Inference of species histories in the presence of gene flow

Nicola F. Müller[a,b,1], Huw A. Ogilvie[c,d], Chi Zhang[e,f], Alexei Drummond[c] and Tanja Stadler[a,b]

[a] ETH Zurich, Department of Biosystems Science and Engineering, 4058 Basel, Switzerland

[b] Swiss Institute of Bioinformatics (SIB), Switzerland

[c] Centre for Computational Evolution, University of Auckland, New Zealand

[d] Division of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australia

[e] Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China

[f] Center for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, Beijing 100044, China

[1]Corresponding author

Abstract: When populations become isolated, members of these populations can diverge genetically over time. This leads to genetic differences between individuals of these populations that increase over time if the isolation persists. This process can be counteracted when genes are exchanged between populations. In order to study the speciation processes when gene flow is present, isolation-with-migration methods have been developed. These methods typically assume that the ranked topology of the species history is already known. However, this is often not the case and the species tree is therefore of interest itself. To infer it is currently only possible when assuming no gene flow. This assumption can lead to wrongly inferred speciation times and species tree topologies.

Building on a recently introduced structured coalescent approach, we introduce a new method that allows inference of the species tree while explicitly modelling the flow of genes between coexisting species. By using Markov chain Monte Carlo sampling, we co-infer the species tree alongside evolutionary parameters of interest. By using simulations, we show that our newly introduced approach is able to reliably infer the species trees and parameters of the isolation-with-migration model from genetic sequence data. We then infer the species history of six great ape species including gene flow after population isolation. By using this dataset, we are able to show that our new methods is able to infer the correct species tree not only on simulated but also on a real data set where the species history has already been well studied. In line with previous results, we find some support for some gene flow between bonobos and common chimpanzees.

1 Introduction

Populations can diverge genetically and become separated over time, due to geographical or other factors. This divergence is captured in the genome of sampled individuals of those species. In turn, the genetic sequences of sampled species allow us to draw conclusions about their common history (the species phylogeny) by modelling the speciation process. Gene flow after populations become genetically isolated can counteract this process (Sousa and Hey, 2013), and if not accounted for, this can lead to biased inferences about the ancestral history of the species (Leaché *et al.*, 2014).

To reconstruct the speciation process from genetic sequence data, the multispecies coalescent model (MSC) can be used (Rannala and Yang, 2003; Liu *et al.*, 2009; Heled and Drummond, 2010). This allows for reconstruction of the species tree while accounting for discordance between gene trees from due to incomplete lineage sorting. It however assumes complete isolation after speciation and can thus be biased in the presence of gene flow (Leaché *et al.*, 2014). To account for gene flow after speciation, isolation-with-migration (IM) models have been developed (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004; Wilkinson-Herbots, 2008) (see also Sousa and Hey, 2013, for review). Initial Bayesian implementations of the IM model were applicable to only two populations. Further, they could suffer from poor Markov chain Monte Carlo (MCMC) performance due to an extremely diffuse parameter space (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004). This difficulty was partly overcome by analytically integrating out some model parameters (that is population sizes and migration rates) to sample from the posterior distribution of gene trees and speciation times. The gene trees and speciation times are then used to estimate the rest of the parameters (Hey and Nielsen, 2007). This approach was then extended to deal with more than two populations (Hey, 2010). Currently however, these methods require the ranked species tree topology and node rank to be known *a priori*.

One of the challenges that restricts joint inference of the species history is the need to infer the migration histories. Doing so requires inferring which species each lineage was in the past and can lead to computational issues (De Maio *et al.*, 2015; Müller *et al.*, 2018). For the related structured coalescent process (Takahata, 1988; Hudson, 1990; Notohara, 1990), some of us previously tackled this challenge by deriving an approximation that integrates over every possible migration history (Müller *et al.*, 2017).

Here, we introduce a novel isolation-with-migration model (AIM) that allows joint inference of the species history. We do so by extending the marginal approximation of the structured coalescent (Müller *et al.*, 2017) to the isolation-with-migration model. Modelling the movement of lineages

between speciation events as a structured coalescent process, we evaluate the probability of a gene tree given a species tree, a set of rates of gene flow and effective population sizes. By using MCMC sampling, we then jointly infer species tree topologies and divergence times, as well as the rates of gene flow and effective population sizes for extinct and extant species.

By using simulations, we show that the AIM model is able to infer migration rates, effective population sizes and species trees reliably from molecular sequences directly. In contrast to AIM, the MSC can strongly support wrong species tree topologies and systematically underestimate speciation times. We then apply AIM to jointly infer the species trees and the probability of gene flow between individual species in a previously described dataset (Yu *et al.*, 2002, 2003, 2004).

