CoMuS: Simulating coalescent histories and polymorphic data from multiple species

Journal:	*Molecular Biology and Evolution*
Manuscript ID:	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Papadantonakis, Stefanos; University of Crete, Biology Poirazi, Panayiota; Foundation for Research and Technology - Hellas (FORTH), Institute for Molecular Biology and Biotechnology (IMBB) Pavlidis, Pavlos; Foundation of Research and Technology - Hellas (FORTH), Institute of Molecular Biology and Biotechnology (IMBB)
Key Words:	Coalescent, Multi-Species Coalescent, Simulations, Approximate Bayesian Computation, Parameter Inference

Article
Methods

CoMuS: Simulating coalescent histories and polymorphic data from multiple species

Stefanos Papadantonakis[1], Panayiota Poirazi[2], Pavlos Pavlidis[2*]

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

*Corresponding author:

Pavlos Pavlidis
pavlidisp@gmail.com; pavlidis@imbb.forth.gr
Institute of Molecular Biology and Biotechnology (IMBB)
Foundation of Research and Technology - Hellas (FORTH)
Nikolaou Plastira 100, Vassilika Vouton
GR - 711 10, Heraklion, Crete, Greece

Keywords: Coalescent, Multi-Species Coalescent, simulations, Approximate Bayesian Computation, parameter inference

Running Title: CoMuS

Nonstandard abbreviations: MSC, ABC

Abstract

The advent of modern DNA sequencing technology is the driving force in obtaining complete intra- and inter-specific genomes. The simultaneous analysis of intra- and inter-species variation is difficult mainly because our knowledge about patterns of polymorphisms in samples where both intra- and inter-species sequences co-exist is limited. In the present study, we implement CoMuS (Coalescent of Multiple Species), a multi-species coalescent software that can simulate intra- and inter-species polymorphisms. CoMuS supports a variety of birth-death models for speciation events and demographic scenarios related to the history of each species. In CoMuS, speciation can be characterized either by instant isolation or by partial isolation between sister species. Our software supports both the infinite and the finite site model. As an application, we estimate parameters from simulated data, as well as the speciation time and mutation rate using human-chimpanzee alignments. Furthermore, we study the relations of summary statistics in multi-species scenarios. 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.

Availability: CoMuS is implemented in C programming language and source code is available on www.pop-gen.eu. Most updated and experimental versions of the code are available from an online repository at bitbucket.org (git@bitbucket.org:idaios/comus.git).

Introduction

Coalescent simulation represents a Monte Carlo process to generate samples drawn from a Wright-Fisher model of evolution. The most famous 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 to simulate a multitude of evolutionary scenarios, *ms* has been used in Approximate Bayesian Computation (ABC) inference methods to construct distributions of summary statistics under various evolutionary models (e.g. Pavlidis et al. 2010; Saminadin-Peter et al. 2012).

Here, we extend Hudson's *ms* to simulate multiple species, thus implementing the multi-species coalescent (MSC). Multi-species coalescent (Hobolth et al. 2007; Heled and Drummond, 2010) represents a retrospective model of evolution when speciation events have occurred in the ancestry of the sample. Thus, similar to single-species coalescent it effectively models random genetic drift within species, but it also allows for speciation events at a specific point in time, after which the two sub-species evolve in total isolation. Recently, Heled et al. (2013) proposed a modification of MSC to model speciation more realistically. Thus, Heled et al. (2013) allowed lineages from different species to remain in partial contact for a period after the speciation event. The study by Heled et al. (2013) focuses on phylogenetic applications of MSC and specifically on the effect of gene flow on phylogenetic inference and migration rate estimation. MSC has been exploited in several different applications of evolutionary biology. In phylogenetics, it has been used to study the effect of incomplete lineage sorting (Degnan and Rosenberg 2006; Hobolth et al. 2007; Heled and Drummond 2010; Mossel and Roch (unpublished work at *arxiv*) 2007), which is the failure of gene copies to coalesce within the species time boundaries and, thus, they coalescent in

