A new approach to estimate parameters of speciation models with application to apes

Celine Becquet[1] and Molly Przeworski[1]

Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA

How populations diverge and give rise to distinct species remains a fundamental question in evolutionary biology, with important implications for a wide range of fields, from conservation genetics to human evolution. A promising approach is to estimate parameters of simple speciation models using polymorphism data from multiple loci. Existing methods, however, make a number of assumptions that severely limit their applicability, notably, no gene flow after the populations split and no intralocus recombination. To overcome these limitations, we developed a new Markov chain Monte Carlo method to estimate parameters of an isolation-migration model. The approach uses summaries of polymorphism data at multiple loci surveyed in a pair of diverging populations or closely related species and, importantly, allows for intralocus recombination. To illustrate its potential, we applied it to extensive polymorphism data from populations and species of apes, whose demographic histories are largely unknown. The isolation-migration model appears to provide a reasonable fit to the data. It suggests that the two chimpanzee species became reproductively isolated in allopatry ~850 Kya, while Western and Central chimpanzee populations split ~440 Kya but continued to exchange migrants. Similarly, Eastern and Western gorillas and Sumatran and Bornean orangutans appear to have experienced gene flow since their splits ~90 and over 250 Kya, respectively.

[Supplemental material is available online at www.genome.org.]

Although central to evolutionary biology, the question of how species form remains largely open. In fact, the very definition of species is a subject of active debate (Hey 2006). The most common definition is the "biological" one, in which species are defined as groups of interbreeding organisms that are reproductively isolated from other populations. The introduction of this concept >60 yr ago transformed the study of speciation into a research program to examine the conditions under which reproductive isolation emerges and to uncover its genetic architecture (Mayr 1963).

Accumulating evidence suggests that incipient species arise primarily in populations with restricted gene flow, as alleles (or combination of interacting alleles) that contribute to reproductive isolation reach fixation (e.g., Wittbrodt et al. 1989; Sawamura et al. 1993; Ting et al. 1998; Wang et al. 1999; Fossella et al. 2000; Barbash et al. 2003; Presgraves et al. 2003; Coyne and Orr 2004a). The speciation process initiates after two populations become completely isolated from one another (i.e., are in allopatry) or as they continue to exchange migrants (i.e., in parapatry).

Under a model of allopatric speciation, the process occurs through the homogeneous divergence of the genome. Shortly after the split, the two populations share alleles due to the persistence of ancestral polymorphism (more so if the ancestral population size is large). Eventually, however, the shared alleles are lost or reach fixation and the two populations start to accumulate fixed differences, either by genetic drift or due to differential adaptation (Coyne and Orr 2004a). Under a simple allopatric model with no selection, it will take approximately 9–12N generations (where N is the effective size of the descendant population) for the genealogies of >95% of loci to be reciprocally monophyletic and, hence, for the two populations not to share alleles that are identical by descent (Hudson and Coyne 2002). Given these assumptions, humans and common chimpanzees should almost never share alleles (as they are thought to have diverged ~ 20–25N generations ago) (Wall 2003; Hobolth et al. 2006; Patterson et al. 2006), while bonobos and common chimpanzees are expected to share alleles at ~50% of loci (since they are estimated to have diverged ~ 4N generations ago; Won and Hey 2005).

If the incipient species are in parapatry, however, divergence is not believed to occur homogeneously across the genome but instead in a number of stages (Wu 2001). First, alleles that cause a decrease in hybrid fertility or viability reach fixation in the parental populations. The populations may continue to exchange migrants, but in the genomic regions carrying functionally divergent or incompatible alleles, gene flow is selected against and hence effectively restricted. By contrast, in unlinked (or loosely linked) genomic regions, alleles can be brought in by migrants with no associated fitness costs. Thus, at neutral loci, populations share alleles longer than expected under allopatric speciation. Eventually, reproductive isolation factors accumulate in sufficient numbers as to prevent gene flow throughout the genome—the final stage of speciation. This model predicts variation in the number of shared alleles and levels of divergence along the genomes of closely related species. While shared alleles are also expected under a model of recent allopatric speciation, greater variation is expected along the genome under parapatry, such that, with enough data, the two scenarios should be distinguishable.

In these simple speciation models, the salient parameters are the split times, effective population sizes, and, in the case of parapatry, the gene flow rates. Thus, learning about these parameters should greatly deepen our understanding of speciation. This realization has motivated the development of statistical methods

[1]Corresponding authors.
E-mail cbecquet@uchicago.edu; fax (773) 834-0505.
E-mail mfp@uchicago.edu; fax (773) 834-0505.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6409707. Freely available online through the *Genome Research* Open Access option.

to estimate the parameters from multilocus patterns of polymorphism in closely related species.

Existing methods all assume that genetic variation data are available from both populations, at a number of independently-evolving loci, but differ in their assumptions about gene flow and recombination, and in whether they use all the polymorphism data or summaries. Loosely, they can be classified into two groups. The first set assumes an extreme model of allopatry, in which a panmictic (i.e., randomly mating) ancestral population instantaneously splits into two panmictic descendant populations, with no subsequent gene flow. In this model, there are four parameters: the three effective population sizes and the split time. The parameters are estimated using summaries of the polymorphism data, either by a moment estimator (Wakeley and Hey 1997; Kliman et al. 2000) or by maximum likelihood (Leman et al. 2005; Putnam et al. 2007). While, in its current version, the method of Leman et al. (2005) can only be applied to one, non-recombining locus, other methods can be applied to multiple loci and incorporate recombination (Wakeley and Hey 1997; Putnam et al. 2007). They use highly summarized versions of the data, however, at the potential cost of much information. Moreover, in the presence of gene flow after the split, their estimates will be biased—the ancestral effective population size will tend to be overestimated (Wall 2003) and the split time underestimated (Leman et al. 2005).

The second set of methods considers a more general model, often called the "isolation-migration" model, in which there is gene flow between incipient species throughout the genome, either at fixed (Hey and Nielsen 2004) or locus-specific rates (Won and Hey 2005). The parameters are estimated from all the polymorphism data at a single locus (Nielsen and Wakeley 2001) or at multiple loci (Hey and Nielsen 2004), using Markov Chain Monte Carlo (MCMC). The Hey and Nielsen method, henceforth called IM, has been applied to a number of species, from *Heliconius* (Bull et al. 2006) to cichlids (Hey et al. 2004; Won et al. 2005). These applications suggest that speciation often occurs in the presence of some gene flow (Hey 2006).

While IM considers a wide range of models, it assumes that haplotypes are known and that there is no intralocus recombination. Although not ideal, the first assumption is not restrictive, as a two-step procedure can be used in which haplotype phase is inferred (e.g., using the program PHASE; Stephens et al. 2001) and then IM is run on the phased data. In contrast, the assumption of no recombination is more limiting, because the method can only be applied to autosomal loci by excluding segments or haplotypes with evidence for recombination. This practice is likely to bias estimates of the parameters, as excluding segments with visible recombination will tend to lead to shorter genealogical histories (Hey and Nielsen 2004). Moreover, if intralocus recombination is not taken into account, a small variance in divergence times across segments may be confounded with a small ancestral effective population size (Takahata and Satta 2002). The assumption of no intralocus recombination represents an especially severe limitation in species in which the ratio of recombination to mutation is thought to be high (e.g., *Drosophila melanogaster*, Andolfatto and Wall 2003; or *Papilio glaucus*, Putnam et al. 2007). In such species, any segment with polymorphisms in a sample is likely to have experienced numerous recombination events in its genealogical history, making recombination hard to ignore (Hudson and Kaplan 1985; Nordborg and Tavare 2002).

To overcome this limitation, we developed a new Bayesian approach to estimate parameters of an isolation-migration model from recombining loci. We have in mind data sets similar to the ones most commonly collected to date: short noncoding sequences distributed throughout the genome. Our approach is to summarize the polymorphism data at each locus by four statistics known to be sensitive to the parameters of interest (Wakeley and Hey 1997; Leman et al. 2005). We then estimate the posterior probability of the parameters given these summaries using MCMC. Simulations suggest that, in the absence of recombination, our method performs as well or almost as well as the full likelihood approach. Moreover, the approach presents the advantage of being quite flexible in the demographic model that it can consider and in allowing for intralocus recombination.

We illustrate the potential of our method by applying it to multilocus polymorphism data from noncoding loci in chimpanzees, gorillas, and orangutans. Very little is known about the evolutionary history of great apes, in part because of a poor fossil record. Chimpanzees, the closest living relatives of humans, are classified into two species, common chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*), both found exclusively in Africa. The two chimpanzee species were thought to have diverged as a result of the formation of the River Congo 1.5–3.5 million years ago (Mya) (Beadle 1981; Myers Thompson 2003), but recent estimates of their split time based on genetic data appear to be too recent for this scenario to be plausible (Fischer et al. 2004; Won and Hey 2005). Common chimpanzees are usually subdivided further into three (or sometimes four) "subspecies," including Eastern (*P. troglodytes schweinfurtheii*), Central (*P. troglodytes troglodytes*), and Western (*P. troglodytes verus*) (Hill 1969). The meaning of the term "subspecies" is unclear, at least to us, but the labels are thought to correspond to the most pronounced population structure within the species. This view is supported by a recent analysis of 310 microsatellites, which found three populations within common chimpanzees, which correspond to the three subspecies, and little evidence of recent gene flow between them (Becquet et al. 2007).

Gorillas, in turn, are classically subdivided into two subspecies: Western (*Gorilla gorilla*) and Eastern gorilla (*Gorilla beringei*), found in western and central African forest, respectively (Groves 1970). Some controversy surrounds this classification: The range of the two populations does not currently overlap in the wild; but on the basis of morphological and genetic diversity, it has been proposed that the subspecies should be elevated to the rank of species (e.g., Grubb et al. 2003). Here, we refer to Western and Eastern gorillas as subspecies or populations. A recent application of IM to polymorphism data from the two gorilla populations suggests that they split between 0.08 and 1.6 Mya and experienced low levels of gene flow since (Thalmann et al. 2006).

Even less is known about the history of orangutans (*Pongo pygmaeus*), currently found only in Indonesia and Malaysia, but whose range is thought to have spanned much of southeast Asia until recently (Smith and Pilbeam 1980). Some taxonomies consider Sumatran (*P. p. abelii*) and Bornean (*P. p. pygmaeus*) orangutans to be subspecies (e.g., Groves 1971), and others to be species (e.g., Zhi et al. 1996). Again, these populations do not overlap in their range, so that the classification is based on morphology and behavior, as well as on patterns of genetic diversity. The islands of Sumatra and Borneo were fully formed 500 thousand years ago (Kya) but were reconnected by land bridges during the two last glaciations, ~130–200 Kya and ~10–100 Kya, respectively (Muir et al. 2000; Hughes et al. 2006). Estimates of the

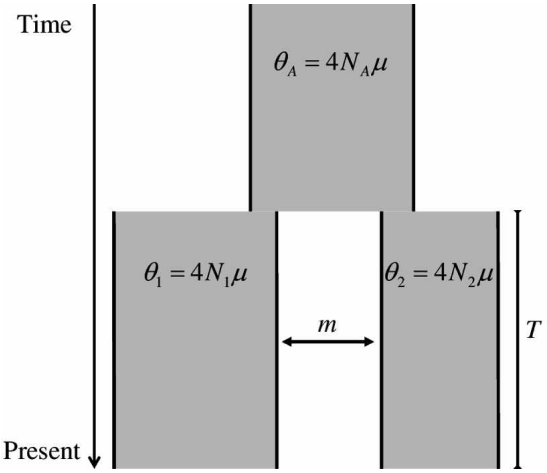

Figure 1. The "isolation-migration" model, in which two populations diverged T generations ago from a common ancestral population. The parameters θ_1, θ_2, and θ_A are the population mutation rates per base pair for populations 1 and 2 and the ancestral population, respectively. The split time in generations is T, and m is the symmetrical migration rate between populations per generation (for details, see Methods).

average time to the most recent common ancestor for both populations based on mitochondrial DNA (mtDNA) and a small number of microsatellites and autosomal and X-linked loci are ~1.5–2.5 Mya (Zhi et al. 1996; Kaessmann et al. 2001; Zhang et al. 2001), but to our knowledge, there are no published estimates of the population split time.

