MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2016) **16**, 1435–1448 doi: 10.1111/1755-0998.12544

CoMuS: simulating coalescent histories and polymorphic data from multiple species

S. PAPADANTONAKIS,* P. POIRAZI† and P. PAVLIDIS†

**Department of Biology, University of Crete, PO Box 2208, 71409 Heraklio, Greece, †Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology-Hellas (FORTH), 70013 Heraklio, Greece*

Abstract

The simultaneous analysis of intra- and interspecies variation is challenging mainly because our knowledge about patterns of polymorphisms where both intra- and interspecies samples coexist is limited. In this study, we present CoMuS (Coalescent of Multiple Species), a multispecies coalescent software that can simulate intra- and interspecies polymorphisms. CoMuS supports a variety of speciation models and demographic scenarios related to the history of each species. In CoMuS, speciation can be accompanied by either instant or gradual isolation between sister species. Sampling may also occur in the past, and thus, we can study simultaneously extinct and extant species. Our software supports both the infinite- and the finite-site model, with substitution rate heterogeneity among sites and a user-defined proportion of invariable sites. We demonstrate the usage of CoMuS in various applications: species delimitation, software testing, model selection and parameter inference involving present-day and ancestral samples, comparison between gradual and instantaneous isolation models, estimation of speciation time between human and chimpanzee using both intra- and interspecies variation. We expect that CoMuS will be particularly useful for studies where species have been separated recently from their common ancestor and phenomena such as incomplete lineage sorting or introgression still occur.

Keywords: approximate Bayesian computation, coalescent, population genetics, simulations, speciation

Received 8 June 2015; revision received 30 April 2016; accepted 6 May 2016

Introduction

Coalescent simulation represents a Monte Carlo process to generate samples drawn from a Wright–Fisher model of evolution. In its simplest form, the coalescent (Kingman 1982; Hein *et al*. 2004; Wakeley 2008) is a retrospective model of random genetic drift that traces all sampled alleles back to their most recent common ancestor (MRCA). Inheritance is usually depicted as a bifurcating ultrametric tree; recent extensions allow also for multifurcations (Λ-coalescent; Pitman 1999; Birkner & Blath 2008). During the last 15 years, numerous extensions have been proposed, allowing to model processes such as recombination, gene conversion, past population size changes, population subdivision, gene flow and positive selection (Hudson 2002; Hein *et al*. 2004; Wakeley 2008; Ewing & Hermisson 2010). The coalescent model tracks only the ancestors of sampled alleles; thus, tracking the whole population is not necessary. This property makes coalescent attractive for computer simulations that can be used to: (i) verify analytical results, (ii) explore complex evolutionary models that are mathematically intractable (Kessner & Novembre 2014) and (iii) construct empirical distributions for summary statistics in a specific evolutionary framework. The most widely used implementation of single-species coalescent is HUDSON'S MS software (Hudson 2002). *ms* can simulate samples from neutrally evolving populations, allowing for migration, recombination, gene conversion and ancestral changes of the population size. Due to its efficiency and flexibility, *ms* has been used in Approximate Bayesian Computation (ABC) inference methods to generate distributions of summary statistics under various evolutionary models (e.g. Pavlidis *et al*. 2010; Saminadin-Peter *et al*. 2012).

The multispecies coalescent (MSC) (Hobolth *et al*. 2007; Heled & Drummond 2010) is a retrospective model of evolution when speciation events have occurred in the ancestry of the sample. Thus, similar to the single-species coalescent, it effectively models random genetic drift within species, but it also accounts for speciation events at specific time points, after which the two sister species evolve independently. The MSC has been exploited in

Correspondence: Pavlos Pavlidis, Fax: +30-2810-39110; E-mail: pavlidisp@gmail.com

several different applications of evolutionary biology. In phylogenetics, it has been used to study the effect of incomplete lineage sorting (Hobolth *et al.* 2007; Mossel & Roch 2007; Heled & Drummond 2010). Degnan & Rosenberg (2006) discuss thoroughly the effect of discordant genealogical histories at multiple loci on phylogenetic inference, as an effect of incomplete lineage sorting, mainly between recently diverged species (but also in deep phylogenies for some combinations of branching patterns and lengths), horizontal gene transfer or gene duplication. They conclude by suggesting that discordant genealogies between loci can provide insights into the (population genetics) processes that take place in the ancestry of present-day species and help us examine species divergence processes or infer parameters such as ancestral population sizes and divergence times. In population genetics, Hobolth *et al.* (2007) used the MSC to infer ancestral effective population sizes and speciation times from a sample of human, chimpanzee and gorilla. Furthermore, Zhang *et al.* (2013) applied MSC simulations to assess the performance of their species delimitation method under various speciation rates. Recently, Heled *et al.* (2013) proposed a more realistic implementation of the speciation in MSC: lineages from different species remain in partial contact after the speciation event and gene flow between them is gradually reduced. The study by Heled *et al.* (2013) focuses on the effect of gene flow on phylogenetic inference and migration rate estimation.

Here, we present an open-source software, CoMuS, for multiple-species coalescent simulations. CoMuS couples the core simulation machinery of HUDSON'S MS with Rambaut and Grassly's Seq-Gen (Rambaut & Grassly 1997) as well as a variety of algorithms to generate sequences under various phylogenetic models (Stadler 2009, 2011; Hartmann *et al.* 2010). The simplest usage of CoMuS allows the simulation of samples from diverged species, the simulation of the phylogenetic model describing the evolution of species as well as the implementation of the finite-site model. We have implemented additional features enabling the study of (i) speciation model parameters such as birth, death, sampling rates; (ii) gradual isolation between diverging species; (iii) simultaneous sampling of present-day and archaic sequences; and (iv) divergent species and diverging populations within species. CoMuS also implements population size changes, gene flow between populations or introgression between species, and population subdivision.

In addition, we provide an accompanying software, CoMuStats, that can calculate summary statistics directly from the output of CoMuS, for each population/species individually as well as for the entire sample. CoMuStats can calculate many summary statistics, such as, θ value estimators, Tajima's D (Tajima 1989), Wall's statistics (B and Q) (Wall 1999), F_{ST} values (Hudson *et al.* 1992), site frequency spectrum (Fisher 1930; Wright 1938) and others. CoMuS, accompanied by CoMuStats can be used naturally in an ABC framework, where CoMuS generates sequence data and CoMuStats calculates summary statistics. Then, R packages (e.g. 'abc', Csilléry *et al.* 2012) can be used to estimate parameter values.

We demonstrate the capabilities of CoMuS in several applications. We show that it can be used to study the distribution of summary statistics to infer parameters in an ABC framework and/or to provide insights into the effects of the finite-site model on summary statistics (compared to the infinite-site model), the rate heterogeneity between sites, the different mutation models (e.g. HKY; Hasegawa *et al.* 1985 or GTR; Tavaré 1986) and speciation models (e.g. gradual or direct isolation after speciation). First, we test the results of a species delimitation software (Zhang *et al.* 2013) and examine the parameter values (e.g. speciation rate) that make species delimitation perform poorly. Species delimitation, based on genomic information, is currently widely used due to the availability of massive amounts of genomic sequences and the need to classify them into known or new species. Second, we study how the gradual speciation model affects commonly used summary statistics. Gradual speciation has been recently proposed by Heled *et al.* (2013), and its effects on summary statistics have not been studied yet. Third, we illustrate how to infer parameter values when ancestral sampling is involved. With the availability of sequences from extinct species, we can now infer parameters related to the evolution of extinct species and their interactions with extant species. Here, using simulated data and the Approximate Bayesian Computation (ABC) framework, we study whether gene flow between an extinct and an extant species was present after speciation. Fourth, we use CoMuS in ABC to infer the speciation (birth) rate of a sample of species. Fifth, we use data from human and chimpanzee multiple sequence alignments and estimate their speciation time as well as the mutation rate. This application demonstrates the usage of CoMuS on real data. CoMuS and CoMuStats as well as their manual and examples are available at http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species and at bitbucket (https://bitbucket.org/idaios/comus).