ancestral species. In population genetics, Hobolth et al. (2007) analyzed a sample of human, chimpanzee, and gorilla sequences by considering the MSC of these three species and applying a hidden Markov model to infer ancestral effective population sizes and speciation times. Furthermore, Zhang et al. (2013) applied MSC simulations to assess the performance of their species delimitation method under various speciation rates. In this article, we propose CoMuS, an open-source MSC simulation software written in C programming language. CoMuS combines Hudson's ms with Rambaut's Seq-Gen (Rambaut and Grassly, 1997) and various birth-death speciation models presented by T. Stadler (Stadler 2009; Hartmann et al. 2010). Since it is based on Hudson's *ms*, it can efficiently implement recombination, population size changes, multiple-populations per species, gene flow (or species introgression). By implementing CoMuS, our aim is to provide a tool that can generate simultaneously intra- and inter-species samples. Thus, it can be used to infer parameters when both intra- and inter-species variation is available (for example, in an ABC framework), to study the effect of incomplete lineage sorting in phylogenetics tree reconstruction or to test species delimitation methods (e.g. Zhang et al. 2013).

New Approaches

CoMuS extends coalescent simulations to multiple species. It allows simulations under either the infinite site or the finite site model. Furthermore, most of the flexibility of Hudson's *ms* has been preserved. Thus, complicated evolutionary scenarios can be simulated efficiently. Simulations are guided by a phylogenetic tree, which specifies the species boundaries. The guide phylogenetic tree can be either simulated, under a variety of speciation models, or it is provided by the user. By using simulated guide trees, uncertainty about the speciation process is incorporated in the results. One important addition in CoMuS is the implementation of the gradual isolation model after a speciation event. Thus, two species can be generated at a

certain time point, but decreasing gene flow can continue until complete isolation. Finally, we have developed CoMuStats, a C++ software that can calculate summary statistics from the output of CoMuS, or from user's single- or multi-FASTA alignment files. For more details please consult the Materials and Methods section, as well as the manual and the examples that are provided with the software.

Results

CoMuS can be used in studies where random intra- and inter-species samples are required. We focus on two related applications: first, using an Approximate Bayesian Computation (ABC; Beaumont et al. 2002) framework we estimate the speciation time either on real data, between human and chimpanzee or on simulated data. Second, we study the pairwise relations of summary statistics that are commonly used in population genetics studies. We show that even though we simulate inter-species variation, summary statistics that are commonly used in population genetics studies are informative for the inference of the simulation parameters. ABC is used widely in population genetics studies (e.g. Excoffier et al. 2005; Duchen et al. 2013; Gray et al. 2014). The second application demonstrates the usability of CoMuS on studying the relations of summary statistics using simulations. All scripts and command lines used for the generation of simulated data and analysis can be downloaded from http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species.

Estimation of parameters from simulated data

For this application, we have simulated two scenarios: (i) inference of the birth rate of the speciation process using data from two species (10 sequences per species); and (ii) inference of the birth rate and first speciation time (i.e. time of the most recent common ancestor, TMRCA) using data from 10 species (10 sequences per species).

Scenario (i): For the first scenario, mutation and recombination rates were fixed and assumed to be known. The prior distribution of the birth rate *b* follows a uniform distribution, with parameters 0 and 500 (i.e. $b \sim U(0,500)$). To assess the accuracy of the method we produced 1,000 pseudo-observed datasets with $b = 5$ for each of them. In the ABC framework, a parameter can be inferred by using the mean, the median or the mode of the posterior distribution. Figure 1 shows the density of the inferred values when the mean, the median or the mode of the posterior distribution is used. Note that *b* is overestimated by all approaches. This overestimation reflects the prior U(0,500), which puts most of the density in large values of *b*.

Scenario (ii): For the second scenario, mutation and recombination rates were fixed and assumed known. Unknown parameters comprise the time of the root of the phylogenetic tree (TMRCA) and the birth rate of the speciation process. The prior for the TMRCA is log-uniform within the range [0.0001, 1]. Log-uniform priors are used in ABC for distributions that span several orders of magnitude and we wish to assign the same density on each order of magnitude. The prior for the birth rate *b* follows a uniform distribution U(0, 100). As in scenario (i), we generated 1,000 pseudo-observed datasets; *b* and TMRCA are 80 and 0.033, respectively, for each dataset. Figure 2 shows the distribution of the median, mean and mode values for each of the parameters.

