] = \sum_{k=2}^n \frac{2}{k(k-1)} = 2 \sum_{k=2}^n \left(\frac{1}{k-1} - \frac{1}{k} \right) \\
 &= 2 \left(1 - \frac{1}{n} \right).
 \end{aligned}$$

Note that $\lim_{n \rightarrow \infty} E[W_n] = 2$.

The variance of W_n is

$$\begin{aligned}
 \text{Var}[W_n] &= \sum_{k=2}^n \text{Var}[T_k] = \sum_{k=2}^n \binom{k}{2}^{-2} \\
 &= 8 \sum_{k=1}^{n-1} \frac{1}{k^2} - 4 \left(1 - \frac{1}{n} \right) \left(3 + \frac{1}{n} \right).
 \end{aligned}$$

Note that $\lim_{n \rightarrow \infty} \text{Var}[W_n] = \left(\frac{8\pi^2}{6} - 12 \right) = 1.16$.

The consequences of these calculations are that when we only sample within a population we are limited to relatively recent events. The expected time for a large sample to find their MRCA is approximately $2 \times (2N) = 4N$ generations with standard deviation $\sqrt{1.16} \times (2N) = 2.15N$ generations. As a consequence, a neutral sample within a population contains little information beyond $6N$ generations.

Humans have a generation time of approximately 20 years and an effective population size of approximately $N = 10,000$ (see [21, p. 251]), and therefore $6N$ generations correspond to approximately 1.2 million years (My) for humans. Therefore human

diversity at neutral loci contains little demographic information beyond 1.2 My.

2.2 Adding Mutations to the Standard Coalescent Model

Now suppose mutations occur at a rate μ per locus per generation. In a lineage of r generations, we then expect $r\mu$ mutations or in the coalescent time units with $r = 2Nt$ we expect $2Nt\mu$ mutations. We let $\theta = 4N\mu$ be the mutation rate parameter. Since μ is small we can make a Poisson approximation of the binomial number of mutations in a lineage of r generations

$$\text{Bin}(r, \mu) = \text{Bin}(2Nt, \theta/(2 \cdot 2N)) \approx \text{Pois}(t\theta/2).$$

We have thus arrived at the following two-step process for simulating samples under the coalescent: (a) simulate the genealogy by merging lineages uniformly at random and with waiting times exponentially distributed with rate $\binom{n}{2}$ when n lineages are present; (b) on each lineage in the tree add mutations according to a Poisson process with rate $\theta/2$.

Another possibility is to scale the coalescent process such that one mutation is expected in one time unit. In this case the exponentially distributed waiting times in (a) have rate $\binom{n}{2}(2/\theta)$, and in (b) the mutations are added with unit rate. We use the latter version of the coalescent-with-mutations process below.

2.3 Taking Recombination into Account

For species where recombination occurs, different parts of the genome come from distinct ancestors, and therefore have a distinct history. Figure 3 exemplifies this phenomenon for two species. It displays the genealogical relationships for two sequences which underwent a single recombination event. In the presence of recombination, each position of a genome alignment therefore has a specific genealogy, and close positions are more likely to share the same one (recall Fig. 1). The genome alignment can therefore be described as an ordered series of genealogies, spanning a variable amount of sites, and then changing because of a recombination event [4]. The genealogy is therefore depicted as a complex graph with nodes representing both coalescence and recombination events, the ancestral recombination graph (ARG, Fig. 3c). A single genome thus contains different samples from the distribution of the age of the MRCA, and the distribution contains information about the ancestral population size and speciation time. The coalescent with recombination serves as a basis for modeling genome-wide genealogy, a point that we will further develop in Subheading 4.

3 Adding Genetic Barriers and Gene Flow to the Picture: The Structured Coalescent

In this section we extend the standard coalescent model. We consider coalescent models with multiple species and introduce population splits or speciation events. The models that we describe are

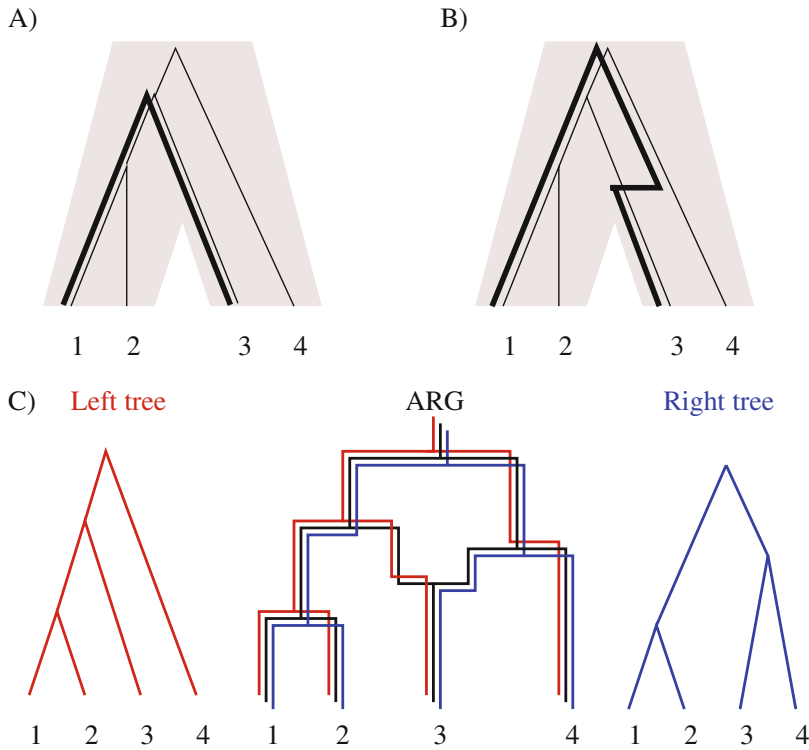


Fig. 3 Ancestral recombination graph for two species. (a) Genealogy of four sampled sequences from two species. The bold line shows the divergence of two sequences of interest. (b) A single recombination event happened between the lineages of sequences 3 and 4 (horizontal line), so that in a part of the sequences, the genealogy is as depicted by the bold line and therefore displays an older divergence. (c) The corresponding ancestral recombination graph (in black) with the trees of each side of the recombination break point superimposed (red: left tree; blue: right tree). When going backward in time, a split corresponds to a recombination event and a merger to a coalescence event

shown in Fig. 4 (see also Table 1) and include: (a) The two species isolation model; (b) The two species isolation-with-migration models; (c) The three species isolation model (and incomplete lineage sorting); and (d) The three species isolation-with-migration model. We also discuss the general multiple species isolation-with-migration model. The two species isolation model was introduced in [24] and the isolation-with-migration model was introduced in [25].

3.1 Isolation Model with Two Species

If the sequences are sampled from two distinct species that have diverged a time T ago (see Fig. 4a), then the distribution of the age of the MRCA is shifted to the right with the amount T , resulting in the distribution

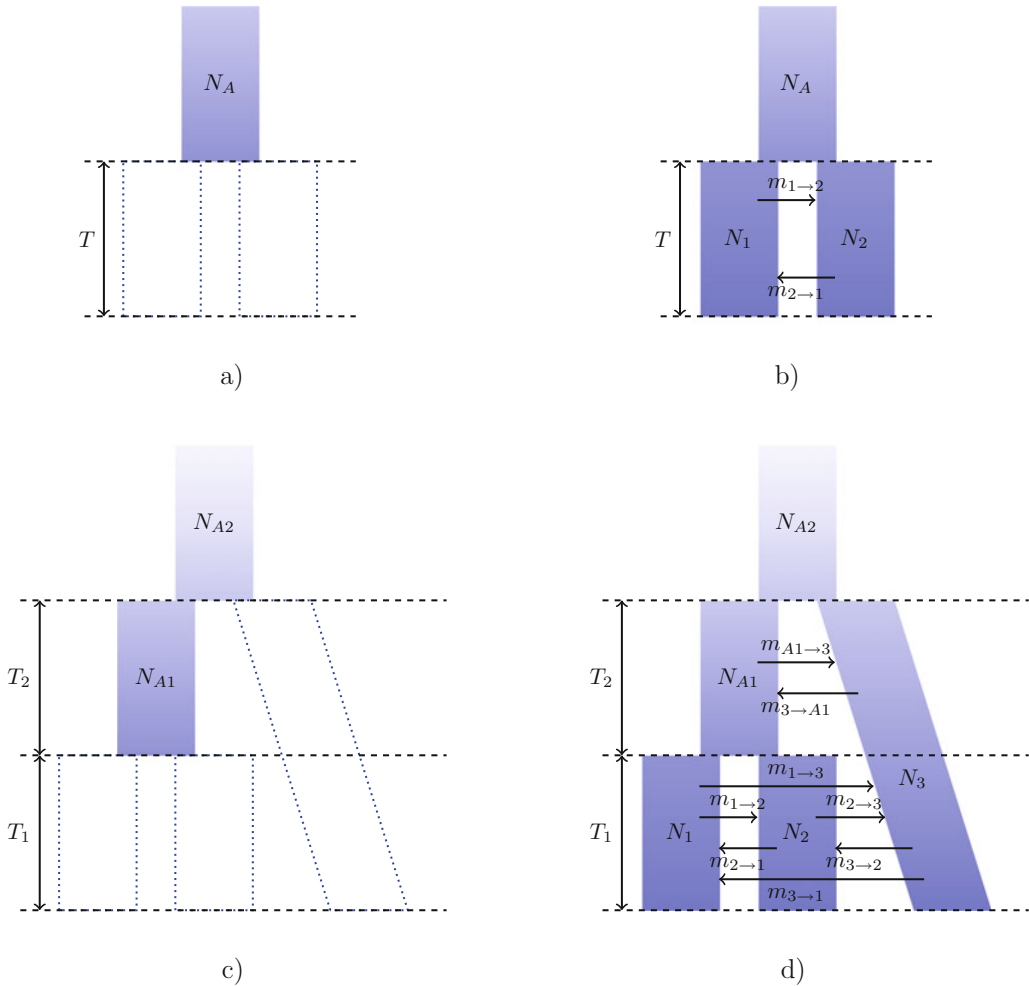


Fig. 4 Speciation models and associated parameters. In all exemplified models effective population size is constant between speciation events, represented by dash lines. The timing of the speciation events, noted T are parameters of the models, together with ancestral effective population sizes, noted N_A . In some cases, contemporary population sizes can also be estimated, and are noted N_i , where i is the index of the population. Models with post-divergence genetic exchanges have additional migration parameters labeled $m_{\text{from} \rightarrow \text{to}}$. The number of putative migration rates increases with the number of contemporary populations under study, and some models might consider some of them to be equal or eventually null to reduce complexity. **(a)** Isolation model with two species. **(b)** Isolation-migration model with two species. **(c)** Isolation model with three species. **(d)** Isolation-Migration model with three species

$$f_{T_2}(t) = \begin{cases} 0 & \text{if } t < T \\ \frac{2}{\theta_A} e^{-2(t-T)/\theta_A} & \text{if } t > T \end{cases}$$

where $\theta_A = 4N_A \cdot u$ is the ancestral mutation rate. The mean time to coalescent is $E[T_2] = T + \theta_A/2$ and the average divergence time between two sequences is twice this quantity, that is, $2T + \theta_A$. Since $\theta_A = 4N_A u$ it follows that the larger the size of the ancestral

population, the bigger the difference between the speciation time and the divergence time.

