Lineage Sorting in Apes

Thomas Mailund,[1] Kasper Munch,[1] and Mikkel Heide Schierup[1,2]

[1]Bioinformatics Research Centre, Aarhus University, DK-8000 Aarhus C, Denmark; email: mailund@birc.au.dk, kaspermunch@birc.au.dk, mheide@birc.au.dk

[2]Department of Bioscience, Aarhus University, DK-8000 Aarhus C, Denmark

Annu. Rev. Genet. 2014. 48:519–35. Downloaded from www.annualreviews.org

First published online as a Review in Advance on September 19, 2014

The *Annual Review of Genetics* is online at genet.annualreviews.org

This article's doi: 10.1146/annurev-genet-120213-092532

Keywords

population genomics, great apes evolution, speciation, incomplete lineage sorting, phylogenetics

Abstract

Recombination allows different parts of the genome to have different genealogical histories. When a species splits in two, allelic lineages sort into the two descendant species, and this lineage sorting varies along the genome. If speciation events are close in time, the lineage sorting process may be incomplete at the second speciation event and lead to gene genealogies that do not match the species phylogeny. We review different recent approaches to model lineage sorting along the genome and show how it is possible to learn about population sizes, natural selection, and recombination rates in ancestral species from application of these models to genome alignments of great ape species.

SPECIES DIVERGENCE AND GENOME DIVERGENCE

DNA sequencing has had a dramatic effect on the field of molecular phylogenetics. The rate of DNA substitutions often adheres quite closely to a molecular clock, i.e., differences accumulate linearly since the point at which a pair of sequences had a common ancestor. It is therefore tempting to sample pieces or entire genomes from a set of species of interest, build phylogenies, enforce the molecular clock, and assign the dates of speciation events from the resulting tree. The literature is abundant with examples of such practice. For examples of application to primates, see References 16, 17, 45, and 65. Divergence times, however, do not represent speciation times. Assuming for now that speciation occurs as a simple split of one species into two, with no subsequent gene flow, then the most recent common ancestor of sequences from two species is always further back in time than the speciation event. Tracing the ancestry of a human and a chimpanzee sequence back in time to the exact time of speciation, their ancestral sequences will be found in separate individuals in the ancestral population. We would thus need to look further back in time to find their common ancestor.

The difference between species divergence and genome divergence depends on the population size and generation time in the ancestor, as detailed below. In primates, the relative difference between genome divergence time and speciation time is appreciable, as shown in **Table 1**. Generally, more ancient speciation times result in a smaller relative difference between the species split time and the time to most recent common ancestry inferred from genome sequences, as the speciation time takes up more of the total time. However, the divergence time of human and orangutan genome sequences may still exceed the speciation time by 50% because of a large ancestral population size. Using fossil calibration points to date speciation times and then equate speciation and divergence times can thus lead to grossly inaccurate dates.

The divergence time of two sequences from separate species varies along the genome because each sequence consists of many segments with different ancestry separated by recombination. **Table 1** shows that human and chimpanzee genome divergence is on average further back in time than the species divergence of human and gorilla, implying that the common ancestor of many such segments of human and chimpanzees genomes are found in the ancestor to human, chimpanzee, and gorilla. In these cases, a human segment may find a most recent common ancestor

Table 1 The difference between speciation time (population divergence time) and average genomic divergence time for pairs of great ape species[a]

Species 1	Species 2	Generation time	Ancestral population size	Speciation time	Divergence time	Reference
Human	Neanderthal	29	15,000	0.55	1.42	(49)
Human	Chimpanzee	25	86,000	5.6	9.9	(46) ILS model
Human	Gorilla	25	67,000	8.6	12.0	(46) ILS model
Human	Orangutan	25	167,000	18.5	26.9	(46) CTMC model
Human	Macaque	25	196,667	44.0	53.8	(8)
Chimpanzee	Bonobo	25	32,000	1.2	2.8	(46) ILS model
Bornean orangutan	Sumatran orangutan	23	63,000	1	3.9	(46) CTMC model
West gorilla	East gorilla	20	64,000	0.4	2.9	(46) CTMC model

[a]All values assume a mutation rate of 0.6×10^{-9} per base pair per year. The ancestral population size estimate for the human-Neanderthal ancestor is visually estimated from Reference 49.

Abbreviations: CTMC, continuous-time Markov chain; ILS, incomplete lineage sorting.