Estimation of the speciation time between human and chimpanzee

We used 50 homologous gene regions between human and chimpanzee (kindly provided by Qihui Zhu, personal communication; Supplementary Table 1). Using all 50 fragments we calculated common population genetics summary statistics (Supplementary Table 2). 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 (1,763

bp). Simulation parameters are given in Supplementary Text 1.

In this demo application, we have been 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 since they may affect the inference 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 the current 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 is the usual phylogenetic unit (i.e. expected substitutions per site). The prior distribution for both parameters was uniform on the log-scale. However, we have conditioned on the existence of SNPs in the simulations 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. The total mutation rate parameter 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, though, N_e reflects a measurement of the 'total' population size. The median, mean and mode of θ is 7.6×10^{-5}, 2.1×10^{-4}, and 3.5×10^{-5} per base pair, respectively (Figure 3). In other words, θ reflects the rate at which mutations occur on the ancestral lineages of both human and chimpanzee from the present until their most recent common ancestor.

Generating scatterplots of summary statistics under various evolutionary models

CoMus, together with summary statistics calculation software, can be used to produce distributions and scatterplots of summary statistics under various evolutionary scenarios. We have implemented a C++ software, CoMuStats, based on Libsequence (Thornton 2003) that is able to calculate commonly used population genetics summary statistics. Thus, it is possible to study marginal distributions or the relations between summary statistics under various inter- and intra-species models. CoMuStats can read the output of CoMuS, i.e. a file with multiple FASTA alignments separated by '//'. It produces a table of summary statistics where each line is associated with one dataset and each column with one summary statistic. The output of CoMuStats can be readily processed by R and produce either marginal distributions or pairwise scatterplots. R scripts are available from http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species. CoMuStats can be also used in the ABC framework for the calculation of summary statistics. Supplementary Figure 1 demonstrates both CoMuS and CoMuStats on producing summary statistics scatterplots. The simulated scenario refers to 5 species with 15, 10, 10, 2, and 20 sampled sequences. The length of the simulated genomic region is 10,000 bp, recombination rate is 100 and mutation rate 20. The birth rate for the speciation process is 20.

Limitations and Discussion

The current version of CoMuS has some limitations that we will be improved in future versions of the software. CoMuS can run in a single CPU and is not able yet 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 since 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 multi-core computers and simulate indel events.

Even though it is possible to generate inter- and intra-species datasets 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 substitute CoMuS. CoMuS has several advantages over the simple pipeline approach: (i) it is simpler since 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) it automatically performs the time scale conversion; (iv) it 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. The rate of gene flow is maximum at speciation time and decreases linearly until the complete isolation. This model is more realistic than an instantaneous isolation after a speciation event (Heled et al. 2013).

Heled and Drummond (2010) have implemented *BEAST (StarBEAST) to infer species trees from multilocus data. *BEAST implements the multi-species coalescent but their study focuses on the inference of the species tree. In CoMuS we have focused on generating polymorphic data from multiple species, allowing the full flexibility of Hudson's *ms* software. Thus, CoMuS allows recombination, exponential size changes, introgression between different species (i.e. horizontal gene transfer), gene flow between different populations.

CoMus can simulate efficiently intra- and inter-species genomic variability under different evolutionary scenarios and mutation models. It is written in C and it is freely available under the GNU GPLv3 license. Its core machinery is based on Hudson's ms (Hudson 2002) and Seq-Gen (Rambaut and 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. We demonstrate the usage of CoMuS and CoMuStats in ABC and pairwise scatterplots of summary statistics. Both softwares as well as scripts used in the analysis are available from http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species.

Material and Methods

Outline of the workflow

The workflow during a typical simulation process performed by CoMuS is described in Figure 4.

The guide phylogenetic tree

CoMuS implements the multispecies coalescent 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 with birth and death events. As in INDELible (Fletcher and Yang 2009), 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 the number of species sampled.