We implemented this approach as an update to StarBEAST2 (Ogilvie *et al.*, 2017), which is available as a package for the phylogenetic software platform BEAST2 (Bouckaert *et al.*, 2014).

Results

Inference of the species tree from genetic sequences

We first test if AIM is able to infer the true species tree, gene flow and effective population sizes of extant and extinct species. To so so, we simulated 50 gene trees given the four taxa species tree in Figure 1 by using MCcoal (Yang, 2015). We simulated these trees using 20 individuals from species A, 10 from B, 5 from C and 15 from D, resulting in 50 individuals, each with 50 independent gene trees. Rates of gene flow were allowed to be asymmetric between individual species and sampled randomly from an exponential distribution with $\lambda = 1$. The rates were then scaled using eq 4 to account for species that are more distantly related to have a lower probability of gene flow. The relative effective population sizes between extinct and extant species were sampled randomly from a log normal distribution with mean= 0 and $S = 0.2$. We then used the gene trees to simulate genetic sequences under an HKY model for each individual and loci by using SeqGen (Rambaut and Grassly, 1997).

We next used AIM to jointly infer the species tree, rates of gene flow and all effective population sizes of extant and ancestral species from the simulated sequences. We further inferred the species tree from the same sequence data using MSC in StarBEAST2 without accounting for gene flow between species. We repeated this 100 replicates.

Figure 1 shows the ability of AIM to infer gene flow and effective popula-

tion sizes and compares how well species trees are inferred to the MSC. The approximate isolation-with-migration model is able to infer the species tree topologies and speciation times well. Effective population sizes and migration rates are estimated well for extant species, but can be highly imprecise for ancestral species. This is to be expected due to relatively little information in the gene trees for these rates in the presence of phylogenetic uncertainty. After removing phylogenetic uncertainty by fixing the species and gene tree topologies and times, all rates are estimated well (see figure S1). In contrast to AIM, the MSC is in some cases unable to retrieve the correct species tree topology. One of the most striking differences is the underestimation of speciation times by the MSC, compared to AIM which does not show any bias.

Inference of the common history of chimpanzee species

We next apply the approximate isolation-with-migration model to study the evolutionary history of the great apes. We used sequences of gorillas, humans, bonobos and western, eastern and central chimps from 50 non-coding loci. We then inferred the evolutionary history, i.e. the species tree topology and speciation times, while allowing for gene flow between the different sampled and ancestral species. To estimate which gene flows are significant, we sampled the probability of each migration rate being zero, similar to (Lemey *et al.*, 2009), within the MCMC. The probability of each rate being zero was *a priori* assumed to be 0.95. We then computed the Bayes Factor for gene flow between different clades as the Bayes Factor of any migration rate from clade a to clade b being non zero.

Figure 2 shows the inferred species tree of great apes. We infer the order of speciation event to be consistent with previous results (Prado-Martinez *et al.*, 2013). We infer possible gene flow between the chimp clade and bonobos, as well as between the central/eastern chimpanzees and bonobos (see figure S3). Gene flow between clades was assumed to be active if at least one rate from clade a to b was inferred to be active. We performed the same analysis only using common chimpanzees and bonobo sequences (see figure S2 and figure S4). The inferred effective population sizes for the bonobos and common chimpanzee are consistent with (Hey, 2010) and between the analyses with and without gorilla and human data. We also infer gene flow between the same species or clades using the reduced data set.

Discussion

The AIM model and implementation introduced here is able to jointly estimate the species tree topology and times, effective population sizes, and migration rates, from multilocus molecular sequence data. This combination of parameters is relevant to many biological systems including the great apes, but has not been possible to infer jointly until now. Our approach is implemented in a new version of an open source package (StarBEAST2) which is an extension to the phylogenetics software platform (BEAST2). This means that users of AIM can take advantage of the flexibility of BEAST2, including the large number of available molecular clock models and substitution models.

Using simulations, we demonstrate the validity of our approach as well as the problems that can occur when gene flow is not accounted for. The species tree topologies and node heights are inferred accurately in all scenarios we simulated. Not accounting for gene flow can lead to underestimated speciation times as well as incorrect species tree topologies. This is consistent with previous observations (Leaché *et al.*, 2014). The estimates of migration rates are unbiased but can be of low precision due to the a lack of observed migration events in the sampled geneological histories. Potential extensions to the model only allow migration to occur for a defined period of time after speciation (Wilkinson-Herbots, 2012). Estimates of effective population sizes of ancestral species are problematic to estimate in the presence of phylogenetic uncertainty since they are directly linked to the distance between speciation times and coalescent times of gene trees. Genetic sequence data typically informs only poorly about these differences, and strong prior assumptions have to be made.

Analysis of non-coding sequence of 6 great ape species did reveal some evidence for gene flow between central/eastern common chimpanzees and bonobos ancestral branch of the chimp clade and bonobos. These results are consistent with the ones obtained when only using bonobo and common chimpanzee sequences.