The variance of the divergence time is $\text{Var}[T_2] = \theta_A^2/4$. With access to the distribution of divergence times, we could estimate the speciation time and population size from the mean and variance of the distribution. Unfortunately we do not know the complete distribution of divergence times and it is not immediately available to us, because long regions are needed for precise divergence estimation but have experienced one or more recombination events.

3.2 Isolation Model with Three or More Species and Incomplete Lineage Sorting

Now consider the isolation model with three species depicted in Fig. 4c. Such a model is often used for the human–chimpanzee–gorilla (HCG) triplet (e.g., [10–12]).

The density function for the time to coalescence between sample 1 and sample 2 is given by

$$f_{T_2}(t) = \begin{cases} 0 & \text{if } t < T_1 \\ \frac{2}{\theta_{A1}} e^{-2(t-T_1)/\theta_{A1}} & \text{if } T_1 < t < T_{12} \\ P_{12} \frac{2}{\theta_{A2}} e^{-2(t-T_{12})/\theta_{A2}} & \text{if } t > T_{12}, \end{cases} \quad (1)$$

where

$$T_{12} = T_1 + T_2 \quad \text{and} \quad P_{12} = e^{-2(T_{12}-T_1)/\theta_{A1}}$$

is the probability of the two samples *not* coalescing in the ancestral population of sample 1 and sample 2. In the upper right corner of Fig. 5 we plot the density (Eq. 1) with parameters that resemble the HCG triplet.

If sample 1 and sample 2 do not coalesce in the ancestral population of sample 1 and sample 2, then the three trees ((1,2),3), ((1,3),2), and ((2,3),1) are equally likely. The probability of the gene tree being different from the species tree is thus

$$\text{Pr}(\text{incongruence}) = \frac{2}{3} P_{12} = \frac{2}{3} e^{-2(T_{12}-T_1)/\theta_{A1}}. \quad (2)$$

The event that the gene tree is different from the species tree is called incomplete lineage sorting (ILS). ILS is important because species tree incongruence often manifests itself as a relatively clear signal in a sequence alignment and thereby allows for accurate estimation of population parameters. In Fig. 6 we show the (in)congruence probability Eq. 2. We also refer to Exercise 1 (see Subheading 8.1) and Exercise 2 (see Subheading 8.2) for more discussion of ILS.

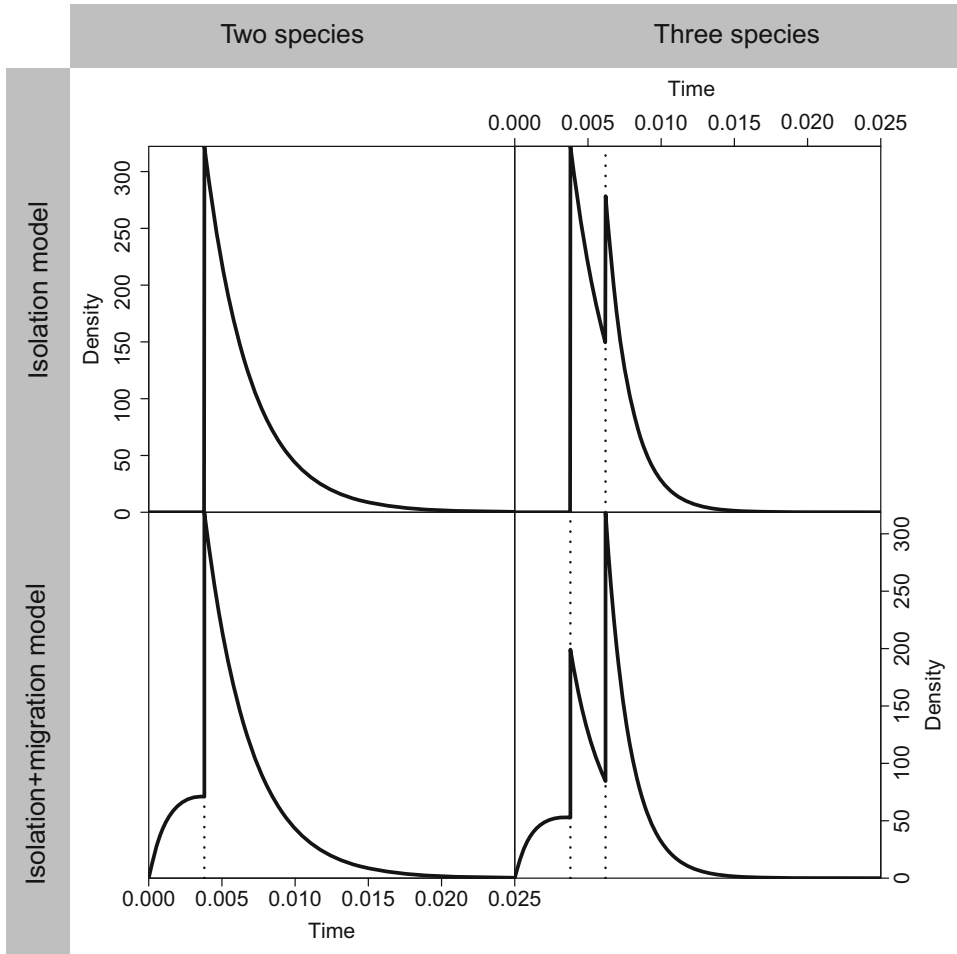


Fig. 5 Illustration of the density for coalescent in various models and data layout. The curves are the probability density functions. In the most simple case with two species, a constant ancestral population size and a punctual speciation (top left panel), more genomic regions find a common ancestor close to the species split (the vertical line), while a few regions have a more ancient common ancestor, distributed in an exponential manner (*see* Eq. 1). If speciation is not punctual and migration occurred after isolation of the species, then some sequences have a common ancestor which is more recent than the species split and the distribution in the ancestor becomes more complex (bottom left panel, *see* Eqs. 4 and 6). When a third species is added (right panel), then another discontinuity appears and all distributions depend on additional parameters, particularly when migration is allowed. We use $\theta_{A1} = 0.0062$, $\theta_{A2} = 0.0033$ and $\tau_1 = 0.0038$ (the first vertical line), $\tau_2 = 0.0062$ (the second vertical line) corresponding to the HCG triplet. Ancestral population sizes are taken from the simulation study in Table 6 in Wang and Hey [8]: $\theta_1 = 0.005$ and $\theta_2 = 0.003$. Migration parameters are all set to 50

In the three species isolation model the mean coalescent time for a sample from population 1 and a sample from population 2 is given by

$$E[T_2] = T_1 + (1 - P_{12}) \frac{\theta_{A1}}{2} + P_{12} \frac{\theta_{A2}}{2}. \quad (3)$$

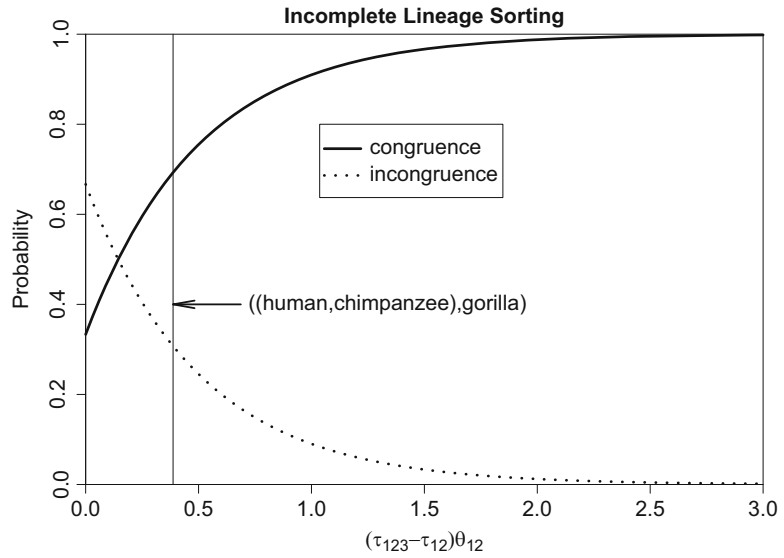


Fig. 6 Probability (Eq. 2) of gene tree and species tree being incongruent. In case of the HCG triplet we obtain $(T_{12} - T_1)/\theta_{A1} = (0.0062 - 0.0038)/0.0062 = 0.39$ which corresponds to an incongruence probability of 30%

Burgess and Yang [9] describe the speciation process for human, chimpanzee, gorilla, orangutan (O), and macaques (M) using an isolation model with five species. The HCGOM model contains four ancestral parameters θ_{HC} , θ_{HCG} , θ_{HCGO} , and θ_{HCGOM} . In this case (Eq. 3) extends to

$$\begin{aligned}
 E[T_2] = & T_{HC} + (1 - P_{HC}) \frac{\theta_{HC}}{2} P_{HC} (1 - P_{HCG}) \frac{\theta_{HCG}}{2} \\
 & + P_{HC} P_{HCG} (1 - P_{HCGO}) \frac{\theta_{HCGO}}{2} \\
 & + P_{HC} P_{HCG} P_{HCGO} (1 - P_{HCGOM}) \frac{\theta_{HCGOM}}{2}.
 \end{aligned}$$

3.3 Isolation-with-Migration Model with Two Species and Two Samples

The isolation-with-migration (IM) model with two species is shown in Fig. 4b. The IM-model has six parameters: The mutation rates θ_1 , θ_2 , and θ_A , the migration rates m_1 and m_2 , and the speciation time T . We let $\Theta = (\theta_1, \theta_2, \theta_A, m_1, m_2, T)$ be the vector of parameters.

Wang and Hey [8] consider a situation with two genes. Before time T the system is in one of the following five states:

- S_{11} : Both genes are in population 1.
- S_{22} : Both genes are in population 2.
- S_{12} : One gene is in population 1 and the other is in population 2.
- S_1 : The genes have coalesced and the single gene is in population 1.
- S_2 : The genes have coalesced and the single gene is in population 2.

