Chapter 12

Ancestral Population Genomics

Julien Y. Dutheil and Asger Hobolth

Abstract

The full genomes of several closely related species are now available, opening an emerging field of investigation borrowing both from population genetics and phylogenetics. 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. We discuss the basic speciation models for closely related species, including the isolation and isolation-with-migration models. 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 different approaches for understanding ancestral species from analyses of genomic data from closely related species. In particular, we emphasize core assumptions and working hypothesis. Finally, we discuss computational and statistical challenges that arise in the analysis of population genomics data sets.

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

1. Introduction

We are on the edge of the population genomics era, but the majority of population genomics data sets, such as the 1000 human genomes project (1) and the 1001 arabidopsis genomes project (2), are still in the production stage. The current data available consists of alignments of fully sequenced and closely related genomes. In some cases, the genomes are consensus genomes obtained by pooling sequences from several individuals. Under these conditions, the recent history of species is not available to the investigator (although

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods, Volume 2*,
Methods in Molecular Biology, vol. 856, DOI 10.1007/978-1-61779-585-5_12,

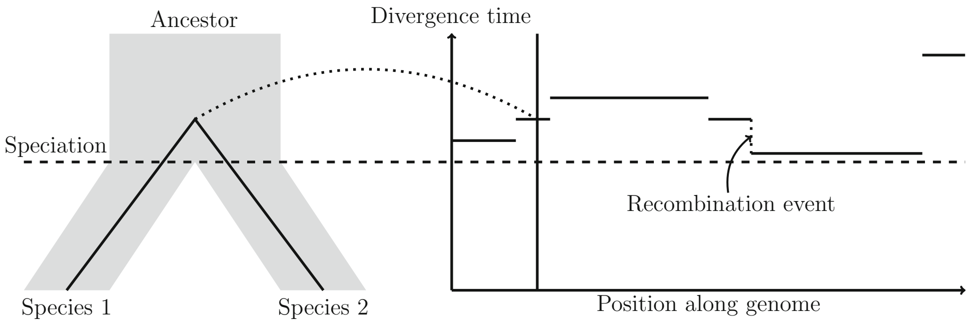


Fig. 1. *Left*: Isolation model of two species. *Right*: The coalescent process along the genomes of the two species. By comparing the two genomes, we obtain information about the split time of the species and the ancestral population size. Furthermore, the break points along the genomes correspond to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

in some cases information is available from heterozygous positions (3)). By comparing genomes from closely related species, we can, however, obtain information about split times, ancestral population sizes, ancestral recombination events, and selection in ancestral 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 the ones taken 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 is 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 30 years of specific methodological developments. The convergence of these two fields and their key mathematical 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).

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 refs. 5 and 6. 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 (IM) 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 nonoverlapping 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

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

where the approximation is valid for large N. In coalescent time units, the waiting time $T_2 = R_2 / (2N)$ before coalescent 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 MRCA is given by

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

Table 1
Methods comparison. This table summarizes and compares existing ancestral population genomics methods. Parameters correspond to the one in figure 4

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 nonrepetitive DNA segments of 2–20 kb	(28)
Count alignment patterns, fit EM model to infer parameters		I	T_1, N_{A}	Correction with outgroup	Primates	(17)
Likelihood calculation under a demographic model, numerical integration over genealogies	Independent loci	I	T_1, T_2, N_{A}	Independent estimate of rate	Primates	(25)
	Independent loci	IM	T_1, T_2, N_1, N_2, N_{A}, $m_{1\to 2}$, $m_{2\to 1}$	RAS	*Drosophila*	(14)
	Independent loci	IM	T_1, T_2, N_1, N_2, N_{A}, $m_{1\to 2}$, $m_{2\to 1}$	Independent estimate of rate	Primates: Same data as 12 restricted to human, chimpanzee, gorilla, and orangutan	(11)
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)	(12)
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	(9)
	Markov process	I	T_1, T_2, N_{A1}, N_{A2}, r RAS		Primates: 1 Mb alignment	(10)
Integrating over the discretized distribution of divergence for a pair of genomes	Markov process	I	T, N_{A}, r		Orangutans: Two full genomes	(20)

RAS Rate Across Site model, assuming an a priori distribution of evolutionary rate (usually a discretized gamma distribution) over alignment positions *I* Isolation model *IM* Isolation with migration model