Overall, we provide an approach that is able to jointly infer species histories and gene flow in an MCMC frameworks. It's implementation as part of StarBeast2 further allows to include additional sources of data such as fossil data to infer the species tree (Ogilvie *et al.*, 2018) using the Fossilize-Birth-Death model framework (Stadler, 2010). Additionally, since the underlying structured coalescent theory has been developed explicitly for serially sampled data (Müller *et al.*, 2017, 2018), accounting for temporal information such as with using ancient DNA will be possible in the future.

2 Materials and Methods

The location of a gene over time

Isolation-with-migration models are closely related to the structured coalescent process. These models generally require the state or location of every single lineage to be inferred backwards in time. A recently introduced approximation to the structured coalescent avoids this by formally integrating over every possible history (Müller *et al.*, 2017). For each gene tree $T_{\mathcal{G}}$, we calculate the probability of a lineage L_i at time t being in a particular state a (i.e. the gene being in a particular species a), given $T_{\mathcal{G}}$, $P_t(T_{\mathcal{G}})$. Between speciation, coalescent, or fossil sampling events, this probability can be described via differential equations as described in **?** eqn. 1:

$$\begin{aligned}\frac{d}{dt}P_t(L_i = a|T_{\mathcal{G}}) = \sum_{b=1}^{s}\Big(m_{ba}P_t(L_i = b|T_{\mathcal{G}}) - m_{ab}P_t(L_i = a|T_{\mathcal{G}})\Big)\\ + P_t(L_i = a|T_{\mathcal{G}})\sum_{b=1}^{s}\frac{1}{N_{e_b}}P_t(L_i = b|T_{\mathcal{G}})\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_k = b|T_{\mathcal{G}})\\ - P_t(L_i = a|T_{\mathcal{G}})\frac{1}{N_{e_a}}\sum_{\substack{k=1\\k\neq i}}^{n}P_t(L_k = a|T_{\mathcal{G}})\end{aligned} \quad (1)$$

m_{ba} describes the backwards in time rate at which migration events from species b to a happen and N_{e_a} is the effective population size of species a and s denotes the number of species. At a coalescent event between lineage i and j, the probability of the parent lineage can be calculated using the following equation (14, page 2979):

$$P_t(L_{parent} = a, T_{\mathcal{G}}) = \frac{1}{N_{e_a}}P_t(L_i = a|T_{\mathcal{G}})P_t(L_j = a|T_{\mathcal{G}})P_t(T_{\mathcal{G}}) \quad (2)$$

Then we can proceed solving eqn. 1 again backward in time with the initial value $P_t(L_{parent} = a|T_{\mathcal{G}}) = P_t(L_{parent} = a, T_{\mathcal{G}})/P_t(T_{\mathcal{G}})$.

Calculation of the probability of a gene tree under the approximate isolation-with-migration model

To calculate the probability of a gene tree under the approximate isolation-with-migration (AIM) model, the following probability has to be calculated:

$$P(T_{\mathcal{G}}|T_{\mathcal{S}}, S, \mathcal{M}, \mathcal{N})$$

with $T_{\mathcal{G}}$ being the gene tree, $T_{\mathcal{S}}$ being the species tree, S being the species to which each sampled individual belongs and $\mathcal{M}$ and $\mathcal{N}$ being the set of migration rates and effective population sizes. To allow for multi-locus data, we assume that each loci independently evolved from the same isolation-with-migration process.

$P(T_{\mathcal{G}}|T_{\mathcal{S}}, S, \mathcal{M}, \mathcal{N})$ can be calculated similar to the probability of a tree under the structured coalescent. Between speciation events, $P(T_{\mathcal{G}}|T_{\mathcal{S}}, S, \mathcal{M}, \mathcal{N})$ is updated as shown in the previous section. The backwards in time analogue to a speciation event is the combination of two species. If species a and species b have parent species c, the probability of each remaining gene i being in species c can be calculated as follows:

$$P_t(L_i = c|T_{\mathcal{G}}) = P_t(L_i = a|T_{\mathcal{G}}) + P_t(L_i = b|T_{\mathcal{G}}) \tag{3}$$

Starting from the present and going back in time, $P_t(L_i = a|T_{\mathcal{G}})$ can now be calculated using equations 1 and 2 between speciation events and using equation 3 at speciation events up to the root, $P_{root}(L_i = rootspecies|T_{\mathcal{G}})$.