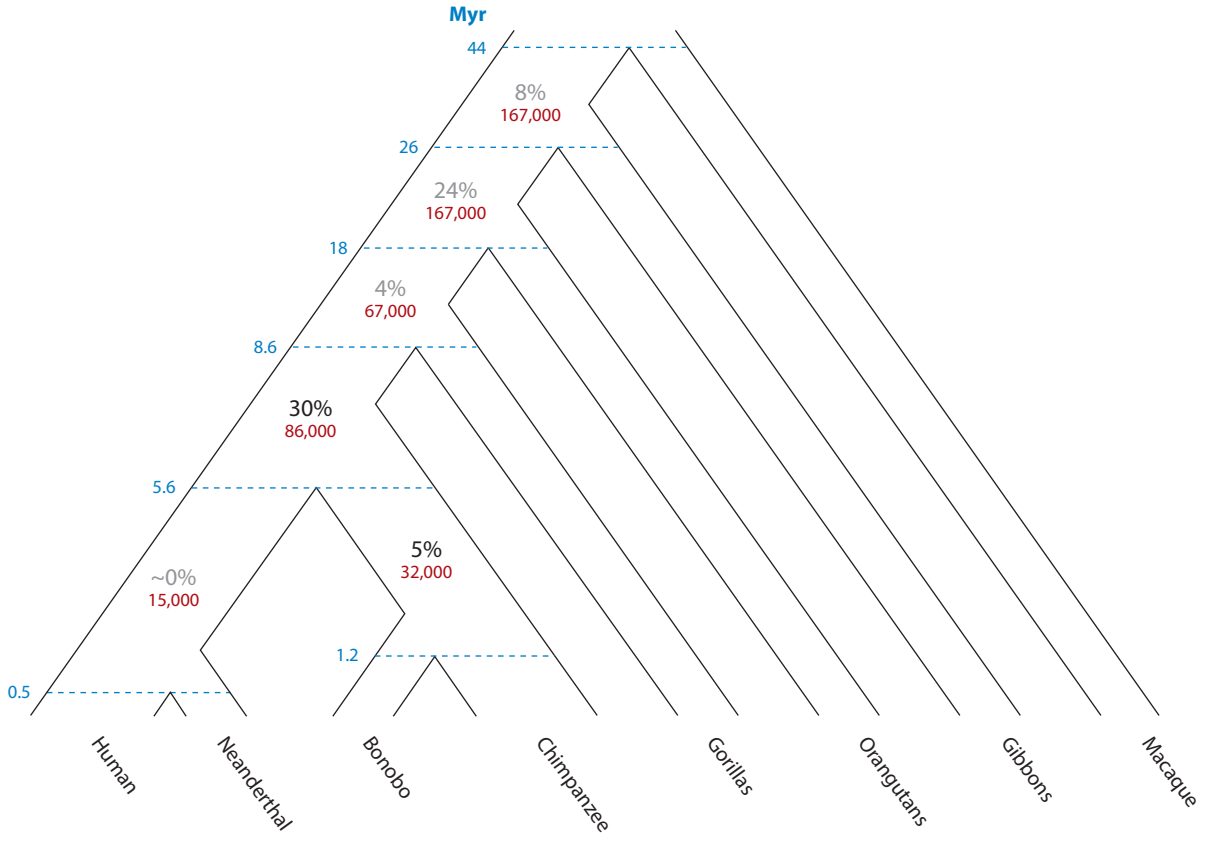

Figure 1

Species tree with speciation times (*blue numbers*) and ancestral population sizes (*red numbers*) as in **Table 1**. Human-gibbon speciation time is derived from an estimated population size shared with the human-orangutan ancestor and the human-gibbon divergence (35a). Black numbers are empirical estimates of incomplete lineage sorting from CoalHMM (coalescent hidden Markov model) analyses of species alignments, and gray numbers represent theoretical expectations from the estimated species divergence times and ancestral population sizes. These numbers are associated with uncertainty. All values assume a substitution rate of 0.6×10^{-9} per base pair per year.

with a gorilla sequence and thus display a genealogy different from the species phylogeny. This phenomenon is termed incomplete lineage sorting (ILS). In the ancestral lineage to human, ILS is expected to affect all the internal branches to various extents, as shown in **Figure 1**. The extent of ILS warrants caution in relying on small DNA fragments, such as single genes, for phylogenetic analysis. However, ILS also provides valuable information about speciation processes and ancestral population processes, including natural selection, population size changes, and recombination patterns from an alignment of genomes from different apes. Before we show how this can be done we introduce the relevant population genetics theory.

GENE TREES WITHIN SPECIES TREES: THE MULTISPECIES COALESCENT

Coalescence theory (18, 70) describes the ancestry of a sample of genes and assigns probabilities to the timing and order in which sampled sequences find common ancestors. This framework

has been successful because it keeps track of the ancestry of only a sample and not the whole population backward in time. In the following, we consider time in this backward fashion, starting at present and reaching into the past. In a sample of n genes in a randomly mating population of constant size N, we let T_n denote the number of generations into the past when the first pair of genes shares a common ancestor. This time is well approximated by the exponential distribution with density

$$f(T_n = t) = \frac{\binom{n}{2}}{2N} e^{-t\binom{n}{2}/2N}$$

Any of the $\binom{n}{2}$ possible pairs in the sample are equally likely to be the first (most recent) to coalesce into one lineage. After the first pair coalesces, the waiting time until the next pair coalesces is given by the same formula for a sample of $n-1$ genes. In many populations, the assumptions of a random mating is violated by population structure or mating patterns. Nevertheless, the framework usually provides a very good description of the genealogical process if the census population size is substituted by the effective population size (N_e), which represents the size of a random mating population giving rise to the pattern of ancestry observed in the population. For humans this effective population size is usually estimated to be approximately 10,000, despite a census population size in the billions. This is because the human population only very recently rose to large numbers and the effective size represents the human population size over the complete genealogical history (~0.5 million years), which includes periods with a small number of breeding individuals (population bottlenecks).

Coalescence theory allows us to compute how many genes in the sample of size n have coalesced at some time τ or, equivalently, how many lineages represent the original sample at that time. **Figure 2** shows the probability that k lineages remain at time τ when the original sample was of size n. The rate at which coalescences occur when k lineages are left is of the order k^2. This means that while k is large, coalescence happens very quickly, and therefore the number of lineages is rapidly reduced to a few lineages. In fact, for any sample size, the amount of time during which only two lineages remain is expected to take up more than half of the time to the most recent common ancestor of the entire sample. The probability that only a single lineage is left as a function of time and initial sample size is shown in **Figure 3**, showing that initial sample size primarily affects this probability for recent times, whereas the probability that all lineages have coalesced is largely independent of n as time increases.

By analyzing sequences sampled from multiple species, we find that lineage sorting is determined by the number of lineages that have not coalesced at the time of a particular speciation event. If a species is represented by only a single lineage at the time of a speciation event then lineage sorting is complete. However, when more than one lineage remains at the time of speciation, ILS may result, producing genetic relationships that do not reflect the species relationship. Consider samples of n_1 and n_2 genes from two different species that split apart at time τ in the past and let these be coalesced into k_1 and k_2 lineages at the time of speciation. If $k_1 = 2$ and $k_2 = 1$ this corresponds to a sample of three lineages from the ancestral species. The ancestry of these lineages is determined by the coalescence process in the ancestral population and may result in three different gene-tree topologies, only one of which corresponds to the species relationship. In the other two cases, sequences from different species are most closely related.

At the modern human-Neanderthal split [~0.5 Myr ago (49)] or the bonobo-chimpanzee split [1–1.5 Myr ago (20, 34, 48, 73, 76)], we do not expect to observe complete lineage sorting but rather that some polymorphism is shared between the different species (**Figure 3**). At the human-chimpanzee split, however, where even the most recent estimates put the event more than four million years in the past (22), we would not expect any shared polymorphism.