Material and methods

Outline of the workflow

The workflow during a typical multispecies simulation process performed by CoMuS is described in Fig. 1. The user either provides the phylogenetic tree or the

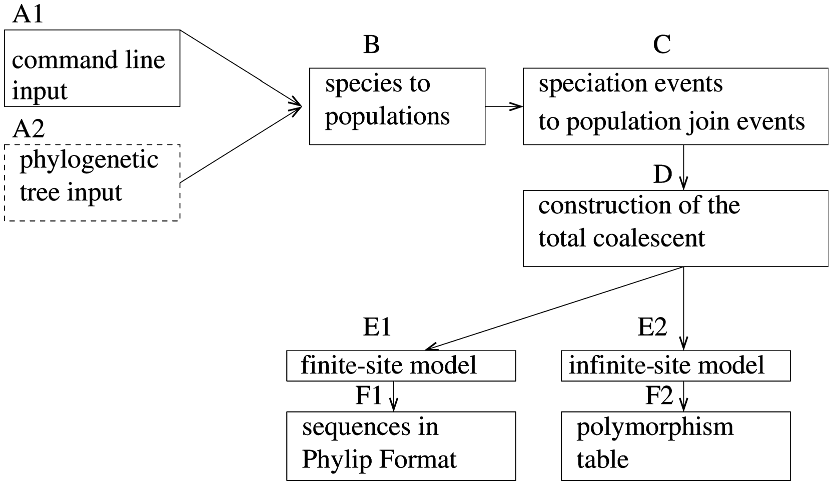

Fig. 1 The workflow during a typical multispecies simulation with CoMuS. Dashed-line boxes denote optional steps. Splits in the workflow denote alternative paths that a user may follow.

necessary parameters for its simulation. Then, species (as they are defined by the phylogenetic tree) are treated as populations and speciation events as population-merge events (backwards in time) and the construction of the total coalescent tree follows. When the infinite-site model is assumed, the result is a polymorphic table similar to the output of HUDSON'S MS. When the finite-site model is assumed, sequences are generated either in PHYLIP or FASTA format.

The guide phylogenetic tree

CoMuS implements the MSC using a guide phylogenetic tree as input. The phylogenetic tree delimits the species boundaries; specifically, it determines the speciation time points. After a speciation event, the user defines whether isolation happened instantaneously or gradually, as in Heled *et al.* (2013). In case of gradual isolation, the gene flow between sister species is reduced linearly until species are completely isolated. The guide phylogenetic tree is provided either by the user as input or it can be simulated. To simulate a phylogenetic tree, we assume a Yule process (Yule 1925) with extinction. The following parameters are needed to simulate a phylogenetic tree: birth rate b, death rate d, the number of species n and the proportion r of species sampled (Fletcher & Yang 2009).

Generating the guide phylogenetic tree. To simulate the guide phylogenetic tree, the following assumptions are made: (i) all sequences are sampled simultaneously at the present time point; (ii) both the birth and death rates are constant; (iii) as the birth–death process does not require a definite start or an end, *one of the following* conditions must hold to be able to simulate it according to our needs: (a) the number of sampled species is fixed; (b) the process starts at a specified time in the past; (c) the age of the most recent common ancestor (TMRCA) of the sampled species is known; (d) the time that the process starts is fixed at T_{origin}, and we condition on the sampled number of species; and (e) the process is not older than a certain time T_{oldest}. Condition (e) is implemented via a rejection algorithm, where only times younger than T_{oldest} are kept. This can be useful when the exact time point that the birth–death process begins is unknown, but we know that it cannot be older than a certain time point. Also, note that the time when the process starts (T_{origin}; case (b)) is different conceptually than the time of the most recent common ancestor of the sample (TMRCA; case (c)). The latter denotes the root of the tree, whereas the former denotes the initial point of the process which is older than the root of the phylogenetic tree (the root is the first branching event). Considering (d), we have also implemented a special case where T_{origin} is fixed to the value 1.0 expected substitutions per site. This condition has been developed first by Yang & Rannala (1997) and is implemented in INDELIBLE software by Fletcher & Yang (2009). The guide phylogenetic tree can also be provided by the user in newick tree format, and it is required to be rooted and ultrametric.

Mutation model

HUDSON'S MS assumes the infinite-site model to simulate mutation events on ancestral lineages. Consequently, each polymorphic site hosts two states. The infinite-site model is justified in *ms* because its goal is to model events that happen within a species boundaries, and thus, events that are relatively young in the evolutionary time scale. Indeed, the probability of more than one mutation at a given site is negligible for realistic mutation rate values. Contrary to single-species coalescent, multiple mutation events may occur frequently in a multispecies setup. Thus, CoMuS is able to simulate DNA sequences given an evolutionary mutation model (JC, F81, HKY, GTR), as well as two-state SNPs (infinite-site model). To obtain mutations under the finite-site model (and thus to generate nucleotide sequences), CoMuS adopts the following approach (see also below in Time unit conversion section): the user provides the guide phylogenetic tree (or the parameters to construct it) in typical phylogenetic units (expected substitutions/site). Then, time is converted to coalescent units to perform the usual coalescent process. Finally, time is converted back to phylogenetic units in order to put mutations on the branches of the total multispecies coalescent tree.

Time unit conversion

The usual time unit in coalescent theory is the number of generations divided by the effective population size *Ne* (or by 4*Ne* in *ms*). In other words, if the effective population size is *Ne*, then a time period t = 1 corresponds to *Ne* generations. Consequently, the generation of a coalescent tree becomes independent of the effective population size. On the other hand, in phylogenetics, time is measured in expected numbers of substitutions per site. For example, a branch of length 0.1 corresponds to a time period in which on average 0.1 substitutions per site occur. To model coalescent processes on a phylogenetic tree, it is necessary to use identical units for both the phylogenetic and the coalescent tree. Assume a branch of length β expected substitutions per site. If the mutation rate per site and per generation is μ, and the number of generations that correspond to the branch is γ, then:

$$\beta = \mu\gamma \rightarrow \beta/\theta = \mu\gamma/\theta \rightarrow \beta/\theta = \mu\gamma/4N_e\mu \rightarrow /|\theta = \gamma/4N_e$$

As $\gamma/4Ne$ represents time in $4N_e$ generations, we can divide the branch length (in phylogenetic units; denoted by β) by θ to obtain the branch length in coalescent units (Degnan & Rosenberg 2009).

Implementation of speciation events as population-merge events

CoMuS is using the *ms* machinery to build genealogies. As, in *ms,* the concept of species is absent, we treat species as distinct populations of a common origin. Thus, a speciation event involving species A and B is translated to a merge event between populations. When each species (A and B) contain a single population, the process of merging them is simple and equivalent to a single population-merge event in accordance with ms 'j' events (implemented by the −ej flag; see Fig. 2A). The situation is more complicated when a species consists of more than one population. In this case, by convention, we simultaneously merge each population to the population with the largest index (Fig. 2B). For example, species A with three populations (pop_1, pop_2, and pop_3) is generated after a speciation event. Backward in time, pop_1 and pop_2 will be fused into pop_3.