The instantaneous rate matrix Q is given by

	S_{11}	S_{12}	S_{22}	S_1	S_2
S_{11}	\cdot	$2m_2$	0	$2/\theta_1$	0
S_{12}	m_1	\cdot	m_2	0	0
S_{22}	0	$2m_1$	\cdot	0	$2/\theta_2$
S_1	0	0	0	\cdot	m_2
S_2	0	0	0	m_1	\cdot

Starting in state a , the density for coalescent in population 1 at time $t < T$ is given by [26]

$$f_1(t) = (e^{Qt})_{aS_{11}}(2/\theta_1), \tag{4}$$

the density for coalescent in population 2 at time $t < T$ is

$$f_2(t) = (e^{Qt})_{aS_{22}}(2/\theta_2), \tag{5}$$

and the total density for a coalescent at time $t < T$ is

$$f(t) = f_1(t) + f_2(t). \tag{6}$$

Here $e^A = \sum_{i=0}^{\infty} A^i/(i!)$ is the matrix exponential of the matrix A and $(e^A)_{ij}$ is entry (i, j) in the matrix exponential.

After time T the system only has two states: S_{AA} corresponding to two genes in the ancestral population and S_A corresponding to one single gene in the ancestral population. The rate of going from state S_{AA} to state S_A is $2/\theta_A$. The density for coalescent in the ancestral population at time $t > T$ is therefore

$$f(t) = \left[(e^{QT})_{aS_{11}} + (e^{QT})_{aS_{12}} + (e^{QT})_{aS_{22}} \right] \frac{2}{\theta_A} e^{-(2/\theta_A)(t-T)}. \tag{7}$$

In Fig. 5 we illustrate the coalescent density in the two species isolation-with-migration model.

The likelihood for a pair of homologous sequences X is given by

$$P(X|\theta) = L(\theta|X) = \int_0^{\infty} P(X|t)f(t|\theta)dt \tag{8}$$

where $f(t) = f(t|\theta)$ given by Eqs. 6 and 7 is the density of the two sequences finding a MRCA at time t and $P(X|t)$ is the probability of the two sequences given that they find a MRCA at time t . The latter term is calculated using a distance-based method. One possibility is to use the infinite sites model where it is assumed that substitutions

happen at unique sites, i.e., there are no recurrent substitutions. In this case the number of differences between the two sequences follows a Poisson distribution with rate 1.

For an application of the isolation-with-migration model with two sequences, we refer to [8]; a discussion of their approach can be found in [27].

3.4 Isolation-with-Migration Model with Three or More Species and Three or More Samples

Hey [28] considered the multipopulation isolation-with-migration (IM) model. Recall from Fig. 4b that the two-population IM model has six parameters: two present population sizes, one ancestral population size, one speciation time, and two migration rates. The three-population IM model in Fig. 4d has fifteen parameters: three present population sizes, two ancestral population sizes, two speciation times, and eight migration rates. In general a k -population IM model has $3k - 2 + 2(k - 1)^2$ parameters:

- k present population sizes,
- $(k - 1)$ ancestral population sizes,
- $(k - 1)$ speciation times, and
- $2(k - 1)^2$ migration rates.

See Fig. 5 for an example of divergence distribution with three species and migration and Exercise 3 (see Subheading 8.3) for a derivation of the number of migration rates in the general k -population model. For $k = 5, 6,$ and 7 we obtain 45, 66, and 91 parameters. Because the number of parameters becomes very large even for small k , Hey [28] suggests adding constraints to the migration rates, e.g., setting some rates to zero or introducing symmetry conditions where rates between populations are the same.

4 Approximating the Coalescent with Recombination Along Genomes

Before the genomic era, multilocus population genetics models were addressing a small fraction of the complete ancestral recombination graph (ARG) by considering independent loci. As sequencing technologies evolved and allowed access to larger samples of genomic diversity, this independence assumption had to be relaxed and more explicit modeling of the ARG was required. Yet the complexity of the coalescent with recombination process makes its application to genome-scale data sets very challenging. Two directions of analysis methods have emerged: simulation-based or spatial approximations along the genome. In this chapter we focus on the latter and refer to Kelleher et al. [29] and Staab et al. [30] for the former. Simonsen and Churchill [31] described the first model of the joint distribution of genealogies at two loci for two genomes. Wiuf and Hein [32] extended this approach and described the coalescent as a spatial process along the genome. McVean and

Cardin [33] further approximated the description with a Markov process. In this section we describe and discuss these types of approximations.

4.1 The Independent Loci Approach: Free Recombination Between, No Recombination Within

The simplest way to handle issues relating to the ancestral recombination graph is to divide the data into presumably independent loci. Such analyses are therefore restricted to candidate regions that are not too large (to avoid including a recombination point) and not too close (to ensure several recombination events happened between loci). Each region can then be described by a single underlying tree, reducing the analytical and computational load.

Using 15,000 loci distant from 10 kb totaling 7.4 Mb and the isolation model introduced above, Burgess and Yang [9] (Table 2, model (b) sequencing errors) find the following ancestral population sizes and speciation times estimates for human (H), chimpanzee (C), gorilla (G), orangutan (O), and macaque (M) ancestors: $\theta_{HC} = 0.0062$, $\theta_{HCG} = 0.0033$, $\theta_{HCGO} = 0.0061$, $\theta_{HCGOM} = 0.0118$ and $T_{HC} = 0.0038$, $T_{HCG} = 0.0062$, $T_{HCGO} = 0.0137$, $T_{HCGOM} = 0.0260$. Converting these estimates into time units requires an estimate of the substitution rate, either absolute or deduced from a scaling point. Using $\mu = 10^{-9}$ as an estimate for substitutions per year, this leads to an estimate of 3.8 My for the human–chimpanzee speciation, a very recent estimate. Using the same data, Yang [10] showed that the isolation-with-migration model was preferred. Yang finds a more ancient speciation time $T_{HC} = 0.0053$ (5.3 My with $\mu = 10^{-9}$) when migration is accounted for.

4.2 State-Space Model: Simonsen–Churchill Framework

The coalescent with recombination for two loci and two sequences is originally described in Simonsen and Churchill [31] as a continuous-time Markov chain backward in time with eight states as shown in Fig. 7. This Markov chain is given a careful treatment in the textbooks by Durrett [34, Section 3.1.1] and Wakeley [21, Section 7.2.4], and we therefore only briefly explain the basic properties of the model here.

A single sequence is either linked ($\bullet\text{---}\bullet$, $\times\text{---}\bullet$, $\bullet\text{---}\times$, or $\times\text{---}\times$) meaning that it contains material ancestral to the sample at both loci, or it is unlinked ($\bullet\text{---}$, $\text{---}\bullet$, $\text{---}\times$, or $\times\text{---}$) when it contains material ancestral to the sample at only one locus. The coalescent rate is one for any two sequences, and the recombination rate is $\rho/2$ for any linked sequence. The chain begins at time zero in state 1 with two linked sequences. After an exponential waiting time with rate $1 + \rho$ the chain enters state 8 with probability $1/(1 + \rho)$ or state 2 with probability $\rho/(1 + \rho)$. The transition from state 1 to state 8 is a coalescent event, and the left and right tree heights are identical. The transition from state 1 to state 2 is a recombination event that breaks apart one of the two sequences. All other transitions have similar interpretations. Common ancestry for a locus is marked

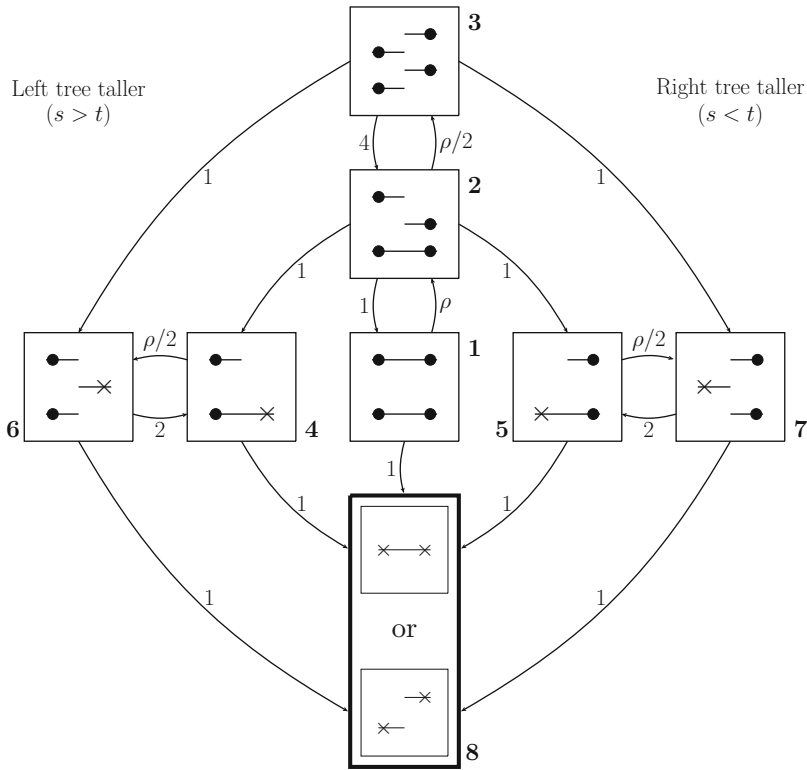


Fig. 7 State transition diagram for two loci and two sequences described as a continuous-time Markov chain backward in time. The figure is adapted from Figure 7.7 in Wakeley [21]. A line with a bullet or a cross at both ends is a linked sequence (ancestral material to the sample at both loci), whereas a line with a bullet or a cross at one end only is a sequence with ancestral material at one locus only. A cross denotes common ancestry. s and t denote the heights of the left and right trees, respectively

with a \times , so the transition from, e.g., state 1 to state 8 is a transition to the state $\times \rightarrow \times$.

The height S of the left tree is the first time at which the process enters one of the states 5, 7, or 8 (states with a left \times), and the height T of the right tree is the first time at which one of the states 4, 6, or 8 is entered (states with a right \times). When state 8 is entered from state 1 the two tree heights are identical. State 8 is absorbing because only the tree heights are of interest.

The two key ingredients for the state-space model are the conditional probability for staying in a state $P(T = s | S = s)$ and the conditional density $q(t|s)$ of a new tree height t conditional on a change and a previous tree height s . Hobolth and Jensen [35] show that the conditional probability of no change from the left to the right tree is

$$P(T = s | S = s) = e^s [e^{A_s}]_{11}, \tag{9}$$

and the conditional density $q(t|s)$ of T given $S = s$ and given $T \neq S$ is

$$q(t|s) = \begin{cases} e^{-(s-t)} \frac{[e^{\Lambda t}]_{12} + [e^{\Lambda t}]_{13}}{e^{-s} - [e^{\Lambda s}]_{11}} & t < s, \\ e^{-(t-s)} \frac{[e^{\Lambda s}]_{12} + [e^{\Lambda s}]_{13}}{e^{-s} - [e^{\Lambda s}]_{11}} & t > s, \end{cases} \tag{10}$$

where Λ denotes the 8×8 rate matrix from Fig. 7.