Prior assumptions

Instead of parametrizing the gene flow between species directly, we assume that the migration rate $m_{a,b}$ between species A and B is the product of a migration rate scaler $\alpha_{a,b}$ and the inverse of the difference in time between their most recent common ancestor t_{AB} and the oldest daughter node of A and B, denoted by $max(t_A, t_B)$.

$$m_{a,b} = \alpha_{a,b} \frac{m_{tot}}{t_{AB} - max(t_A, t_B)} \tag{4}$$

This is to account for the rate of gene flow being likely smaller between more distant species. This parametrization allows to assume the same priors for all $\alpha_{a,b}$ independent of the pairwise distance of these species A, B, while accounting for more distant species to be less likely to exchange genes and to therefore have an *a priori* smaller migration rate. $\alpha_{a,b}$ denotes a scaler that is estimated between every pair of coexisting species. m_{tot} is an estimated migration rate that allows to put a prior distribution on the magnitude of gene flow that one expects. Further, this allows to put a prior distribution on the $\alpha_{a,b}$ values to describe how much variation we expect between them.

In this study, we assume an exponential distribution with $mean = 1$ for $\alpha_{a,b}$. The effective population sizes of extant and ancestral species are assumed to be log-normally distributed with $S = 0.25$. The mean of the log-normal distribution is estimated. Exact specifications for all parameters

as well as MCMC operators are provided in the BEAST2 xml input files here https://github.com/nicfel/Isolation-With-Migration. Bayes Factors for the support of gene flow between species were computed as $BF = \frac{p(1-q)}{q(1-p)}$ with p being the posterior and q the prior support for gene flow.

Sequence data

Sequences from four species - human, chimpanzee, bonobo and gorilla - corresponding to 50 non-coding nuclear loci from Genbank were used. These were originally published in (Yu *et al.*, 2002, 2003, 2004). The accessions numbers of human sequences are AY286547-AY288046. Chimpanzee and bonobos sequences have accession numbers AY275957-AY277244 and AY463943-AY463951, and gorilla sequences have accession numbers AY447205-AY447950.

Sequences from all species from the same locus were aligned manually, including using the reverse complement where necessary to ensure the same strandedness for all species. 5' and 3' ends were trimmed so that all species for a given locus had the same start and end positions.

For each locus, sequences were separated into one alignment per species. For any per-species alignment with one or more heterozygous sequences, haplotypes were imputed using PHASE v2.1.1 (Stephens and Scheet, 2005). The default model was used but with a recombination rate prior centered on $4N_ec$, where N_e is the estimate for a given species from (Yu *et al.*, 2004), and c is the default value 10^{-8}. A deviation factor f of 100 was used for the recombination rate prior, implying that the true recombination rate should be within two orders of magnitude of $4N_e$. Five independent chains were run for 500 burn-in and 5000 final iterations. The best fitting chain was used to obtain haplotypes. SEQPHASE (Flot, 2010) was used to convert FASTA alignments to PHASE files and vice versa.

Sequences sourced from non-African humans, and from four chimpanzees of unknown subspecies were discarded. Phased haplotypes from all remaining individuals belonging to all species were reassembled into one alignment per locus.

Software and Data Availability

Simulation of gene trees given a species tree and migration rates were performed using the software MCcoal in BPP (Yang, 2015). Simulations of genetic sequences of length 1000 were performed using Seq-Gen 1.3.3 (Rambaut and Grassly, 1997), using a HKY site model with a transition/transversion ratio of 3 and base frequency of 0.3,0.2,0.2 and 0.3.

Data analyses were performed using BEAST 2.4.6 (Bouckaert *et al.*, 2014). The source code of the BEAST2 package AIM can be downloaded here: https://github.com/genomescale/starbeast2/tree/aim. All the scripts used in this study are publicly available at https://github.com/nicfel/Isolation-With-Migration. Analyses were done using Matlab R2015b. Plotting was done in R 3.2.3 using ggplot2 (Wickham, 2009). Trees were analysed by using ape 3.4 (Paradis *et al.*, 2004) and phytools 0.5-10 (Revell, 2012).

3 Acknowledgement

NM and TS are funded in part by the Swiss National Science foundation (SNF; grant number CR32I3_166258). TS is supported in part by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: grant agreement number 335529).