Here, we analyze a compilation of multilocus polymorphism data recently published in the three great ape species (Yu et al. 2003; Fischer et al. 2006; Thalmann et al. 2006), refining population parameter estimates for chimpanzees and gorillas and providing the first estimates for orangutans.

Results

We developed a method that estimates the demographic parameters of an "isolation-migration" model from recombining loci (Fig. 1). There are five parameters of interest: the population mutation rates for the two descendant populations, θ_1 and θ_2, and the ancestral population, θ_A; the time since the populations split in generations, T; and the migration rate, m. To estimate these parameters, the method requires resequencing data from two populations (or closely related species) at independently-evolving loci, and an outgroup sequence. Briefly, the polymorphism data *for each locus* are summarized by the four statistics studied by Wakeley and Hey (1997), as these carry information about the divergence time and other parameters of interest (Wakeley and Hey 1997; Leman et al. 2005). We choose the parameters of the model from prior distributions, and for each locus, we generate a set of genealogies under a model with those parameters. We then estimate the likelihood by calculating the probability of the data summaries at all the loci given the set of genealogies and the parameters. Finally, we obtain a sample from the posterior distribution of the parameters given the data summaries using MCMC (see Methods). Thus, our method follows a number of Bayesian approaches that use summaries of the data but differs in that we update the parameters using MCMC (for more details, see Methods). Hereafter, we refer to our method as MIMAR: MCMC estimation of the isolation-migration model allowing for recombination.

Performance of MIMAR under the allopatric model

In order to assess the performance of our method, we generated 30 simulated data sets, each consisting of 20 non-recombining loci, with parameter values applicable to *Drosophila* species in which related studies have been conducted (Llopart et al. 2005; see Methods). Supplemental Figure S1 shows the 30 posterior distribution samples for the four parameters of interest. As can be seen, the posterior distributions estimated by MIMAR for θ_1, θ_2, and T are centered around their true values with relatively little variance, suggesting that the summaries that we use contain enough information to estimate these parameters precisely and accurately. However, for these parameters, 20 non-recombining loci do not seem to contain as much information about the ancestral effective population size, leading to a wider posterior distribution estimate for θ_A. This does not appear to be a feature of our statistics, since the use of IM yields similar results, even though it is based on the full polymorphism data set (data not shown). As expected, our estimates of θ_A become more precise with larger data sets (data not shown).

Comparison to IM for the case of no recombination

Next, we studied the performance of MIMAR by generating 30 simulated data sets under the allopatric model for 20 non-recombining loci, but this time drawing the parameters from prior distributions (for details, see Methods); the parameter values are, as above, applicable to *Drosophila* species. The results confirm that the estimates of θ_1, θ_2, and T are precise and have very little bias, while the estimates of θ_A are less precise (Table 1; Supplemental Fig. S2).

We analyzed the same simulated data sets with IM to compare the estimates from MIMAR, which are based on summaries, with a full likelihood approach (since IM does not allow for recombination, we set the intralocus recombination rate to 0 when generating the 30 data sets). We found that the two methods perform similarly well, in terms of both accuracy and precision (see Table 1). For example, the mean absolute error over the 30 simulated data sets for the estimate of T is 5.19×10^5 using MIMAR and 5.94×10^5 using IM. Similarly, if we consider the estimate divided by the true value as a measure of bias, the mean over the 30 data sets is 1.004 for MIMAR and 0.980 for IM. Similar results were obtained for all parameters, with the possible exception of the current effective population sizes, for which MIMAR appears to yield slightly more reliable estimates (see Table 1). Moreover, we found that the two methods have similar coverage:

Table 1. Performance of MIMAR and IM

	Mean absolute error		Mean of the estimate divided by the true value	
Parameters	MIMAR	IM	MIMAR	IM
θ_1	0.0003	0.0002	1.000	0.983
θ_2	0.0004	0.0003	1.001[a]	0.968[a]
θ_A	0.0027	0.0037	0.927	0.875
T	5.19×10^5	5.94×10^5	1.004	0.980

Precision and accuracy for the four parameters of the allopatric model (using the mode as a point estimate). MIMAR and IM were applied to 30 simulated multilocus data sets under the allopatric model (for details, see Methods).

[a]The biases in θ_2 estimates from IM and MIMAR are significantly different at the 5% level, after Bonferroni correction ($P = 0.006$ using a Wilcoxon signed rank test).

For both, the central 95th percentiles of the marginal posterior distribution sample for T included the true value in 29 out of 30 cases; for θ_A, this occurred in 29 out of 30 cases for IM and 30 out of 30 cases for MIMAR. We also compared the results of MIMAR and IM on larger simulated data sets of 100 loci and found that, in this case, IM outperformed MIMAR. However, with such large data sets, both methods provided highly accurate and precise estimates (data not shown).

In the comparison, we ran both methods long enough for them to appear to have reached convergence (Supplemental Fig. S3). For the same number of iterations of the MCMC, IM was two to three times faster than MIMAR (data not shown).

Assessing the evidence for gene flow

To assess our ability to distinguish a model with gene flow from one without, we generated 20 simulated data sets (each consisting of 40 recombining loci) under both an allopatric and a parapatric model, with parameter values applicable to apes. In the parapatric model, we fixed the expected number of migrants $M = 4N_1m$ to 1, which corresponds to an average of 11 migration events in the history of the sample. We applied MIMAR to the 40 data sets, allowing for recombination and sampling the expected number of migrants from the prior $\ln(M) \sim U[-5, 2]$ (for details, see Methods). When applied to data sets generated under a model with no gene flow, MIMAR suggested no migration (using the criterion that the mode of the marginal posterior distribution, $\hat{M}$, be < 0.1) in 14 out of 20 cases; moreover, in one out of the six cases in which $\hat{M} \geq 0.1$, most of the posterior probability mass for M was close to 0 (data not shown). For the data sets simulated with gene flow, there was evidence of migration (i.e., $\hat{M} \geq 0.1$) in 17 out of 20 cases. Other parameter estimates were generally accurate and precise (see Table 2; Supplemental Fig. S4), although the estimates of θ_A were slightly underestimated in data sets generated with $M = 0$, and the estimates of T were slightly underestimated in data sets generated with $M = 1$ (for possible explanations, see Table 2).

When we applied either MIMAR and IM to smaller simulated data sets (i.e., 20 loci and no intralocus recombination), estimates of the split times and migration rates provided by both methods were much less reliable (data not shown).

Table 2. Performance of MIMAR when detecting gene flow

	Mean absolute error		Mean of the estimate divided by the true value	
Parameters	$M > 0$	$M = 0$	$M > 0$	$M = 0$
θ_1	0.0008	0.0005	1.144	1.153
θ_2	0.0008	0.0005	1.092	1.085
θ_A	0.0003	0.0004	1.000	0.880[a]
T	1.81×10^4	5.66×10^3	0.721[a]	0.965
M	1.0436	0.487	1.293	NA

Precision and accuracy for the five parameters of the isolation-migration model (using the mode as a point estimate). MIMAR was applied to 20 simulated multilocus data sets under parapatric and allopatric models (for details, see Methods). When $M = 0$, the mean estimate of θ_A is significantly lower than the true value ($P = 0.0003$, using a Wilcoxon signed rank test). This can be explained as follows: The prior on M does not include 0 (the true value) so $\hat{M}$ is necessarily an overestimate and consequently, θ_A tends to be underestimated slightly. This problem is likely to apply to IM as well, since the prior on M is likewise exclusive of 0. When $M = 1$, the mean estimate of T is significantly lower than the true value ($P = 0.005$, using a Wilcoxon signed rank test). This can be explained by the fact that, whenever M and/or θ_A are slightly underestimated, T tends to be underestimated (see Supplemental Fig. S4)
[a]A significant bias in the estimates at the 5% level, after Bonferroni correction.

Sensitivity to intralocus recombination rates

Intralocus recombination rates are often unknown or are estimated with substantial error. To assess how this might affect the reliability of MIMAR, we generated 16 data sets under an allopatric model, with parameter values applicable to *Drosophila* (see above). Each data set consisted of 10 recombining loci, with the locus-specific recombination rates chosen from an exponential distribution with mean $c/\mu = 10$. These data sets were analyzed using MIMAR by fixing all the parameters but T to their true values, and (1) setting the locus-specific recombination rates to their true values, (2) sampling the recombination rates from the same prior as used when generating the simulated data, and (3) setting the intralocus recombination rates to 0 (for details, see Methods). The results from steps 1 and 2 were virtually identical, suggesting that error in the locus-specific recombination rates does not have much effect on the results so long as intralocus recombination is taken into account. In contrast, when we assumed no recombination in our analysis of recombining loci, the estimates of the split time were significantly less accurate and precise (see Supplemental Fig. S5). These results highlight the importance of allowing for intralocus recombination when estimating demographic parameters.

Application to ape data

We compiled a set of recently published resequencing data in bonobo and common chimpanzee, gorilla, and orangutan populations (Yu et al. 2003; Fischer et al. 2006; Thalmann et al. 2006). Won and Hey (2005) had previously reported evidence for intralocus recombination at some of the loci included in this study, and we found further evidence of recombination, in spite of low power to do so (given the small sample sizes). We therefore analyzed these data sets with MIMAR, allowing for intragenic recombination (see Methods). For these analyses, we assumed that the recombination rate is exponentially distributed across loci but constant within a locus. This model seems sensible for the short fragments (~650 bp on average) that we considered but may not be appropriate for longer loci.