Wakeley [21, Section 7.2.4] noted that the transitions between state 4 and 6 and the transitions between state 5 and 7 can be removed from the chain if we are only interested in the tree heights. Actually, even more transitions can be removed from the chain. Note from Eqs. 9 and 10 that we only need the entries (1, 1), (1, 2), and (1, 3) in $e^{\Lambda t}$ for calculating the probability of the same tree height in the next position and the transition density conditional on a change. These entries can be found from a reduced rate matrix where states 4, 5, 6, and 7 are removed and the rate from states 2 and 3 to a new absorbing state equals 2. In other words, define the reduced rate matrix

$$\tilde{\Lambda} = \begin{pmatrix} -(1 + \rho) & \rho & 0 & 1 \\ 1 & -(3 + \rho/2) & \rho/2 & 2 \\ 0 & 4 & -6 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

where states are numbered 1, 2, 3, and 4. The holding time and transition density for the model are now given by Eqs. 9 and 10 with Λ substituted by $\tilde{\Lambda}$.

In the left plot in Fig. 8 we illustrate the probability (Eq. 9) of the same tree height in the left and right loci conditional on the tree height in the left locus and different recombination rates.

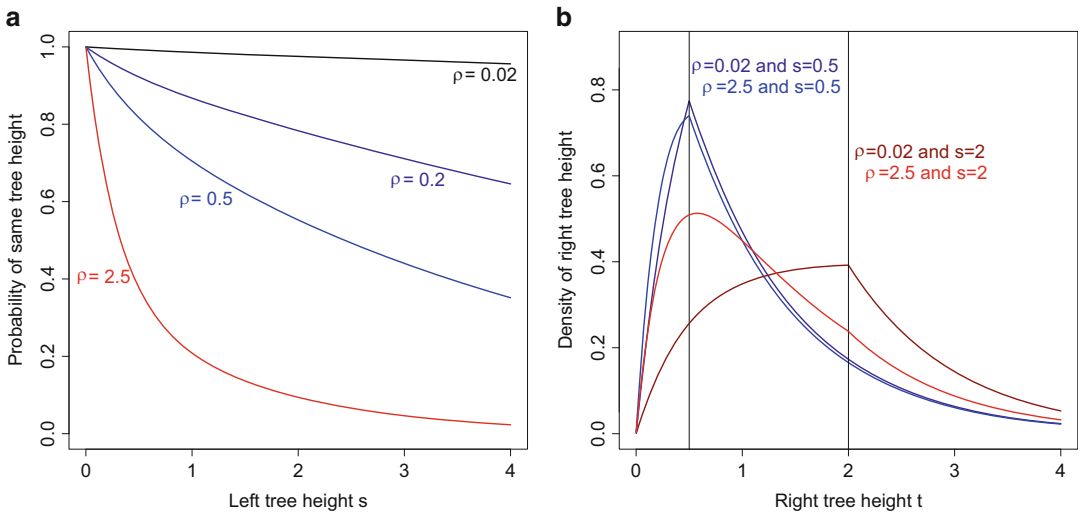


Fig. 8 (a) Probability of same tree height. (b) Density for right tree height conditional on the left tree height being equal to s and a recombination rate equal to ρ

As expected the probability for identical tree heights decreases with the height of the left tree and with the recombination rate.

In the right plot in Fig. 8 we illustrate the density (Eq. 10) of the right tree height conditional on the left tree height and a change in tree height. When the recombination rate increases, the density for the right tree height moves toward smaller tree heights. The reason is that at least one recombination is needed for having a change in tree height. We also observe that the density is continuous but not differentiable in the position of the left tree height.

**4.3 Time
Discretization: Setting
Up the Finite State
HMM**

Li and Durbin [14] and Mailund et al. [13] analyze pairs of sequences using a hidden Markov model (HMM). The hidden states are tree heights (times to the most recent common ancestor), and the tree height is discretized to obtain a finite hidden state space. The observed states of the HMM are alignment columns, with probabilities corresponding to a substitution process on the tree (see Fig. 9). In the Li and Durbin model, an infinite site model is assumed and observed states are converted to binary data, telling whether the site is heterozygous (one mutation) or homozygous (no mutation).

We now describe how we discretize time for the case of two sequences considered in the previous section. The discrete version of the Markov process is used to build a finite Markov chain along the two sequences. When the finite Markov chain is combined with a substitution process, we obtain an HMM as in Li and Durbin [14].

Let the discrete time points (backward in time) of the Markov chain be $d_0 = 0 < d_1 < d_2 < \dots < d_{M-1} < d_M = \infty$ and denote the corresponding states by $1, 2, \dots, M$. State m ($m \in \{1, \dots, M\}$) then corresponds to a tree height in the interval between d_{m-1} and d_m . The continuous stationary distribution is $\pi(t) = \exp(-t)$, and therefore the discrete times are chosen such that

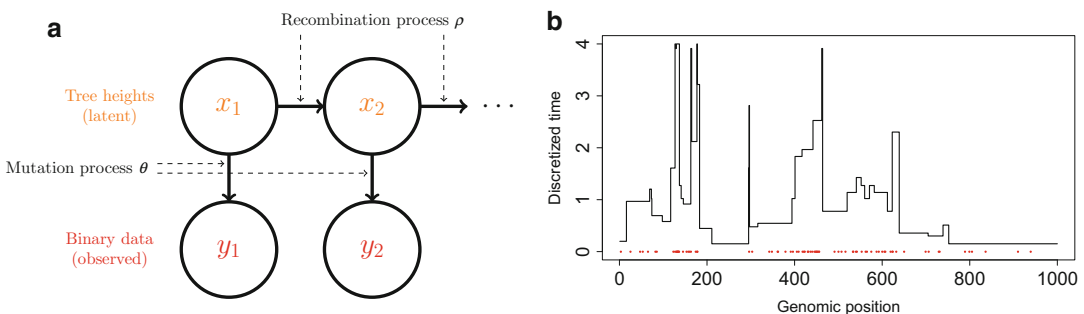


Fig. 9 (a) Graphical structure of the hidden Markov Model. (b) Simulation from the hidden Markov model

$1 - \exp(-d_m) = m/M$, or $d_m = -\log(1 - m/M)$, where we define $\log(0) = -\infty$.

We now get for $1 \leq \ell, r \leq M$ the joint density

$$\begin{aligned}
 &P(L = \ell, R = r) \\
 &= \begin{cases} \sum_{k \in \{5,7\}} \sum_{j \in \{5,7\}} \sum_{i \in \{1,2,3\}} [e^{\Lambda d_{\ell-1}}]_{1i} & \\ [e^{\Lambda(d_{\ell}-d_{\ell-1})}]_{ij} [e^{\Lambda(d_{r-1}-d_{\ell})}]_{jk} [e^{\Lambda(d_r-d_{r-1})}]_{k8} & \text{if } \ell < r \\ \sum_{i \in \{1,2,3\}} [e^{\Lambda d_{\ell-1}}]_{0i} [e^{\Lambda(d_{\ell}-d_{\ell-1})}]_{i8} & \text{if } \ell = r \\ P(L = r, R = \ell) & \text{if } \ell > r. \end{cases}
 \end{aligned}
 \tag{11}$$

The reason for the first case is that in order for the left tree height to be in state $\ell < r$, it must be in state 1, 2, or 3 at time $d_{\ell-1}$ and in state 5 or 7 at time d_{ℓ} (i.e., there have been no coalescent events before time $d_{\ell-1}$ and a left coalescent event between time $d_{\ell-1}$ and d_{ℓ}), and similarly it must still be in state 5 or 7 at time d_{r-1} and in state 8 at time d_r (i.e., there have been no coalescent events between time d_{ℓ} and time d_{r-1} and a right coalescent event between time d_{r-1} and time d_r). The next case corresponds to no coalescent events before time $d_{\ell-1}$ and both a left and a right coalescent event between time $d_{\ell-1}$ and d_{ℓ} . The last case is due to symmetry of the chain.

From the joint tree states (ℓ, r) we easily get the conditional tree states

$$P_{(\ell,r)} = P(r|\ell) = P(R = r|L = \ell) = \frac{P(L = \ell, R = r)}{P(L = \ell)},$$

where $P(L = \ell) = \sum_r P(R = r, L = \ell)$. These probabilities are used in the HMM.

4.4 Careful Treatment of Mutation Process

A careful treatment of the mutation process allows for a more coarse binning procedure and is needed to avoid biasing the results. In continuous time the probability for a mutation given a tree height t is given by $\mu(t) = 1 - \exp(-\theta t)$, and the stationary tree height distribution is $\pi(t) = \exp(-t)$. The probability of a mutation conditionally on the hidden state m becomes

$$\begin{aligned}
 \mu_m &= p(y_i = 1 | x_i = m) \\
 &= p(y_i = 1 | t \in (d_{m-1}, d_m)) = \frac{p(y_i = 1 | t \in (d_{m-1}, d_m))}{p(t \in (d_{m-1}, d_m))} \\
 &= \frac{\int_{d_{m-1}}^{d_m} p(y_i = 1 | t) \pi(t) dt}{\int_{d_{m-1}}^{d_m} \pi(t) dt} = \frac{\int_{d_{m-1}}^{d_m} (1 - e^{-\theta t}) e^{-t} dt}{\int_{d_{m-1}}^{d_m} e^{-t} dt} \tag{12} \\
 &= 1 - e^{-\theta d_{m-1}} \frac{(1 - e^{-(1+\theta)(d_m - d_{m-1})})}{(1 + \theta)(1 - e^{-(d_m - d_{m-1})})}.
 \end{aligned}$$

Note that with a fine discretization we have that the interval $d_m - d_{m-1}$ is small and the first-order Taylor expansion $\exp(-az) \approx 1 - az$ for z small gives

$$p(y_i = 1 | x_i = m) \approx 1 - e^{-\theta d_{m-1}},$$

as perhaps expected. We are, however, discretizing the interval $[0, \infty[$, so it is not possible to avoid one or more large bins. Generally we have found that a careful treatment of the mutation process is crucial for accurate inference [36].

4.5 Statistical Inference of Population Parameters from Sequences

Here we choose to focus on three inference methods for estimating the recombination rate. The first method is based on the full likelihood obtained from the classical forward (or backward) algorithm for HMMs. The second is based on the distribution of the distance between segregating sites. This summary statistics was used in Harris and Nielsen [37] for demographic inference. It is sometimes also described as the distribution of the distance between heterozygote sites, runs of homozygosity, or the nearest-neighbor distribution. The third summary statistics is the probability that two sites at certain distance apart are both heterozygote sites. This probability is closely related to the pair correlation function from spatial statistics [36] and to the zygosity correlation introduced in [38].