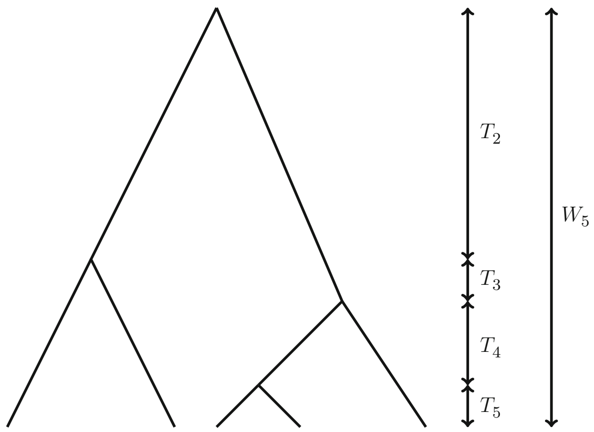


Fig. 2. Illustration of the coalescent process. The waiting time before two out of n individuals coalesce is T_n and the time before a sample of n individuals find common ancestry is W_n.

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

$$EW_n = \sum_{k=2}^{n} ET_k = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\left(\frac{1}{k-1} - \frac{1}{k}\right) = 2\left(1 - \frac{1}{n}\right).$$

Note that $\mathrm{E}W_n \uparrow 2$ for $n \to \infty$.

The variance of W_n is

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

Note that $\mathrm{Var}W_n \uparrow (8\pi^2 / 6 - 12) = 1.16$ for $n \to \infty$.

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 its MRCA is approximately $2(2N) = 4N$ 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 $N = 10{,}000$, 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 u per locus per generation. In a lineage of r generations, we then expect ru mutations or in the coalescent time units with $r = 2Nt$ we expect $2Ntu$ mutations. We let $\theta = 4Nu$ be the mutation rate parameter. Since u is small, we can make a Poisson approximation of the Binomial number of mutations in a lineage of r generations

$$\mathrm{Bin}(r, u) = \mathrm{Bin}(2Nt, \theta/(2 \times 2N)) \approx \mathrm{Pois}(t\theta/2).$$

We have, thus, arrived at the following two-step process for generating samples under the coalescent: (a) generate 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). A single genome, thus, contains different samples

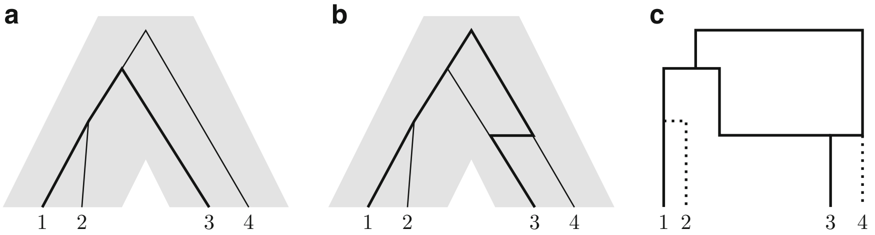
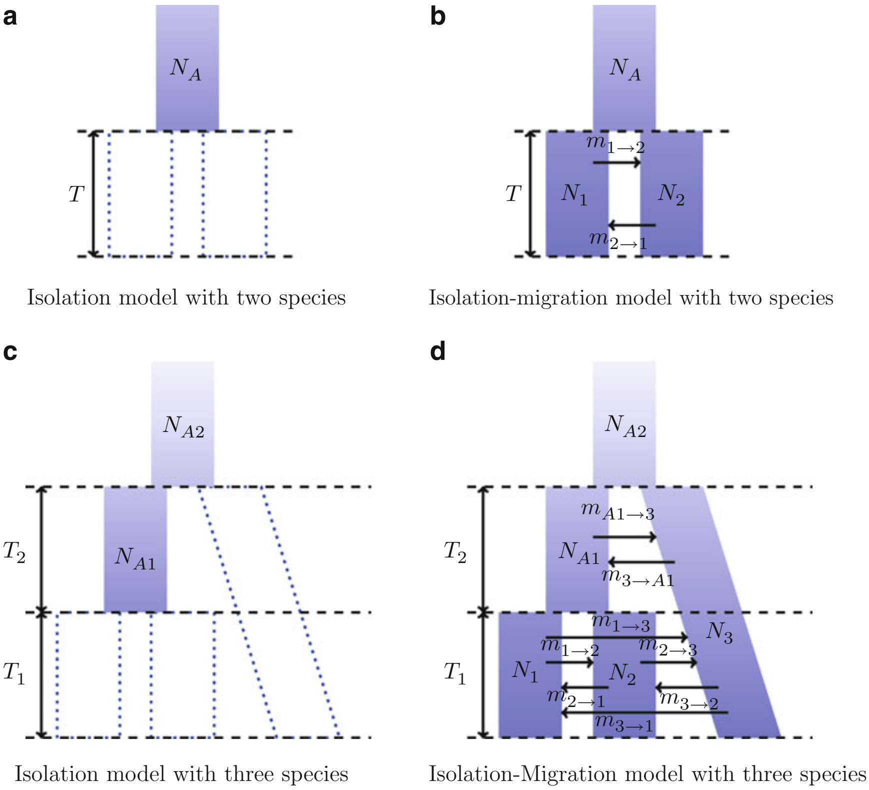