Gradual isolation after a speciation event

Recently, Heled *et al.* (2013) presented a model where speciation may occur over an extended time period. During this period, sister species are able to exchange genetic material, resulting in a gradual isolation model. Gradual isolation may be a more realistic model for describing the speciation of several species, including the one between human and chimpanzee (Patterson *et al.* 2006). Backward in time, there is a time period (specified by the user) that ends at the speciation time point, in which gene flow between different species is allowed. To sample times until the next lineage migration event, we model migration as a nonhomogeneous Poisson process with linear intensity (geneflow rate changes as a linear function of time from 0 to $\lambda_{\max}$). The rate of the Poisson process corresponds to the intensity (rate) of the migration. In brief, we first simulate a Poisson process with a maximum arbitrary intensity $\lambda_{\max}$ (provided by the user; by default the maximum rate value equals to 1.0). This means that we first propose a time t of the next migration event from an exponential distribution with rate $\lambda_{\max}$. Then, if t lies between speciation and complete isolation time points, we accept t as the time of the next migration event with a probability $p(t) = \lambda(t)/\lambda_{\max}$, where $\lambda(t) = \lambda_{\max}\ (t{-}t_{\text{isolation}})/(t_{\text{speciation}}{-}t_{\text{isolation}})$, where $t_{\text{isolation}}$ is the time where the two sister species become completely isolated; $t_{\text{speciation}}$ is the speciation time, that is the age of the node of the guide phylogenetic tree. It has been shown (Ross 2006) that the process of counted events corresponds to a nonhomogeneous Poisson process with intensity function $\lambda(t) = \lambda_{\max}p(t)$. In our case,

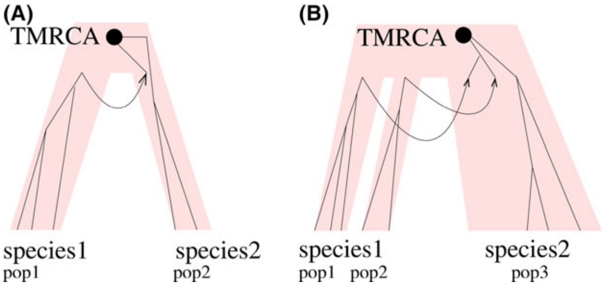

Fig. 2 Speciation model as implemented in CoMuS. In (A), each species comprises one population. Therefore, all lineages from population 1 (pop1) 'migrate' to population 2 (pop2) at the speciation time. In (B), species one consists of two populations, pop1 and pop2. Thus, two events take place at speciation time: The lineages of pop1 migrate to pop3, and at the same time, lineages from pop2 migrate to pop3. By convention, at each node of the guide tree (nodes define speciation events), lineages migrate to the population with the largest index (here, pop3).

we are interested only in the first event and the time needed for the first event to occur. This event represents gene flow during the postspeciation period, when the sister species have not been completely isolated. Further details, as well as the pseudocode of drawing times for the nonhomogeneous migration process, are provided in the manual of CoMuS (in Section: Partial isolation after a speciation event).

Ancestral sampling

Modern population genetics analyses often consist of both present-day and ancestral samples; for example, Neandertal sequences together with modern human sequences (Green *et al.* 2006; Wall & Hammer 2006; Noonan 2010; Hammer *et al.* 2011; Sankararaman *et al.* 2012; Prüfer *et al.* 2014; Fu *et al.* 2015; Pääbo 2015). CoMuS allows the simultaneous simulation of both modern and ancestral samples. The implementation starts at the present day with the whole data set (modern and ancestral). However, all evolutionary events (coalescent, recombination, gene flow, mutations) that involve ancestral samples or their population are forbidden until sampling (time proceeds backwards). After sampling (backwards in time), all evolutionary events take place according to the model parameters. Further details are provided both in the manual and the Supplement (Section: Ancestral sampling in CoMuS, Figs S1 and S2, Supporting information). Importantly, our implementation of ancestral sampling time refers to a whole (sub) population sample (e.g. a sample drawn from a subpopulation of a certain species). This implementation allows the recombination, migration, coalescent or mutation rates of other (sub) populations to remain unaffected.

Processing CoMuS output with CoMuStats

CoMuStats is based on libsequence (Thornton 2003) implemented in C++ and is able to calculate commonly used population genetics summary statistics (available from http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species). Thus, it allows the study of marginal distributions or the relations between summary statistics under various inter- and intraspecies models. CoMuStats can read the output of CoMuS (i.e. a file with multiple FASTA alignments separated by '//' or an ms-like file) and produces a table of summary statistics, where each line is associated with one data set and each column with one summary statistic. When more than one population is generated, summary statistics are calculated for each population separately, as well as for the total sample. Besides summary statistics, CoMuStats can calculate the site frequency spectrum for the total sample or for each population. Finally, given a window length and an offset value, CoMuStats can calculate the values of summary statistics in a sliding window. CoMuStats outputs result in a table format that can be readily processed by R and produce either marginal distributions or pairwise scatter plots. CoMuStats can be also used in the ABC framework for the calculation of summary statistics.

CoMuS as software testing tool

In evolutionary biology, simulations are often used to test inference tools. If the goal is to infer the value of a parameter of interest, we simulate data given the value of the parameter. We can assess how well the inference tool performs by counting and how often it is able to report the correct (known) value of the parameter. Here, we use CoMuS to test the performance of a species delimitation program, named Poisson Tree Processes (PTP) (Zhang *et al.* 2013), with respect to speciation (birth rate) and population genetics parameters (migration rate in population substructure). As PTP uses a phylogenetic tree to classify samples into species, we use both the true coalescent genealogy as it is produced by CoMuS and the inferred RAxML (Stamatakis 2014) phylogeny.

Using CoMuS as a part of the approximate Bayesian computation pipeline

Approximate Bayesian Computation methods generate samples from the posterior distribution of some parameter of interest. However, in ABC, the posterior distribution is just an approximation of the true posterior distribution. In the ABC methodology used in population genetics, we first simulate data from an evolutionary model (e.g. a species with two populations), assuming a prior distribution for the parameters of interest (e.g. a uniform distribution for the gene flow between the two populations). Next, both the observed and the simulated data are summarized by the same summary statistics. Simulated data (i.e. the set of their summary statistics) are sorted according to the (euclidean) distance from the observed data. Then, based on a degree of tolerance, we keep the simulated data that are most similar to the observed data and discard the rest. We use these data to estimate the posterior distribution utilizing simulated data using a regression step that corrects for the fact that they are not identical to the observed data (Beaumont *et al.* 2002; Csilléry *et al.* 2012; Duchen *et al.* 2013; Excoffier *et al.* 2005; Gray *et al.* 2014). Here, we use the 'abc' R package (Csilléry *et al.* 2012) for ABC inferences. The 'abc' package can perform both parameter inference and model selection. For model selection, the goal is to estimate the posterior probability of each competing model. For parameter inference, the goal is to estimate the posterior distribution of parameters of interest. The 'abc' package performs the model selection using three alternative approaches. The simpler approach is to use the rejection method, that is, to approximate the posterior distribution of each model by counting how often a certain parameter set (i.e. model) produces simulations similar to the observed data. The other two approaches (multinomial logistic regression and neural networks) treat the model selection as a discrete response variable which has to be predicted by the summary statistics (independent variables). For the inference of parameter values, 'abc' is using a rejection, a local linear regression or a neural network approach. In all cases, the entire data set is used as input. However, posterior distributions are estimated by only a (user-defined) fraction (tolerance level) of the data that is closer to the observation. Statistical tests such as cross-validation, goodness-of-fit and posterior predictive checks are implemented in the 'abc' package to assess the quality of the results.