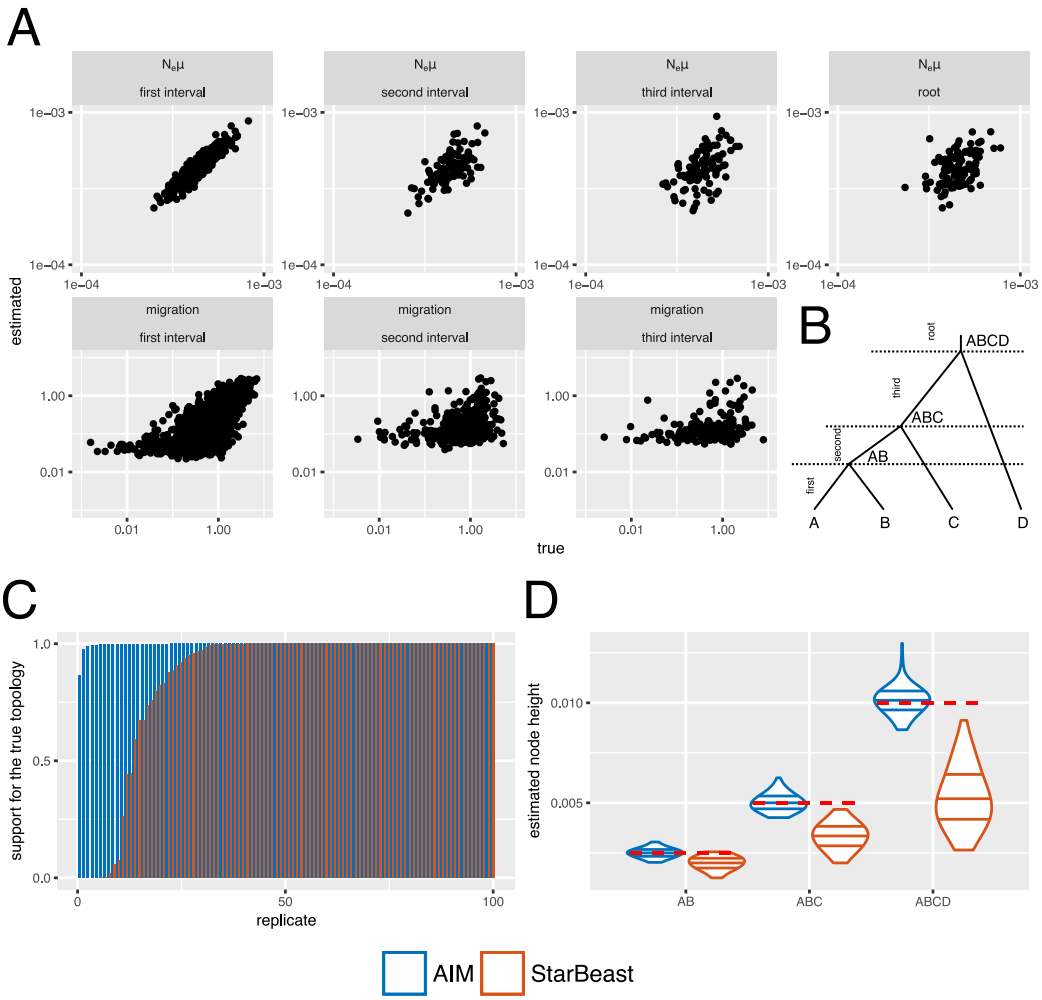

Figure 1: **Inference of the species trees from genetic sequences. A** Estimates of effective population sizes and migration rates separated by the different intervals of the species tree. The x-axis of each figure denotes the true rate under which simulations were performed. The y-axis denotes the median estimated rate. Each point in a subplot denotes one estimated value for an effective population size of any species that occurs first in that interval or of a migration rate between any two species in that interval. **B** Species tree for which gene trees were simulated. **C** Posterior support for the true species tree topology using AIM and StarBeast. The x-axis denotes the different replicates of the simulations ordered by their posterior support of the true topology independently for both methods. **D** Distribution of estimated median node heights conditioned on the correct topology using AIM and StarBeast. The dashed lines represent the true value.