Fig. 3. Ancestral recombination graph for two species, (**a**) genealogy of four sampled sequences from two species. The *bold line* shows the divergence of two sequences of interest, (**b**) a single recombination event happened between the lineages of sequences 3 and 4 (*horizontal line*) so that in 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. *Dotted lines* show the portions of lineages which are not present in the sample composed of sequences 1 and 3. When going backward in time, a split corresponds to a recombination event and a merger is a coalescence event.

from the distribution of the age of the MRCA, and the distribution contains information about the ancestral population size and speciation time.

3. Models of Speciation

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 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



Isolation model with two species

Isolation-migration model with two species

Isolation model with three species

Isolation-Migration model with three species

Fig. 4. Speciation models and associated parameters. In all exemplified models, effective population size is constant between speciation event, 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 postdivergence 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.

model. We also discuss the general multiple-species isolation-with-migration model. The two-species isolation model is introduced in ref. 7 and the isolation-with-migration model is introduced in ref. 8.

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

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

The mean time to coalescent is $\mathrm{E}[T_2] = T + \theta_{\mathrm{A}}/2$ and the average divergence time between two sequences is twice this quantity, that is, $2T + \theta_{\mathrm{A}}$. Since $\theta_{\mathrm{A}} = 4N_{\mathrm{A}}u$, 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 $\mathrm{Var}[T_2] = \theta^2/4$. 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 long regions 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., refs. 9–11).

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

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

where

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

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

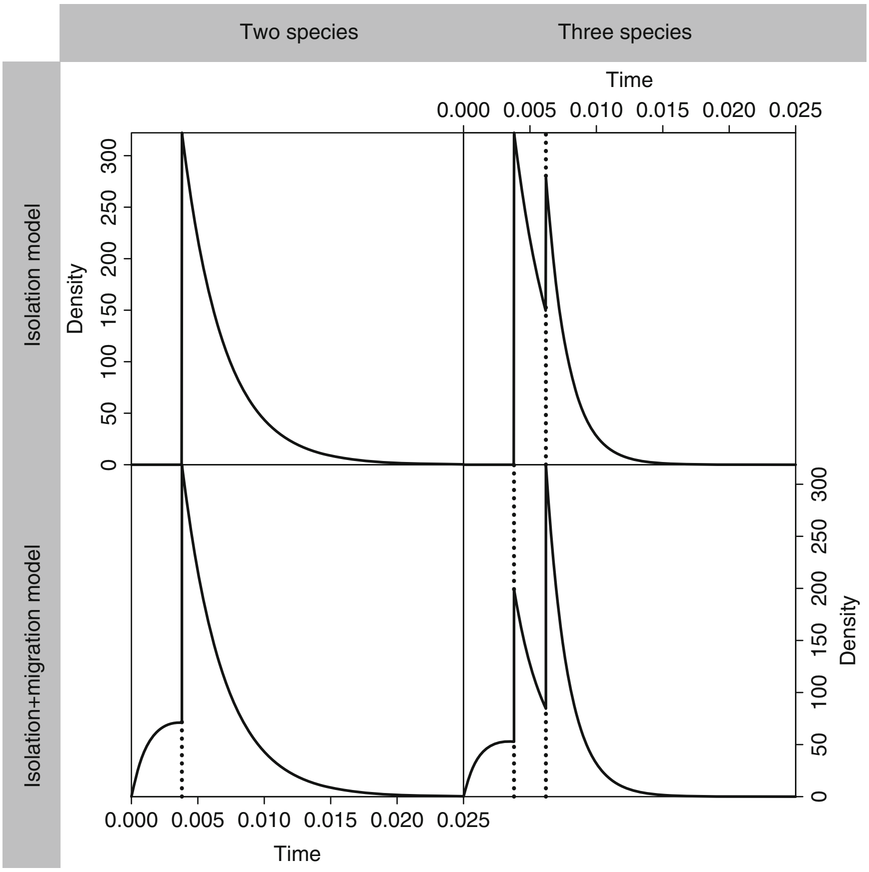


Fig. 5. Illustration of the density for coalescent in various models and data layout. The *curves* are the probability density functions. In the most simple case with two species, a constant ancestral population size, and a punctual speciation (*top left panel*), more genomic regions find a common ancestor close 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 ref. 14: $\theta_1 = 0.005$ and $\theta_2 = 0.003$. Migration parameters are all set to 50.