Generating the guide phylogenetic tree: To simulate the guide phylogenetic tree, the following assumptions are made (as in TreeSim; Stadler 2011): (i) all sequences are sampled simultaneously at the present time point; (ii) both the birth and death rates are constant; (iii)

since 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 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 (e) the number of sampled species is fixed, and we condition on the fact that the process is not older than a certain time T_{oldest}. Considering the case (d) we have also implemented a special case where $T_{origin} = 1.0$. This condition has been developed firstly by Yang and Rannala (1997) and is implemented in INDELible software by Fletcher and Yang (2009). The guide phylogenetic tree can also be provided by the user in newick tree format, and it is required to be rooted.

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, 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 multi-species setup. Thus, CoMuS is able to simulate DNA sequences (and not just polymorphisms as in ms) given an evolutionary mutation model (JC, Jukes and Cantor 1969; F84, Felsenstein 1984; HKY, Hasegawa et al. 1985; GTR, Rodríguez et al. 1990).

Time unit conversion

The usual time unit in coalescent theory is the number of generations divided by the effective population size N_e (or by $4N_e$ in *ms*). In other words, if the effective population size is N_e, then

a time period $t = 1$ corresponds to N_e 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 0.1 substitutions per site occur on average. 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 l 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 are γ, then:

$l = \mu\gamma \rightarrow \frac{\gamma}{4N_e} = \frac{l}{4Ne\mu} = \frac{l}{\theta}$. Since, $\frac{\gamma}{4N_e}$ represents time in $4N_e$ generations, we can divide the branch length (in phylogenetic units) by θ to obtain the branch length in coalescent units.

Implementation of speciation events as population merge events

CoMuS is using the *ms* machinery to build genealogies. Since 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 to ms 'j' events (implemented by the -ej flag; see Figure 5A). On the other hand, the situation is more complicated when a species contains more than one populations. In this case, by convention, we merge simultaneously each population to the population with the largest index (Figure 5B).

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, thus creating a gradual isolation model after a speciation event. Gradual isolation may be a more

realistic model for speciation for several species including the speciation 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. The rate of gene flow is increasing linearly (backward in time). To sample times until the next lineage migration event we use a non-homogeneous Poisson process with linear intensity (gene flow rate changes as a linear function of time from 0 to λ_{max}). In brief, we first simulate a Poisson process with a maximum arbitrary intensity λ_{max} (provided by the user; by default the maximum rate value equals to 1.0). Then at time t, we accept values with a probability $p(t) = \lambda(t)/\lambda_{max}$, where $\lambda(t) = \lambda_{max} (t - t_{isolation}) / (t_{speciation} - t_{isolation})$, where $t_{isolation}$ is the time where the two sister species become completely isolated; $t_{speciation}$ is the speciation time, i.e. the age of the node. It has been shown (Ross, 2006) that the process of counted events corresponds to a non-homogeneous Poisson process with intensity function $\lambda(t) = \lambda_{max}p(t)$. In our case, we are interested only in the first event and the time needed to occur. This event represents gene-flow during the post-speciation period that the sister species have not been completely isolated.

Availability

CoMus as well as CoMuStats are freely available under the GPLv3.0 license and they are available from http://pop-gen.eu/wordpress/software/comus-coalescent-of-multiple-species. CoMuS is available also from bitbucket.org (git clone git@bitbucket.org:idaios/comus.git) with public access level. Also, we have constructed a public group forum on Google (https://groups.google.com/forum/#!forum/popgen-comus) where users can ask questions or report problems.

Acknowledgements

We would like to thank Dr. Qihui Zhu (Harvard Medical School) for providing 50 gene alignments between human and chimpanzee. This study was supported by the FP7 REGPOT-InnovCrete (No. 316223) grant, as well as by the FP7-PEOPLE-2013-IEF EVOGREN (625057) to P. Pavlidis.