CoMuS can be used as the simulation machinery of the ABC pipeline. As it can simulate data from multiple species, it can be used to infer parameters in models involving more than a single species in complex evolutionary scenarios. In this study, we demonstrate the usage of CoMuS in ABC by (i) inferring potential ancestral gene flow between an extinct species sample and an extant present-day sample. Also, (ii) we estimate the sampling time of the extinct species, that is to date the ancestral sample. We (iii) estimate the speciation time either on real data, between human and chimpanzee, or on simulated data.

Results and discussion

The examples provided below demonstrate the usage of CoMuS as (i) a software testing tool and (ii) a part of the ABC pipeline. All examples include both within- and between-species data sets. Scripts and commands used for the generation of simulated data and analysis can be downloaded from the website of CoMuS (http://popgen.eu/wordpress/software/comus-coalescent-of-multiple-species).

Testing species delimitation software

We used CoMuS to simulate (i) a sample of 20 sequences from two species and (ii) a sample of 50 sequences from five species. We sampled 10 sequences from each species. For each of the following scenarios, we modified a simulation parameter (birth rate in scenario I, population substructure in scenario II and III). Then, we applied PTP (Zhang *et al*. 2013) to perform species delimitation. PTP requires a phylogenetic tree to perform species delimitation. We either provide the correct rooted coalescent tree produced by simulations or the unrooted phylogenetic tree inferred by RAxML. In the latter, we use the option $-r$ of PTP that roots the tree on the longest branch. All the following results refer to the two-species case. Results from the five-species scenario are discussed at the end of the section, as well as in the Supplement (Section: Species delimitation with five species).

Scenario I—effect of birth rate. Speciation is represented by a Yule process with birth rate $b = \{0.002, 0.01, 0.1, 1, 10, 50, 100, 500\}$. Birth rate represents the rate with which lineages can split forward in time to generate a new species. Extinction rate, d, equals to 0. As we simulate a speciation process for two species, the process terminates just before the generation of the third species. Thus, the lower the birth rate, the more divergent the species are. The length of the simulated locus is 1000 bp, and we assume a simple mutation model with equal substitution rates (Jukes & Cantor 1969). The population mutation rate for the simulated locus is $\theta = 4N_e\mu = 5$, and there is no recombination.

Scenario II—effect of population structure. For scenario II, the goal is to test how PTP performs when there is population subdivision with various migration rates between the populations. The simulation parameters are the same as in scenario I. Here, however, we fix the birth rate to 0.01 (i.e. the phylogenetic model is described by a birth–death process with birth rate = 0.01 and death rate = 0) and assume that there is population subdivision in the first group of sequences (two subpopulations or demes with five sequences each). Migration rate values determine the gene flow between the two demes. The smaller the migration rate, the more isolated the demes are, thus, increasing the possibility of defining them as separate species. In this example, migration is symmetrical and the migration rate is defined by $M = 4N_em_{12} = 4N_em_{21}$, where N_e is the effective population size and m_{12}, m_{21} are the fraction of the population 1 that were made by immigrants from population 2 and the fraction of the population 2 made by immigrants from population 1, respectively. In general, m_{12} might be different than m_{21}. For our simulations, $M = \{0, 0.0001, 0.001, 0.005, 0.01, 0.05, 0.5, 1\}$.

Scenario III. The phylogenetic model in the previous scenario is a random process. Thus, if speciation takes place recently and populations merge relatively late, PTP may produce erroneous results that are difficult to interpret. That is, they can be explained either because of late population merge or because of recent speciation. To simplify the model, in Scenario III, we test the performance of PTP when the phylogenetic model has been fixed (divergence time between two species is set to $t_d = 0.3$ expected substitutions per site). As in Scenario II, species

1 consists of two populations. The migration rate between the two populations is given by $M = \{0, 0.0001, 0.001, 0.005, 0.01, 0.05, 0.5, 1\}$.

Table 1 reports the performance of PTP for scenario I (effect of birth rate), and Table 2 reports the performance of PTP for scenarios II and III (effect of population substructure with random simulated phylogenies and given phylogeny, respectively). Misclassification rate is a function of the proportion of sequences that have been classified erroneously. Details, as well as numerical examples, are given in the Supplement (Section: Testing species delimitation software; Formula 1 and Box 1). As a general rule, we observed that PTP tends to interpret populations as different species when the migration rate is low, regardless whether PTP is given the correct coalescent tree or the RAxML inferred phylogenetic tree. Interestingly, PTP reports more than 2 species for higher birth rate values (4.76 on average when birth rate is 500), even though, in this case, we would expect to underestimate the number of species. When birth rate value is small, groupings of leaves in the tree are clear and PTP can fit two different branching rates along the tree (for speciation and coalescent processes). However, when birth rate is large, there is no clear distinction between species. Thus, any grouping of leaves with similar length separated by longer branches could be interpreted as a separate species. Notably, in the five-species scenario, PTP overestimates the number of species when the RAxML tree is used (Table S3, Supporting information). A potential explanation for this is that RAxML assigns a very small branch length, b_0, between identical sequences. PTP interprets b_0 branches as within-species ancestral lineages and any longer branches as between-species lineages. Thus, it overestimates the number of species.

Table 1 Implementation of speciation with various birth rate values to test species delimitation software PTP

	PTP with true coalescent tree	PTP; inferred RAxML tree
Birth rate	av. # species (CI); av. Error	av. # species (CI); av. Error
0.002	2.04, (1.98,2.09), 0.005	2.02, (1.98,2.06), 0.003
0.01	2.04, (1.98,2.09), 0.002	2.02, (1.98,2.06), 0.001
0.1	2.33, (2.07,2.6), 0.037	2.08, (2,2.16), 0.012
1	2.47, (2.26,2.68), 0.065	2.12, (2.01,2.23), 0.014
10	4.71, (4.03,5.38), 0.262	2.6, (2.36,2.84), 0.078
50	6.02, (5.24,6.8), 0.393	3.7, (3.23,4.17), 0.271
100	6.41, (5.49,7.33), 0.479	4.56, (3.78,5.34), 0.42
500	4.76, (3.85,5.68), 0.612	4.36, (3.58,5.14), 0.661

The gradual speciation model

Testing the effect of gradual speciation on summary statistics. We assessed the effect of gradual speciation, (gradual isolation between two diverging species) on common summary statistics, and we compared their values with a model of instantaneous speciation, with or without subsequent gene flow. Two present-day species were simulated and 10 sequences were sampled from each species. The birth rate for the speciation model for all scenarios is 0.1, the mutation rate for a region of 1000 bp is 100 and recombination is absent. For the gradual isolation after speciation model, geneflow rate drops from a maximum value of 100 at the moment of speciation to 0 after 0.3 phylogenetic time units. We have also fixed the time for the TMRCA to 0.5 phylogenetic units. For the model

Table 2 Implementation of speciation with various geneflow levels between demes to test species delimitation software PTP