((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,

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

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)

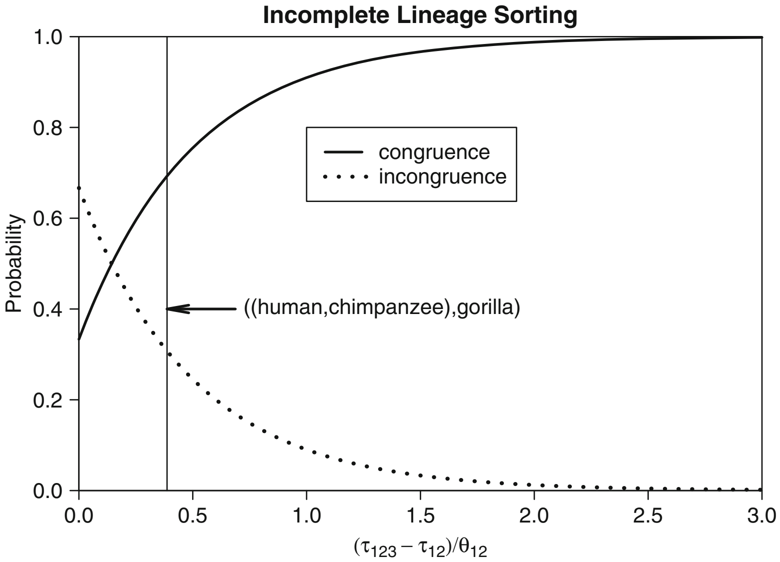


Fig. 6. Probability (Eq. 2) of gene tree and species tree being incongruent. In case of the HCG triplet, we obtain $(T_{12} - T_1)/\theta_{A1} = (0.0062 - 0.0038)/0.0062 = 0.39$ which corresponds to an incongruence probability of 30%.

congruence probability (Eq. 2). We also refer to Subheadings 7.1 and 7.2 for more discussion of ILS.

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

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

Burgess and Yang (12) describe the speciation process for humans (H), chimpanzees (C), gorillas (G), orangutans (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

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

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

The isolation-with-migration 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 (14) 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

$$\begin{array}{c||ccc|cc} & S_{11} & S_{12} & S_{22} & S_1 & S_2 \\ \hline\hline S_{11} & \cdot & 2m_2 & 0 & 2/\theta_1 & 0 \\ S_{12} & m_1 & \cdot & m_2 & 0 & 0 \\ S_{22} & 0 & 2m_1 & \cdot & 0 & 2/\theta_2 \\ \hline S_1 & & & & \cdot & m_2 \\ S_2 & & & & m_1 & \cdot \end{array}.$$

Starting in state a, the density for coalescent in population 1 at time $t < T$ is given by (13)

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

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

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

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

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

Here, $e^A = \sum_{i=0}^{\infty} A^i i!$ 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 $2/\theta_{\mathrm{A}}$. The density for coalescent in the ancestral population at time $t > T$ is, therefore,

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

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

$$P(X|\Theta) = L(\Theta|X) = \int_0^{\infty} P(X|t) f(t|\Theta) \mathrm{d}t, \tag{8}$$

where $f(t) = f(t|\Theta)$ given by Eqs. 6 and 7 is the density of the two sequences finding an MRCA at time t and $P(X|t)$ is the probability of the two sequences given that they find an 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 ref. 14; a discussion of their approach can be found in ref. 15.

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

Hey (16) considers the multipopulation isolation-with-migration 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 15 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
- $2(k - 1)^2$ migration rates

See Subheading 7.3 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, respectively. Because the number of parameters becomes very large even for small k, Hey (16) 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 Ancestral Recombination Graph

In this section, we discuss the three methods of taking recombination into account. The three methods are visualized in Fig. 7c–e and correspond to (1) independent loci, (2) site patterns, and (3) hidden Markov model (HMM).

