

# Approximate Bayesian Computation (ABC) in practice

Katalin Csilléry<sup>1</sup>, Michael G.B. Blum<sup>1</sup>, Oscar E. Gaggiotti<sup>2</sup> and Olivier François<sup>1</sup>

<sup>1</sup> Laboratoire Techniques de l'Ingénierie Médicale et de la Complexité, Centre National de la Recherche Scientifique UMR5525, Université Joseph Fourier, 38706 La Tronche, France

<sup>2</sup> Laboratoire d'Ecologie Alpine, Centre National de la Recherche Scientifique UMR5553, Université Joseph Fourier, 38041 Grenoble, France

**Understanding the forces that influence natural variation within and among populations has been a major objective of evolutionary biologists for decades. Motivated by the growth in computational power and data complexity, modern approaches to this question make intensive use of simulation methods. Approximate Bayesian Computation (ABC) is one of these methods. Here we review the foundations of ABC, its recent algorithmic developments, and its applications in evolutionary biology and ecology. We argue that the use of ABC should incorporate all aspects of Bayesian data analysis: formulation, fitting, and improvement of a model. ABC can be a powerful tool to make inferences with complex models if these principles are carefully applied.**

## Inference with simulations in evolutionary genetics

Natural populations have complex demographic histories: their sizes and ranges change over time, leading to fission and fusion processes that leave signatures on their genetic composition [1]. One 'promise' of biology is that molecular data will help us uncover the complex demographic and adaptive processes that have acted on natural populations. The widespread availability of different molecular markers and increased computer power has fostered the development of sophisticated statistical methods that have begun to fulfill this expectation. Most of these techniques are based on the concept of likelihood, a function that describes the probability of the data given a parameter.

Current approaches derive likelihoods based on classical population genetics or coalescent theory [2,3]. In recent years, likelihood-based inference has frequently been undertaken using Markov chain Monte Carlo (MCMC) techniques [2,4]. Many of these methods are, however, limited by the difficulty of computing the likelihood function, thus restricting their use to simple evolutionary scenarios and molecular models. Additionally, even with ever-increasing computational power, these techniques cannot keep up with the demands of the large amounts of data generated by recently developed, high-throughput DNA sequencing technologies. Both of these factors have stimulated the development of new methods that approximate the likelihood [4].

One of the most recent approaches is Approximate Bayesian Computation (ABC [5]). ABC approaches bypass

exact likelihood calculations by using summary statistics and simulations. Summary statistics are values calculated from the data to represent the maximum amount of information in the simplest possible form. Their use goes

## Glossary

**Bayes factor:** the ratio of probabilities of two models that is used to evaluate the relative support of one model in relation to another in Bayesian model comparison.

**Bayesian statistics:** a general framework for summarizing uncertainty (prior information) and making estimates and predictions using probability statements conditional on observed data and an assumed model.

**Coalescent theory:** a mathematical theory that describes the ancestral relationships of a sample of 'individuals' back to their common ancestor. Individuals may represent molecular marker loci, genes, or chromosomes depending on the context.

**Credible interval:** a posterior probability interval used in Bayesian statistics that can be directly constructed from the posterior distribution. For example, a 95% credible interval for the parameter  $\theta$  means that the posterior probability that  $\theta$  lies in the interval is 0.95.

**Deviance Information Criterion (DIC):** an information theoretic measure used to determine if improvement in model fit justifies the use of a more complex model whereby model complexity is expressed with a quantity related to the number of parameters.

**Dimension reduction:** the mathematical process of transforming a number of possibly correlated variables into a smaller number of variables. A well-known example of such methods is principal component analysis, but other methods such as feed-forward neural networks (FFNNs) or partial least-squared regression (PLS) are used in the analysis of complex genetic data. FFNNs are flexible non-linear regression models. PLS is a regression method for constructing predictive models in the presence of many factors.

**Effective population size ( $N_e$ ):** the size of an idealized Wright-Fisher population that has the same level of genetic drift as the population in question.

**Hierarchical models:** models in which the parameters of prior distributions are estimated from data rather than using subjective information. Hierarchical models are central to modern Bayesian statistics and allow an objective approach to inference.

**High-throughput genotyping:** a common name for recently developed, next-generation sequencing technologies that provide massively parallel sequencing at low cost and without the requirement for large, automated facilities.

**Markov Chain Monte Carlo (MCMC):** an iterative Bayesian statistical technique that generates samples from the posterior distribution. Well-designed MCMC algorithms converge to the posterior distribution, which is independent of the starting position.

**Posterior distribution:** the conditional distribution of the parameter given the data, which is proportional to the product of the likelihood and the prior distribution.

**Posterior predictive distribution:** the distribution of future observations conditional on the observed data.

**Prior distribution:** the distribution of parameter values before any data are examined.

**Statistical phylogeography:** an interdisciplinary field that aims to understand the processes underlying the spatial and temporal dimensions of genetic variation by combining information from genetic, ecological and paleontological data.