Chimpanzee species (*P. paniscus and P. troglodytes*) and subspecies (*P. t. verus, P. t. troglodytes*, and *P. t. schweinfurthii*)

Figures 2 and 3 show the marginal posterior distributions for the parameters of the model, averaging the results for two independent runs (for details, see Methods). We considered each pair of populations in turn. Running MIMAR under a model that allows for gene flow strongly suggests that the bonobo and the common chimpanzee populations split without subsequent migration (Table 3). In contrast, there is evidence of gene flow since the split of Western, Central, and Eastern chimpanzee populations (Table 4; see also Won and Hey 2005). Figure 2 shows the posterior distribution estimates for the parameters of the model for bonobo and common chimpanzee populations and Figure 3, for Western, Central, and Eastern chimpanzee populations. We note slight support for gene flow between Eastern chimpanzees and bonobos (see Fig. 2C), whose ranges are closer together than those of bonobos and other chimpanzee subspecies. However, more data and more precise geographic information are needed to evaluate this possibility, especially in light of the relatively

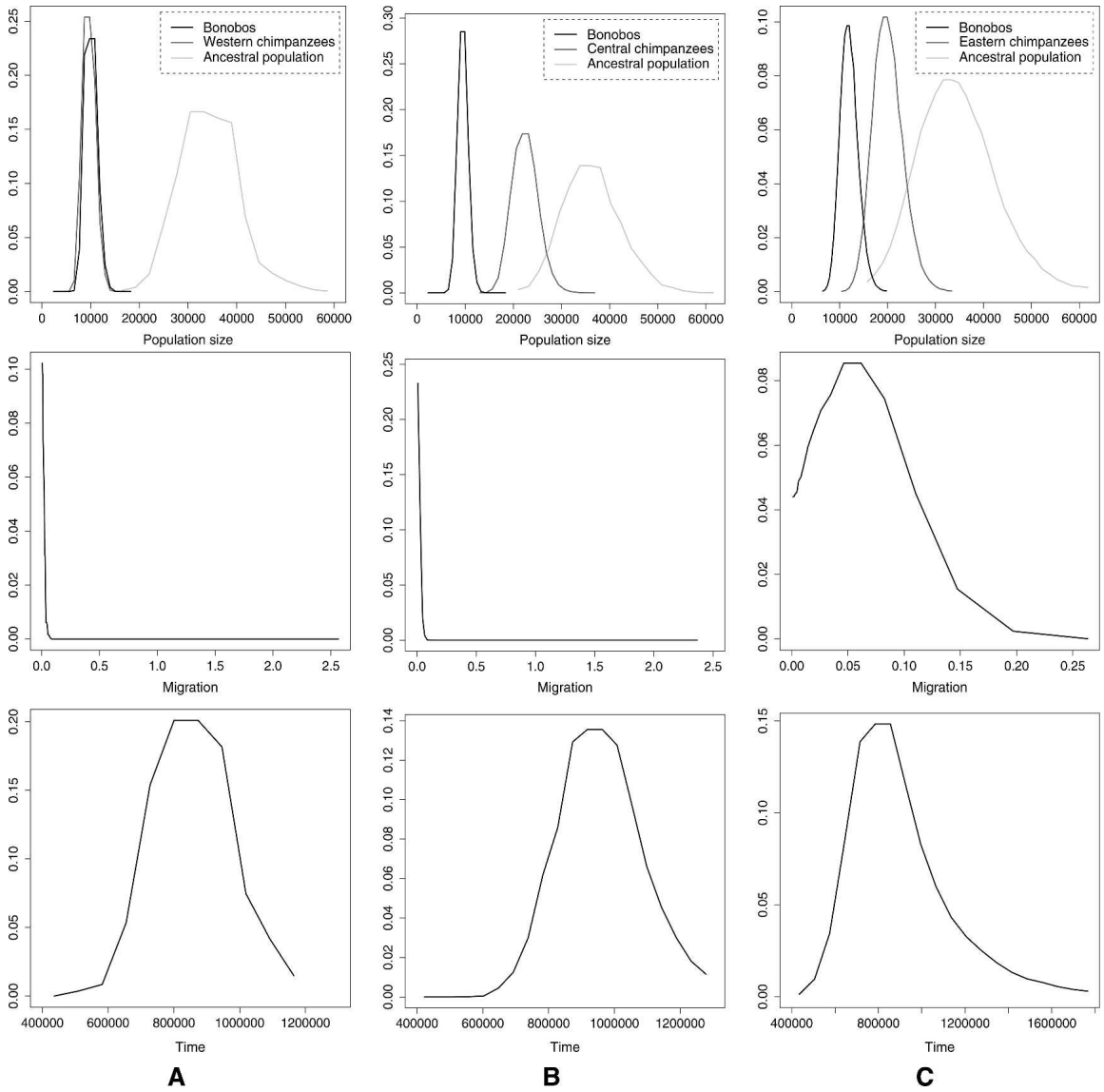

Figure 2. Smoothed marginal posterior distributions estimated by MIMAR from bonobo and common chimpanzee polymorphism data (for details, see Methods). The range of the *X*-axis corresponds to the support of the prior. The distributions are for the analyses of bonobos and Western chimpanzees (*A*), bonobos and Central chimpanzees (*B*), and bonobos and Eastern chimpanzees (*C*).

unreliable estimates of migration from small data sets (see simulation results above).

Assuming 20 yr per generation and a mutation rate of 2×10^{-8} per base pair per generation (see Methods), the estimates of the effective population sizes of bonobos and Western chimpanzees are ~10,000 in all analyses involving these populations. In turn, the estimates of split time for bonobos and common chimpanzee populations range from 790–920 Kya, and the estimates of the ancestral effective population size are ~30,000. These estimates are consistent with those obtained by Won and Hey (2005), who applied IM to a smaller data set, which overlaps with ours. The only exception is that they estimated a smaller ancestral effective population size than we did, but the confidence intervals overlap slightly. These results confirm that polymorphism data from bonobos and common chimpanzees are consistent with an allopatric speciation model and that the divergence between the chimpanzee species occurred more recently than the estimated formation of the River Congo.

In the analyses of Western and Central chimpanzees and Western and Eastern chimpanzees, the time estimates range from 280–440 Kya, the ancestral effective population sizes from 11,000–15,000, and the migration rate, $M = 4N_1m$, from 0.32–0.43 (where N_1 is the effective population size of Western chimpanzees). These results are roughly consistent with those of Won and Hey (2005): Using a model that allowed for asymmetric migration rates, they estimated that $M \sim 0.28$ from Western to Central chimpanzees but did not find evidence for gene flow in the opposite direction.

For the analyses of Central and Eastern chimpanzees, the split time estimate is ~220 Kya, the ancestral effective population size is ~46,000, and the migration rate, $M = 4N_1m$, is ~0.80 (where N_1 is the effective population size of Central chimpanzees). Thus,

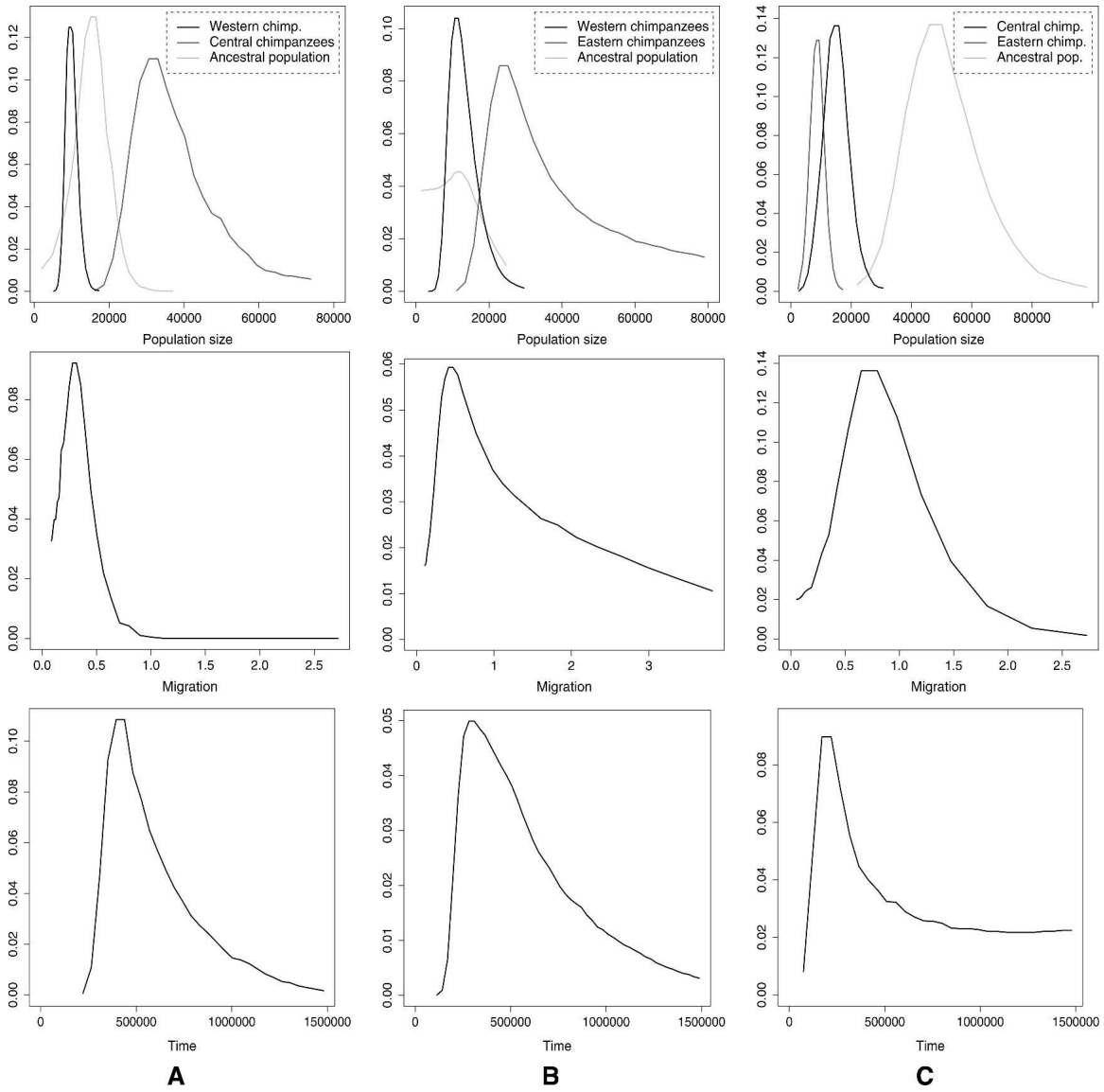

Figure 3. Smoothed marginal posterior distributions estimated by MIMAR from the common chimpanzee subpopulation polymorphism data (for details, see Methods and legend of Fig. 2). The distributions are for the analyses of Western and Central chimpanzees (*A*), Western and Eastern chimpanzees (*B*), and Central and Eastern chimpanzees (*C*).

it appears that the split time for Central and Eastern chimpanzees is about half that of Western and Central (or Eastern) chimpanzees.

While the estimates are generally consistent across pairwise analyses, the effective population size estimates for Central and Eastern chimpanzees are not. In both analyses of Central chimpanzees and bonobos and of Central and Eastern chimpanzees, the effective population size of Central chimpanzees is estimated to be 15,000–22,000 (consistent with the results of Won and Hey 2005). However, a larger population size estimate is obtained from the analysis of Western and Central chimpanzees. Similarly, in both analyses of bonobos and Eastern chimpanzees and of Western and Eastern chimpanzees, estimates of the effective population size of Eastern chimpanzees are 20,000–25,000, while in the analysis of Eastern and Central chimpanzees, the estimate is smaller. These discrepancies may reflect complex histories of chimpanzee populations not captured by the model (see the goodness-of-fit test below). For example, analyses of other data sets suggest that Central chimpanzees may have experienced a recent population expansion (Fischer et al. 2004; D. Reich, pers. comm.).