4.1. The Independent Loci Approach: All 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 that several recombination events

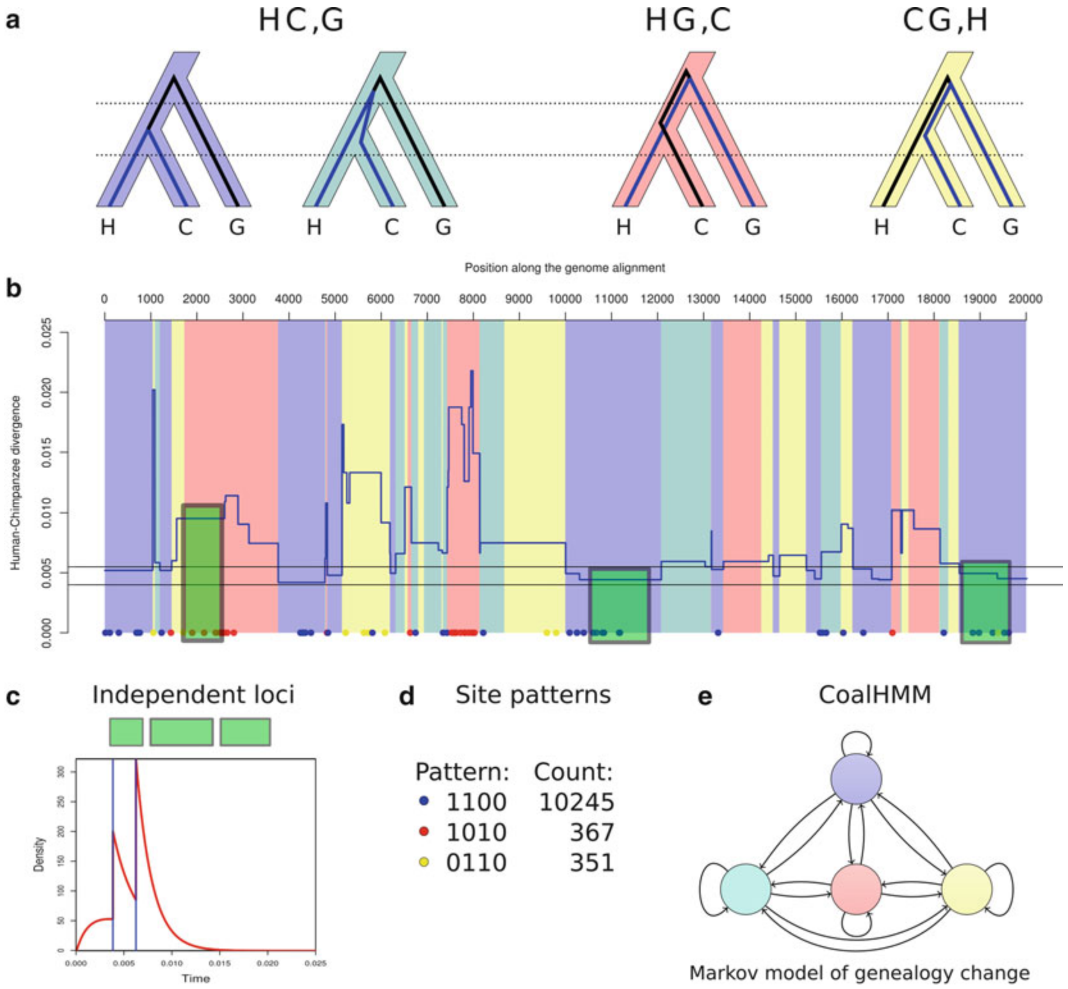


Fig. 7. The coalescent process along genomes, (**a**) four archetypes of coalescence scenarios with three species, exemplified with human, chimpanzee, and gorilla. In the first scenario, human and chimpanzee coalesce within the human–chimpanzee common ancestor. In the three other scenarios, all sequences 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. Every change in color or break of the *blue line* is the result of a recombination event. (**c**–**e**) Three possible ways of approximating the ancestral recombination graph. In (**c**), a number of small loci are analyzed independently under an assumption of no recombination within loci, which allows to estimate the probability distribution of sequence divergence. In (**d**), the alignment is summarized in terms of counts of site patterns, and in (**e**) the data is analyzed in terms of a hidden Markov model along the sequence, with distinct genealogies featuring various divergence times as hidden states. The underlying model includes transition probabilities between genealogies along the genome. See Subheading 4 for more details.

happened between loci). Each region can, therefore, be described by a single underlying tree, reducing the analytical and computational load. This approach cannot be used when the species under study are too distantly related, as recombination events will have fragmented the ARG up to a point, where no single region size without recombination can be defined.