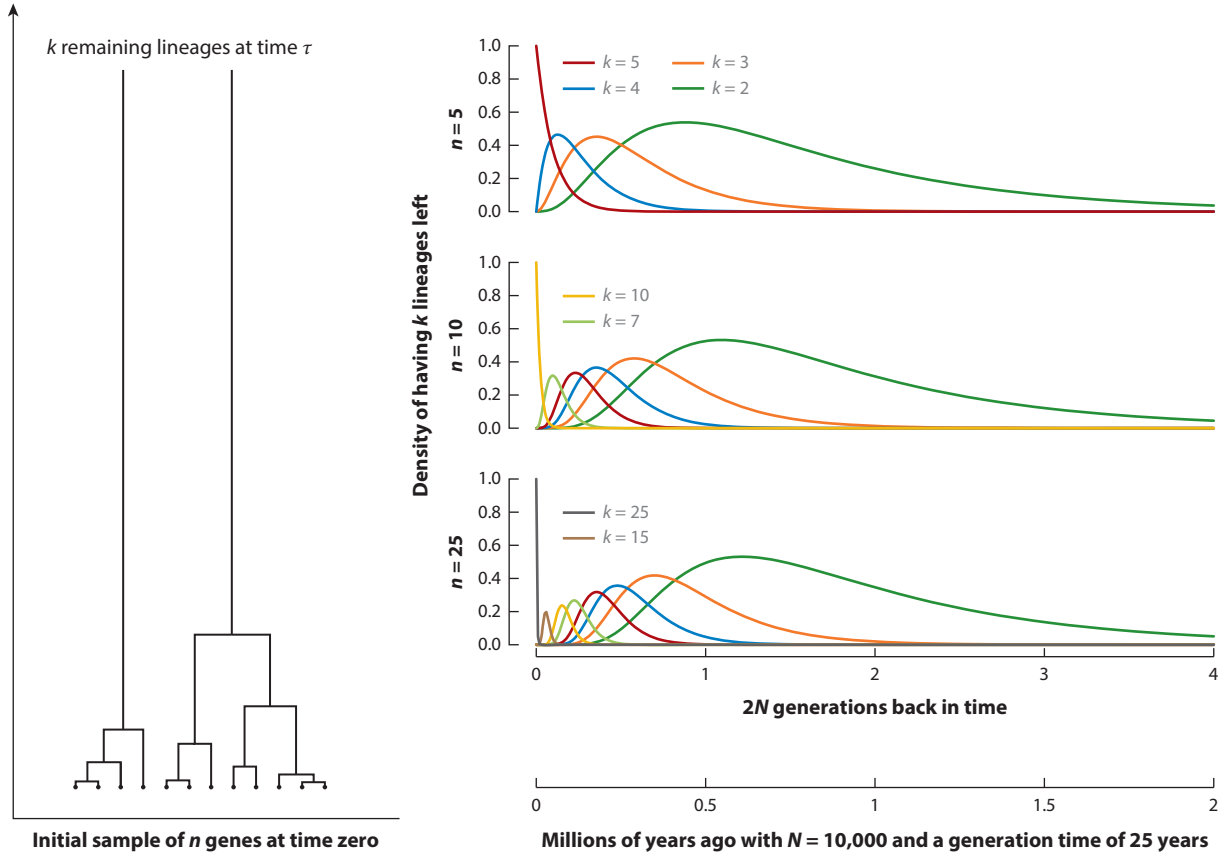

Figure 2

The probability of having exactly k lineages remaining at time τ for a coalescence process that starts with a sample of size n at time 0. Time is shown in two units, so-called coalescence units corresponding to $2N_e$ generations and the corresponding number of years if the effective population size is 10,000 and the generation time is 25 years, with the values roughly corresponding to modern humans.

If a speciation event follows so shortly after another that lineage sorting does not have time to complete, ILS between the three species becomes fixed for some segments of the genome. Gorilla split from the ancestor of human and chimpanzee approximately two to three millions years before human and chimpanzee separated. When the human and chimpanzee ancestors split into distinct non-interbreeding species, some genes in the emerging populations of one species were more closely related to sequences from the other species. Most of these have been lost by genetic drift, but some are fixed in humans, making them more closely related to gorilla than to chimpanzee. So even when ancestral polymorphism is no longer shared, ILS is a permanent part of our genome.

A sample of one genome from each of three species allows us to compute the probability that a gene tree is incongruent with the species tree. The probability that a lineage from species A and one from species B have not coalesced at time t_2 is $e^{-(t_2-t_1)/2N}$, where t_1 is the number of generations since speciation of species A and B, and t_2 is the number of generations between C and the ancestral species AB. In this event, lineages from all three species coexist in one species and coalesce with equal probability, which results in three different gene-tree topologies, two of

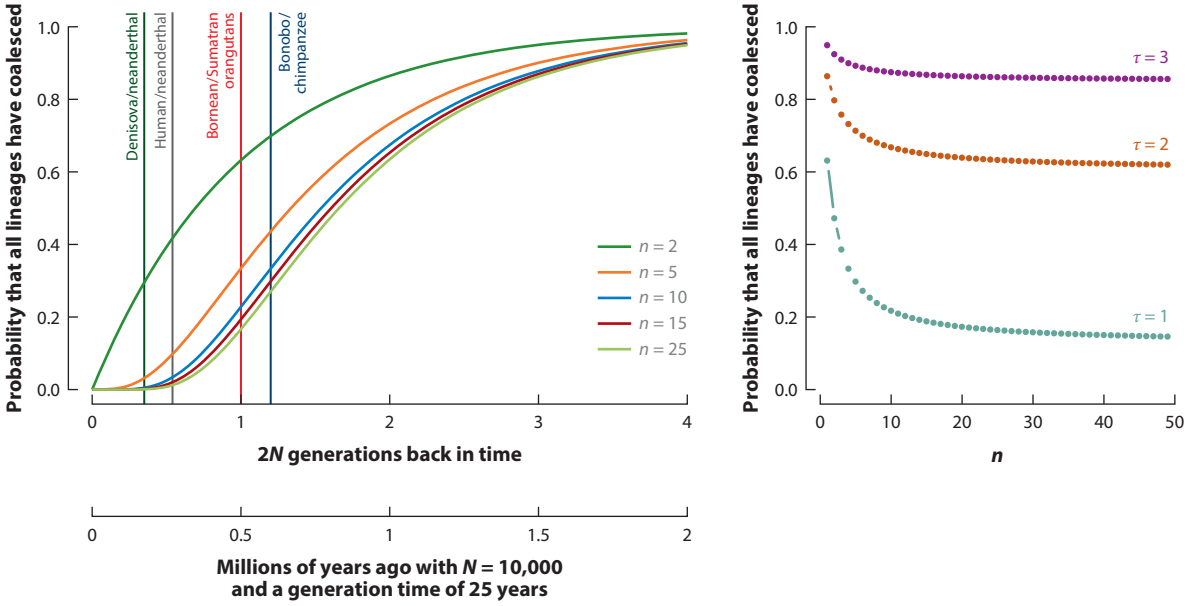

Figure 3

The probability of having completed lineage sorting for five different sample sizes assuming an effective population size and generation time close to modern humans, i.e., $N_e = 10{,}000$ and a generation time of 25 years. The figure shows the probability of having exactly one lineage back at time τ when starting with a sample of size n. On the left the x axis is time and the different lines are different initial sample sizes, whereas on the right the x axis depicts a range of initial sample sizes and the different lines show different time points. On the left are also shown recent speciation times taken from **Table 1**.

which represent ILS. The proportion of genomic sequences that are incongruent with the species tree is simply the above formula multiplied by two-thirds (**Figure 4**).

INCOMPLETE LINEAGE SORTING AND PHYLOGENETIC INFERENCE

ILS renders phylogenetic inference more complex because local gene trees in a genome are likely to produce conflicting results. A phylogenetic tree based on a single or a few genes may therefore not reflect the true species tree. This may be true for all internal nodes in a phylogeny (**Figure 1**) and should also be of concern when constructing phylogenies for distantly related species. Although this complication has long been acknowledged (32, 41, 77), it has been fully appreciated only since the recent availability of whole-genome sequences. Sampling more taxa paradoxically exacerbates this problem, as each new taxon splits a branch in two, further reducing distances between speciation events in the taxon group.