	Simulated phylogenetic tree (scenario II)		Provided phylogenetic tree (scenario III)	
	PTP with true coalescent tree	PTP; inferred RAxML tree	PTP with true coalescent tree	PTP; inferred RAxML tree
Migration rate	av. # species (CI); av. Error	av. # species (CI); av. Error	av. # species (CI); av. Error	av. # species (CI); av. Error
0	3.06, (2.99,3.13), 0.251	3.48, (3.17,3.79), 0.263	4, (3.65,4.35), 0.308	3.3, (3.15,3.45), 0.261
0.0001	3.06, (2.99,3.13), 0.25	4.06, (3.67,4.45), 0.281	3.96, (3.66,4.26), 0.288	3.28, (3.13,3.43), 0.257
0.001	3.12, (2.92,3.32), 0.255	4, (3.62,4.38), 0.267	4.34, (4.02,4.66), 0.299	3.34, (3.14,3.54), 0.261
0.005	2.96, (2.91,3.02), 0.24	4.12, (3.77,4.47), 0.232	3.98, (3.71,4.25), 0.299	3.38, (3.21,3.55), 0.266
0.01	3.02, (2.95,3.09), 0.241	3.74, (3.32,4.16), 0.169	3.6, (3.36,3.84), 0.272	3.28, (3.12,3.44), 0.252
0.05	2.8, (2.67,2.94), 0.175	2.84, (2.51,3.17), 0.102	3.96, (3.62,4.3), 0.28	3.28, (3.05,3.51), 0.215
0.1	2.53, (2.39,2.67), 0.117	2.26, (2.1,2.42), 0.043	4.26, (3.88,4.64), 0.30	3.28, (3.07,3.49), 0.23
0.5	2.27, (2.14,2.41), 0.043	2.08, (1.95,2.21), 0.006	3.68, (3.35,4.01), 0.206	2.64, (2.47,2.81), 0.103
1	2.1, (2.01,2.18), 0.014	2, (2,2), 0	3.4, (3.14,3.66), 0.173	2.8, (2.59,3.01), 0.11
10	2.2, (2.08,2.31), 0.028	2.06, (1.99,2.13), 0.008	3.78, (3.36,4.2), 0.19	2.7, (2.47,2.93), 0.092
100	2.18, (2.05,2.3), 0.023	2.12, (2.03,2.21), 0.013	3.32, (2.92,3.72), 0.143	2.46, (2.28,2.64), 0.047

Columns 1 and 2 refer to a model with unknown phylogenetic tree that is the phylogenetic tree simulated by CoMuS. In columns 3 and 4, the phylogenetic tree is assumed known and given in the command line.

with constant gene flow after speciation, the geneflow rate was set to 1. Density plots of eight widely used summary statistics: θ_W (Watterson 1975), θ_π (Tajima 1983), *Tajima's D* (Tajima 1989), *Fu and Li D**, *Fu and Li F** (Fu & Li 1993), *Wall's B*, *Wall's Q* (Wall 1999) and F_{ST} (Hudson *et al.* 1992) are illustrated in Fig. 3.

As shown in Fig. 3, most of the density plots for the gradual isolation after speciation are intermediate between the constant geneflow model and the instantaneous isolation after speciation model. For several summary statistics, such as θ_π, θ_W, *Wall's B* and *Wall's Q*, the density plot is more similar to that of the constant geneflow model. Interestingly, however, the F_{ST} density plot is closer to the respective plot from the isolated species. Thus, simulations show (at least with the parameter values tested) that summary statistics do not behave consistently for the gradual isolation model. This might complicate studies where ABC is used to infer parameter values or perform model selection; different summary statistics may support contradictory models.

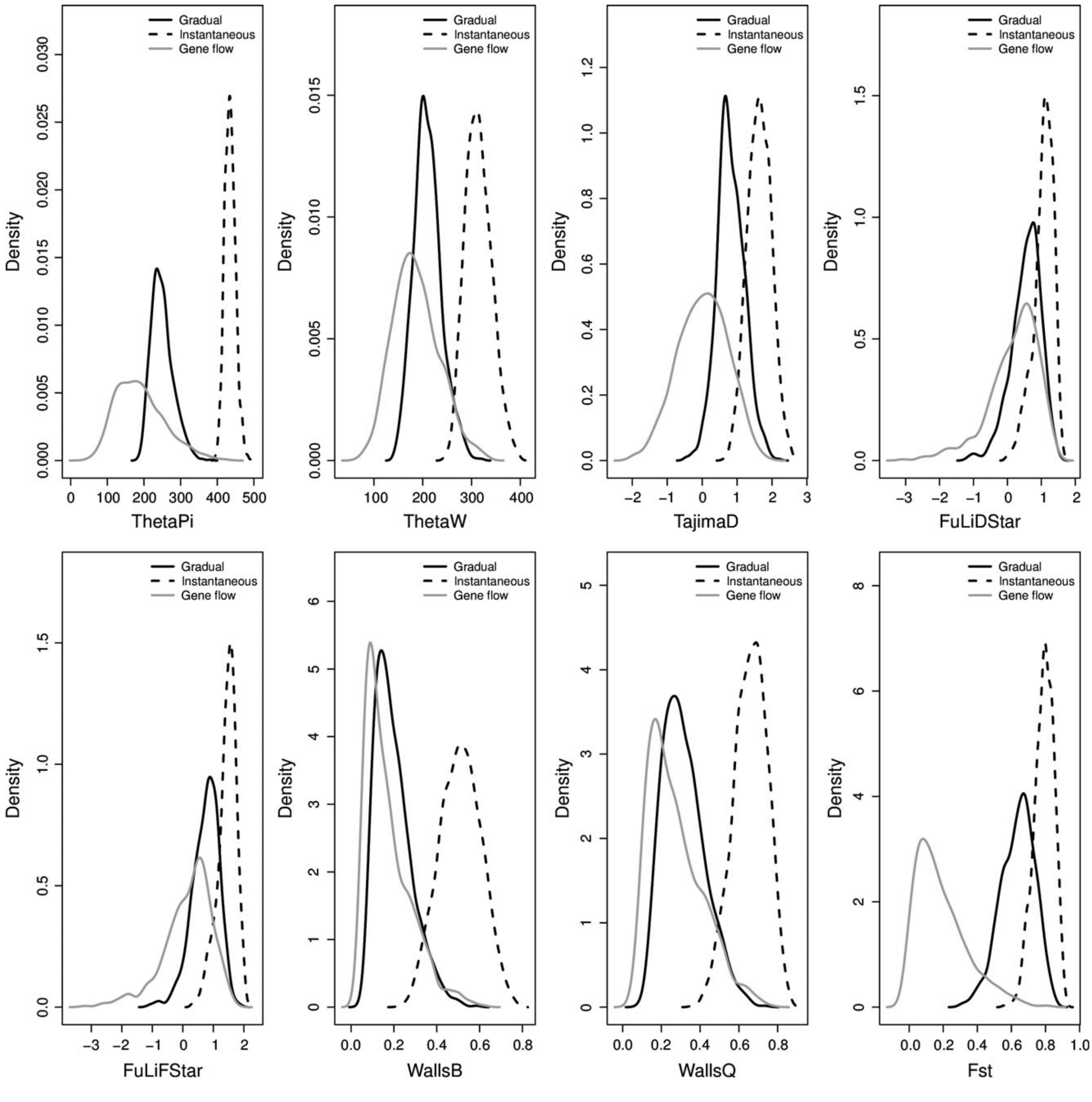

Fig. 3 Density plots of summary statistics values under a model of gradual isolation (solid line), instantaneous speciation and isolation (dashed line), and speciation with subsequent continuous gene flow (grey line). Details for the simulation parameters are given in the text.