Gorilla subspecies, Western (*G. gorilla*) and Eastern gorillas (*G. beringei*)

Figure 4A shows the posterior distributions of the five parameters of the parapatric model of speciation. Assuming 15 yr per generation and a mutation rate of 2×10^{-8} per base pair per generation (see Methods), the estimates of the effective population sizes for Western and Eastern gorillas and their ancestral population are ~9000, ~8000, and ~27,000, respectively (see Table 5). The divergence time estimate between Western and Eastern gorilla

Table 3. Results for chimpanzee species

Analysis[a]	Loci[b]	n_1[c]	n_2	N_1[d]	N_2	N_A	T^*[e]	M[f]
Bonobos × Western chimpanzees	69	18 (16)	20 (12)					
Mode				9,790	9,790	33,300	873,000	0.007
2.5th percentile				8,360	7,820	25,200	681,000	0.007
97.5th percentile				12,000	11,700	44,300	1,070,000	0.031
Bonobos × Central chimpanzees	68	18 (16)	20 (10)					
Mode				9,900	21,900	33,800	918,000	0.008
2.5th percentile				7,870	18,300	27,300	759,000	0.007
97.5th percentile				11,300	27,000	46,800	1,170,000	0.036
Bonobos × Eastern chimpanzees	26	18	20					
Mode				11,500	19,900	31,600	785,000	0.062
2.5th percentile				9,150	15,300	22,200	616,000	0.001
97.5th percentile				15,200	25,600	48,700	1,350,000	0.100

[a]Estimates are obtained from two independent runs (see Methods).
[b]Number of loci used in the analyses.
[c]n_1 and n_2 are the number of chromosomes in the first and second population of the analysis, respectively (the sample size varies because we pooled loci from multiple studies and because of missing data).
[d]N_A, N_1, and N_2 are the estimates of the effective population size for the ancestral, first and second population of the analysis, respectively.
[e]T^* is the estimate of the time since the populations split in years.
[f]$M = 4N_1m$, where N_1 is the effective population size of the first population of the analysis.

subspecies is ~92 Kya, and the migration rate, $M = 4N_1m$, is ~0.87 (where N_1 is the effective population size of Western gorillas).

To compare our estimates to those recently obtained by Thalmann et al. (2006) using IM, we considered their mutation rate estimate (1.44×10^{-8} per base pair per generation). Our estimates of the effective population sizes of Western gorillas and ancestral population and the split time are of the same order (~13,000 vs. 17,500, ~37,000 vs. 42,000, and 92 vs. 78 Kya), but our estimate of the effective population size of Eastern gorillas is larger (11,000 vs. 3000). Whether this discrepancy reflects differences in the use of summaries vs. the whole data or in the prior distributions is unclear.

Orangutan subspecies, Sumatran (*P. p. abelii*) and Bornean orangutans (*P. p. pygmaeus*)

The posterior distributions of the five parameters of the parapatric model of speciation are shown in Figure 4B. Assuming 20 yr per generation and a mutation rate of 2×10^{-8} per base pair per generation (see Methods), the estimates of the effective population sizes for Sumatran and Bornean orangutans and their ancestral population are ~17,000, ~10,000, and ~87,000, respectively (see Table 5). The estimate of the symmetrical migration rate, $M = 4N_1m$, is ~0.87 (where N_1 is the effective population size of Sumatran orangutans). The data further suggest that the split time for Sumatran and Borneon orangutan populations is likely to be older than 250 Kya. However, the data (19 loci) do not appear to carry much information about this parameter (see the posterior distribution estimate in Fig. 4B), and in particular, the mode of the posterior distribution, 1.4 Mya, is likely to be an unreliable estimate of the split time.

Since the islands of Borneo and Sumatra were connected during the two last glaciations ~130–200 Kya and ~10–100 Kya ago, it is not surprising to find evidence of gene flow between those two populations. Our results further suggest that the Sumatran and Bornean orangutan populations diverged before the second to last Ice Age. To our knowledge, this analysis provides the first estimates of population parameters for the two orangutan subspecies.

Goodness-of-fit test

To examine whether the isolation-migration model is an appropriate description of the history of the ape species and subspecies, we generated simulated data sets for parameters sampled from the posterior distributions estimated by MIMAR, and compared the simulated data to what is observed for a number of statistics. Encouragingly, the isolation-migration model appears to provide a reasonable fit to the four statistics used in the inferences of MIMAR as well as to the mean F_{ST}, π, and Tajima's D

Table 4. Results for chimpanzee subspecies

Analysis	Loci	n_1	n_2	N_1	N_2	N_A	T^*	M[a]
Western × Central chimpanzees	68	20 (12)	20 (10)					
Mode				9,750	33,000	15,000	439,000	0.315
2.5th percentile				7,690	24,200	6,140	325,000	0.097
97.5th percentile				12,900	59,700	22,400	1,100,000	0.523
Western × Eastern chimpanzees	26	20	20					
Mode				10,800	24,700	11,000	282,000	0.425
2.5th percentile				8,040	18,600	2,270	230,000	0.143
97.5th percentile				21,100	71,800	21,900	1,210,000	2.622
Central × Eastern chimpanzees	26	20	20					
Mode				14,400	8,590	46,000	219,000	0.797
2.5th percentile				8,560	5,070	33,500	143,000	0.084
97.5th percentile				22,300	12,700	75,100	1,400,000	1.389

[a]For details, see legend of Table 3.

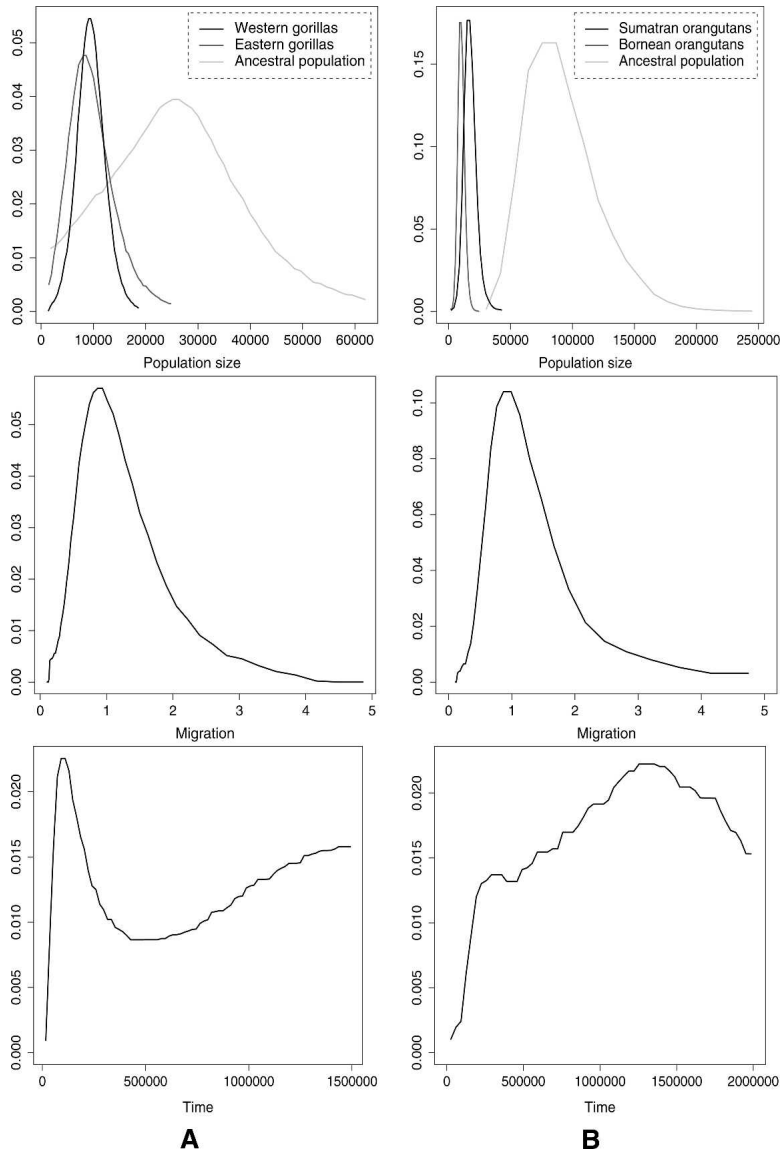

Figure 4. Smoothed marginal posterior distributions estimated by MIMAR from the gorilla (*A*) and orangutan (*B*) subspecies polymorphism data (for details, see Methods and legend of Fig. 2). (*A*) Distributions for the analysis of Western and Eastern gorillas. The apparent multimodality of the marginal posterior distribution estimated for the split time was also noted by Thalmann et al. (2006). (*B*) Distributions for the analysis of Sumatran and Bornean orangutans. Note that the posterior distribution for the split time is rather flat, suggesting that the data do not carry much information about this parameter.

across loci (Supplemental Fig. S6; for details, see Methods). The one exception is for Central and Eastern chimpanzees (Supplemental Fig. S6f): There is a poor fit to F_{ST} and to Tajima's *D* for the Central chimpanzees (see also Supplemental Fig. S6b). This suggests either that an isolation-migration model is not appropriate for these subspecies or that a crucial demographic feature is missing from the model. Given the proximity of Central and Eastern chimpanzees and their low F_{ST}, one possibility is that, rather than a split model, a model of isolation by distance is more appropriate (Fischer et al. 2006). Interestingly, though, there does not appear to be substantial gene flow between the Eastern and Central ranges (see the estimates of the migration rate in this study and Becquet et al. 2007). We also find that, while the model fits most aspects of the bonobo data quite well, the observed Tajima's *D* is lower than expected (Supplemental Fig. S6a–c), perhaps reflecting recent demographic events in bonobos not included in the model.

Discussion

Advantages and limitations of MIMAR

We have developed a new method to estimate parameters of simple allopatric and parapatric speciation models. It considers summaries of the polymorphism data from each locus, rather than the entire data set. Extensive simulations, and comparisons to IM for the case of no recombination, suggest that the use of these summaries provides accurate and precise estimates of parameters of interest from data sets comparable in size to those analyzed to date (see Table 1).

The method presents the important advantage of allowing for intralocus recombination. This feature makes the approach applicable to autosomal data, even in species where the ratio of recombination to mutation events is high ($\rho/\theta \gg 1$), such as in *Drosophila* (Andolfatto and Wall 2003) and *Papilio* (Putnam et al. 2007) and hence where any segment containing polymorphisms is likely to have experienced recombination in its genealogical history. In contrast, when applied to recombining regions, IM requires one to exclude loci that show evidence of recombination and assumes that no recombination occurred at the other loci, potentially biasing the estimates.

In other respects, the model of speciation that we consider is more restrictive than the one used in IM. Mutation rates for each locus are estimated from divergence data and then fixed, rather than coestimated along with other parameters (see Methods). We set the migration rate, *m*, to be symmetric between populations, which may be inappropriate. Finally, we assume that the distribution of coalescent times only varies across loci due to differences in the mode of inheritance and, therefore, that it can be specified a priori. In contrast, IM allows one to estimate inheritance scalars for each locus from the data, which may be important if a subset of loci have experienced recent selection (Hey and Nielsen 2004). Our model could readily be extended to allow for these features, notably for asymmetric migration rates (in fact, the MIMAR program that we make available already allows for this feature). However, the data from a given locus carry limited information, and it is unclear how many parameters can reliably be estimated, even using all the information. Indeed, our simulations suggested that IM and MIMAR estimates of the migration rate from a small data set can be unreliable even in the absence of these complications (see Results).