Fitting a single tree to data with ILS results in excess homoplasies because informative sites supporting discordant gene trees mimic the effects of multiple mutations on individual branches. Even rare unique genomic events such as LINE (long interspersed element) insertions and specific gene duplications cannot be used as proof for a specific phylogenetic relationship because of ILS. False homoplasies bias the number of nonsynonymous and synonymous events, possibly to different degrees, leading to biases in estimates of dN/dS, the rate of nonsynonymous to synonymous substitutions, and to biased estimates of positive selection through branch-specific models (84). Likewise, ancestral genome sequence reconstructions based on a single tree are biased (7).

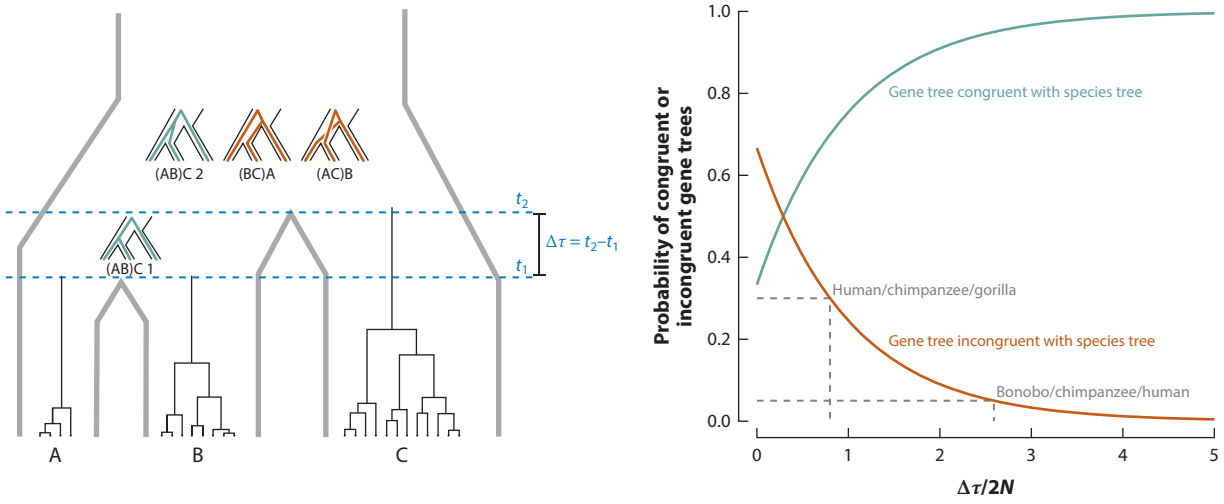

Figure 4

The probability that a gene tree is congruent or incongruent with the species tree (see also Reference 15a). Consider three species, A, B, and C, with A and B being more closely related by a time period $\Delta\tau$ and with all speciation events far enough back in time that we can ignore ancestral polymorphism. Regardless of which gene we pick in the present day species, at the time of the speciation events we have one from each species. The possible gene trees can be split into four categories, two of which are congruent with the species tree and two of which are incongruent. Either the genes from A and B coalesce in the time between the two speciation events, a case we call (AB)C1, or the two lineages reach the common ancestor of all three species, in which case either A and B coalesce first [(AB)C2], B and C coalesce first [(BC)A], or A and C coalesce first [(AC)B]. (AB)C1 and (AB)C2 are congruent with the species tree, whereas (BC)A and (AC)B are not. The probability of seeing (AB)C1 depends on $\Delta\tau$, which is determined by the actual time between the two speciation events divided by the effective population size in the A and B ancestor. The other three cases all have the same probability. The probability of seeing a congruent or incongruent gene tree is shown on the right as a function of this $\Delta\tau$. The human-chimpanzee-gorilla triplet, in which approximately one-third of the gene trees are incongruent (48), and the bonobo-chimpanzee-human triplet, in which approximately 5% of the gene trees are incongruent (58), are shown.

A common approach is to assume that the most frequently occurring gene tree supports the true phylogeny, although this is not always true (12, 13). Widely used approaches include reconciliation by concatenation of data (54; but see also 55), voting procedures (9), and modeling frameworks (51, 79). A recent suggestion is to choose data where ILS is expected to occur less frequently because of a smaller expected effective population size in the ancestral species (44). This can be the X chromosome, with an expected population size that is three-quarters that of autosomes, or mitochondria or the Y chromosome, with an expected population size one-quarter that of autosomes. However, mitochondria and Y chromosomes do not recombine and are therefore single manifestations of an evolutionary relationship. In recent human evolution, mitochondria appear to not be very reliable for phylogenetic inference. Mitochondrial sequences group human and Neanderthals, with Denisovans as an outgroup, but the nuclear genome puts Neanderthals and Denisovans as sister taxa (28, 38, 53). A recently published, more divergent hominoid mitochondrial sequence (37) should be interpreted cautiously until nuclear sequences are reported from the same sample.

LINEAGE SORTING AND ANCESTRAL POPULATION GENOMICS

Although ILS can be considered a nuisance that must be treated carefully in phylogenetic analysis, it is a blessing for our ability to infer evolutionary processes in ancestral species. Incorporating

multiple loci in phylogenetic inference improves the inference but requires explicit modeling of ILS to be generally applicable. More importantly, such models describe population genetics processes of ancestral populations and allow inference of the parameters of these processes. In an early example of explicit modeling of ILS to resolve a trichotomy, Wu (77) derived the probabilities of congruent versus incongruent gene trees for a three-species model (**Figure 4**). He developed a maximum likelihood test to resolve the species tree from the number of gene trees, exposing each of the three topologies, and provided a formal method for testing whether the fraction of different gene trees significantly prefers one species tree. Chen & Li (9) later used this method to resolve the human-chimpanzee-gorilla phylogeny using sequence data from 53 autosomal segments.

Takahata et al. (67) took this beyond merely resolving phylogenies in the presence of incongruent gene trees. They developed a maximum likelihood method that could not only resolve the phylogeny but also separate species divergence times from gene divergence times and infer the effective population size of ancestral species. The method considers the four cases of gene trees from **Figure 4**, (AB)C1, (AB)C2, (BC)A, and (AC)B, and computes the probabilities of observing them, conditional on the branch lengths of the trees and the effective population sizes in the two ancestral species, and then computes the likelihood of the data given these gene trees. The method was used to infer the speciation time between humans, chimpanzees, and gorillas, the effective size of the human-chimpanzee ancestor, and the ancestor of all three species.

Building upon Takahata et al. (67), Yang and colleagues developed both maximum likelihood and Bayesian methods for inferring speciation times and ancestral effective population sizes from three species (8, 50, 81, 82). These methods also consider the four cases from Takahata et al. (67) and integrate over the probabilities of each gene tree and over the branch lengths to compute the parameters of interest, speciation times, and population sizes. Wall (71) took a similar approach, using Monte Carlo simulations of the coalescence process in the ancestral species to compute the likelihood of speciation times and ancestral population parameters.