Using 15,000 loci distant from 10 kb totaling 7.4 Mb and isolation model introduced above, (Table 2, "Model (b) Sequencing errors" in (12)) find $\theta_{\text{HC}} = 0.0062$, $\theta_{\text{HCG}} = 0.0033$, $\theta_{\text{HCGO}} = 0.0061$, and $\theta_{\text{HCGOM}} = 0.0118$ and $T_{\text{HC}} = 0.0038$, $T_{\text{HCG}} = 0.0062$, $T_{\text{HCGO}} = 0.0137$, and $T_{\text{HCGOM}} = 0.0260$. They get $\mathrm{E}T_{\text{HCG}} = 0.0062$ (corresponding to a 1.2% divergence between human and chimpanzee) and $T_{\text{HC}} = 0.0038$. Therefore, $38/62 = 0.61 = 61\%$ of the divergence between humans and chimpanzees is due to speciation and 39% is due to ancestral polymorphism. Converting those estimated in time units requires an estimate of the substitution rate, either absolute or deduced from a scaling point. Using $u = 10^{-9}$ 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 (11) showed that the isolation with migration model was preferred. Yang finds a more ancient speciation time $T_{\text{HC}} = 0.0053$ (5.3 My with $u = 1e - 9$) when migration is accounted for (was $T_{\text{HC}} = 0.0044$ without migration).

4.2. Site Pattern Analysis

Patterson et al. (17) used a different approach based on site patterns. They sequenced fragments of DNA from a western lowland gorilla and a spider monkey, which they combined with whole-genome reads from the orangutan and macaque, and built a genome alignment using the human scaffold. The resulting 20-Mb data set was extended and/or used thereafter by refs. 9–12. Patterson et al. counted the frequencies of all possible site patterns in the resulting HCGOM alignment. These patterns can be sorted depending on which genealogy they support: ((H,C),G),O, ((H,G),C),O, ((C,G), H),O, etc. They introduced a model that allowed them to estimate speciation time and ancestral population sizes from the frequencies of the observed patterns, independently of the recombination rate. The only requirement is that recombination occurred to enable the various patterns to be observed, which is warranted by the large genomic region they used. This method makes very little assumption on the data, particularly regarding recombination, and uses ILS as its only source of signal for estimating population parameters. However, it ignores alternative sources of signal, like singletons, which carry information about the local sequence divergence. Such an approach is, therefore, limited to simple models of speciation, and cannot easily be extended to more complex scenarios like isolation with migration.

Patterson et al. inferred a recent speciation time for human and chimpanzee, below 5.4 My. They also found a most recent divergence on the X chromosome, which they interpret in terms of complex speciation event with hybridization. Alternative explanations for this observation were provided (18, 19).

4.3. The Markov Assumption Along Sites

The work by Hobolth et al. (9) used site patterns in a different way. With a hidden Markov model, they used the correlation of patterns along the genome to reconstruct the site-specific genealogy, including divergence times. They further used these divergence estimates together with the inferred amount of incomplete lineage sorting to compute the speciation times and ancestral population sizes. In this approach, the recombination rate is embedded into the transition matrix of the hidden Markov chain, which specifies the probabilities of transition from one genealogy to the other along the genome. Hobolth et al. showed that this matrix is constrained by symmetric relationships, and estimated the remaining three parameters together with the divergence parameters. Dutheil et al. (10) extended this approach by identifying further constraints on the parameters and fully expressing the divergence times and probabilities of transition between genealogies as function of the speciation times, ancestral population sizes, and recombination rate, therefore allowing their direct estimation. The analytical expressions of the parameters as function of populational quantities are, therefore, difficult to obtain, notably for the transition probabilities, even in the simplest case.

Mailund et al. (20) used a different approach to compute these for the two-species isolation model. They used a continuous Markov chain to model the evolution of a pair of contiguous positions. This model features two types of events: when going backward in time, the two positions can either coalesce (with a rate proportional to the effective population size) or split (with a rate equal to the recombination rate). The transition probabilities between genealogies are immediately available from the joint pair of contiguous positions and the Markov assumption. This approach can be generalized to more species are and potentially allows for more realistic demographic scenarios, for instance allowing migration between populations.

The coalescent HMM framework, thus, models recombination, which is assumed to be constant in all lineages and along the alignment. The model further assumes that the probability of switching from one genealogy to another when we walk along a genome alignment only depends on the genealogy at the previous position, that is, the process of genealogy change along the genome is Markovian. This is an approximation of the true coalescent process that greatly simplifies calculation (21). Dutheil et al. (10) and Mailund et al. (20) used simulated data sets under a coalescent process with recombination to show that this assumption had, however, little influence on the parameter estimates. 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. (10) 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.

5. Specific Issues Faced When Dealing with Genomic Data

In previous sections, we discussed population genetic models for between-species comparisons and methods for parameter estimation. We now describe several pitfalls encountered when analyzing whole-genome data sets, including sequencing errors and alignment errors, but also computational and statistical issues related to the data sets of large dimension that are underlying genomics analyses.