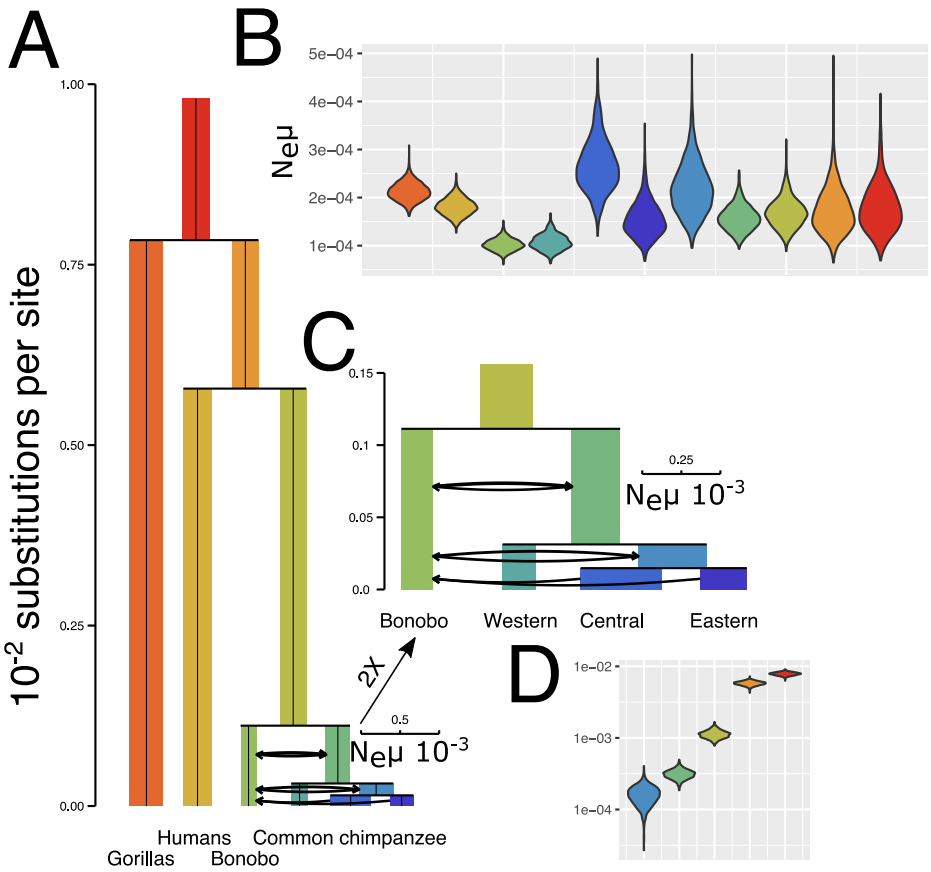

Figure 2: **Inferred species history of great apes. A** Here we show the inferred species history of different great ape species in units of substitutions per site. The tree denotes the most likely inferred species tree topology. The width of the coloured rectangles denotes the inferred median effective population sizes of the extant and ancestral species. The arrows denote the Bayes Factor for un-directional gene flow between the different clades subtracted by its prior probability. This is the support for gene flow in any direction between any species of each clade involved. Arrows are shown between clades for which the BF is 2 or greater. **B** Violin plots of the effective population sizes of each extant and ancestral species. The colors of the violin plots correspond to the colors of the species tree. **C** Inferred phylogenetic tree of the common chimpanzee and bonobos zoomed in by a factor of 2. Here, the arrows denote inferred possible directional gene flow between the different clades **D** Inferred distribution of speciation times in substitutions per site. The color of the violin plots correspond to the species tree.

References

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, **10**(4), e1003537.

De Maio, N., Wu, C.-H., O'Reilly, K. M., and Wilson, D. (2015). New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS genetics*, **11**(8), e1005421.

Flot, J.-F. (2010). Seqphase: a web tool for interconverting phase input/output files and fasta sequence alignments. *Molecular Ecology Resources*, **10**(1), 162–166.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, **27**(3), 570–580.

Hey, J. (2010). Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, **27**(4), 905–920.

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**(2), 747–760.

Hey, J. and Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(8), 2785–2790.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, **7**(1), 44.

Leaché, A. D., Harris, R. B., Rannala, B., and Yang, Z. (2014). The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology*, **63**(1), 17–30.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS computational biology*, **5**(9), e1000520.

Liu, L., Yu, L., Kubatko, L., Pearl, D. K., and Edwards, S. V. (2009). Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, **53**(1), 320–328.

Müller, N. F., Rasmussen, D. A., and Stadler, T. (2017). The structured coalescent and its approximations. *Molecular biology and evolution*, **34**(11), 2970–2981.

Müller, N. F., Rasmussen, D., and Stadler, T. (2018). Mascot: Parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*, page bty406.

Nielsen, R. and Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**(2), 885–896.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, **29**(1), 59–75.

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution*.

Ogilvie, H. A., Vaughan, T. G., Matzke, N. J., Slater, G. J., Stadler, T., Welch, D., and Drummond, A. J. (2018). Inferring species trees using integrative models of species evolution. *bioRxiv*, page 242875.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England)*, **20**(2), 289–90.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O?Connor, T. D., Santpere, G., *et al.* (2013). Great ape genetic diversity and population history. *Nature*, **499**(7459), 471.

Rambaut, A. and Grassly, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences : CABIOS*, **13**(3), 235–238.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, **164**(4), 1645–1656.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**(2), 217–223.

Sousa, V. and Hey, J. (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, **14**(6), 404–414.

Stadler, T. (2010). Sampling-through-time in birth–death trees. *Journal of theoretical biology*, **267**(3), 396–404.

Stephens, M. and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *The American Journal of Human Genetics*, **76**(3), 449–462.

Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetical research*, **52**(3), 213–222.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilkinson-Herbots, H. M. (2008). The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. *Theoretical Population Biology*, **73**(2), 277–288.

Wilkinson-Herbots, H. M. (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. *Theoretical Population Biology*, **82**(2), 92–108.

Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Current Zoology*, **61**(5), 854–865.