These methods assume allopatric speciation, in which a panmictic population splits instantaneously into two populations with no subsequent gene flow. Recent evidence, however, suggests that in great apes (34, 46, 58, 68) and archaic hominins (38, 49, 53, 57), this scenario is the exception rather than the rule.

Innan & Watanabe (26) constructed a test for comparing a clean population split to a split followed by gene flow by modeling the coalescence density as a continuous-time Markov model and integrating over this density. This method is also applied in References 21 and 72 and later generalized to larger sample sizes (2, 74). Innan & Watanabe (26) applied their method to the human-chimpanzee speciation but were unable to reject the null hypothesis of no gene flow after an initial split.

Patterson et al. (42) analyzed the mutation patterns and the spatial patterns of segregating sites supporting different gene topologies in alignments of human, chimpanzee, gorilla, orangutan, and macaque. They concluded that the variation in coalescence time of the human-chimpanzee ancestor was extremely large, spanning up to four million years. For the X chromosome they observed a much smaller variation in coalescence time as well as near absence of gene-tree incongruence with the species tree, and suggested that this would be compatible with a complex speciation process in which proto-humans and chimpanzees initially split apart but after one or two million years hybridized before splitting into the two species we have today. This scenario was met with critiques arguing that the variation in coalescence times would also be consistent with an allopatric speciation provided that the ancestral effective population size was sufficiently large (5, 69). Others suggested that the reduced variation in the X chromosome could be caused by male-biased mutation and sperm competition (47) and that the observations could be caused by selection on X (58). Yamamichi et al. (80) built and tested an explicit model of the hybridization

scenario in a maximum likelihood framework and did not reject the null model of instantaneous speciation.

Wu & Ting (78) argued that speciation genes in the presence of gene flow could leave a signal in the divergence patterns along the genome. Genes adapting in different directions in two ancestral populations would cause reproductive isolation by slowly decreasing heterozygotic fitness as they evolve, and this should be visible as a deeper divergence between the two species. Osada & Wu (40) compared coding and noncoding SNPs and rejected the hypothesis that they had the same (speciation) divergence time, and Yang also rejected a genome-wide split time using an extension of his previous model to allow different speciation times along the genome (83). Zhu & Yang (85) extended this further with an explicit model of gene flow and again found a preference for the gene-flow model. All of this suggests that the human and chimpanzee speciation could have occurred in the presence of gene flow.

Incorporating Recombination in Full-Genome Analyses

The methods reviewed in the previous section model coalescence times and changing gene trees by assuming that the loci considered are sufficiently far apart to be practically independent and that they are sufficiently small for intralocus recombination to be ignored. These assumptions are problematic because the rates of recombination and mutation are on the same order of magnitude. Loci without recombination are either too short to have enough mutations to inform us of the divergence or are biased toward more recent ancestry.

Becquet & Przeworski (6) introduced intralocus recombination in a Markov chain Monte Carlo framework to model speciation in the presence of gene flow. The model estimates the likelihood at any parameter point by simulating genealogies from the coalescent process with recombination and computing the probability of the genealogies producing the same summary statistics as observed in the real data. Simulating data makes models with recombination more computationally tractable and for short loci the method is able to include intralocus recombination while still assuming free recombination between loci. Simulating data with recombination for very long sequences, however, is not a feasible approach, as the simulation time is expected to grow exponentially as a function of recombination rate. For long sequences, some approximation seems necessary.

Focusing on independent loci imparts a great reduction in the amount of data available to explore the speciation events and ancestral population genetics. Effective analysis of complete genomes cannot ignore dependencies between loci, requiring that models take recombination fully into account. The sequential Markov coalescent has recently received a lot of attention. The coalescence process with recombination was originally described as a process running backward in time (24), but Wiuf & Hein (75) showed that it could also be modeled as a process running along a sequence alignment. This process, however, is computationally intractable for long sequences because it requires keeping track of all local gene trees along the sequence.

The sequential process can be approximated by one that is Markovian along the sequence. McVean & Cardin (36) constructed such a model, the sequential Markov coalescence (SMC), by restricting the model to allow only coalescence events between lineages that share ancestral material. This implies that two lineages arising from a recombination event cannot coalesce back to remove any effect of the recombination. A large fraction of the recombination and coalescence events deep in the genealogy are, however, of this type, and Chen et al. (10) and Marjoram & Wall (35) removed the restriction creating the alternative sequential Markov coalescence SMC′. The theoretical differences between these two models were recently explored by Hobolth & Jensen (23).

Hobolth et al. (22) built a model to analyze four species alignments that combine the coalescence with hidden Markov models (HMMs) to infer the speciation times of human and chimpanzee

and of human and gorilla. The HMM used the four genealogies of Takahata et al. (**Figure 4**) as the hidden states in the HMM and inferred coalescence parameters indirectly from fitted transition probabilities of the HMM. Dutheil et al. (15) later combined the SMC model with the HMM from Reference 22 with a model deriving transition probabilities between gene trees from coalescence theory. This coalescent HMM has been applied to genome alignments of the great apes to disentangle speciation time from divergence time and obtain many of the estimates in **Table 1** and **Figure 1**. Using an alignment of human, chimpanzee, and gorilla (and orangutan as an outgroup), the human-chimpanzee and human-gorilla speciations were dated to 5.2 and 8.6 Myr, respectively, using full-genome data (58). This analysis showed that the ancestral population sizes of humans and chimpanzees were much larger than those of living great apes (86,000), producing 30% ILS between these three species. Applying the same methodology, the bonobo-chimpanzee split (48) was dated to 1.2 Myr, but these species show only 3% ILS owing to the much longer time between the bonobo-chimpanzee and human-chimpanzee speciation events and to a smaller population size of the bonobo-chimpanzee ancestor. A more recent study (46), however, suggests that the proportion of ILS between these three species may be closer to 5%. Only a small amount of ILS in a full-genome alignment is required to obtain confident estimates of split times and ancestral population sizes, evidenced by an analysis of human, chimpanzee, and orangutan showing only ~1% ILS (31).