4.5.1 Summary Statistics: Runs of Homozygosity and Pair Correlation

Recall that in continuous time the probability for a mutation given a tree height t is given by $\mu(t) = 1 - \exp(-\theta t)$, and the stationary tree height distribution is $\pi(t) = \exp(-t)$. The marginal probability for a mutation is therefore given by

$$\int_0^\infty \mu(t) \pi(t) dt = \theta / (1 + \theta). \tag{13}$$

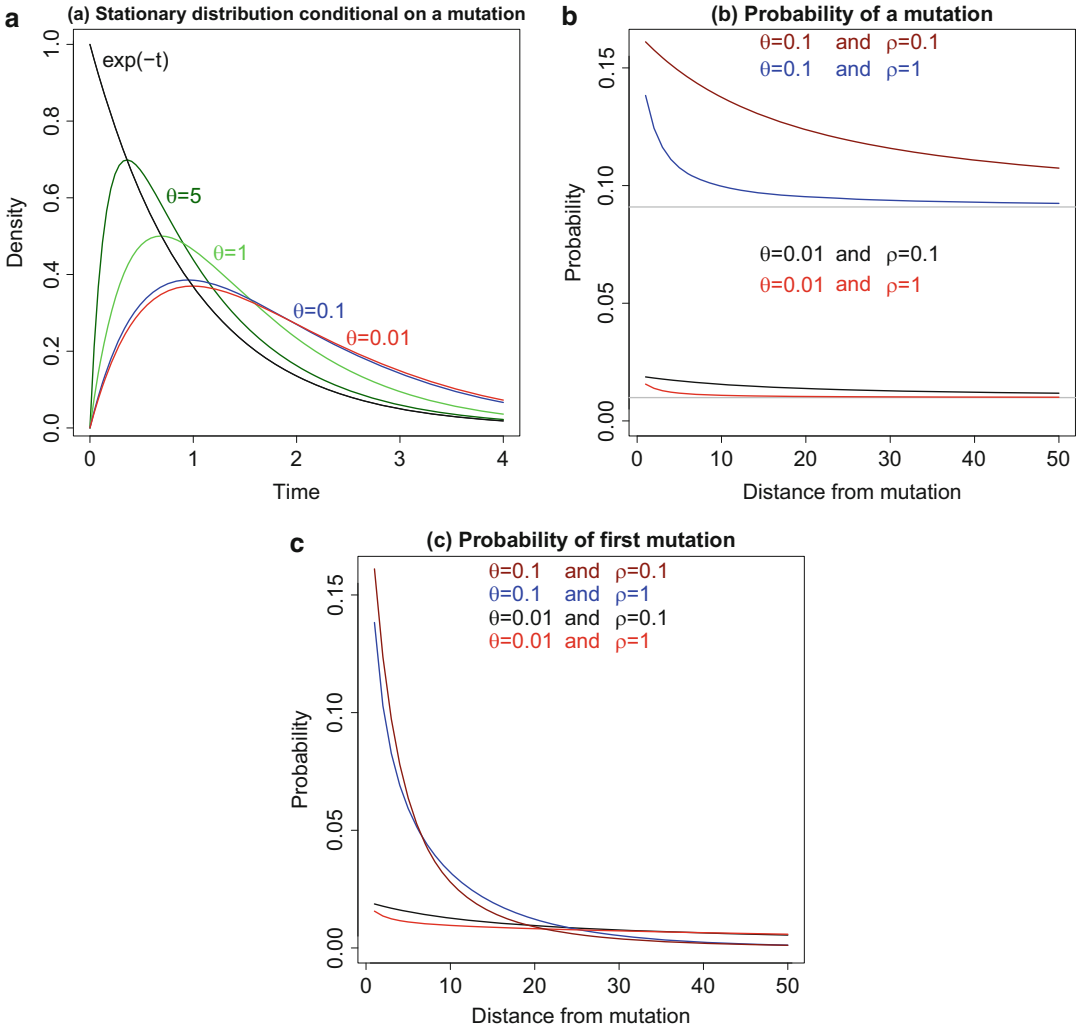


Fig. 10 (a) Stationary distribution of tree height conditional on a mutation. (b) Probability of a mutation at various distances away from a mutation. (c) Probability of the first mutation at various distances away from a mutation

We also get the stationary distribution

$$\phi(t) = \frac{\mu(t)\pi(t)}{\int_0^\infty \mu(t)\pi(t)dt} = \frac{1 + \theta}{\theta} e^{-t}(1 - e^{-\theta t})$$

for a tree height t conditional on a mutation. Figure 10a shows $\phi(t)$ for different values of θ . Note that small mutation rates imply a higher tree height when we condition on a mutation. In discrete time the probability for a mutation given a tree height m was given by Eq. 12. Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)$ be the vector of mutation

probabilities. The stationary distribution $\boldsymbol{\phi} = (\phi_1, \dots, \phi_M)$ for a state m conditional on a mutation is given by

$$\phi_\ell = \frac{\mu_\ell \pi_\ell}{\sum_{m=1}^M \mu_m \pi_m},$$

where $\pi_m = 1/M$ because this is how the time discretization was chosen.

The probability for a mutation at a distance r from a typical mutation is then given by

$$\kappa(r) = \boldsymbol{\phi}' P^r \boldsymbol{\mu},$$

where $'$ denotes vector transpose. In Fig. 10b we show $\kappa(r)$ as a function of ρ and θ . Note that the curves converge to $\theta/(1 + \theta)$ and that the behavior for small r is determined by the recombination rate.

The distribution of runs of homozygosity is given by

$$\nu(r) = \boldsymbol{\phi}' [P \text{diag}(e - \boldsymbol{\mu})]^{r-1} P \boldsymbol{\mu}.$$

Here $e = (1, \dots, 1)$ is the vector of length M with 1 in every entry and $\text{diag}(e - \boldsymbol{\mu})$ is the diagonal matrix with $e - \boldsymbol{\mu}$ on the diagonal. In Fig. 10c we show $\nu(r)$ as a function of ρ and θ .

4.5.2 Parameter Estimation

We estimate the mutation rate using an estimating equation based on the marginal probability for a mutation (Eq. 13). If the observed frequency of a mutation is \hat{p} , then the mutation rate is $\hat{\theta} = \hat{p}/(1 - \hat{p})$ (see left plot in Fig. 11). The recombination rate is estimated using maximum likelihood for the HMM and goodness of fit for the pair correlation (see middle plot in Fig. 11) and runs of homozygosity (see right plot in Fig. 11).

We simulated 50 sequences of length 20,000 base pairs and with mutation rate $\theta = 0.1$ and recombination rate $\rho = 0.1$. We estimated the mutation rate using the estimating equation and the recombination rate using maximum likelihood and the HMM, and goodness of fit for the pair correlation and nearest neighbor (Fig. 12) [35]. As expected the HMM procedure shows the best results because here we are using all the available information. It seems, however, that we are not losing too much power when applying the pair correlation function. This is in contrast to the nearest-neighbor summary statistics that perform much worse than the other two methods.

We have provided a detailed treatment of the main components involved in an analysis of pair of DNA sequences based on an HMM derived from coalescent theory. Pairwise sequentially Markov coalescent (PSMC) models have been extensively applied to various organisms, see, for instance [39–43].

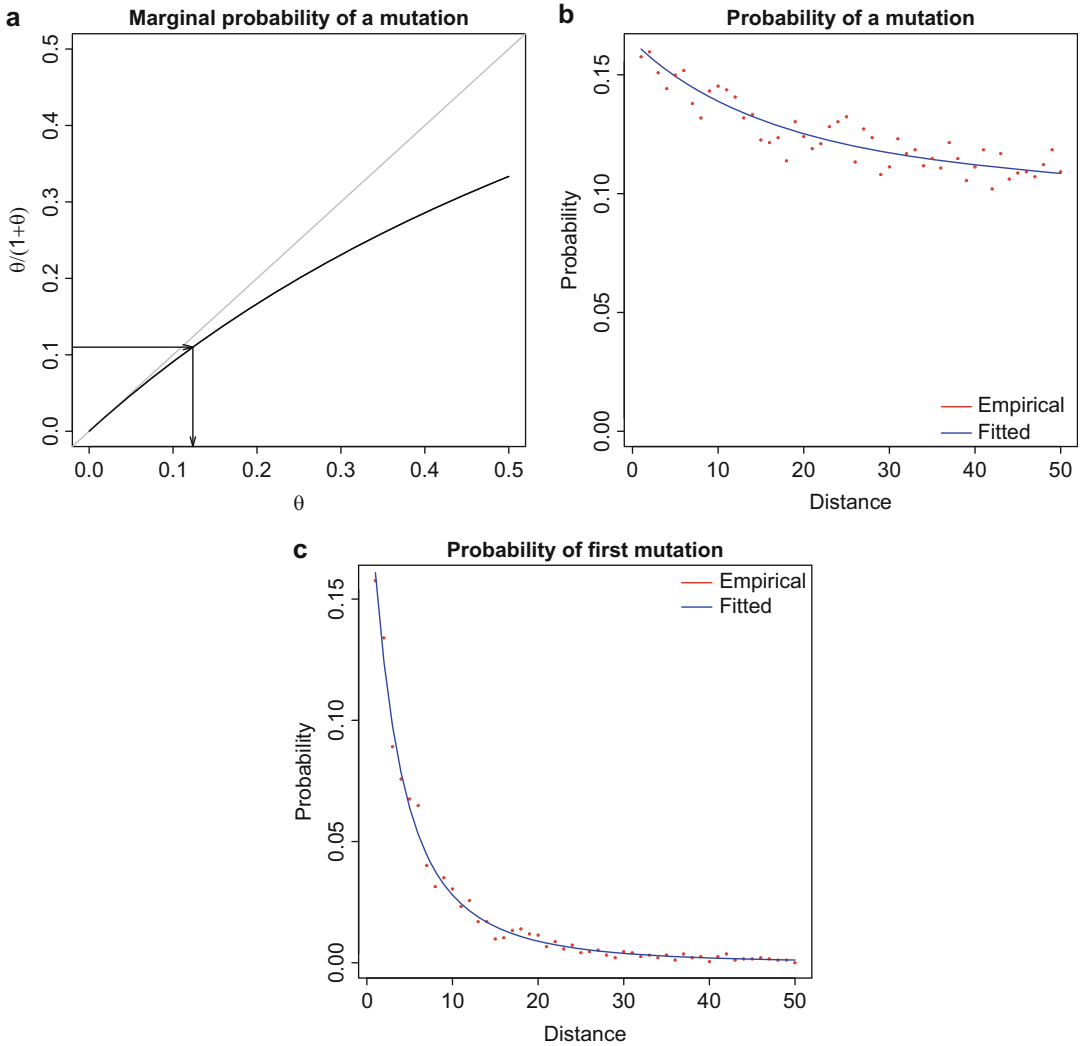


Fig. 11 Parameter estimation for summary statistics. (a) The mutation rate θ is estimated from the observed number of mutations and length of the region. (b) The recombination rate ρ is estimated using the empirical distribution of a mutation at various distances from a mutation. (c) The recombination rate is estimated using the empirical distribution of the first mutation from a mutation

5 Extending the Pairwise Sequentially Markov Coalescent

Extending the SMC to more than two genomes has proved to be challenging. The number of hidden states becomes prohibitive, as several divergence times have to be modeled and combined with distinct possible topologies. Further simplifications are therefore needed to account for an increasing number of genomes.