Bibliography

1. Beaumont, MA., Zhang, W., & Balding, DJ. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025–35.
2. Degnan, J.H., Rosenberg, N. (2006) Discordance of species trees with their most likely gene trees. *PLoS Genet* 2:e68.
3. 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.
4. Excoffier, L., Estoup, A., Cornuet, J-M. (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* 169:1727–38.
5. Felsenstein, J. (1984) Distance methods for inferring phylogenies - a justification. *Evolution* 38:16-24.
6. Fletcher, W., & Yang, Z. (2009). INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*, 26:1879–88.
7. Gray, MM., Wegmann, D., Haasl R.J., White, M., Gabriel, S.I., Searle, J.B., Cuthbert, R.J., Ryan, P.G., Payseur, B.A. (2014) Demographic History of a Recent Invasion of House Mice on the Isolated Island of Gough. *Mol Ecol* 23:1923–39.
8. Hartmann, K., Wong, D., Stadler, T. (2010) Sampling trees from evolutionary models. *Syst Biol* 59:465–76.
9. Hasegawa, M., Kishino, H., Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22:160–74.
10. Heled, J., Bryant, D., Drummond, AJ. (2013) Simulating gene trees under the multispecies coalescent and time-dependent migration. *BMC Evol Biol* 13:44.
11. Heled, J., Drummond, AJ. (2010) Bayesian inference of species trees from multilocus data. *Mol Biol Evol* 27:570–80.
12. Hobolth, A., Christensen, OF., Mailund, T., Schierup, MH. (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet* 3:e7.
13. Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–8.
14. Jukes. H,, and Cantor, CR. (2006), *Evolution of Protein Molecules,* New York: Academic Press
15. Mossel, E., Roch, S. (2013) Incomplete Lineage Sorting : Consistent Phylogeny Estimation From Multiple Loci. *arxiv.org*.
16. Patterson, N., Richter, DJ., Gnerre, S., Lander, ES., Reich, D. (2006) Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441:1103–8.
17. Rambaut A, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* 13:235–8.

18. Rodríguez, F., Oliver, J. L., Marín, A., & Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142:485–501.

19. Ross, SM. (2006) *Simulation,* New York, Academic Press.

20. 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. *Mol Biol Evol* 29:1167–74.

21. Stadler, T. (2009) On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J Theor Biol* 261:58–66.

22. Stadler, T. (2011) Simulating trees with a fixed number of extant species. *Syst Biol* 60:676–84.

23. Thornton, K. (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–7.

24. Yang, Z., & Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol*, *14*(7), 717–24.

25. Zhang, J., Kapli, P., Pavlidis, P., Stamatakis, A. (2013) A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* 29:2869–76.

Figure Legends

Figure 1: Densities of the inferred mean, median and mode of the birth rate. We generated 1,000 pseudo-observed datasets with birth rate $b = 5$. To infer *b,* we constructed 100,000 simulated data in which $b \sim U(0, 500)$. For each of the 1,000 pseudo-observed datasets the posterior distribution of *b* was computed using the ABC framework, and the mean, median and mode values were extracted and used to generate the plot. All three measurements overestimate the value of *b*. This is because the prior distribution of *b* heavily favors large values of *b* ($b \sim U(0, 500)$).

Figure 2: Densities of the inferred mean, median and mode of TMRCA and birth rate. We generated 1,000 pseudo-observed datasets 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 $\log_{10}(\text{TMRCA}) \sim U(0.0001, 1)$, i.e. log-uniform in [0.0001, 1] . For each of the 1,000 pseudo-observed datasets 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.

Figure 3: Estimation of the speciation time (TMRCA) and the total mutation rate parameter (Theta). 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 or Theta). The x-axis in both plots are in log-scale. A) Inference of the speciation time. Dotted-

line is the prior distribution of the speciation time, dashed-line 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, though, N_e reflects a measurement of the 'total' population size. Since the species are isolated for a long period of time, N_e is very large (since coalescent is not allowed). The median, mean and mode of θ is 7.6×10^{-5}, 2.1×10^{-4}, and 3.5×10^{-5} per base pair, respectively.

Figure 4: The workflow during a typical multi-species simulation with CoMuS. Dashed-line boxes denote optional steps. Splits in the workflow denote alternative paths that a user may follow.

Figure 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).