The sequential Markov coalescent approximation allows a much wider range of different population genetics applications recently developed in parallel by many groups. Li & Durbin (30) presented the now widely used pairwise sequential Markov coalescent (PSMC) approach, which models changing coalescence times for a diploid genome to infer changing population sizes over time. This approach can also indirectly infer speciation times when comparing the trajectories of population sizes for individuals from different species or by analyzing an artificial diploid individual consisting of haploid genomes from individuals of two different species (46). Song and others (43, 61, 63, 64) combined the sequential Markov coalescent with conditional sampling approaches to extend the model beyond what is computationally feasible if all possible gene trees are modeled. Rasmussen et al. (52) combined the sequential Markov coalescence with a Markov Chain Monte Carlo approach to sample a distribution of gene trees. For inference of speciation events, Mailund et al. (33, 34) modeled changing coalescence times. They used a continuous-time Markov chain similar to Simonsen & Churchill's (62) to construct HMM models of simple divergence either with no gene flow or that allowed for a period of restricted gene flow, thus modeling allopatric and sympatric speciation, respectively. Applying these models to pairs of great ape species, speciation with a period of gene flow provides the best fit to the divergence of the eastern and western gorilla and to the Bornean and Sumatran orangutan, but not to the split between bonobo and chimpanzee, which appears to be instantaneous (34). The same study also finds evidence for a slow divergence process of human and chimpanzee in line with References 40, 83, and 85, and discussed above.

Patterns of ILS inferred along a genomic alignment can identify the genomic segments where pairs of species are more closely related and the genomic breakpoints where the shift in gene genealogy approximately occurs. An HMM allows posterior decoding of the probability for each possible gene genealogy for each base pair in the alignment. **Figure 5** shows posterior decoding for a typical example for the human, chimpanzee, and gorilla analysis based on the coalescent HMM approach (15). The definition of regions sharing the same genealogy is subsequently possible by defining an appropriate threshold of, e.g., 50% support. In general, for the case of human, chimpanzee, and gorilla, the median lengths of segments where gorillas are more closely related to either humans or chimpanzees are approximately 200 bp, whereas segments supporting the species phylogeny are longer (median: approximately 2 kb). The mixed ancestry induced by ILS may also be reflected in functional differences between species. A comparison of transcriptome

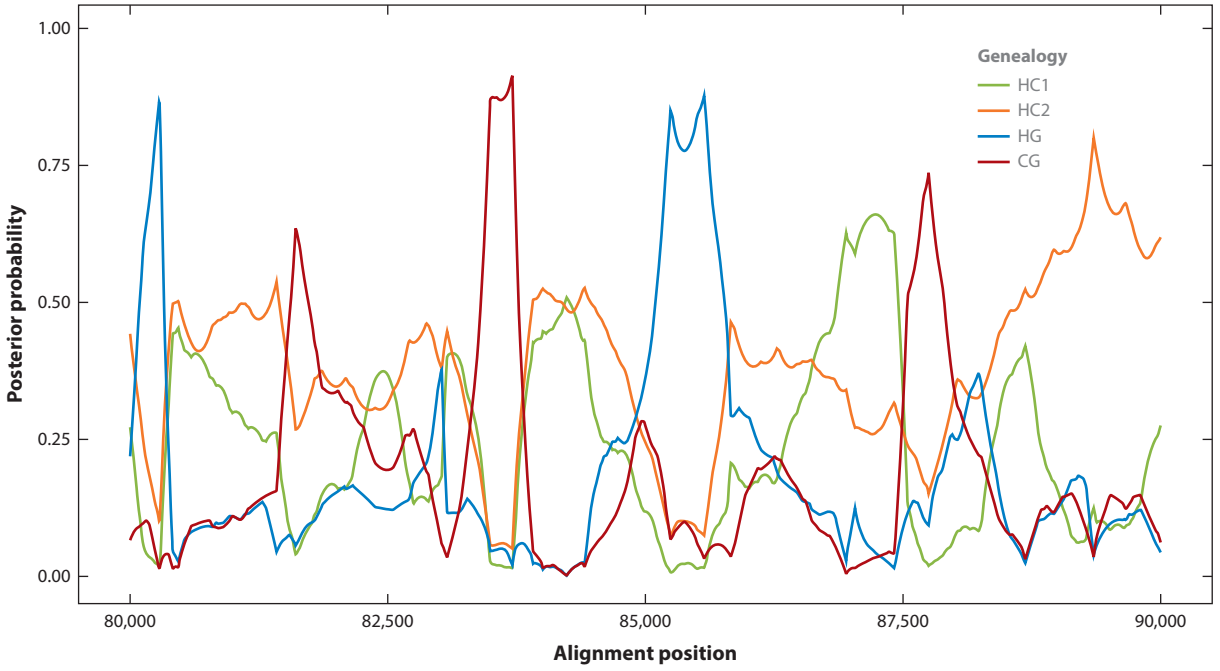

Figure 5

Posterior probabilities of the four genealogies representing states in the CoalHMM (coalescent hidden Markov model) shown for a segment of aligned chromosomes 2 from human, chimpanzee, gorilla, and orangutan. HC, human-chimpanzee; HG, human gorilla; and CG, chimpanzee-gorilla gene genealogy. HC1 and HC2 denote the two different cases for the human-chimp genealogy, see **Figure 4**.

variation among 1:1:1 orthologous genes among human, chimpanzee, and gorilla (58) showed that patterns in relative divergence of expression and splicing significantly overlapped the patterns of inferred ILS.

DETECTING SELECTION IN ANCESTRAL POPULATIONS

Speciation times generally apply equally to the entire genome, although exceptions include regions involved in speciation or subsequent admixture. In contrast, the effective size of an ancestral population varies along the genome as a result of population genetics processes such as selection. Coalescent HMMs can estimate N_e but require at least one megabase of genomic alignment to fit all model parameters. However, the pattern of ILS offers information about N_e on a finer scale. By identifying individual segments of the genomic alignment showing ILS, the proportion of ILS can be identified for any genomic region. Given two speciation times $\Delta\tau$ generations apart, the effective population size expected to produce an observed proportion of ILS, p, can be computed as $N_e = \Delta\tau/2\ln(2/3p)$.

Both positive and negative selection in the ancestral population is expected to reduce the effective population size in regions linked to selected sites. Indeed, the proportion of ILS is reduced near coding genes (22, 48, 58). Consecutive regions devoid of ILS may be evidence of strong positive selection given that a selective sweep can force all lineages to coalesce, leaving none available for ILS. One such region was found on chromosome 3 in the bonobo-chimpanzee ancestor spanning 6.1 Mb, twice the length of the second longest such region found on autosomes