Ancestral sampling

CoMuS is able to simultaneously simulate data from present-day samples and ancestral samples. Thus, it can be used to infer parameter values in studies where ancestral sampling is involved. Here, we demonstrate the usage of CoMuS to infer potential ancestral gene flow between an extant population and an extinct sample. The simulation scenario is as follows: we assume a sample of 10 sequences from species A sampled at present, and a sample of 10 sequences from an extinct species B sampled at time 0.2 (phylogenetic time units). The time of the MRCA has been set to 0.5 (phylogenetic time units). We assume no gene flow after speciation between species A and B. The dendrogram for this scenario is shown in Fig. S2 (Supporting information). Assuming that the above scenario represents the true evolutionary history for extant species A and extinct species B, our goal is to infer: (i) whether gene flow between A and B is absent or present and (ii) the time of sampling of species B. We first assume that the time of the MRCA (=0.5 phylogenetic units) is known. Then, we show that in this specific example, the success of the ABC model choice process depends critically on the right choice of TMRCA. We tried four values for the TMRCA: (i) 0.5, which is the 'true' value used in the simulations; (ii) 0.63, which is the value inferred to reconstruct the phylogenetic tree of the data and the sampling time 0.2 for species B is known; we then used two overestimated times (iii) 0.9 and (iv) 1.0, to demonstrate when model choice may fail. The population mutation rate value, θ (=100), is assumed to be known. The length of the simulated region is 1 kb, and we assumed a mutation model with equal mutation rates between each pair of bases (Jukes & Cantor 1969). We performed 100 000 simulations under each competing model (with and without gene flow between A and B). In the simulations where gene flow is present, we assume that it lasts for a time period equal to the time between sampling species B and the TMRCA. Then, we applied the 'abc' R package (Csilléry *et al.* 2012) to select the model that best explains the data and to infer the parameters' values. The total number of summary statistics (for each species and for the total sample) is 64. For model selection, we used the 'mnlogistic' algorithm and a tolerance threshold of 1%. We ran a cross-validation (100-fold) to assess whether the two models can be separated.

Table 3 summarizes the results. For each value of the TMRCA, we performed a cross-validation test to estimate the true-positive rate for each model (with and without gene flow). The greater the TMRCA, the greater the true-positive rate, because gene flow between the two population lasts for a longer period of time, and thus, it impacts the summary statistics more extensively. Therefore, it is easier to distinguish the two competing models. However, the greater the TMRCA, the more probable becomes the model *with* gene flow (i.e. the wrong model). In the (pseudo-) observed data, there is a certain degree of divergence between A and B. Divergence is captured by the summary statistics, and the two models compete to explain it. The greater the assumed value of TMRCA, the greater the divergence between A and B is in the model without gene flow. If its value is greater than a critical value, then the model without gene flow fails to fit the data. Therefore, the model with gene flow is favoured (Table 3, column 3).

When TMRCA is 0.5 (i.e. the correct value), and given the selected model (no gene flow between A and B), the sampling time of species B is inferred. We used the 'logit' transformation of the data with boundaries 0 and 0.5 to avoid meaningless inference (as we know that sampling time should be a positive value smaller than 0.5, which is the time of MRCA). The inferred mean value of sampling time was about 0.09. This means that the inference is not very accurate (as the true value was 0.2). It is possible to increase the accuracy of ABC by selecting more appropriate tolerance values and summary statistics and/or by increasing the number of simulations. Here, however, our goal is only to demonstrate the usage of CoMuS as a simulation tool for inferring parameter values and not the optimization of ABC protocol. Scripts (PERL and R) used for the analysis are provided at the CoMuS webpage http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species.

Estimation of parameters from simulated data

To demonstrate an application of CoMuS on inferring parameter values in the ABC framework, we simulated inter- and intraspecies data. We simulated two scenarios: (i) inference of the birth rate of the speciation process

Table 3 Results of the model choice process. Cross-validation as well as posterior probabilities of both models are provided. Model 1 is with gene flow, whereas model 2 is without gene flow. The correct (pseudo-observed) model is without gene flow

Time of the MRCA (TMRCA) in expected substitutions per site (phylogenetics units)	Cross-validation (true positive for model 1; true positive for model 2)	Posterior probabilities (model 1, model 2)
0.5	0.768; 0.729	0.111; 0.888
0.63[†]	0.802; 0.821	0.144; 0.855
0.8	0.864; 0.881	0.239; 0.760
0.9	0.894; 0.883	0.698; 0.301
1.0	0.926; 0.911	0.992; 0.007

[†] The value 0.63 was chosen because it corresponds to the root of the RAxML phylogenetic tree assuming that species A was sampled at the present and species B at time 0.2.

using data from two species (10 sequences per species) and (ii) inference of the birth rate and first speciation time (i.e. TMRCA) using data from 10 species (10 sequences per species).

Scenario I. For the first scenario, population mutation and recombination rates were fixed and assumed to be known (20 and 100, respectively). The length of the region was set to 1000 bp, and the mutation model assumes equal rates between all nucleotides (Jukes & Cantor 1969). The prior distribution of the birth rate b followed a log-uniform distribution (i.e. the logarithm of birth rate was distributed uniformly; $\log(b) \sim U(0, \log(500))$). The log-uniform distribution is useful when the parameter value spans several orders of magnitude, and we aim at weighting each order of magnitude equally *a priori*. To assess the accuracy of the method, we produced 1000 pseudo-observed data sets with birth rate $b = 5$ for each of them. Both the phylogeny as well as the coalescent was regenerated for each of the 1000 data sets (with the given $b = 5$). In the ABC framework, a parameter can be inferred using the mean, the median or the mode of the posterior distribution. For each of the (pseudo-) observed data set, we used the mode as a point estimator of the birth rate value. Figure 4 shows the (empirical) density of the inferred modes.

Scenario II. For the second scenario, mutation and recombination rates were fixed and assumed known. We generated 1000 (pseudo-) observed data set; each was consisted of 10 sequences sampled from each of the 10 species. The length of the region was 1000 bp, the population mutation rate for the region, θ, was 20 and the population recombination rate, ρ, was 100. Birth rate was set to 80, and we used the JC mutation model (Jukes & Cantor 1969). The TMRCA (root) of the (pseudo-) observed phylogenetic tree was 0.033 expected substitutions per site (phylogenetic units). Assuming the above scenario, we re-estimated the TMRCA and the birth rate of the speciation process (i.e. b and TMRCA were assumed unknown) for each of the (pseudo-) observations. The remaining parameters were assumed to be known. The prior distribution for the TMRCA was log-uniform within the range [0.0001, 1]. The prior for the birth rate b followed a uniform distribution U(0, 100). Figure 5 shows the empirical distribution of the median, mean and mode values for each of the parameters as they were inferred for all the (pseudo-) observed data sets. All command lines and results are available at http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species.

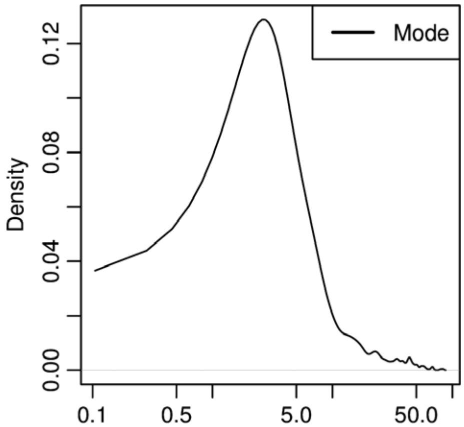

Fig. 4 Density distribution for the modes of the distributions of inferred birth rate values. We generated 1000 pseudo-observed data sets with birth rate $b = 5$. To infer b, we constructed 5000 simulated data in which $\log(b) \sim U(0, \log(500))$. For each of the 1000 pseudo-observed data sets, the posterior distribution of b was computed using the ABC framework (with the 'loclinear' method), and the mode values were extracted and used to generate the plot.

For both scenarios mentioned above, the inference is quite precise. The TMRCA in the case of 10 species can be estimated precisely by the median, the mean or the mode of the posterior distributions as the vast majority of the inferred values are between 0.01 and 0.04. Similarly, for the birth rate, the vast majority of the values are between 60 and 100. Thus, in the scenario with 10 species, both of the parameters can be estimated accurately for most of the (pseudo-) observed data sets. For the two species scenario, inference is quite accurate, even though the variance of the estimation is greater. Of 1000 mode values that we used for inferring the birth rate, 292 are below 2.5, 533 values are between 2.5 and 10, and 175 values are above 10; true value is 5. Thus, in about half of the inferences (467 of 1000), the birth rate value is estimated erroneously by at least a factor of 2.