Table 5. Results for gorilla and orangutan subspecies

Analysis	Loci	n_1	n_2	N_1	N_2	N_A	T^*	M^a
Western × Eastern gorillas	15	30	6 (2)					
Mode				9,130	8,140	26,400	91,500	0.867
2.5th percentile				5,090	3,570	5,990	84,300	0.282
97.5th percentile				14,100	18,100	49,100	1,440,000	2.059
Sumatran × Bornean orangutans	19	12	20 (18)					
Mode				17,200	10,200	86,900	1,390,000	0.868
2.5th percentile				10,200	6,230	52,400	254,000	0.361
97.5th percentile				26,600	15,000	143,000	1,900,000	2.235

[a]For details, see legend of Table 3.

In its current implementation, our method is also limited in the type of data that it can consider, as it is not applicable to surveys of variation that suffer from ascertainment bias. Moreover, it assumes an infinite site model, so only two alleles can be present at a given site. As long as the ascertainment bias and mutation model are known, however, it should be reasonably straight-forward to extend the model to consider these cases (Nielsen and Signorovitch 2003). MIMAR is further intended for use on resequencing data from short, independently-evolving loci, in which there is little information about how genealogies change along the genome or, viewed another way, about linkage disequilibrium (McVean et al. 2004), and for which it is reasonable to assume that recombination rates are uniform. Applying MIMAR to longer stretches of sequence may require a change in the model of recombination to capture fine-scale heterogeneity in recombination rates. In that setting, it may also be helpful to consider summaries of linkage disequilibrium in addition to the four statistics used here. More generally, our approach could be extended to consider a number of other aspects of the data. For instance, one could consider the number of singletons in each population (in addition to the four current statistics) or the joint frequency spectrum in two population samples.

In addition to improving the inference method, it will also be important to consider more realistic models of speciation. For example, detailed studies of closely related species reveal that many apparent cases of parapatry may in fact reflect allopatric speciation followed by secondary contact (Coyne and Orr 2004b; Llopart et al. 2005). One approach to distinguishing between the two scenarios might be to allow migration between diverging populations to stop at different time points, and estimate which times are most likely given the polymorphism data. Similarly, for sets of species (or populations) that split over a short time period, it may be important to consider more than two species at a time (Wall 2000; Degnan and Rosenberg 2006; Pollard et al. 2006).

Another salient feature, ignored in existing methods, may be population structure in the ancestral population. Indeed, in many of our analyses of ape data, as well as in most analyses of the isolation-migration model published to date (e.g., Hey et al. 2004; Hey 2005; Won and Hey 2005; Thalmann et al. 2006), the estimate of the ancestral effective population size is larger than that of the descendant populations. Since it seems unlikely that so many populations have shrunk over time, this suggests that a salient and fairly common demographic feature is being ignored. One possibility is that the assumption of a panmictic ancestral population is inappropriate. If so, it may be relevant to consider a model of population structure in which a geographic barrier becomes stronger over time (e.g., Innan and Watanabe 2006). In this respect, an attractive feature of our method is that it is easy to generalize to other demographic settings (see Methods).

Finally, our approach could also be extended to scan the genome for regions that contribute to reproductive isolation (Won et al. 2005; Bull et al. 2006; Geraldes et al. 2006; Miller et al. 2006). Indeed, models of parapatric speciation predict that loci involved in the formation of species will experience no or little gene flow since the split and therefore have more fixed differences and fewer shared alleles than do background loci. Moreover, theoretical results suggest that, in this setting and unless selection in very strong, regions of marked differentiation should be relatively short (Barton and Bengtsson 1986). Thus, identifying regions with evidence for decreased gene flow should be an effective way to find the specific loci that contribute to reproductive isolation. This idea has been implemented by estimating gene flow for each locus separately (Won et al. 2005). However, this approach may have limited power to detect loci with reduced gene flow. An alternative may be to use the goodness-of-fit test results for individual loci to identify outliers that behave as expected if they contributed to reproductive isolation.

Analyses of ape polymorphism data

Analyses of genetic polymorphism data from apes can help to characterize the geographic distribution of variation (e.g., Becquet et al. 2007), shed light on their demographic history, and place the evolutionary history of humans in context (Stone and Verrelli 2006). Here, we considered the largest set of polymorphism data to date for all three species of nonhuman great apes, and estimated parameters of a simple isolation-migration model. Using a goodness-of-fit test, we find that this model provides a reasonable point of departure for analyzing ape data, other than for Eastern and Central common chimpanzees.

The use of the model suggests that the effective populations sizes of the ape populations range from 8000–33,000, on the same order as estimates for human populations (10,000–15,000; Frisse et al. 2001; Voight et al. 2005). In contrast, the subspecies split times appear to be older than those of human populations (Cavalli-Sforza and Feldman 2003; Goebel 2007), ranging from 92–440 Kya.

We find no evidence for gene flow since the split for chimpanzee species (with the possible exception of Eastern chimpanzees and bonobo), consistent with the results of Won and Hey (2005), but do detect limited migration ($\hat{M} \leq 1$) for all ape subspecies. The split time estimate for chimpanzee species is 790–920 Kya, suggesting that speciation occurred after the formation of the River Congo, 1.5–3.5 Mya. These estimates do not take into account possible error in the mutation rate per year. But even if we consider a time to the most recent common ancestor between human and chimpanzee at the upper limit of what has been estimated so far, 8 Mya, and a generation time of only 15 yr, the

central 95th percentile for the split time is 1–2.3 Mya. Moreover, the recent finding of a chimpanzee fossil in Kenya indicates that common chimpanzees may have occupied a much wider range than inferred on the basis of their current distribution (McBrearty and Jablonski 2005). Thus, existing data support a more recent speciation time for common chimpanzees and bonobos, which may have occurred outside of their current habitats.

More generally, this application illustrates how the increasing availability of multilocus polymorphism data sets, together with development of novel statistical approaches, can yield insights into speciation, both in apes and in other organisms.

Methods

Model

We consider a neutral model in which an ancestral population suddenly splits into two populations, which either diverge in isolation or continue to exchange migrants (Fig. 1). We further assume that n_1 and n_2 chromosomes have been sampled from two populations and fully resequenced at Y randomly chosen, independently-evolving loci.

The population model, often called "isolation-migration", is described by the population split time in generations, T, and three population mutation rates, $\theta_1 = 4N_1\mu$, $\theta_2 = 4N_2\mu$, and $\theta_A = 4N_A\mu$ (Fig. 1). Throughout, the subscripts 1, 2, and A refer to parameters that describe populations 1 and 2 and the ancestral population, respectively. Following IM, we assume that there is an independent estimate of the average mutation rate across loci, μ, which can be used to estimate the effective population sizes from the population mutation rates (e.g., as $N_1 = \theta_1/4\mu$). In addition, there is a symmetric migration rate, m, which corresponds to the fraction of a population that is replaced by migrants from the other population each generation.

The parameters θ_1, θ_2, and θ_A are defined per base pair and are chosen from uniform distributions; the time in generations, T, is also chosen from a uniform distribution. The prior for the migration rate is on the expected *number* of individuals in population 1 replaced by migrants (backward in time), $M = 4N_1m$, where N_1 is obtained from θ_1 by dividing by $4\hat{\mu}$ ($\hat{\mu}$ is the estimate of μ). Specifically, $\ln(M)$ is chosen from a uniform distribution.

In addition to the five demographic parameters, there are a number of locus-specific parameters. We assume that each locus follows the infinite sites mutation model (Kimura 1969), then define an inheritance scalar, u, which, for example, is equal to 1 for autosomal, ¾ for X-linked, and ¼ for Y-linked and mtDNA-linked loci. To allow for mutation rate variation among loci with the same mode of inheritance, we introduce an additional scalar, v, for each locus. Given this parameterization, the locus-specific mutation rate in population 1 is given by $uvZ\theta_1$, where Z is the length of the locus in base pairs; the locus-specific population mutation rates for other populations are defined analogously.

The population recombination rate per base pair is defined as $\rho = 4N_1c$, where c is the per base pair per generation recombination rate. We ignore gene conversion, treating all recombination as crossovers alone. We also define an inheritance scalar for recombination, w ($w = 0$ for the mtDNA and Y, 2/3 for X, and 1 for autosomes). We then consider three options to specify the locus-specific population recombination rate. We either fix ρ across loci, such that the population recombination rate at a locus is $wZ\rho$. Alternatively, if an estimate, $\hat{\rho}$, of the population recombination rate is available for each locus, we set the scalar w to the inheritance scalar for recombination multiplied by $\hat{\rho}$ to incorporate this knowledge in the estimation. The final option is to allow rates to vary for each locus, in which case the locus-specific population recombination rate is $r \cdot wZ\theta_1$, and we draw the ratio $r = \rho/\theta_1$ from an exponential distribution with mean λ for each locus. Thus, we allow for rate variation among loci but assume a constant rate within a locus. This model should be a sensible description of the rate variation if the loci are short (e.g., 1 kb), as in most data sets collected to date. The set of locus-specific population recombination rates, $(\rho_1, \cdots, \rho_Y)$, is referred to as $\mathbf{P}$.

Data summaries

Our goal is to estimate the parameters of the isolation-migration model illustrated in Figure 1. We do so by estimating the posterior distribution $\pi(\mathbf{\Theta}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{\Theta})p(\mathbf{\Theta})$, where $\mathbf{\Theta} = (\theta_1,\theta_2,\theta_A,T,M,\mathbf{P})$, $\mathbf{D}$ is the data, and $p(\mathbf{\Theta})$ denotes the prior distribution. Unfortunately, when $\mathbf{D}$ is the entire polymorphism data set under our model, estimating the likelihood of the data given the parameters, $p(\mathbf{D}|\mathbf{\Theta})$, is computationally extremely intensive and becomes prohibitive when recombination is included (Nielsen and Wakeley 2001; Hey and Nielsen 2004). In their program IM, Hey and Nielsen (2004) address this problem by considering the full data set and using a MCMC approach but restricting themselves to a model with no intralocus recombination (i.e., $\mathbf{P} = \mathbf{0}$). Instead, we focus on a model with intralocus recombination but summarize the polymorphism data from each locus with the summary statistics described below. To do so, we initially explored an importance sampling approach, which provided reliable estimates but was inefficient. We then implemented an MCMC algorithm, which is more efficient than our initial algorithm when the prior and posterior distributions differ substantially.

To summarize the data, we use the statistics introduced by Wakeley and Hey (1997) for this type of inference problem: For each locus, we consider the number of polymorphisms unique to the samples from populations 1 and 2 (S_1 and S_2, respectively), the number of shared alleles between the two samples (S_3), and the number of fixed alleles in either sample (S_4). Previous work has shown that these statistics contain considerable information about the demographic parameters of the isolation-migration model (e.g., Clark 1997; Wakeley and Hey 1997; Hudson and Coyne 2002; Leman et al. 2005). In what follows, we refer to the vector of summaries, S_k, $k \in [1,4]$, for locus y as $\mathbf{D_y}$. In turn, we refer to the set of statistics for the Y loci as $\mathbf{D} = (\mathbf{D}_1, \cdots, \mathbf{D}_Y)$.