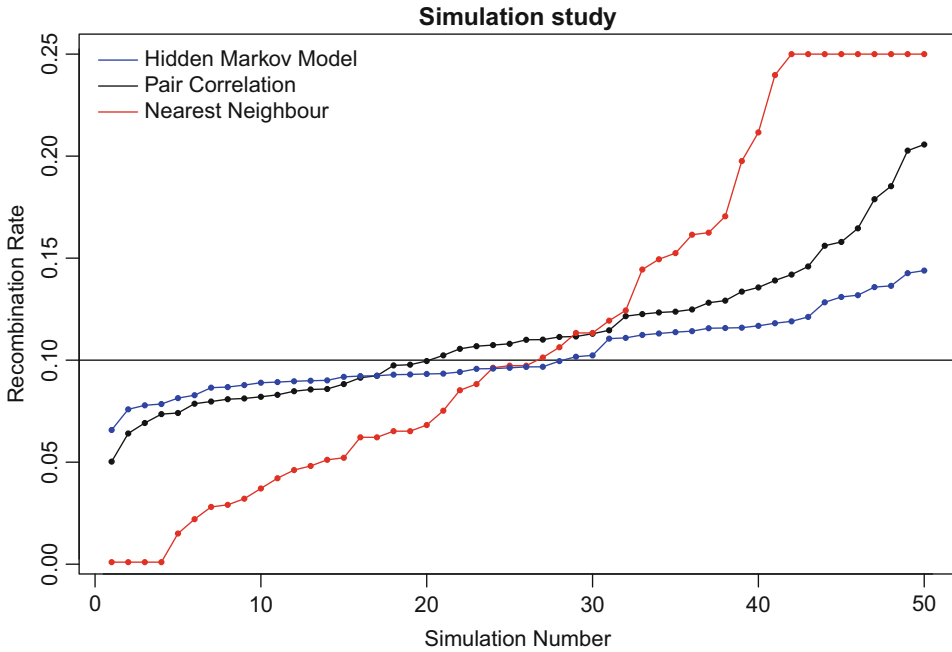


Fig. 12 Results of parameter estimation for simulation study. The pair correlation summary performs rather well compared to the full HMM data analysis. Nearest neighbor is a poor summary statistics

5.1 From 2 to n Genomes

5.1.1 The Multiple Sequentially Markov Coalescent (MSMC)

Schiffels and Durbin [15] proposed to extend the PSMC model [14] to more than two haploid genomes by modeling the most recent coalescence event in the sample. In this framework, the hidden states of the model are a combination of divergence times, taken from a discretized distribution, and identity of the corresponding haplotypes involved. The rationale for such simplification was that the PSMC showed poor resolution in the recent past [14], and considering more genomes would bring additional signal. The drawback of this implementation is that the more genomes are considered, the more “shifted” toward the present is the timeframe where population parameters can be inferred. As a result, the authors reported that with more than 8 diploid individuals (16 haploid genomes), parameters can virtually not be estimated (see also [44] for an illustration of this effect with simulations). Another consequence of this approach is that the recombination rate parameter cannot be reliably estimated [15]. The MSMC was used to infer the recent history of human population. In particular, the authors introduced the possibility to label individuals and look at cross-coalescence rate between groups, a way to get a fine-tuned view of population divergence [15, 45].

5.1.2 The Demographic Inference with Composite Approximate Likelihood (diCal)

An alternative approach was introduced by Song and colleagues [16–18]. The demographic inference with composite approximate likelihood (diCal) approach is based on the conditional sampling distribution, which computes the likelihood of one genome conditioned on the observation of others. Using the so-called composite likelihood formula, it is therefore possible to compute the likelihood of the data for n genomes as the product of the likelihood of one genome given the $n - 1$ other ones and the likelihood the remaining $n - 1$ genomes:

$$P(D_{1\dots n}|\theta) = \Pr(D_1|D_{2\dots n}, \theta) \times P(D_{2\dots n}|\theta),$$

where θ is the set of model parameters and $D_{1\dots n}$ denotes the data set with n genomes. By further noting that

$$P(D_{2\dots n}|\theta) = P(D_2|D_{3\dots n}, \theta) \times P(D_{3\dots n-1}|\theta)$$

the likelihood of the full data set can be computed by recursion. The terms $P(D_i|D_{i+1\dots n})$ form the conditional sampling distribution (CSD). Paul et al. [16] proposed a way to compute the CSD at the cost of introducing several additional hypotheses: (a) the haplotypes upon which the sample is conditioned are considered independent, that is, no coalescence events involving these haplotypes are allowed and (b) mutations can only occur once in any lineage (infinite site hypothesis). The likelihood resulting from this approximated CSD is therefore not exact. This approach was introduced by Li and Stephens [46] and is referred to as the product of approximate conditionals (PAC) model. Under the PAC model, the likelihood depends on the order by which the data is conditioned, which can be circumvented with permutation procedures. While the CSD-based SMC does not have the same drawbacks as the MSMC of Schiffels and Durbin [15], its computational efficiency decreases as the number of haplotypes considered increases and becomes impractical for more than 10 genomes [19]. An elegant feature of the diCal approach is that it can be extended to more complex demographic models, including population structure and gene flow [18, 45]. Such extension is of interest as the SMC approximation has been shown to be sensitive to strong population structure [47].

5.1.3 Extending the SMC with Conditional Site Frequency Spectra (CSFS)

In order to use the large amount of data available in “1000 genomes” projects, Terhorst et al. [19] extended the PSMC in a different direction. Instead of modeling the genealogy of the complete sample, the authors proposed to model the divergence of two haplotypes (the PSMC model) as *hidden states*, yet considering the full set of genomes as *observed states*. In this approach, the transition probabilities of the coalescent HMM are similar to the PSMC (or to be more precise, similar to the MSMC with two haplotypes, as the original PSMC uses the SMC of McVean and Cardin [33] and not

the SMC' of Marjoram and Wall [48]), but the emission probabilities are extended to account for the full site frequency spectrum of hundreds of genomes. This *conditional site frequency spectrum* (CSFS) is computed using coalescence theory, offering a generalization of the Poisson random field (PRF) model introduced by Sawyer and Hartl [49]. Just like the original PRF, however, the CSFS ignores linkage of observed states, only linkage between the two conditioned haplotypes is modeled via the SMC. Additional data reduction steps are therefore required to ensure that the independence condition of sampled sites is met.

5.1.4 *Explicit Reconstruction of the Ancestral Recombination Graph*

While the ARG contains all historical information about a sample of genomes, genomes themselves contain very little information regarding the underlying ARG. As a result, in most statistical inference methods the ARG is treated as a variable accounted for, but not directly inferred. In the SMC models presented above, this is taken care of by the hidden Markov methodology, which computes a likelihood for a given sample by summing over all possible ARG (via the so-called *forward* algorithm). The Viterbi algorithm and the posterior decoding procedure are HMM algorithms that allow to reconstruct a posteriori the most likely ARG for a sample, such procedures are notably used for the inference of patterns of incomplete lineage sorting along genomes [11, 12, 50, 51]. Yet the variance in such estimation is typically very large [12].

Rasmussen et al. [20] proposed a different approach: they developed a Bayesian sampler of ARGs conditioned on a set of genome sequences. Similar in principle to the PAC and CSD approaches, the authors proposed to generate the ARG of n genomes conditioned on the ARG of $n - 1$ genomes, a procedure they refer to as *threading*. The generated ARGs can then be used to infer evolutionary processes of interest. Palacios et al. [52] developed a non-parametric method that allows to estimate the variation in time of the effective population size based on such reconstructed ARG. Rasmussen et al. further showed that while the model used for inference is purely neutral, the a posteriori inferred ARG contains signature of selection, visible for instance as a decrease of the time of the most common ancestor of two samples in the data close to coding sequences. Such approaches offer promising avenues for the development of new statistical methods to detect genomic regions with unusual history.

5.2 *The Case of Multiple Species*

Hobolth et al. [11] developed a hidden Markov model (HMM) to infer the ancestral recombination graph between three closely related species. Because this model only contains one haploid genome per species, it only allows to infer population parameters in the ancestral species. Dutheil et al. [12] reparametrized this model in the context of the sequentially Markov coalescent. In contrast to the previous approaches, only four hidden states were

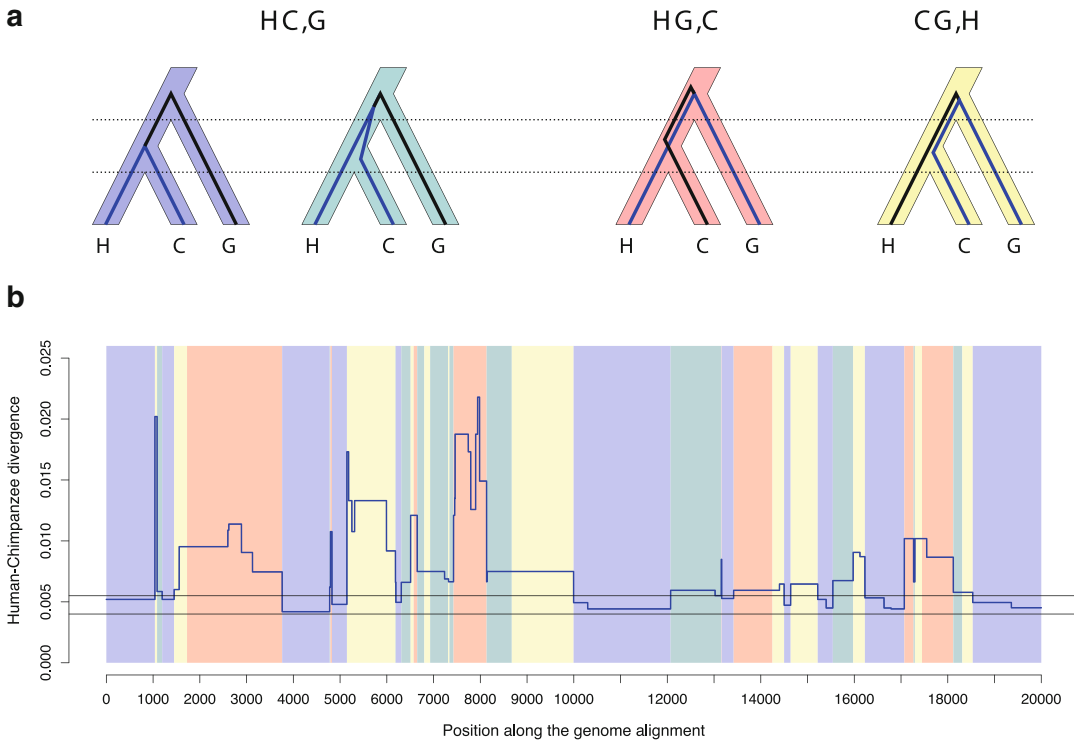


Fig. 13 The coalescent process along genomes of three closely related species. **(a)** Four archetypes of coalescence scenarios with three species, exemplified with human, chimpanzee, and gorilla. In the first scenario, human and chimpanzee coalesce within the human–chimpanzee common ancestor. In the three other scenarios, all sequences coalesce within the common ancestor of all species, with probability 1/3 depending on which two sequences coalesce first. **(b)** Example of genealogical changes along a piece of an alignment. The alignment was simulated using the true coalescent process and parameters corresponding to the human–chimpanzee–orangutan history. The blue line depicts the variation along the genome of the human–chimpanzee divergence. The background colors depict the change in topology, red and yellow corresponding to incomplete lineage sorting. Each change in color or break of the blue line is the result of a recombination event