Yu, N., Chen, F.-C., Ota, S., Jorde, L. B., Pamilo, P., Patthy, L., Ramsay, M., Jenkins, T., Shyue, S.-K., and Li, W.-H. (2002). Larger genetic differences within africans than between africans and eurasians. *Genetics*, **161**(1), 269–274.

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**(4), 1511–1518.

Yu, N., Jensen-Seaman, M. I., Chemnick, L., Ryder, O., and Li, W.-H. (2004). Nucleotide diversity in gorillas. *Genetics*, **166**(3), 1375–1383.

4 Supplementary Material

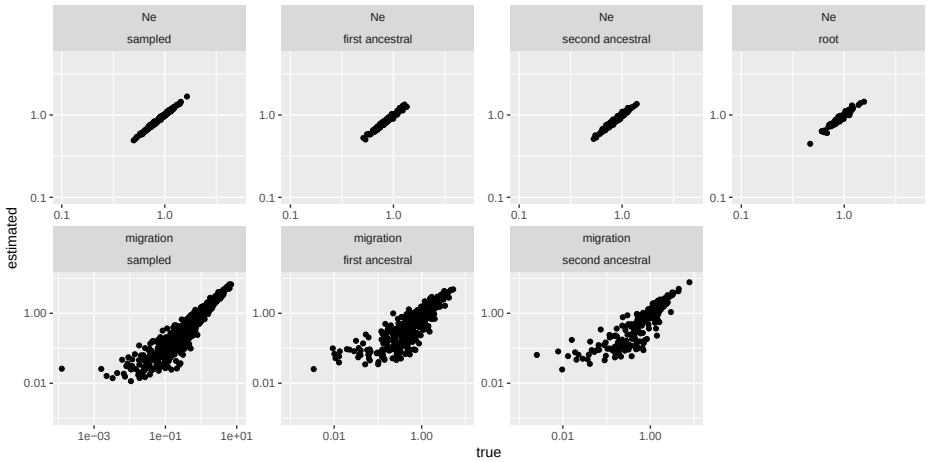

Figure S1: **Estimates of effective population sizes and migration rates from simulated data with fixed gene and species trees.** Estimates of effective population sizes and migration rates separated by the different intervals of the species tree. The x-axis of each figure denotes the true rate under which simulations were performed. The y-axis denotes the median estimated rate. The rates were estimated from 500 simulated gene trees with random effective population sizes and migration rates. The species tree was the same for every of 100 repetitions.

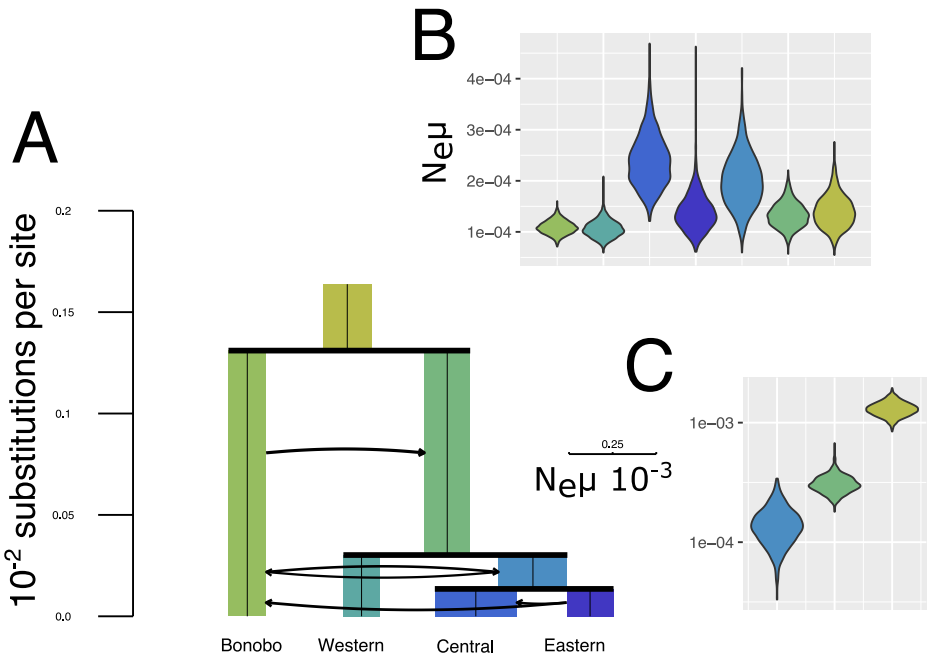

Figure S2: **Inferred species history of common chimpanzees and bonobos. A** Here we show the inferred species history of the common chimpanzee species and bonobos in units of substitutions per site. The tree denotes the most likely inferred species tree topology. The width of the coloured rectangles denotes the inferred median effective population sizes of the extant and ancestral species. The arrows Bayes Factor for gene flow between the different clades. **B** Violin plots of the effective population sizes of each extant and ancestral species. The colors of the violin plots correspond to the colors of the species in the tree. **C** Inferred distribution of speciation times in substitutions per site. The color of the violin plots correspond to the ancestral species.