Figures

Figure 1

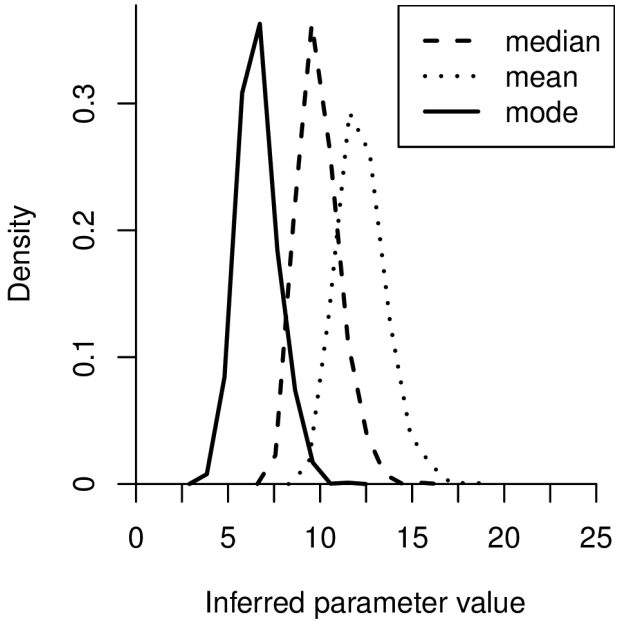

Figure 2

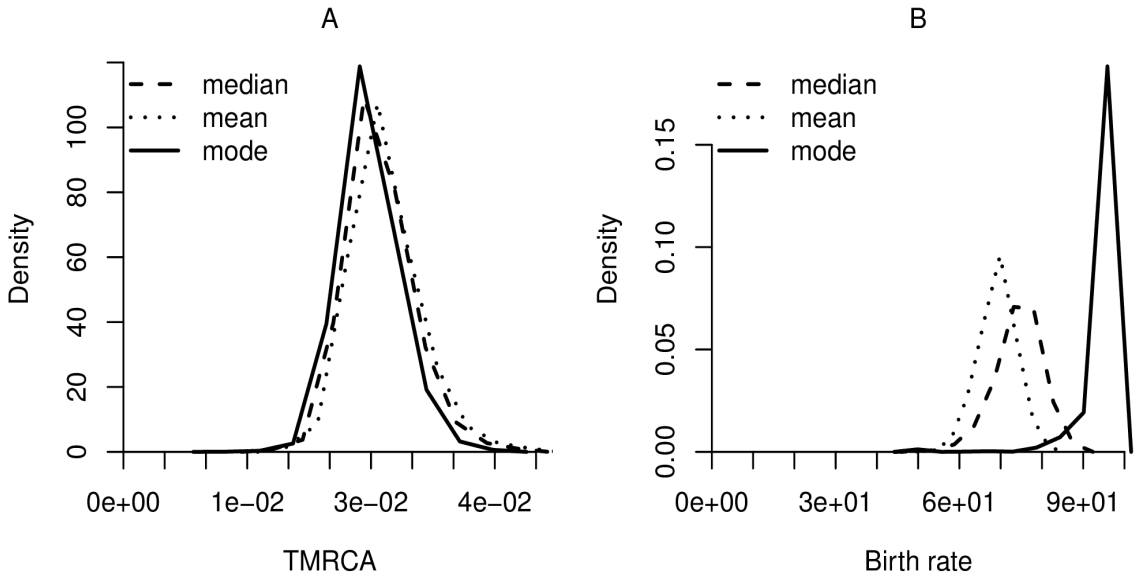