considered, corresponding to four alternative scenarios of lineage segregation (Fig. 13). In states 1 and 2, the genealogy is consistent with the phylogeny and lineages segregate in the same order as the species. In states 2, 3, and 4, allele divergence predates the first speciation event and ancestral polymorphism persists between the two speciation events, leading to incomplete lineage sorting. The scenarios depicted by states 2, 3 and 4 are equally likely, and in the case of states 3 and 4, the resulting topology is inconsistent with the phylogenetic tree. This model therefore does not rely directly on divergence variation along the genome alignment but uses patterns of topology variation instead to compute the speciation times and ancestral population sizes.

Using this approach, Hobolth et al. estimated a speciation time between human and chimpanzee around 4.1 My and a large

ancestral effective population size of 60,000 for the human–chimpanzee ancestor. Dutheil et al. [12] found similar estimates with the same data set while accounting for substitution rate variation across sites and estimated an average recombination rate of 1.7 cM/Mb. With sequencing of more great ape genomes, this approach allowed to estimate population size in several ape ancestors ([27, 50, 53], reviewed in [54]). As ILS is a proxy for ancestral effective population size, a major result of these studies is that the distribution of ILS is not uniform along the genome. For instance, it is reduced in proximity of genes, a pattern that can be explained by background selection [27, 50]. Large regions of the X chromosome were also found to be devoid of ILS, a pattern resulting from recurrent selective sweeps along the chromosomes [55].

6 Specific Issues Faced When Dealing with Genomic Data

In previous sections we discussed population genetic models and methods for parameter estimation. We now describe several challenges encountered when analyzing whole-genome data sets, at the intra- and interspecific levels.

6.1 Sequencing Errors and Rate Variation

Sequencing errors are a well-described source of bias in population genetics analyses, resulting in an excess of singletons [56]. At both the intra- and interspecific/population level, such error therefore leads to incorrect estimates of local divergence, in particular for recent times. When more divergent sequences are compared, for instance, from distinct species, the issue becomes more complex as the error rate differs between and within sequences due to coverage variation, but also properties of the genome (base composition, repeated elements, etc.). Such errors result in a departure from the molecular clock hypothesis, thus potentially leading to biases in parameter estimates, such as asymmetries in genealogy frequencies [57, 58]. In this respect, data preprocessing becomes a crucial step in any genomic analysis. Methods would also benefit in many cases of inclusion of a proper modeling of such errors. Burgess and Yang noticed that sequencing errors can be seen as a contemporary acceleration in external branches, resulting in an extra branch length [9]. Such an extra length can be easily accommodated in many models. It has to be noted that only a differential in error rates between lineages results in a departure from molecular clock, and in such approaches, one still has to consider that at least one sequence is error-free. In addition, as noted by the authors, assuming a constant error rate over all genomic positions may also turn out to be inappropriate, and better models should allow this rate to vary across the sequence. Such approaches still have to be explored. Moreover, sequencing errors are not distinguishable from lineage-specific acceleration (or deceleration in another species). In that

respect, sequence quality scores can be a valuable source of information. They are currently used to preprocess the data by removing doubtful regions, but can ultimately be used in the modeling framework.

The substitution rate also varies along the genome, which potentially affects the reconstruction of sequence genealogy, a phenomenon well known by phylogeneticists. In such case the tools developed for phylogenetic analysis can be applied with a reasonable cost. This generally consists in assuming a prior distribution of the site-specific rate and integrating the likelihood over all possible rates [8, 9, 12]. Alternatively, one can also use one or more out-group sequences to calibrate the rate, as in [6, 7].

6.2 Diploid Data and Phasing

While sequencing of diploid individuals allows to infer the two alleles present at heterozygous positions, establishing how these alleles are combined on each homologous chromosome requires an additional, error-prone step calling *phasing*. Analyses based on the comparison of individuals from distinct species do not require such information, as the coalescence time of two alleles from the same species is expected to have happened much after the speciation time of the compared species. In such case alleles at each heterozygous position can be sampled randomly [13] in order to build a composite haploid genome. The same rationale applies with respect to the use of the human reference genome, a composite genome obtained from multiple individuals. Conversely, inferences at the population level typically rely on the modeling of haploid genomes and therefore require phased data. A notable exception is the PSMC [14], as well as its extension SMC++ [19], which, when applied to one diploid individual, only requires the knowledge of the position of heterozygous positions.

6.3 Structural Variation and Genome Alignment

Genome data are intrinsically fragmented, firstly because of chromosomal organization, but also because of rearrangements that prevent molecule-to-molecule alignment from one species to another. A genome data set is therefore a set of distinct alignments, one per synteny block. Synteny information can only be extracted when individual genomes are available, which is typically not the case for most “re-sequencing” data sets. At the population level, however, such large-scale variation is considered negligible (but see, for instance, [59] for an exception), while it becomes more prominent when genomes from distinct species are compared. In such cases, a genome alignment is constructed with potential errors ultimately leading to the comparison of nonhomologous regions. So far, the only way to deal with such errors is to restrict the analysis on regions where orthology can be unambiguously resolved, mostly by removing short synteny blocks and regions that contain a high proportion of repeated elements, gaps, and duplications.

7 Discussion

Studying the speciation process with genome data implies new modeling challenges, as the basic configuration of a population genetics data set is drastically changed: instead of having a few loci sequenced in several individuals, we have an (almost) exhaustive set of loci sequenced in several individuals for multiple closely related species. The change involves the spatial dimension, but also time, as the process under study occurred much further back in time than the ones that are commonly studied with a “standard” population genetics data set. The use of the spatial signal has a major consequence, namely, that recombination has to be taken into account, even if it is not directly modeled.

Apart from these considerations, ancestral population genomics, as population genetics, heavily relies on the study of sequence genealogy, its shape, but also its variation. The underlying models build on existing intraspecies population modeling, as they only need to add the species divergence process, that is, a moment in time where two populations stop exchanging genetic material and evolve fully independently. The simplest isolation model assumes that the speciation is instantaneous, while the isolation-with-migration model assumes that the two neo-species can still exchange some material, at least for a certain time after the split. Such a model is not different from a pure isolation model where the ancestral population is structured into two subpopulations: in the first case the speciation time is defined as the time of the split, while in the second case it is the time of the last genetic exchange. Recent work on primates [10] suggests that the speciation of human and chimpanzee was not instantaneous. If the average divergence of the human and chimpanzee is a bit more than 6 My (using widely accepted mutation rate), then the split of the two species initiated around 5.5 My ago, and the last genetic exchange can be dated around 4 My.

The fact that we sample a large number of positions in the genome thus appears to have the power to counterbalance the reduced sampling of individuals within population, allowing the estimation of demographic parameters in the ancestor. Nonetheless, complexity limits are rapidly reached, when considering, for example, three closely related species that can exchange migrants. More complex demographic scenarios, incorporating, for instance, variation in population sizes, will also add additional parameters that might not all be identifiable.

If the ancient speciation processes have left signatures in the contemporary genomes, we do not know yet how far back in time this is true. Intuitively, the signal is maximal when the variation in divergence due to polymorphism is large enough compared to the total divergence. The divergence due to polymorphism is

proportional to the ancestral population size, while the divergence of species is only dependent on the time when it happened. So the further back in time we are looking at, the bigger the population sizes need to be so that the ancient polymorphism leaves a signature in the total divergence time. In addition to this, one has to take into consideration sequence saturation due to the too large number of substitutions that accumulated since ancient splits, and the fact that demographic scenarios complexity increases with time. For instance, when considering the evolution of a species over several millions of generations, the probability that a bottleneck, resetting the signal from past events, occurred once is not negligible.

We are in the population genomics era. Data sets are available that allow us to understand the evolutionary processes that are associated with the formation and evolution of species. Analyzing such data sets with the current methodologies however offers major challenges: (1) developing the appropriate computational tools able to handle such data sets with current machines (both in terms of processor speed and memory usage) and (2) design realistic models with enough complexity to capture the most important historical events while remaining computationally tractable.

8 Exercises

8.1 ILS in Primates

Assuming that there are 5 My between the speciation times of human with the gorilla and the orangutan, that the HG ancestral effective population size was 50,000, what is the expected amount of ILS between human, gorilla, and orangutan? Assuming that another 2.5 My separates the speciations of human with chimpanzee and gorilla, with an HC effective ancestral population size of 50,000, what is the expected amount of ILS between human, chimpanzee, and orangutan? We assume a generation time of 20 years for all extant and ancestral primates.

8.2 Estimating Ancestral Population Size from the Observed Amount of ILS

Given that 30% of incomplete lineage sorting is observed between human, chimpanzee, and gorilla and assuming a generation time of 20 years and a that 2.5 My separate the splits between human/chimpanzee and human—chimpanzee/gorilla, what is the effective ancestral population size compatible with this observed amount? Using Burgess and Yang's method [9], a researcher finds a higher estimate of N_e than expected. What could explain this discrepancy?

8.3 Number of Migration Rates in the General k -Population IM Model

In this exercise we show that a k -population IM model has $2(k-1)^2$ migration rates.

1. Starting at the bottom of the k -population IM model argue that the number of migration rates at the level of k populations is $k(k-1)$.