In calculating these statistics, we assume that an outgroup sequence is available and can be used to determine which allele is derived without error. We note that, in practice, it may be advisable to use two outgroup sequences to minimize error in inferring the ancestral state. We assign each polymorphic site to one of the statistics depending of the frequency of the derived allele in the population i, f_i. Specifically, if $0 < f_i \leq 1$ in each population sample, the allele is shared, if $f_i = 0$, $f_j = 1$, $i \neq j$, the allele is fixed in the sample j, and if $f_i = 0$ and $f_j < 1$, $i \neq j$, the allele is specific to sample j. The statistics are easy to calculate and do not require determination of haplotypes.

Estimation method

Our goal is to sample from the posterior distribution, $\pi(\mathbf{\Theta}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{\Theta})p(\mathbf{\Theta})$, which is the likelihood of the data summaries given the parameters times the prior distributions of the parameters. The parameters are initially chosen from these prior distributions and subsequently updated using MCMC, which requires information about the likelihood of the data given the parameters. Very briefly, our strategy is to estimate the likelihood of the data summaries at all the loci for a chosen set of parameters by, for each of the Y loci, (1) generating a set of X ancestral recom-

bination graphs (ARGs) (Hudson 1983) given the parameters and (2) calculating the probability of the data summaries given the set of ARGs. Specifically, we estimate the likelihood $p(\mathbf{D}|\Theta)$ as

$$\prod_{y=1}^{Y}\frac{1}{X}\sum_{x=1}^{X}p(\mathbf{D}_y|G_{yx},\Theta)p(G_{yx}|\Theta), \quad (1)$$

where G_{yx} is the x^{th} ARG at locus y (Hudson 1983). In other words, we estimate $p(\mathbf{D}|\Theta)$ by taking the average of $p(\mathbf{D}_y|G_{yx},\Theta)$ over X ARGs, then taking the product over loci (since they are assumed to be independent). The term $p(G_{yx}|\Theta)$ is given by the coalescent, using a modified version of the program ms (Hudson 2002).

We can readily calculate $p(\mathbf{D}_y|G_{yx},\Theta)$. Given a coalescent genealogy, G_{yx}, we compute the sum of the lengths of all the branches (in coalescent units), which would lead to unique polymorphisms in sample 1 and 2 (L_1 and L_2, respectively), alleles shared by both samples (L_3), and alleles fixed in either samples (L_4). Given the infinite site mutation model, the numbers of mutations, S_k, randomly placed along the branches of type $k \in [1,4]$, is Poisson distributed with mean $L_k u v Z \theta_1$. Conditional on a genealogy, the probabilities of observing S_1, S_2, etc. . . are independent, so the probability of the data $\mathbf{D}_y$ for the locus y is given by

$$p(\mathbf{D}_y|G_{yx},\Theta)=\prod_{k=1}^{4}Po(S_k|L_k u v Z\theta_1). \quad (2)$$

Equation 2 also applies to a recombining locus, but in this case, G_{yx} is an ARG and L_k is computed as follows: With recombination, a locus of size Z has R segments of length Z_j, $j \in [1, R]$, with different genealogical histories. The genealogy of a segment has branch length L_{jk}, such that $L_k = \sum_{j=1}^{R} L_{jk} Z_j / Z$ for the ARG.

Our prior distributions for the parameters, $p(\Theta)$, are uniform over a bounded support (except for $\mathbf{P}$ and a uniform prior on $\ln(M)$). For the MCMC, we use random walk Metropolis transition kernels to propose parameter values, so that the proposed value of a parameter is taken from a normal distribution with mean its previous value and variance defined to maximize the acceptance rate (after exploratory simulations) (Gilks et al. 1996). If a parameter value lies outside the support of the prior, the proposed set of parameters is rejected. In turn, $\mathbf{P}$ is a nuisance parameter and its values are either fixed (when ρ is fixed), or drawn from the distribution described above (see Model); in the MCMC, the values of $\mathbf{P}$ are sampled independently at each step from the prior.

Our approach follows a number of Bayesian methods based on summaries of the data, developed in other contexts (e.g., Tavaré et al. 1997; Pritchard et al. 1999; Beaumont et al. 2002; Przeworski 2003). It differs in that we update the parameters using MCMC rather than sampling them independently from the prior. This general approach was described by Beaumont (2003). As pointed out to us by Matthew Stephens (pers. comm.), our approach can also be viewed as a MCMC on the set of all genealogies, $\mathbf{G} = (G_{11},\ldots,G_{1X},\ldots,G_{Y1},\ldots,G_{YX})$, and the parameters. In this case, the X ARGs are independent samples from the coalescent prior across the Y independent loci. Thus, for the MCMC, the set of ARGs is updated using the transition kernel $q(\mathbf{G} \rightarrow \mathbf{G}') = p(\mathbf{G}'|\Theta)$, while the parameters of interest are updated using Metropolis transition kernels. We sample sets $(\mathbf{G}, \Theta)$ from the following target distribution:

$$\pi(\mathbf{G}, \Theta) \propto \prod_{y=1}^{Y}\left(\frac{1}{X}\sum_{x=1}^{X}p(\mathbf{D}_y|G_{yx},\Theta)\right)p(\Theta)p(\mathbf{G}|\Theta). \quad (3)$$

The marginal distribution of sampled values of Θ from the target distribution is $\pi(\Theta|\mathbf{D})$ (as shown in the Supplemental Materials; see also Beaumont 2003: appendix). A nice feature of viewing our approach in this way is that it demonstrates that the stationary distribution of the Markov chain is the correct distribution, i.e., that we are exploring the true posterior distribution rather than an approximation.

MIMAR—MCMC estimation of the isolation-migration model allowing for recombination

To sample from the target distribution $\pi(\mathbf{G}, \Theta)$, we use an MCMC approach (MIMAR). In the initial step, Θ is chosen from the prior, $p(\Theta)$, and $\mathbf{G}$ is sampled from the coalescent with those parameters. Subsequent sets $(\mathbf{G}, \Theta)$ are updated following a Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970). More specifically, we proceed as follows:

I1. If now at $(\mathbf{G}, \Theta)$, propose a move to $(\mathbf{G}', \Theta')$ according to the transition kernels $q(\Theta \rightarrow \Theta')$ and $q(\mathbf{G} \rightarrow \mathbf{G}')$ (i.e., Generate X ARGs given the parameters Θ' for each of the Y loci).
I2. For the y^{th} locus:
   a. Calculate $p(\mathbf{D}_y|G'_{yx}, \Theta')$ for each of the X ARGs using Equation 2.
   b. If the average of $p(\mathbf{D}_y|G'_{yx}, \Theta')$ over the X ARGs is 0, record $(\mathbf{G}, \Theta)$ and go to *I1*; else go to *I2a* for the locus $y + 1$.
I3. Calculate

$$h = \min\left(1, \frac{A'}{A}\right), \quad (4)$$

where

$$A' = \prod_{y=1}^{Y}\left(\frac{1}{X}\sum_{x=1}^{X}p(\mathbf{D}|G'_{yx},\Theta')\right)p(\Theta')p(\mathbf{G}'|\Theta')q(\Theta' \rightarrow \Theta)q(\mathbf{G}' \rightarrow \mathbf{G})$$

$$A = \prod_{y=1}^{Y}\left(\frac{1}{X}\sum_{x=1}^{X}p(\mathbf{D}|G_{yx},\Theta')\right)p(\Theta)p(\mathbf{G}|\Theta)q(\Theta \rightarrow \Theta')q(\mathbf{G} \rightarrow \mathbf{G}')$$

I4. Move to $(\mathbf{G}', \Theta')$ with probability h [i.e., record $(\mathbf{G}', \Theta')$] or else record $(\mathbf{G}, \Theta)$. Return to *I1*.

The choice of proposal distribution for $\mathbf{G}$ and $\mathbf{P}$ and normal kernel distributions and uniform prior distributions for the parameters of interest lead to the following simplification of Equation 4:

$$h = \min\left(1, \frac{\prod_{y=1}^{Y}\left(\frac{1}{X}\sum_{x=1}^{X}p(\mathbf{D}|G'_{yx},\Theta')\right)}{\prod_{y=1}^{Y}\left(\frac{1}{X}\sum_{x=1}^{X}p(\mathbf{D}|G_{yx},\Theta)\right)}\right) \quad (5)$$

In practice, we consider $X = 100$ (or $X = 50$ or 5, see below), thus generating 100 (50 or 5) ARGs given the locus-specific parameters. Generating so many ARGs is computationally demanding, but we find that this approach has improved mixing over $X = 1$.

We note that our approach presents the advantage of being flexible, since it can easily be extended to consider any summaries for which $p(\mathbf{D}_y|G_{yx}, \Theta)$ can be readily calculated, such as the allele frequency spectrum at each locus.

MIMAR and its documentation are available at http://mplab.bsd.uchicago.edu/dataNprograms.htm.

Assessing the performance of the estimator

To assess the performance of our method, we ran MIMAR on simulated data sets with two independent seeds (see below). We considered that MIMAR reached convergence when the posterior distributions from the two independent runs were highly similar (e.g., Supplemental Fig. S3). In the documentation provided with MIMAR, we describe a number of other criteria that can be used to assess convergence and proper mixing. We took the mode and the central 95th percentile of the marginal posterior distribution averaged over the two independent runs as the point estimate and measure of uncertainty, respectively.

Simulated data and performance analyses

We generated simulated data sets under the isolation-migration model using a modified version of the program ms (Hudson 2002). Unless otherwise indicated, we considered 20 loci of 1 kb each, and sampled 20 chromosomes from each of two populations.

Performance of MIMAR under allopatry

We generated 30 simulated data sets with no recombination and fixed parameter values relevant for *Drosophila yakuba* and *D. santomea* (Llopart et al. 2005), assuming a per base pair per generation mutation rate of $\mu = 2 \times 10^{-9}$ and 20 generations per year (Andolfatto and Przeworski 2000). We analyzed the 30 simulated data sets for 60 h with 1×10^5 burnin steps and prior distributions as indicated (see Supplemental Fig. S1).

Comparison to IM under allopatry

In order to compare our estimates with those generated by IM (Hey and Nielsen 2004), which does not allow for intralocus recombination, we set the population recombination rate, ρ, to 0. To be comparable to IM, we also chose uniform prior distributions with 0 as the lower limit. We generated 30 simulated data sets with parameters relevant for *Drosophila* species as above, setting M to 0 and drawing the other parameters from prior distributions: θ_1 and θ_2 from U (0, 0.01) and θ_A from U (0, 0.02) per base pair and T from $U(0, 1.5 \times 10^7)$ generations. We analyzed those 30 simulated data sets with MIMAR and IM using the same prior distributions as used when simulating the data sets, 4×10^6 recorded steps and 5×10^5 burnin steps.

Assessing the evidence for gene flow

We generated 40 data sets, consisting of 40 recombining loci with parameter values relevant for apes (see below). We assumed that $\mu = 2 \times 10^{-8}$ per base pair per generation to translate coalescent time units into generations (Nachman and Crowell 2000). This mutation rate estimate is also obtained assuming a most recent common ancestor of human and chimpanzee of 7 Mya and an average nucleotide divergence of 1.28% (The Chimpanzee Sequencing and Analysis Consortium 2005). The intralocus recombination rate was set for each locus by choosing $r = c/\mu$ from the prior exp(1/0.6) (assuming that the mean c is 1.2×10^{-8}) (Kong et al. 2002). The other parameter values were sampled from the following prior distributions: θ_1, θ_2, and θ_A from U (0.0006, 0.006) per base pair and T from $U(0, 1 \times 10^5)$ generations. M was either fixed to 0 (for 20 data sets simulated under the allopatric model) or to 1 (for 20 data sets simulated with parapatric divergence). We analyzed the 40 simulated data sets with MIMAR, choosing $\ln(M)$ from U (–5, 2) and the other parameters from the same prior distributions as used when simulating the data sets, the number of ARGs per locus set to $X = 50$, 4×10^6 recorded steps, and 5×10^5 burnin steps.