5.1. Sequencing Errors and Rate Variation

Sequencing errors are a well-described source of bias in population genetics analyses, resulting in an excess of singletons (22). When full genome sequences are used, the issue becomes more complex as the error rate differs between and within sequences not only 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 (23, 24). 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 (12). 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 rate of substitution also varies along the genome which potentially affects the reconstruction of sequence genealogy, a phenomenon well known by phylogeneticists. There, things are a bit easier, as the tools developed for phylogenetic analysis can in most cases be applied with a reasonable cost. This generally consists in assuming a prior distribution of the site-specific rate, and integrate the likelihood over all possible rates (10, 12, 14). Alternatively, one can also use one or more outgroup sequences to calibrate the rate, as in refs. 17, 25.

5.2. Aligning Genomes

To sequence errors, one should add assembly errors due to the sequencing technology. Assembling reads can be error prone in case of repeated or duplicated regions, which ultimately can lead to compare nonorthologous regions. In addition to this technical issue, 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 syntheny block. Building the genome alignment, that is, recovering the syntheny structure, is, therefore, performed with potential issues that are close in effect to the assembly errors. Finally, as all comparative methods rely on an input alignment, any artifact affecting the alignment process itself is relevant. As populational methods are based on closely related species, alignment programs are, however, expected to perform accurately, and alignment errors should be negligible compared to other sources. So far, the only way to deal with such errors is to restrict the analysis on regions, where orthology can be unambiguous resolved, mostly by removing short syntheny blocks and regions that contain a high proportion of repeated elements, gaps, and duplications.

5.3. Computational Load

Dealing with genomic data heavily relies on computer performance. Depending on the genome sizes and the method used, the analysis may cover from millions to billions of genomic positions. As most methods rely on maximum likelihood or Bayesian inference, efficient algorithmics and software implementation are much needed. Fortunately, the data structure here comes handy: independent parts of the genomes, like chromosomes, syntheny blocks, or even loci, depending on the methodology used, can be analyzed separately, therefore enabling easy parallelization for use of computer grids. Aside to the computational issue, the genomic area also dramatically changed the structure of the result tables. While analyzing per-gene result sets, consisting of a few dozen thousand rows, is still feasible with statistical software like R, it becomes much more problematic when per-site result sets are considered. As our understanding of genome evolution grows, we are more keen on fishing specific regions with a peculiar demographic or selective history. Such data sets typically reach sizes of several millions rows. While they can still be loaded into the memory of computers with strong configuration, a single pass on the table for retrieving information becomes prohibitive, which becomes problematic when several sets are to be compared (for instance, in order to compare a window-based calculation with gene annotations). The only alternative currently available is to use database engines, with proper indexing algorithms. Such databases are currently used in genome browsers, like the UCSC genome browser. In that respect, cross-information storage and retrieval, as well as Web-based services, will become even more crucial for genome data analysis.

5.4. Statistical Challenges

The genetics to genomics shift also leads to new challenges in data analysis. When tests are performed, for instance when comparing models of speciation like in ref. 11, the global false discovery rate has to be properly controlled for. As genomes are not analyzed in one single analysis (at least full chromosomes are analyzed independently, but in most cases chromosomes are also split into several parts), multiple testing issues occur. Multiple testing also matters when candidate regions are scanned for, for instance for specific selection regime. Verhoven et al. (26) offer a nice tutorial presenting appropriate statistical methods for handling multiple testing. A related matter, when performing several types of tests on a wide set of genomics regions, is the so-called overoptimism issue, also named "data optimization" (27). This concerns the selection of data sets in order to increase the significance of results, resulting in a potential bias. In genomics, the data set selection often takes the form of an extensive filtering of the data in order to exclude regions with potential paralogous sequences, low complexity, or known functional role. It, therefore, appears important to emphasize to which peculiar region of the genome the obtained conclusions apply to, and eventually report how they change when other regions are included (see, for instance, ref. 12).

6. 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 one individual for a few species. The change involve 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 dealt with, 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, as well as 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 (11) suggests that the speciation of human and chimp 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 split 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.

The population genomics era is just ahead, where we will have full individual genomes for closely related species. Such data sets are the key to understand the detailed evolutionary processes that are linked to the formation and evolution of species, as they will open windows to new periods in time. 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.

7. Exercises

7.1. 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 among 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 among human, chimpanzee, and orangutan? We assume a generation time of 20 years for all extent and ancestral primates.

7.2. Estimating Ancestral Population Size from the Observed Amount of ILS

Given that 30% of incomplete lineage sorting is observed among human, chimpanzee, and gorilla and assuming a generation time of 20 years and 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 (12), a researcher finds a higher estimate of N_e than expected. What could explain this discrepancy?