2. Moving up to the next level where $(k - 1)$ populations are present (one of them being an ancestral population, we assume that there two speciation events are never simultaneous) argue that the new ancestral population introduces $2(k - 1)$ new migration rates.
3. Moving up yet another level where $(k - 2)$ populations are present argue that the new ancestral population introduces $2(k - 2)$ new migration rates.
4. Show that the total number of migration rates is $2(k - 1)^2$.

Acknowledgements

We thank an anonymous reviewer for constructive comments and detailed suggestions on the manuscript.

References

1. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. *Nature* 526(7571):68–74
2. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubí C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegmund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilson C, Andrés AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T (2013) Great ape genetic diversity and population history. *Nature* 499(7459):471–475
3. Weigel D, Mott R (2009) The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol* 10(5):107
4. Siepel A (2009) Phylogenomics of primates and their ancestral populations. *Genome Res* 19(11):1929–1941
5. Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* 68(2):444–456
6. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441(7097):1103–1108
7. Yang Z (2002) Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* 162(4):1811–1823
8. Wang Y, Hey J (2010) Estimating divergence parameters with small samples from a large number of loci. *Genetics* 184(2):363–379
9. Burgess R, Yang Z (2008) Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol Biol Evol* 25(9):1979–1994
10. Yang Z (2010) A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome Biol Evol* 2:200–211
11. Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet* 3(2):e7

12. Dutheil JY, Ganapathy G, Hobolth A, Mailund T, Uyenoyama MK, Schierup MH (2009) Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* 183(1):259–274
13. Mailund T, Dutheil JY, Hobolth A, Lunter G, Schierup MH (2011) Estimating speciation time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS Genet* 7(3):e1001319
14. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature* 475(7357):493–496
15. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nat Genet* 46(8):919–925. <http://www.nature.com/ng/journal/v46/n8/full/ng.3015.html>
16. Paul JS, Steinrücken M, Song YS (2011) An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics* 187(4):1115–1128
17. Sheehan S, Harris K, Song YS (2013) Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* 194(3):647–662. <https://doi.org/10.1534/genetics.112.149096>
18. Steinrücken M, Paul JS, Song YS (2013) A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theor Popul Biol* 87:51–61
19. Terhorst J, Kamm JA, Song YS (2017) Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet* 49(2):303–309
20. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014) Genome-wide inference of ancestral recombination graphs. *PLoS Genet* 10(5):e1004342
21. Wakeley J (2008) *Coalescent theory: an introduction*, 1st edn. Roberts and Company Publishers, Arapahoe County
22. Hein J, Schierup MH, Wiuf C (2005) *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, Oxford
23. Tavaré S (2004) *Ancestral inference in population genetics*, vol 1837. Springer, New York, pp 1–188
24. Takahata N, Nei M (1985) Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110(2):325–344
25. Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158(2):885–896
26. Tavaré S (1979) A note on finite homogeneous continuous-time Markov chains. *Biometrics* 35:831–834
27. Hobolth A, Andersen LN, Mailund T (2011) On computing the coalescence time density in an isolation-with-migration model with few samples. *Genetics* 187(4):1241–1243
28. Hey J (2010) Isolation-with-migration models for more than two populations. *Mol Biol Evol* 27(4):905–920
29. Kelleher J, Etheridge AM, McVean G (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol* 12(5):e1004842
30. Staab PR, Zhu S, Metzler D, Lunter G (2015) Scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* 31(10):1680–1682
31. Simonsen N, Churchill N (1997) A Markov chain model of coalescence with recombination. *Theor Popul Biol* 52(1):43–59
32. Wiuf C, Hein J (1999) Recombination as a point process along sequences. *Theor Popul Biol* 55(3):248–259
33. McVean GAT, Cardin NJ (2005) Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci* 360(1459):1387–1393
34. Durrett R (2008) *Probability models for DNA sequence evolution*. Probability and its applications. Springer, New York
35. Hobolth A, Jensen JL (2014) Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theor Popul Biol* 98:48–58. <https://doi.org/10.1016/j.tpb.2014.01.002>
36. Nielsen SV, Simonsen S, Hobolth A (2016) Inferring population genetic parameters: particle filtering, HMM, ripples K-function or runs of homozygosity? In: *Algorithms in bioinformatics*. Lecture notes in computer science. Springer, Cham, pp 234–245
37. Harris K, Nielsen R (2013) Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet* 9(6):e1003521. <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003521>
38. Lynch M, Xu S, Maruki T, Jiang X, Pfaffelhuber P, Haubold B (2014) Genome-wide linkage-disequilibrium profiles from single individuals. *Genetics* 198(1):269–281. <https://doi.org/10.1534/genetics.114.166843>

39. Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H (2016) PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Mol Ecol* 25(5):1058–1072. <https://doi.org/10.1111/mec.13540>
40. Deinum EE, Halligan DL, Ness RW, Zhang YH, Cong L, Zhang JX, Keightley PD (2015) Recent evolution in *Rattus norvegicus* is shaped by declining effective population size. *Mol Biol Evol* 32(10):2547–2558. <https://doi.org/10.1093/molbev/msv126>
41. Thomas CG, Wang W, Jovelín R, Ghosh R, Lomasko T, Trinh Q, Kruglyak L, Stein LD, Cutter AD (2015) Full-genome evolutionary histories of selfing, splitting, and selection in *Caenorhabditis*. *Genome Res* 25(5):667–678. <https://doi.org/10.1101/gr.187237.114>
42. Nadachowska-Brzyska K, Li C, Smeds L, Zhang G, Ellegren H (2015) Temporal dynamics of avian populations during pleistocene revealed by whole-genome sequences. *Curr Biol* 25(10):1375–1380. <https://doi.org/10.1016/j.cub.2015.03.047>
43. Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N, Simões ZLP, Allsopp MH, Kandemir I, De la Rúa P, Pirk CW, Webster MT (2014) A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nat Genet* 46(10):1081–1088. <http://www.nature.com/ng/journal/v46/n10/full/ng.3077.html>
44. Dutheil JY (2017) Hidden Markov models in population genomics. *Methods Mol Biol* 1552:149–164
45. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Ávila Arcos MC, Malaspina AS, Eriksson A, Moltke I, Metspalu M, Homburger JR, Wall J, Cornejo OE, Moreno-Mayar JV, Korneliussen TS, Pierre T, Rasmussen M, Campos PF, Damgaard PDB, Allentoft ME, Lindo J, Metspalu E, Rodríguez-Varela R, Mansilla J, Henriksen C, Seguin-Orlando A, Malmström H, Stafford T, Shringarpure SS, Moreno-Estrada A, Karmin M, Tambets K, Bergström A, Xue Y, Warmuth V, Friend AD, Singarayer J, Valdes P, Balloux F, Lebereiro I, Vera JL, Rangel-Villalobos H, Pettener D, Luiselli D, Davis LG, Heyer E, Zollikofer CPE, Ponce de León MS, Smith CI, Grimes V, Pike KA, Deal M, Fuller BT, Arriaza B, Standen V, Luz MF, Ricaut F, Guidon N, Osipova L, Voevoda MI, Posukh OL, Balanovsky O, Lavryashina M, Bogunov Y, Khusnutdinova E, Gubina M, Balanovska E, Fedorova S, Litvinov S, Malyarchuk B, Derenko M, Moshier MJ, Archer D, Cybulski J, Petzelt B, Mitchell J, Worl R, Norman PJ, Parham P, Kemp BM, Kivisild T, Tyler-Smith C, Sandhu MS, Crawford M, Villems R, Smith DG, Waters MR, Goebel T, Johnson JR, Malhi RS, Jakobsson M, Meltzer DJ, Manica A, Durbin R, Bustamante CD, Song YS, Nielsen R, Willerslev E (2015) Population genetics. Genomic evidence for the Pleistocene and recent population history of native Americans. *Science* 349(6250):aab3884
46. Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165(4):2213–2233
47. Eriksson A, Mahjani B, Mehlig B (2009) Sequential Markov coalescent algorithms for population models with demographic structure. *Theor Popul Biol* 76(2):84–91
48. Marjoram P, Wall JD (2006) Fast “coalescent” simulation. *BMC Genet* 7(1):16
49. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. *Genetics* 132(4):1161–1176. <http://www.genetics.org/content/132/4/1161>
50. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery SH, Schwalia PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball EV, Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird GK, Lunter G, Meader S, Mort M, Mullikin JC, Munch K, O’Connor TD, Phillips AD, Prado-Martinez J, Rogers AS, Sajjadian S, Schmidt D, Shaw K, Simpson JT, Stenson PD, Turner DJ, Vigilant L, Vilella AJ, Whitener W, Zhu B, Cooper DN, de Jong P, Dermitzakis ET, Eichler EE, Flicek P, Goldman N, Mundy NI, Ning Z, Odom DT, Ponting CP, Quail MA, Ryder OA, Searle SM, Warren WC, Wilson RK, Schierup MH, Rogers J, Tyler-Smith C, Durbin R (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature* 483(7388):169–175
51. Munch K, Mailund T, Dutheil JY, Schierup MH (2014) A fine-scale recombination map of the human-chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion. *Genome Res* 24(3):467–474. <https://doi.org/10.1101/gr.158469.113>

52. Palacios JA, Wakeley J, Ramachandran S (2015) Bayesian nonparametric inference of population size changes from sequential genealogies. *Genetics* 201(1):281–304
53. Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, Knight JR, Mullikin JC, Meader SJ, Ponting CP, Lunter G, Higashino S, Hobolth A, Dutheil J, Karakoç E, Alkan C, Sajjadian S, Catacchio CR, Ventura M, Marques-Bonet T, Eichler EE, André C, Atencia R, Mugisha L, Junhold J, Patterson N, Siebauer M, Good JM, Fischer A, Ptak SE, Lachmann M, Symer DE, Mailund T, Schierup MH, Andrés AM, Kelso J, Pääbo S (2012) The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486(7404):527–531
54. Mailund T, Munch K, Schierup MH (2014) Lineage sorting in apes. *Annu Rev Genet* 48:519–535
55. Dutheil JY, Munch K, Nam K, Mailund T, Schierup MH (2015) Strong selective sweeps on the X chromosome in the human-chimpanzee ancestor explain its low divergence. *PLoS Genet* 11(8):e1005451
56. Achaz G (2008) Testing for neutrality in samples with sequencing errors. *Genetics* 179(3):1409–1424
57. Slatkin M, Pollack JLL (2008) Subdivision in an ancestral species creates asymmetry in gene trees. *Mol Biol Evol* 25(10):2241–2246
58. Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T (2011) Incomplete lineage sorting patterns among human, chimpanzee and orangutan suggest recent orangutan speciation and widespread natural selection. *Genome Res* 21(3):349–356
59. Stukenbrock EH, Jørgensen FG, Zala M, Hansen TT, McDonald BA, Schierup MH (2010) Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen *Mycosphaerella graminicola*. *PLoS Genet* 6(12):e1001189

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