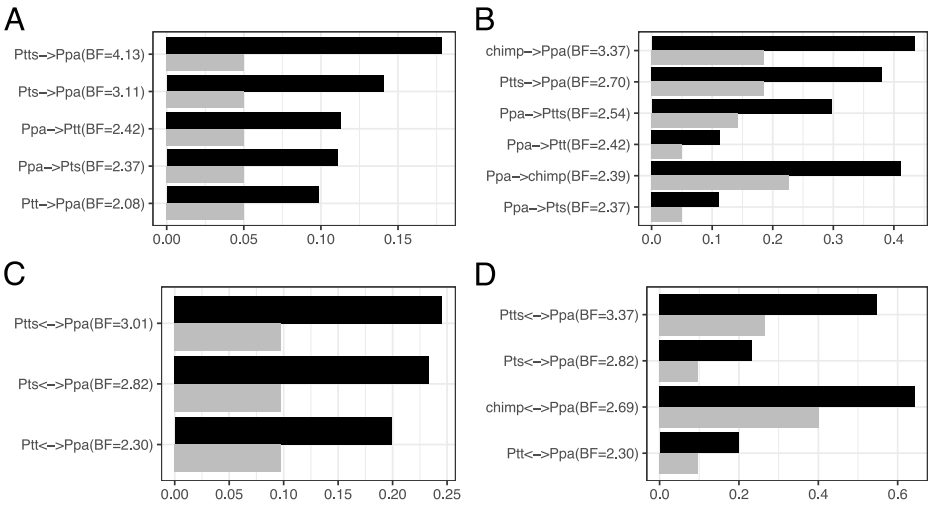

Figure S3: **Inferred gene flow between extant and ancestral species or clades using common chimpanzee, bonobo, human and Gorilla sequences. A** Directional gene flow between species for which the posterior probability of gene flow is inferred to have a Bayes Factor of at least 2. Bayes factor are shown in brackets. Ptts denotes the ancestral species of Central and Eastern chimpanzees. **B** Directional gene flow between clades. Gene flow is assumed to be happening if there is gene flow between at least one of the species belonging to the clade and the destination. The chimp clade includes all extant and ancestral common chimpanzees. **C** Inferred un-directional gene flow between species. **D** Inferred un-directional gene flow between clades.

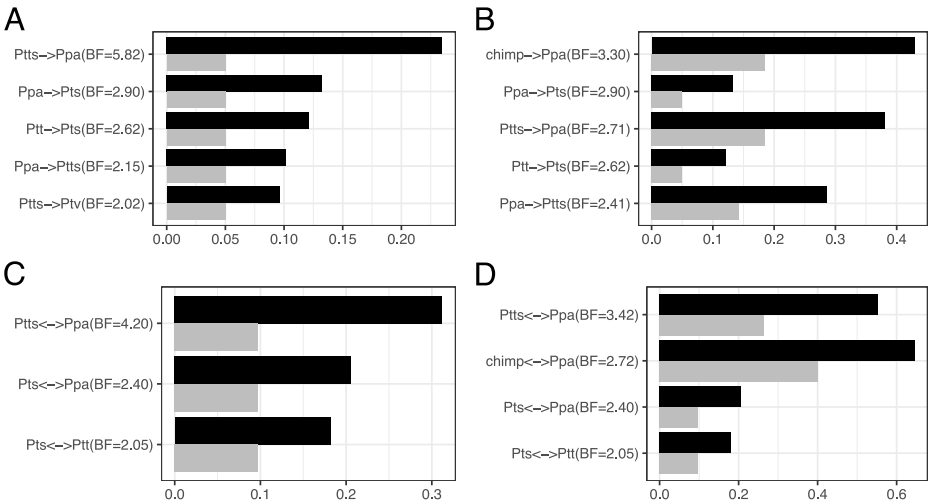

Figure S4: **Inferred gene flow between extant and ancestral species or clades using only common chimpanzee and bonobo sequences.** **A** Directional gene flow between species for which the posterior probability of gene flow is inferred to have a Bayes Factor of at least 2. Bayes factor are shown in brackets. Ptts denotes the ancestral species of Central and Eastern chimpanzees. **B** Directional gene flow between clades. Gene flow is assumed to be happening if there is gene flow between at least one of the species belonging to the clade and the destination. The chimp clade includes all extant and ancestral common chimpanzees. **C** Inferred un-directional gene flow between species. **D** Inferred un-directional gene flow between clades.