Effect of uncertainty in the intralocus recombination rates

We generated 16 simulated data sets, consisting of 10 recombining loci with parameter values relevant for *Drosophila* species. The intralocus recombination rate was set for each locus by choosing $r = c/\mu$ from the prior exp(1/10) (assuming that the mean c is 2×10^{-8}; Andolfatto and Przeworski 2000). M was fixed to 0 and the other parameter values were sampled from the following prior distributions: θ_1, θ_2, and θ_A from U (0.001, 0.01) per base pair and T from $U(0, 1 \times 10^6)$ generation. We then analyzed the data sets with MIMAR in three ways: (1) the locus-specific population recombination rates were fixed to their true values, (2) the locus-specific population recombination rates were sampled from the same prior as used when generating the simulated data, and (3) the locus-specific population recombination rates were set to 0. For the three sets of analysis, we fixed θ_1, θ_2, and θ_A to their true values and used the same prior distribution for T as when generating the simulated data. MIMAR was run with $X = 5$ (cases 1 and 2) or $X = 100$ (case 3), 4.5×10^5 recorded steps, and 5×10^4 burnin steps.

Analysis of ape polymorphism data

Polymorphism data

We analyzed the ape polymorphism data reported in Fischer et al. (2006), Thalmann et al. (2006), and Yu et al. (2003). The first set was kindly provided by A. Fischer (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany); we downloaded the two other data sets from GenBank (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=Nucleotide). The data from Fischer et al. (2006) (and Thalmann et al. 2006) and Yu et al. (2003), consisted of loci of median length ~780 bp and ~470 bp, respectively. The data sets were as follows (see Tables 3–5): 69 loci surveyed in nine unrelated bonobos (pigmy chimpanzee, *P. paniscus*), 26 loci in 10 and 43 loci in six Western chimpanzees (*P. t. verus*), 26 loci in 10 and 42 in five Central chimpanzees (*P. t. troglodytes*), 26 loci in 10 Eastern chimpanzees (*P. t. schweinfurthii*), 15 loci in 15 Western gorillas (*G. gorilla*) and three Eastern gorillas (*G. beringei*), and 19 loci in six Sumatran orangutans (*P. p. abelii*) and 10 Bornean orangutans (*P. p. pygmaeus*).

For each locus, we obtained two outgroup sequences. For the chimpanzee data sets, one orangutan sequence and one human sequence were available for 26 and 19 loci, respectively (Fischer et al. 2006); one human sequence (Yu et al. 2002) and one gorilla sequence (*G. g. gorilla*; Yu et al. 2004) were obtained for 43 loci. We blasted the seven remaining loci and downloaded a homologous human sequence for each of them (BLASTN, http://www.ncbi.nlm.nih.gov/BLAST) (Altschul et al. 1990). For the gorilla data set, one orangutan sequence and one human sequence were available for all loci; for the orangutan data set, one chimpanzee sequence and one human sequence were available for all loci (Fischer et al. 2006). We used CLUSTALW in MEGA3.1 (Thompson et al. 1994; Kumar et al. 2001) to align the resequencing data and the two outgroup sequences. We then wrote a Perl script that requires both outgroup sequences to be consistent to infer the ancestral state at each site, thus minimizing error in the reconstruction of the ancestral state. We ignored sites with gaps, missing data, and more than two variants. (There were only one site with three alleles in the entire gorilla data set, three in the orangutan data, and six in the chimpanzee/bonobo data.) We used a Perl script to calculate for each locus the four statistics S_1, S_2, S_3, and S_4 (see above) and F_{ST} (Hudson et al. 1992) for pairs of populations, as well as the mean pairwise differences, π (Nei and Li 1979) and Tajima's D (Tajima 1989) in each population.

Estimates of mutation rate variation

To allow for variation in mutation rates, we used the scalars v described above. To do so, we calculated the mean pairwise divergence per site between a human sequence and an ape sequence (*div*), using a Perl script. We obtained the expected locus divergence given the number of base pairs, $E(div) \cdot Z$, where $E(div)$ is the mean divergence per base pair over the loci, and performed a goodness-of-fit test using Pearson's χ^2 (Frisse et al. 2001). The gorilla and orangutan data did not deviate significantly from expectation (P-value = 0.24 and 0.85, respectively); therefore, we set $v = 1$ for all loci in the analysis of these two data sets. However, data from the three common chimpanzee populations and the bonobo rejected the null hypothesis of a homogeneous mutation rate across loci (at the 5% level). Thus, for a pair of chimpanzee populations or species, we set v at a locus to the observed divergences per base pair divided by $E(div)$.

Recombination rates

Won and Hey (2005) found evidence of recombination in bonobos, Central and Western chimpanzees in 10 of the 43 short segments surveyed by Yu et al. (2003) used in this study. We estimated the locus-specific recombination rate in the data sets using MAXDIP (http://genapps.uchicago.edu/maxdip/index.html; Hudson 2001), setting 0.005 as the initial value and the gene conversion rate to 0. From each species, we chose the subspecies with the largest estimate of the mean recombination rate across loci, which were Central chimpanzees, Western gorillas, and Sumatran orangutans. Then, to assess whether the point estimates were significantly greater than 0, we simulated 1000 data sets using ms (Hudson 2002), setting the number of segregating sites to what was observed and ρ to 0. We ran MAXDIP on the simulated data sets and calculated how many times $\hat{\rho}$ (i.e., estimated by MAXDIP) was equal or larger than observed under the standard neutral model. By this approach, we rejected $\hat{\rho} = 0$ for four out of 35 loci in Central chimpanzees, one out of seven loci in Western gorillas, and three out of 15 loci in Sumatran orangutans (at the 5% level). Given the small sample sizes, our power to detect recombination was limited. Nonetheless, our results suggest that ignoring recombination will result in a loss of data—even in species in which ρ/θ is relatively small. In the analyses of the ape data, we chose $r = \rho/\theta_1$ for each locus from exp(1/0.6) (see above). We chose this distribution because it has been shown to be a good description of fine-scale recombination rate variation in humans and may also apply to a number of other organisms, notably to other apes (Coop and Przeworski 2007).

Analyses

We ran MIMAR for 2×10^7 recording steps with 1×10^6 burnin steps, $X = 50$, recording the parameters every 50 steps and using prior distributions chosen after preliminary analyses. We repeated our analyses for two independent seeds and considered that convergence was reached when the posterior distributions of both runs were very similar (data not shown). Results reported are for the average from the two independent runs. We obtained estimates of the effective population sizes and split times in years for all the ape species and subpopulations using $\mu = 2 \times 10^{-8}$ per base pair per generation and assuming 20 yr per generation for chimpanzees and orangutans (Gage 1998; Fischer et al. 2004) and 15 yr per generation for gorillas (Thalmann et al. 2006).

Goodness–of–fit test

We investigated how well the data fit the estimated model by generating the posterior predictive distributions of the four statistics S_1, S_2, S_3, and S_4 summed over all loci, the mean F_{ST} (Hudson et al. 1992), and, in each population, the mean pairwise differences, π (Nei and Li 1979) and the mean Tajima's D (Tajima 1989) across loci. To do so, we simulated data sets under the isolation-migration model, sampling the parameters from the posterior distribution estimated by MIMAR. We then compared the observed values of the statistics to the simulated distribution (see Supplemental Fig. S6), conservatively considering the model to be a poor fit if the observed value of a data summary fell in the 2.5th percentile tails of *any* statistic. We note that, since this goodness-of-fit test takes into account the uncertainties associated with the estimates, it is similar to the Bayesian posterior predictive P-value (e.g., Meng 1994).

Acknowledgments

We thank G. Coop, R. Hudson, J. Novembre, J. Pritchard, D. Reich, M. Stephens, K. Teshima, and K. Zeng, as well as three reviewers for helpful discussions and/or comments on earlier versions of the manuscript. This work was supported by an Alfred P. Sloan fellowship in Computational Molecular Biology to M.P. C.B. also acknowledges support from the Summer Institute in Statistical Genetics (2006).

References

Altschul, S., Gish, W., Miller, W., Meyers, E., and Lipman, D. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Andolfatto, P. and Przeworski, M. 2000. A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics* **156:** 257–268.

Andolfatto, P. and Wall, J. 2003. Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. *Genetics* **165:** 1289–1305.

Barbash, D.A., Siino, D.F., Tarone, A.M., and Roote, J. 2003. A rapidly evolving MYB-related protein causes species isolation in *Drosophila*. *Proc. Natl. Acad. Sci.* **100:** 5302–5307.

Barton, N. and Bengtsson, B. 1986. The barrier to genetic exchange between hybridising populations. *Heredity* **57:** 357–376.

Beadle, L.C. 1981. *The inland waters of tropical Africa: An introduction to tropical limnology*, 2nd ed. Longman Group, London.

Beaumont, M. 2003. Estimation of population growth or decline in genetically monitored populations. *Genetics* **164:** 1139–1160.

Beaumont, M.A., Zhang, W., and Balding, D.J. 2002. Approximate Bayesian computation in population genetics. *Genetics* **162:** 2025–2035.

Becquet, C., Patterson, N., Stone, A., Przeworski, M., and Reich, D. 2007. Genomic analysis of chimpanzee population structure. *PLoS Genet.* **3:** e66. doi: 10.1371/journal.pgen.0030066.

Bull, V., Beltran, M., Jiggins, C., McMillan, W., Bermingham, E., and Mallet, J. 2006. Polyphyly and gene flow between non-sibling *Heliconius* species. *BMC Biol.* **4:** 11. doi: 10.1186/1741-7007-4-11.

Cavalli-Sforza, L. and Feldman, M. 2003. The application of molecular genetic approaches to the study of human evolution. *Nat. Genet.* **33:** 266–275.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Clark, A. 1997. Neutral behavior of shared polymorphism. *Proc. Natl. Acad. Sci.* **94:** 7730–7734.

Coop, G. and Przeworski, M. 2007. An evolutionary view of human recombination. *Nat. Rev. Genet.* **8:** 23–24.

Coyne, J.A. and Orr, H.A. 2004a. *Speciation*. Sinauer Associates, Sunderland, MA.

Coyne, J.A. and Orr, H.A. 2004b. Allopatric and parapatric speciation. In *Speciation*, pp. 83–123. Sinauer Associates, Sunderland, MA.

Degnan, J. and Rosenberg, N. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.* **2:** e68. doi: 10.1371/journal.pgen.0020068.

Fischer, A., Wiebe, V., Pääbo, S., and Przeworski, M. 2004. Evidence for a complex demographic history of chimpanzees. *Mol. Biol. Evol.* **21:** 799–808.

Fischer, A., Pollack, J., Thalmann, O., Nickel, B., and Pääbo, S. 2006. Demographic history and genetic differentiation in apes. *Curr. Biol.* **16:** 1133–1138.