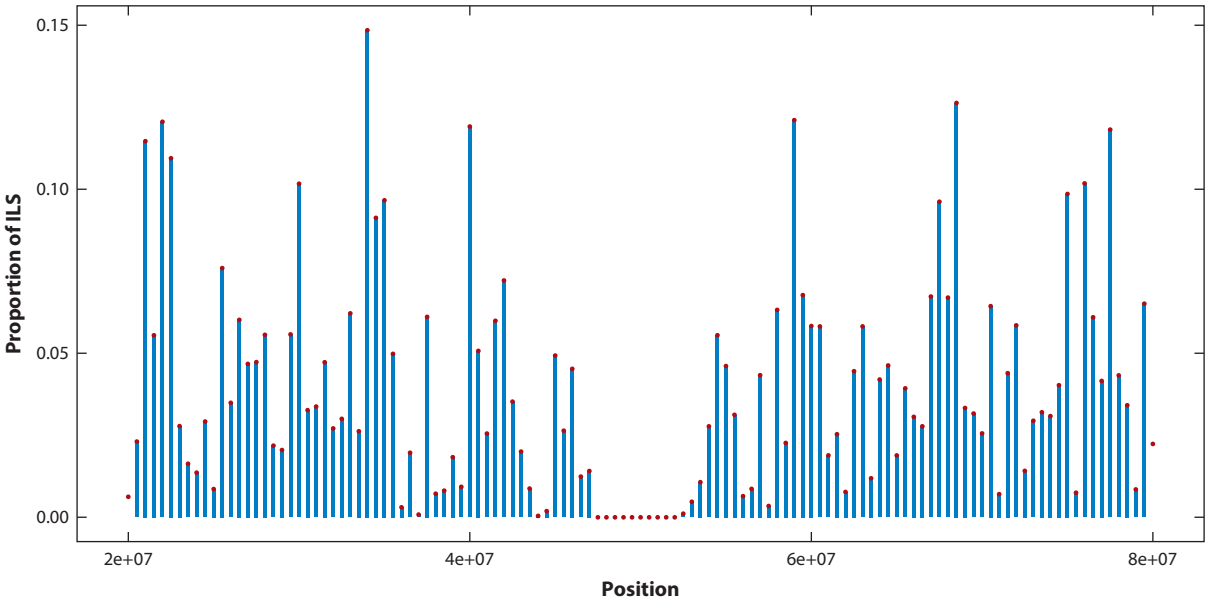

Figure 6

Proportion of incomplete lineage sorting (ILS) between bonobo, chimpanzee, and human in windows of 500 kb around a region of chromosome 3 that is completely devoid of ILS.

(48; see **Figure 6)**. The region contains the *CCR5* gene involved in HIV resistance (56) as well as a cluster of immunity-related genes shown to be under positive selection in humans (4). It is also devoid of polymorphism in central chimpanzees (25) and shows a strong signature of introgression from Neanderthals to modern humans (14). However, the efficacy by which selection reduces the local effective population size, and with it the probability of ILS, is increased in regions where the recombination rate is low, and in the chromosome 3 region the human recombination rate is an order of magnitude lower than the genome average (27). The dependence on recombination rate applies to both positive and negative selection, and a positive correlation between recombination rate and proportion of ILS has been reported in great apes (22, 48, 58), suggesting that selection is a prominent force in shaping patterns of ILS. This example illustrates that patterns of ILS can aid interpretations of patterns of genetic diversity along the genome. It has long been discussed whether the positive correlation between recombination rate and genetic diversity in the human genome (19) can be attributed to an association between recombination and mutation. However, because ILS patterns are not influenced by mutation, the results support the conclusion from the 1,000 Genomes Project that the correlation between recombination and diversity is not caused by mutational processes alone (1).

Balancing selection has the opposite effect on the local effective population size, making lineages coalesce further back in time than under neutrality. Regions targeted by balancing selection should thus be more variable and the variation should be of more intermediate frequency (3, 11). As a result of longer coalescence times, the proportion of ILS should also increase under balancing selection. If sufficiently strong, balancing selection should lead to shared lineages even between human and chimpanzees, and mutations occurring in the common ancestor would potentially manifest themselves as shared polymorphism. Leffler et al. (29) used this idea to scan human

and chimpanzee genome-wide polymorphism data and found 125 cases in which sets of linked variants were polymorphic in both species. These regions tended to show association with proteins presumed to be involved with pathogen interaction. One was even shared with gorilla, showing how extensive ILS can be under balancing selection. These new regions add to well known cases of balancing selection at the MHC (major histocompatibility complex) (59, 66) and the ABO blood group system (60).

ANCESTRAL RECOMBINATION

The spatial pattern of genealogies along an alignment also holds information about past recombination. Segments of genomic alignment representing different genealogies are separated by points of crossover where a historical recombination event has allowed two new lineages to follow individual histories. The rate of change of genealogies along an alignment of many individuals from the same species is summarized by linkage disequilibrium and has been used to estimate genetic maps. In the same way, change in genealogies along an alignment of different species is evidence of past recombination events (39). For species showing ILS, a subset of gene trees represents topologies different from the species tree. Recombination events that mark the transition to such an ILS topology are of particular interest because these events must occur so far back in time that the resulting lineages have the opportunity to coalesce with lineages other than the one representing the sister species and thus form alternative topologies. Such recombination events can be identified by observing transitions between segments of the alignment supporting alternative topologies (see **Figure 5**). An ancestral recombination map can then be derived from the rate of such transitions along an alignment.

Although the detectable number of recombination events causing topology change depends on the amount of ILS, and thus the relative size of $\Delta\tau$ and ancestral N_e, their distribution in time and among the branches of the species tree are determined by the relative size of the three extant populations and the two ancestral populations. The effective size of the human-chimpanzee population is large compared with those of the human, chimpanzee, gorilla, and the human-gorilla ancestor. This serves to concentrate more than 60% of recombination events in the human-chimpanzee ancestor with the remaining events occurring in the proximal part of the three adjoining branches. Ancestral recombination maps can be created from any three species showing ILS. The approach allows differences in recombination rate between species to be resolved into the change that occurred in each species in the course of their divergence, allowing us to study whether the rate of evolution is constant across primate evolution and to directly link evolution of genomic sequence to evolution of recombination rate to better understand how it is controlled.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Freddy B. Christiansen for comments on a previous version of the manuscript. This research was supported by grants from the Danish Council for Independent Research–Natural Sciences to Thomas Mailund and Mikkel H. Schierup.

LITERATURE CITED

1. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–73
2. Andersen LN, Mailund T, Hobolth A. 2014. Efficient computation in the IM model. *J. Math. Biol.* 68(6):1423–51
3. Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, et al. 2009. Targets of balancing selection in the human genome. *Mol. Biol. Evol.* 26(12):2755–64
4. Barreiro L, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat. Rev. Genet.* 11(1):17–30
5. Barton NH. 2006. Evolutionary biology: How did the human species form? *Curr. Biol.* 16(16):R647–50
6. Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* 17(10):1505–19
7. Blanchette M, Diallo AB, Green ED, Miller W, Haussler D. 2008. Computational reconstruction of ancestral DNA sequences. *Methods Mol. Biol.* 422:171–84
8. 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):1979–94
9. 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):444–56
10. Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. *Genome Res.* 19(1):136–42
11. DeGiorgio M, Lohmueller KE, Nielsen R. 2014. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLOS Genet.* 10(8):e1004561
12. Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLOS Genet.* 2(5):e68
13. Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24(6):332–40
14. Ding Q, Hu Y, Xu S, Wang J, Jin L. 2014. Neanderthal introgression at chromosome 3p21.31 was under positive natural selection in East Asians. *Mol. Biol. Evol.* 31(3):683–95
15. 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):259–74