Estimation of the speciation time between human and chimpanzee

We used 50 homologous gene regions between human and chimpanzee (kindly provided by Q. Zhu, personal communication; Table S1, Supporting information). Using all 50 fragments, we calculated the common population genetics summary statistics (Table S2, Supporting information). Simulations were performed with CoMuS, assuming a finite-site model (HKY; Hasegawa *et al.* 1985). The length of *simulated* fragments equals the average length of real fragments (1763 bp). Simulation parameters are given in Text S1 (Supporting information).

In this demo application, we focused on estimating the speciation time between two sister species as well as a total mutation parameter θ; therefore, we neglect the demographic history of each species. In a real application, however, demographic parameters for each species should also be inferred as they may affect the estimation of the speciation time, especially in recent speciation events. Using CoMuS and ABC, it is possible to infer simultaneously both the speciation time and the demographic history of each species. However, inferring such a complicated evolutionary history is out of the scope of this study.

Simulations were performed by drawing random variables (log-uniform) for the speciation time (TMRCA) and the total mutation rate (θ). Note that the units of the speciation time are the usual phylogenetic unit (i.e. expected substitutions per site). The prior distributions of both parameters were uniform on the log scale. However, we have conditioned on the presence of SNPs in the simulations, in order to be able to compute summary statistics. Thus, instances of very recent speciation times that produced no SNPs were not included in the analysis (i.e. the prior density of recent speciation times is lower). The median, mean and mode of the TMRCA is 0.0051, 0.0053 and 0.0026 expected substitutions per site, respectively. Regarding θ, the median, mean and mode values are 7.6×10^{-5}, 2.1×10^{-4} and 3.5×10^{-5} per base pair, respectively (Fig. 6). Assuming that the effective population size for humans is between 10 000 and 100 000, then the mutation rate per individual, per bp and per generation is comparable to the mutation rate estimated by

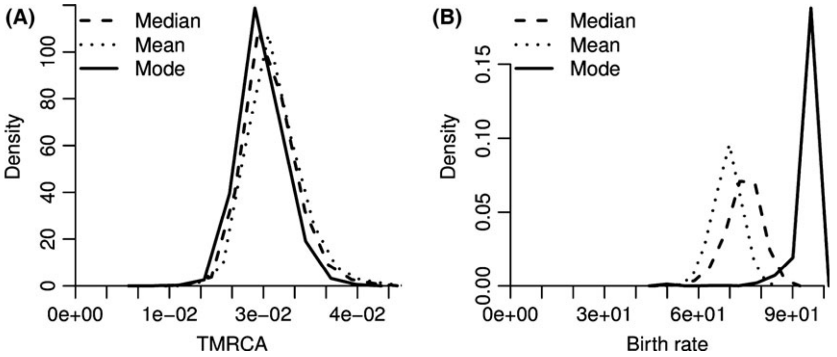

Fig. 5 Densities of the inferred mean, median and mode of (A) TMRCA and (B) birth rate. We generated 1000 pseudo-observed data sets with birth rate b = 80 and TMRCA = 0.033. To infer the parameters b and TMRCA, we constructed 100 000 simulated data in which $b \sim$ U(0, 100) and the $\log_{10}$ (TMRCA) ~ U(0.0001, 1), that is log-uniform in [0.0001, 1]. For each of the 1000 pseudo-observed data sets, the posterior distributions of b and TMRCA were computed using the ABC framework, and the mean, median and mode values were extracted and used to generate the plots. Both of the phylogenetic parameters are precisely estimated using the ABC framework.

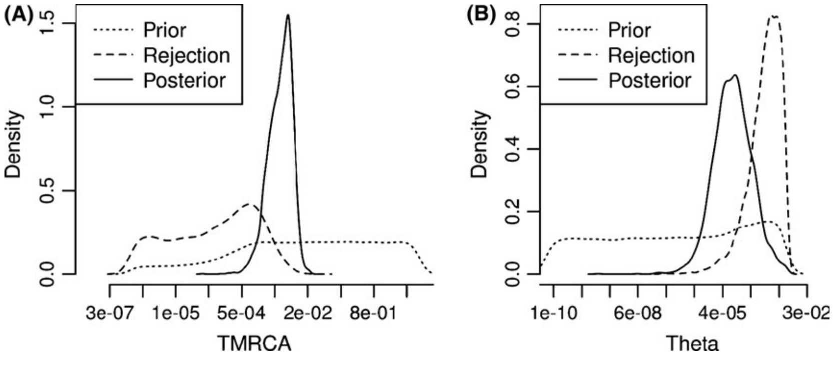

Fig. 6 Estimation of the speciation time (TMRCA) and the total mutation rate parameter (θ). Parameters were inferred using the ABC framework as it is implemented in the ‘abc’ package of the R statistical language. Simulations were performed by drawing random variables (log-uniform) for the speciation time (TMRCA) and the total mutation rate (θ). The x-axis in both plots is in log scale. (A) Inference of the speciation time: Dotted line is the prior distribution of the speciation time, dashed line is the distribution of the speciation time after the rejection step of ABC and the solid line represents the posterior distribution. Prior was uniform on the log scale, but we have conditioned on the existence of SNPs in the simulations. Thus, instances of very recent speciation times that produced no SNPs were not included in the analysis (i.e. the prior density of recent speciation times is lower). The median, mean and mode of the TMRCA is 0.0051, 0.0053 and 0.0026 expected substitutions, respectively. (B) Estimation of the total mutation rate θ: θ corresponds to the parameter $4N_e\mu$, where N_e is the effective population size and μ the mutation rate for the whole genomic region. In our context, however, N_e reflects a measurement of the ‘total’ population size. As the species are isolated for a long period of time, N_e is very large (s coalescent is not allowed). The median, mean and mode of θ are 7.6×10^{-5}, 2.1×10^{-4} and 3.5×10^{-5} per base pair, respectively.

other studies (e.g. about 2.0×10^{-8} in Nachman & Crowell (2000)). We should stress, however, that the total mutation rate parameter θ corresponds to the rate at which mutations occur on the ancestral lineages and therefore are responsible for both diversity (within-species variability) and divergence (between-species variability). Thus, we do not need to make any assumption about the divergence time between human and chimpanzee (TMRCA) in order to estimate the mutation rate. Both parameters are inferred simultaneously based on the within- and between-species patterns of summary statistics.

Limitations

The current version of CoMuS has some limitations that we will improve in future versions of the software. CoMuS can run in a single CPU and is not yet able to exploit multiple cores. Furthermore, the current version cannot simulate insertions and deletions (indels). Indels represent an important class of mutations. However, incorporating them in the current version of the software was challenging as their implementation requires a total redesign of the software to be able to properly handle overlapping indel events. Future versions of CoMuS will be able to fully utilize multicore computers and simulate indel events.

Even though it is possible to generate inter- and intraspecies data sets using *ms* and Seq-Gen in a pipeline (i.e. using the coalescent trees of *ms* as an input for Seq-Gen), this process has some limitations and cannot fully emulate the process that is implemented in CoMuS. CoMuS has several advantages over the pipeline approach: (i) it is simpler, as the user does not need to implement his own scripts to combine *ms* and Seq-Gen; (ii) it can simulate the phylogenetic tree under speciation models with various values of birth rate, death rate and species sampling proportion. It is also possible to read in a species tree from the command line; (iii) CoMuS implements the partial isolation model after a speciation event, thus allowing two sister species to exchange genetic material for some time after the speciation event; (iv) it allows sampling of present-day (extant) and ancestral (extinct) sequences; and (v) it can simulate sequences under a site rate heterogeneity model and a proportion of invariable sites (similar to Seq-Gen).