Figure 3

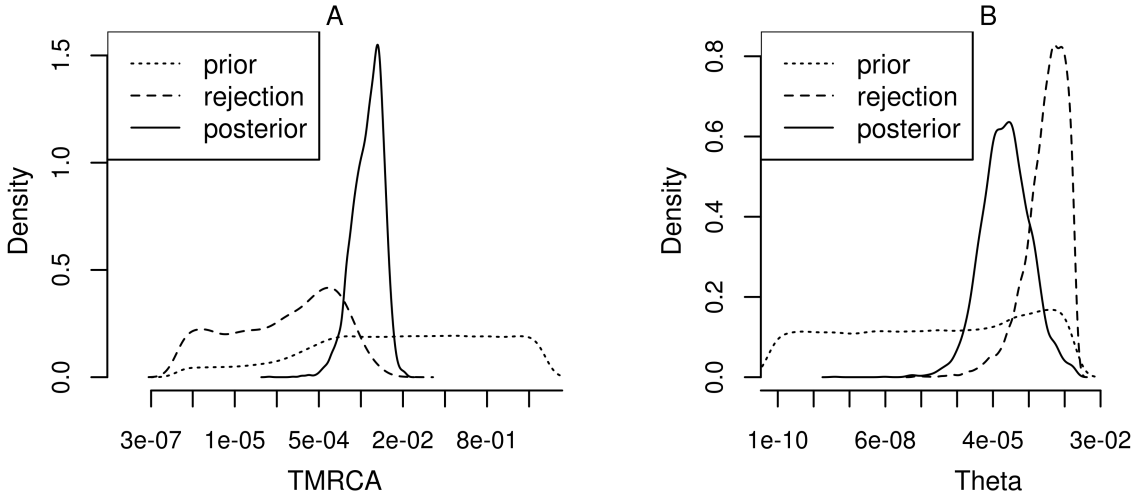

Figure 4

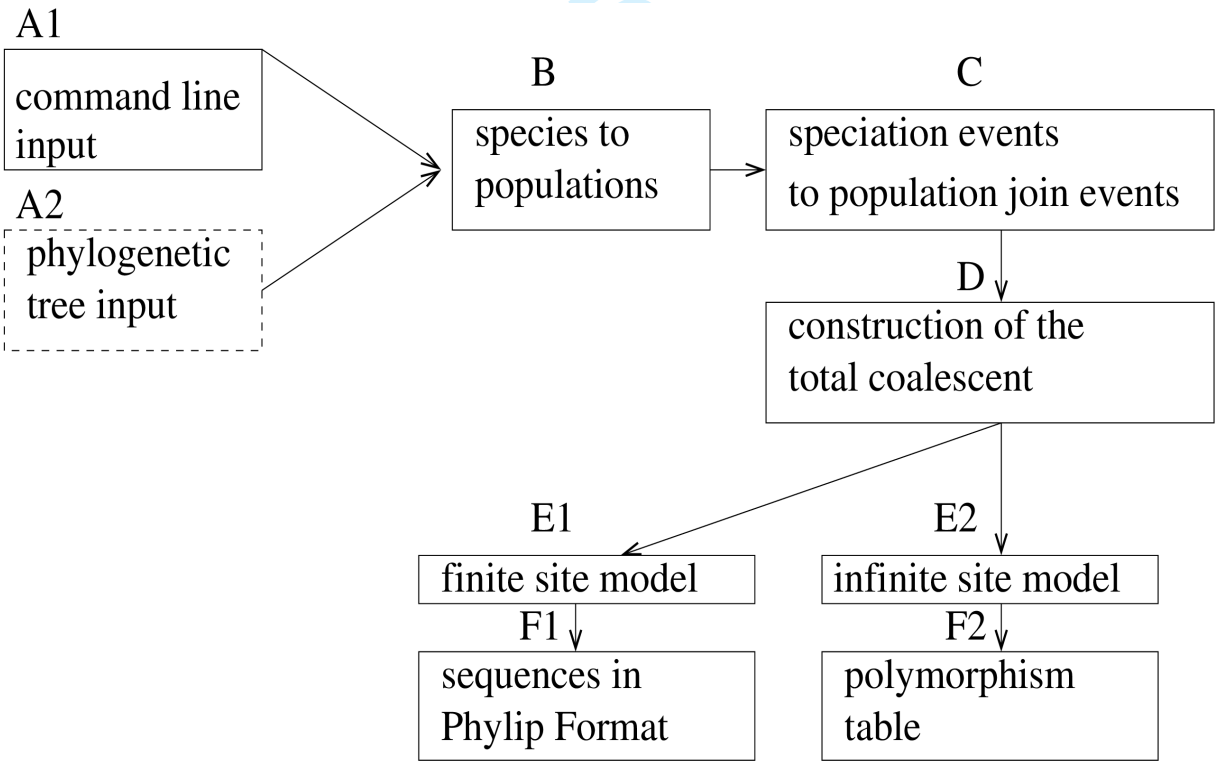

Figure 5