7.3. 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 $2(k - 1)^2$.

Acknowledgments

The authors would like to thank Thomas Mailund for providing useful comments on this chapter. This publication is contribution 2011-035 of the Institut des Sciences de l'Évolution de Montpellier (UMR 5554—CNRS). This work was supported by the French Agence Nationale de la Recherche "Domaines Emergents" (ANR-08-EMER-011 "PhylAriane").

References

1. Siva, N. (2008), 1000 genomes project. Nature Biotechnology **26**(3), 256
2. Weigel, D., Mott, R. (2009), The 1001 genomes project for arabidopsis thaliana. Genome Biology **10**(5), 107+
3. Enard, D., Depaulis, F., Roest Crollius, H. (2010), Human and non-human primate genomes share hotspots of positive selection. PLoS Genet **6**(2), e1000,840+
4. Siepel, A. (2009), Phylogenomics of primates and their ancestral populations. Genome Research **19**(11), 1929–1941
5. Wakeley, J. (2008). Coalescent Theory: An Introduction, 1 edn. Roberts & Company Publishers
6. Tavaré, S. (2004). Ancestral inference in population genetics, vol. 1837, pp. 1–188. Springer Verlag, New York
7. Takahata, N., Nei, M. (1985), Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110**(2), 325–344
8. Nielsen, R., Wakeley, J. (2001), Distinguishing migration from isolation: a markov chain monte carlo approach. Genetics **158**(2), 885–896
9. Hobolth, A., Christensen, O.F., Mailund, T., Schierup, M.H. (2007), Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. PLoS Genet **3**(2), e7+
10. Dutheil, J.Y., Ganapathy, G., Hobolth, A., Mailund, T., Uyenoyama, M.K., Schierup, M. H. (2009), Ancestral population genomics: The coalescent hidden markov model approach. Genetics **183**(1), 259–274
11. Yang, Z. (2010), A likelihood ratio test of speciation with gene flow using genomic sequence data. Genome Biol Evol **2**(0), 200–211
12. Burgess, R,., Yang, Z. (2008), Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. Molecular biology and evolution **25**(9), 1979–1994
13. Tavaré, S. (1979), A note on finite homogeneous continuous-time markov chains. Biometrics **35**, 831–834
14. Wang, Y., Hey, J. (2010), Estimating Divergence Parameters With Small Samples From a Large Number of Loci. Genetics **184**(2), 363–379
15. Hobolth, A., Andersen, L.N., Mailund, T. (2011), On computing the coalescence time density in an isolation-with-migration model with few samples. Genetics **187**(4), 1241–3
16. Hey, J. (2010), Isolation with Migration Models for More Than Two Populations. Mol Biol Evol **27**(4), 905–920
17. Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S., Reich, D. (2006), Genetic evidence for complex speciation of humans and chimpanzees. Nature **441**(7097), 1103–1108
18. Barton, N.H. (2006), Evolutionary biology: how did the human species form? Curr Biol **16**(16)
19. Wakeley, J. (2008), Complex speciation of humans and chimpanzees. Nature **452**(7184), E3–4; discussion E4
20. Mailund, T., Dutheil, J.Y., Hobolth, A., Lunter, G., Schierup, M.H. (2011), Estimating speciation time and ancestral effective population size of bornean and sumatran orangutan subspecies using a coalescent hidden markov model. PLoS Genetics **7**(3), e1001,319
21. Marjoram, P., Wall, J.D. (2006), Fast "coalescent" simulation. BMC Genet **7**(1)
22. Achaz, G. (2008), Testing for neutrality in samples with sequencing errors. Genetics **179** (3), 1409–1424
23. Slatkin, M., Pollack, J.L.L. (2008), Subdivision in an ancestral species creates asymmetry in gene trees. Mol biol Evol **25**(10), 2241–2246
24. Hobolth, A., Dutheil, J.Y., Hawks, J., Schierup, M.H., Mailund, T. (2011), Incomplete lineage sorting patterns among human, chimpanzee and orangutan suggest recent orangutan speciation and widespread natural selection. Genome Research **21**(3), 349–56
25. Yang, Z. (2002), Likelihood and bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics **162**(4), 1811–1823
26. Verhoeven, K.J., Simonsen, K.L., McIntyre, L. M. (2005), Implementing false discovery rate control: increasing your power. Oikos **108**(3), 643–647
27. Boulesteix, A.L. (2010), Over-optimism in bioinformatics research. Bioinformatics **26**(3), 437–439
28. Chen, F.C., Li, W.H. (2001), Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. American journal of human genetics **68**(2), 444–456