In our study, we emphasize the usage of CoMuS in an ABC framework, that is to estimate parameters using simulated data sets and summary statistics. Recently, Excoffier *et al.* (2013) proposed an alternative method to estimate parameters in complex models by simulations. In their method, they use the joint site frequency spectrum (joint-SFS) between different populations to estimate parameter values in a maximum-likelihood framework. Even though CoMuS could be used in a similar framework (by generating sequence data and then using a dedicated software to estimate the joint-SFS), we do not elaborate on this approach in the present study. This would require an extension of CoMuStats or the implementation of another software to calculate the joint-SFS and the likelihoods for each value of the parameters. Such an extension is currently out of scope. Heled & Drummond (2010) have implemented *BEAST (STARBEAST) to infer species trees from multilocus data. *BEAST implements the multispecies coalescent, but their study focuses on the inference of the species tree.

CoMus is written in C and it is freely available under the GNU GPLv3 licence. Its core machinery is based on HUDSON'S MS (Hudson 2002) and Seq-Gen (Rambaut & Grassly 1997). Together with CoMuS, we have developed CoMuStats, a C++ Libsequence software that can calculate common population genetics summary statistics from multi-FASTA-alignment files. Both programs, as well as scripts used in the analysis, are available at http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species.

Acknowledgements

We would like to thank Dr. Qihui Zhu (Harvard Medical School) for providing 50 gene alignments between human and chimpanzee. This project has received funding from the European Union's Seventh Framework Programme for Research, Technological Development and Demonstration under grant agreement No 316223. P. Pavlidis was supported by the FP7-PEOPLE-2013-IEF EVOGREN (625057). P. Poirazi was supported by the ERC Starting Grant dEMORY (ERC-2012-StG-311435).

References

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Birkner M, Blath J (2008) Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *Journal of Mathematical Biology*, **57**, 435–465.

Csilléry K, François O, Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, **3**, 475–479.

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. *PLoS Genetics*, **2**, e68.

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, **24**, 332–340.

Duchen P, Zivkovic D, Hutter S, Stephan W, Laurent S (2013) Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. *Genetics*, **193**, 291–301.

Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, **26**, 2064–2065.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genetics*, **9**, e1003905.

Fisher R (1930) The distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh*, **50**, 205–220.

Fletcher W, Yang Z (2009) INDELible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, **26**, 1879–1888.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Fu Q *et al*. (2015) An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, **524**, 216–219.

Gray MM, Wegmann D, Haasl RJ *et al*. (2014) Demographic history of a recent invasion of house mice on the isolated island of Gough. *Molecular Ecology*, **23**, 1923–1939.

Green RE, Krause J, Ptak SE *et al*. (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature*, **444**, 330–336.

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD (2011) Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 15123–15128.

Hartmann K, Wong D, Stadler T (2010) Sampling trees from evolutionary models. *Systematic Biology*, **59**, 465–476.

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.

Hein J, Schierup MH, Wiuf K (2004) *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, New York.

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

Heled J, Bryant D, Drummond AJ (2013) Simulating gene trees under the multispecies coalescent and time-dependent migration. *BMC Evolutionary Biology*, **13**, 44.

Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genetics*, **3**, e7.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Hudson RR, Boos DD, Kaplan NL (1992) A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution*, **9**, 138–151.

Jukes T, Cantor C (1969) Evolution of protein molecules. In: *Mammalian Protein Metabolism*(ed. Munro HN), pp. 21–132. Academic Press, New York.

Kessner D, Novembre J (2014) forqs: forward-in-time simulation of recombination, quantitative traits and selection. *Bioinformatics*, **30**, 576–577.

Kingman J (1982) On the genealogy of large populations. *Journal of Applied Probability*, **19**, 27–42.

Mossel E, Roch S (2007) Incomplete Lineage Sorting: Consistent Phylogeny Estimation From Multiple Loci. *arXiv*:0710.0262v2.

Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.

Noonan JP (2010) Neanderthal genomics and the evolution of modern humans. *Genome Research*, **20**, 547–553.

Pääbo S (2015) The diverse origins of the human gene pool. *Nature Reviews Genetics*, **16**, 313–314.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, **441**, 1103–1108.

Pavlidis P, Laurent S, Stephan W (2010) msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Molecular Ecology Resources*, **10**, 723–727.

Pitman J (1999) Coalescents with multiple collisions. *Annals of Probability*, **27**, 1870–1902.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S *et al*. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, **505**, 43–49.

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

Ross SM (2006) *Simulation*. Academic Press, San Diego.

Saminadin-Peter SS, Kemkemer C, Pavlidis P, Parsch J (2012) Selective sweep of a cis-regulatory sequence in a non-African population of *Drosophila melanogaster*. *Molecular Biology and Evolution*, **29**, 1167–1174.

Sankararaman S, Patterson N, Li H, Pääbo S, Reich D (2012) The Date of Interbreeding between Neandertals and Modern Humans. *PLoS Genetics*, **8**, e1002947.

Stadler T (2009) On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, **261**, 58–66.

Stadler T (2011) Simulating trees with a fixed number of extant species. *Systematic Biology*, **60**, 676–684.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, **30**, 1312–1313.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, **17**, 57–86.

Thornton K (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*, **19**, 2325–2327.

Wakeley J (2008) *Coalescent Theory: An Introduction*. WH Freeman, New York.

Wall JD (1999) Recombination and the power of statistical tests of neutrality. *Genetical Research*, **74**, 65–79.

Wall JD, Hammer MF (2006) Archaic admixture in the human genome. *Current Opinion in Genetics & Development*, **16**, 606–610.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Wright S (1938) The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences of the United States of America*, **24**, 253–259.

Yang Z, Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Molecular Biology and Evolution*, **14**, 717–724.

Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **213**, 21–87.

Zhang J, Kapli P, Pavlidis P, Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, **29**, 2869–2876.

S.P. and P.P. implemented the software. P.P. implemented the tests and the scripts. P.P., P.P. and S.P. wrote the manuscript.

Data accessibility

CoMuS is implemented in C programming language, and source code is available at http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species. The manual of CoMuS as well as CoMuStats are available within the main CoMuS directory as separate folders. Scripts used to generate the results in the manuscript are available at http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species. Most updated and experimental versions of the code (as well as the manual and CoMuStats) are available from an online repository at bitbucket.org (git clone git@bitbucket.org:idaios/comus.git).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Ancestral sampling in CoMuS.

Fig. S1 The implementation process of ancestral sampling. Initially, the simulations of all samples starts at present-day (A), even though individuals of species 1 (ind1, ind2, . . . , ind5) are actually sampled at the time denoted by the vertical line 'sampling time'.

Fig. S2 Example of a simulated multi-species dendrogram, where species 2 is a fossil (sampled in the past).

Fig. S3 The coalescent tree (A) and the raxml inferred tree (B) for a simulated dataset of two species, when species 1 is subdivided in two populations with gene flow M = 10 immigrants per generation on each population of species 1.

Table S1 List of 50 gene-alignments (human-chimp) used for the inference of mutation rate and speciation time in human-chimp phylogeny.

Table S2 List of summary statistics that were calculated for the ABC analysis.

Table S3 Performance of PTP with five simulated species (10 sequences per species).

Text S1 Command line that generates the simulations used for the estimation of speciation time and mutation rate.