Fossella, J., Samant, S.A., Silver, L.M., King, S.M., Vaughan, K.T., Olds-Clarke, P., Johnson, K.A., Mikami, A., Vallee, R.B., Pilder, S.H., et al. 2000. An axonemal dynein at the hybrid sterility 6 locus: Implications for t haplotype-specific male sterility and the evolution of species barriers. *Mamm. Genome* **11:** 8–15.
Frisse, L., Hudson, R., Bartoszewicz, A., Wall, J., Donfack, J., and Di Rienzo, A. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* **69:** 831–843.
Gage, T. 1998. The comparative demography of primates: With some comments on the evolution of life histories. *Annu. Rev. Anthropol.* **27:** 197–221.
Geraldes, A., Ferrand, N., and Nachman, M. 2006. Contrasting patterns of introgression at X-linked loci across the hybrid zone between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics* **173:** 919–933.
Gilks, W., Richardson, S., and Spiegelhalter, D. 1996. Implementation. In *Markov Chain Monte Carlo in practice*, pp. 8–19. Chapman and Hall/CRC, Boca Raton, FL.
Goebel, T. 2007. Anthropology: The missing years for modern humans. *Science* **315:** 194–196.
Groves, C. 1970. Population systematics of the gorilla. *J. Zool.* **161:** 287–300.
Groves, C. 1971. *Pongo pygmaeus*. *Mamm. Species* **4:** 1–6.
Grubb, P., Butynski, T.M., Oates, J.F., Bearder, S.K., Disotell, T.R., Groves, C.P., and Struhsaker, T.T. 2003. Assessment of the diversity of African primates. *Int. J. Primatol.* **24:** 1301–1357.
Hastings, W. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57:** 97–109.
Hey, J. 2005. On the number of new world founders: A population genetic portrait of the peopling of the Americas. *PLoS Biol.* **3:** e193. doi:10.1371/journal.pbio.0030193.
Hey, J. 2006. On the failure of modern species concepts. *Trends Ecol. Evol.* **21:** 447–450.
Hey, J. and Nielsen, R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167:** 747–760.
Hey, J., Won, Y.-J., Sivasundar, A., Nielsen, R., and Markert, J. 2004. Using nuclear haplotypes with microsatellites to study gene flow between recently separated cichlid species. *Mol. Ecol.* **13:** 909–919.
Hill, W. 1969. *The nomenclature, taxonomy and distribution of chimpanzees*, Vol. 1. Karger, Basel.
Hobolth, A., Christensen, O., Mailund, T., and Schierup, M. 2006. Genomic relationships and speciation times of human, chimpanzee and gorilla inferred from a coalescent Hidden Markov Model. *PLoS Genet.* **3:** e7, doi: 10.1371/journal.pgen.0030007.
Hudson, R. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23:** 183–201.
Hudson, R. 2001. Two-locus sampling distributions and their application. *Genetics* **159:** 1805–1817.
Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18:** 337–338.
Hudson, R.R. and Coyne, J.A. 2002. Mathematical consequences of the genealogical species concept. *Evolution Int. J. Org. Evolution* **56:** 1557–1565.
Hudson, R.R. and Kaplan, N.L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111:** 147–164.
Hudson, R., Slatkin, M., and Maddison, W. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132:** 583–589.
Hughes, P., Woodward, J., and Gibbard, P. 2006. Quaternary glacial history of the Mediterranean mountains. *Prog. Phys. Geogr.* **30:** 334–364.
Innan, H. and Watanabe, H. 2006. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Mol. Biol. Evol.* **23:** 1040–1047.
Kaessmann, H., Wiebe, V., Weiss, G., and Pääbo, S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat. Genet.* **27:** 155–156.
Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61:** 893–903.
Kliman, R., Andolfatto, P., Coyne, J., Depaulis, F., Kreitman, M., Berry, A., McCarter, J., Wakeley, J., and Hey, J. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156:** 1913–1931.
Kong, A., Gudbjartsson, D., Sainz, J., Jonsdottir, G., Gudjonsson, S., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.
Kumar, S., Tamura, K., Jakobsen, I., and Nei, M. 2001. MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics* **17:** 1244–1245.
Leman, S.C., Chen, Y., Stajich, J.E., Noor, M.A.F., and Uyenoyama, M.K. 2005. Likelihoods from summary statistics: Recent divergence between species. *Genetics* **171:** 1419–1436.
Llopart, A., Lachaise, D., and Coyne, J.A. 2005. Multilocus analysis of introgression between two sympatric sister species of *Drosophila*: *Drosophila yakuba* and *D. santomea*. *Genetics* **171:** 197–210.
Mayr, E. 1963. *Animal species and evolution*. The Belknap Press, Cambridge, MA.
McBrearty, S. and Jablonski, N.G. 2005. First fossil chimpanzee. *Nature* **437:** 105–108.
McVean, G., Myers, S., Hunt, S., Deloukas, P., Bentley, D., and Donnelly, P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304:** 581–584.
Meng, X.-L. 1994. Posterior predictive *p*-values. *Ann. Stat.* **22:** 1142–1160.
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. 1953. Equation of state calculation by fast computing machines. *J. Chem. Phys.* **21:** 1087–1092.
Miller, S., Purugganan, M., and Curtis, S. 2006. Molecular population genetics and phenotypic diversification of two populations of the thermophilic cyanobacterium *Mastigocladus laminosus*. *Appl. Environ. Microbiol.* **72:** 2793–2800.
Muir, C., Galdikas, B., and Andrew, T. 2000. mtDNA sequence diversity of orangutans from the islands of Borneo and Sumatra. *J. Mol. Evol.* **51:** 471–480.
Myers Thompson, J.A. 2003. A model of the biogeographical journey from *Proto-pan* to *Pan paniscus*. *Primates* **44:** 191–197.
Nachman, M.W. and Crowell, S.L. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156:** 297–304.
Nei, M. and Li, W. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* **76:** 5269–5273.
Nielsen, R. and Signorovitch, J. 2003. Correcting for ascertainment biases when analyzing SNP data: Applications to the estimation of linkage disequilibrium. *Theor. Popul. Biol.* **63:** 245–255.
Nielsen, R. and Wakeley, J. 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics* **158:** 885–896.
Nordborg, M. and Tavare, S. 2002. Linkage disequilibrium: What history has to tell us. *Trends Genet.* **18:** 83–90.
Patterson, N., Richter, D., Gnerre, S., Lander, E., and Reich, D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441:** 1103–1108.
Pollard, D., Iyer, V., Moses, A., and Eisen, M. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: Evidence for incomplete lineage sorting. *PLoS Genet.* **2:** e173. doi: 10.1371/journal.pgen.0020173.
Presgraves, D.C., Balagopalan, L., Abmayr, S.M., and Orr, H.A. 2003. Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. *Nature* **423:** 715–719.
Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. 1999. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16:** 1791–1798.
Przeworski, M. 2003. Estimating the time since the fixation of a beneficial allele. *Genetics* **164:** 1667–1676.
Putnam, A., Scriber, M., and Andolfatto, P. 2007. Discordant divergence times among Z chromosome regions between two ecologically distcint swallowtail butterfly species. *Evolution Int. J. Org. Evolution* **61:** 912–927.
Sawamura, K., Watanabe, T., and Yamamoto, M. 1993. Hybrid lethal systems in the *Drosophila melanogaster* species complex. *Genetica* **88:** 175–185.
Smith, R.J. and Pilbeam, D.R. 1980. Evolution of the orangutan. *Nature* **284:** 447–448.
Stephens, M., Smith, N., and Donnelly, P. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68:** 978–989.
Stone, A. and Verrelli, B. 2006. Focusing on comparative ape population genetics in the post-genomic age. *Curr. Opin. Genet. Dev.* **16:** 586–591.
Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.
Takahata, N. and Satta, Y. 2002. Pre-speciation coalescence and the effective size of ancestral populations. In *Modern developments in theoretical population genetics* (eds. M. Slatkin and M. Veuille), pp. 52–71. Oxford University Press, Oxford.
Tavaré, S., Balding, D.J., Griffiths, R.C., and Donnelly, P. 1997. Inferring coalescence times from DNA sequence data. *Genetics* **145:** 505–518.

Thalmann, O., Fischer, A., Lankester, F., Paabo, S., and Vigilant, L. 2006. The complex evolutionary history of gorillas: Insights from genomic data. *Mol. Biol. Evol.* **24:** 146–158.
Thompson, J., Higgins, D., and Gibson, T. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Ting, C.-T., Tsaur, S.-C., Wu, M.-L., and Wu, C.-I. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* **282:** 1501–1504.
Voight, B., Adams, A., Frisse, L., Qian, Y., Hudson, R., and Di Rienzo, A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl. Acad. Sci.* **102:** 18508–18513.
Wakeley, J. and Hey, J. 1997. Estimating ancestral population parameters. *Genetics* **145:** 847–855.
Wall, J. 2000. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* **154:** 1271–1279.
Wall, J.D. 2003. Estimating ancestral population sizes and divergence times. *Genetics* **163:** 395–404.
Wang, R.-L., Stec, A., Hey, J., Lukens, L., and Doebley, J. 1999. The limits of selection during maize domestication. *Nature* **398:** 236–239.
Wittbrodt, J., Adam, D., Malitschek, B., Maueler, W., Raulf, F., Telling, A., Robertson, S.M., and Schartl, M. 1989. Novel putative receptor tyrosine kinase encoded by the melanoma-inducing *Tu* locus in *Xiphophorus*. *Nature* **341:** 415–421.
Won, Y.-J. and Hey, J. 2005. Divergence population genetics of chimpanzees. *Mol. Biol. Evol.* **22:** 297–307.
Won, Y.-J., Sivasundar, A., Wang, Y., and Hey, J. 2005. On the origin of Lake Malawi cichlid species: A population genetic analysis of divergence. *Proc. Natl. Acad. Sci.* **102:** 6581–6586.
Wu, C.-I. 2001. The genic view of the process of speciation. *J. Evol. Biol.* **14:** 851–865.
Yu, N., Fu, Y.-X., and Li, W.-H. 2002. DNA polymorphism in a worldwide sample of human X chromosomes. *Mol. Biol. Evol.* **19:** 2131–2141.
Yu, N., Jensen-Seaman, M.I., Chemnick, L., Kidd, J.R., Deinard, A.S., Ryder, O., Kidd, K.K., and Li, W.-H. 2003. Low nucleotide diversity in chimpanzees and bonobos. *Genetics* **164:** 1511–1518.
Yu, N., Jensen-Seaman, M., Chemnick, L., Ryder, O., and Li, W.-H. 2004. Nucleotide diversity in gorillas. *Genetics* **166:** 1375–1383.
Zhang, Y., Ryder, O., and Zhang, Y. 2001. Genetic divergence of orangutan subspecies (*Pongo pygmaeus*). *J. Mol. Evol.* **52:** 516–526.
Zhi, L., Karesh, W., Janczewski, D., Frazier-Taylor, H., Sajuthi, D., Gombek, F., Andau, M., Martenson, J., and O'Brien, S. 1996. Genomic differentiation among natural populations of orangutan (*Pongo pygmaeus*). *Curr. Biol.* **6:** 1326–1336.

Received February 16, 2007; accepted in revised form July 3, 2007.

A new approach to estimate parameters of speciation models with application to apes

Celine Becquet and Molly Przeworski

Genome Res. 2007 17: 1505-1519 originally published online August 21, 2007
Access the most recent version at doi:10.1101/gr.6409707