15a. Dutheil JY, Hobolth A.. 2012. Ancestral population genomics. *Methods Mol. Biol.* 856:293–13

16. Fabre P-H, Rodrigues A, Douzery EJP. 2009. Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA. *Mol. Phylogenet. Evol.* 53(3):808–25
17. Glazko GV, Nei M. 2003. Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* 20(3):424–34
18. Hein J, Schierup M, Wiuf C. 2005. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford: Oxford Univ. Press
19. Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* 72(6):1527–35
20. Hey J. 2010. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Mol. Biol. Evol.* 27(4):921–33
21. 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):1241–43
22. 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
23. Hobolth A, Jensen JL. 2014. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theor. Popul. Biol.* In press
24. Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* 23(2):183–201
25. Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, et al. 2012. Extensive X-linked adaptive evolution in central chimpanzees. *Proc. Natl. Acad. Sci. USA* 109(6):2054–59

26. Innan H, Watanabe H. 2006. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Mol. Biol. Evol.* 23(5):1040–47
27. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467(7319):1099–103
28. Krause J, Fu Q, Good JM, Viola B, Shunkov MV, et al. 2010. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* 464(7290):894–97
29. Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* 339(6127):1578–82
30. Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475(7357):493–96
31. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, et al. 2011. Comparative and demographic analysis of orangutan genomes. *Nature* 469(7331):529–33
32. Maddison WP. 1997. Gene trees in species trees. *Syst. Biol.* 46(3):523–36
33. Mailund T, Dutheil JY, Hobolth A, Lunter G, Schierup MH. 2011. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLOS Genet.* 7(3):e1001319
34. Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, et al. 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLOS Genet.* 8(12):e1003125
35. Marjoram P, Wall JD. 2006. Fast "coalescent" simulation. *BMC Genet.* 7:16
35a. Matsudaira K, Ishida T. 2010. Phylogenetic relationships and divergence dates of the whole mitochondrial genome sequences among three gibbon genera. *Mol. Phylogenet. Evol.* 55(2):454–59
36. McVean GA, Cardin NJ. 2005. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. Ser. B* 360(1459):1387–93
37. Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, et al. 2014. A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* 505(7483):403–6
38. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338(6104):222–26
39. 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):467–74
40. Osada N, Wu C-I. 2005. Inferring the mode of speciation from genomic data: a study of the great apes. *Genetics* 169(1):259–64
41. Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5(5):568–83
42. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441(7097):1103–8
43. Paul JS, Steinrücken M, Song YS. 2011. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics* 187(4):1115–28
44. Pease JB, Hahn MW. 2013. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. *Evol. Int. J. Org. Evol.* 67(8):2376–84
45. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, et al. 2011. A molecular phylogeny of living primates. *PLOS Genet.* 7(3):e1001342
46. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, et al. 2013. Great ape genetic diversity and population history. *Nature* 499(7459):471–75
47. Presgraves DC, Yi SV. 2009. Doubts about complex speciation between humans and chimpanzees. *Trends Ecol. Evol.* 24(10):533–40
48. Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486(7404):527–31
49. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505(7481):43–49
50. Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164(4):1645–56

51. Rasmussen MD, Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* 22(4):755–65
52. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. *PLOS Genet.* 10(5):e1004342
53. Reich D, Green RE, Kircher M, Krause J, Patterson N, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468(7327):1053–60
54. Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425(6960):798–804
55. Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497(7449):327–31
56. Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, et al. 1996. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 382(6593):722–25
57. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, et al. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507(7492):354–57
58. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483(7388):169–75
59. Schierup MH, Mikkelsen AM, Hein J. 2001. Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. *Genetics* 159(4):1833–44
60. Ségurel L, Thompson EE, Flutre T, Lovstad J, Venkat A, et al. 2012. The ABO blood group is a trans-species polymorphism in primates. *Proc. Natl. Acad. Sci. USA* 109(45):18493–98
61. 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):647–62
62. Simonsen KL, Churchill GA. 1997. A Markov chain model of coalescence with recombination. *Theor. Popul. Biol.* 52(1):43–59
63. 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
64. Steinrücken M, Wang YX, Song YS. 2013. An explicit transition density expansion for a multi-allelic Wright-Fisher diffusion with general diploid selection. *Theor. Popul. Biol.* 83:1–14
65. Steiper ME, Young NM. 2006. Primate molecular divergence dates. *Mol. Phylogenet. Evol.* 41(2):384–94
66. Takahata N, Satta Y, Klein J. 1992. Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics* 130(4):925–38
67. Takahata N, Satta Y, Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.* 48(2):198–221
68. Thalmann O, Fischer A, Lankester F, Pääbo S, Vigilant L. 2007. The complex evolutionary history of gorillas: insights from genomic data. *Mol. Biol. Evol.* 24(1):146–58
69. Wakeley J. 2008. Complex speciation of humans and chimpanzees. *Nature* 452(7184):E3–4; discussion E4
70. Wakeley J. 2009. *Coalescent Theory: An Introduction*. Greenwood Village, CO: Roberts & Company Publ.
71. Wall JD. 2003. Estimating ancestral population sizes and divergence times. *Genetics* 163(1):395–404
72. Wang Y, Hey J. 2010. Estimating divergence parameters with small samples from a large number of loci. *Genetics* 184(2):363–79
73. Wegmann D, Excoffier L. 2010. Bayesian inference of the demographic history of chimpanzees. *Mol. Biol. Evol.* 27(6):1425–35
74. Wilkinson-Herbots HM. 2012. The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. *Theor. Popul. Biol.* 82(2):92–108
75. Wiuf C, Hein J. 1999. Recombination as a point process along sequences. *Theor. Popul. Biol.* 55(3):248–59
76. Won YJ, Hey J. 2005. Divergence population genetics of chimpanzees. *Mol. Biol. Evol.* 22(2):297–307
77. Wu CI. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* 127(2):429–35

78. Wu C-I, Ting C-T. 2004. Genes and speciation. *Nat. Rev. Genet.* 5(2):114–22
79. Wu Y-C, Rasmussen MD, Bansal MS, Kellis M. 2013. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome Res.* 24(3):475–86
80. Yamamichi M, Gojobori J, Innan H. 2012. An autosomal analysis gives no genetic evidence for complex speciation of humans and chimpanzees. *Mol. Biol. Evol.* 29(1):145–56
81. Yang Z. 1997. On the estimation of ancestral population sizes of modern humans. *Genet. Res.* 69(2):111–16
82. Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* 162(4):1811–23
83. Yang Z. 2010. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome Biol. Evol.* 2:200–11
84. Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17(1):32–43
85. Zhu T, Yang Z. 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Mol. Biol. Evol.* 29(10):3131–42

Annual Review of Genetics

Volume 48, 2014

Contents

Errata

An online log of corrections to *Annual Review of Genetics* articles may be found at http://www.annualreviews.org/errata/genet