**Sufficient statistics:** a statistic is sufficient for a parameter,  $\theta$ , if the probability of the data, given the statistic and  $\theta$ , does not depend on  $\theta$ . In other words, a sufficient statistic provides just as much information to estimate  $\theta$  as would the full dataset.

**Box 1. Inferring the demographic history of *Drosophila melanogaster*****Selection or demography?**

Understanding the relative importance of selective and demographic processes raises considerable interest, however disentangling the effects of the two processes poses a serious challenge. This is because selection and demography may result in similar patterns of nucleotide variability. Developing models that take into account demographic processes is essential to identify regions of the genome under selection. For *Drosophila melanogaster*, Thornton and Andolfatto [16] propose a simple model of the demographic history and, using ABC and Bayesian evaluation of the goodness of fit, they show that their model can predict the majority of the observed patterns of molecular variability without invoking selective processes.

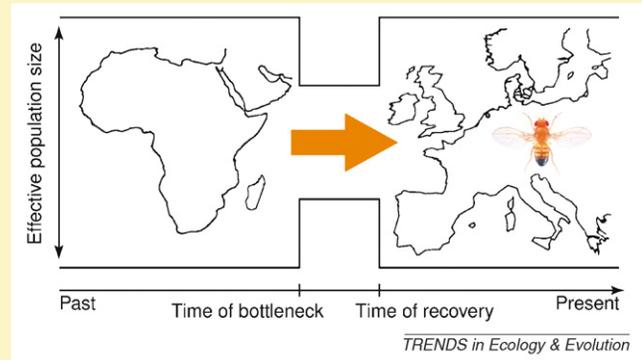
**The bottleneck model**

In the cosmopolitan *D. melanogaster*, molecular variation is substantially reduced outside of Africa. This suggests an African origin and a reduction in population size when colonizing other continents. The population model includes a reduction in population size, and then for simplicity a recovery to the same effective population size as before (bottleneck, Figure 1). Timing of the bottleneck can provide an estimate of the date of colonization of Europe, from where the authors have obtained samples of molecular data.

**Dating the colonization event**

Biogeographical studies suggest that *D. melanogaster* colonized the rest of the world only after the last glaciation period (about

6,000–10,000 years ago) [101]. Using patterns of variability observed at 115 loci scattered across the X chromosome in a European population of *D. melanogaster*, Thornton and Andolfatto estimate the start of the bottleneck to be 16,000 years before the present, associated with a 95% credible interval of 9,000–43,000 years [16]. Thus, molecular data suggest that the colonization of Europe might have occurred before the last glaciation period.



**Figure 1.** The bottleneck model of *D. melanogaster* colonizing Europe. Image of *D. melanogaster* reproduced with permission from Nicolas Gompel ([http://www.ibdml.univ-mrs.fr/equipes/BP\\_NG/](http://www.ibdml.univ-mrs.fr/equipes/BP_NG/)).

back to the origins of population genetics when, for instance, Sewall Wright [6] devised fixation indices to describe correlations among alleles sampled at hierarchically organized levels of a population. The use of simulations, both as artificial experiments in evolution and inference tools, also has a long tradition in population genetics [7].

**The ABC of Approximate Bayesian Computation**

ABC has its roots in the rejection algorithm, a simple technique to generate samples from a probability distribution [8,9]. The basic rejection algorithm consists of simulating large numbers of datasets under a hypothesized evolutionary scenario. The parameters of the scenario are not chosen deterministically, but sampled from a probability distribution. The data generated by simulation are then reduced to summary statistics, and the sampled parameters are accepted or rejected on the basis of the distance between the simulated and the observed summary statistics. The sub-sample of accepted values contains the fitted parameter values, and allows us to evaluate uncertainty on parameters given the observed statistics.

The applications of ABC are often based on improved versions of the basic rejection scheme [5,10–13], and have already yielded valuable insights into various questions in evolutionary biology and ecology (Boxes 1–2). Examples include the estimation of various demographic parameters such as: the effective population size [14]; the detection and timing of past demographic events (e.g. growth or decline of populations [9,15–22]); or the rate of spread of pathogens [23,24]. Other applications have compared alternative models of evolution in humans [25–36]; inferred admixture proportions [37–39]; migration rates [40,41]; mutation rates [9]; rates of recombination and gene conversion

[42–44]; the strength of positive selection [45]; the influence of selection on gene regulation in humans [46]; or the age of an allele [27,47]. ABC has also been used to make inferences at an inter-specific level, for example: dating the divergence between closely related species [48–52]; making inferences about the evolution of polyploidy [53]; and identifying a biodiversity hotspot in the Brazilian forest [54]. In ecology, ABC has been used to infer parameters of the neutral theory of biodiversity in tropical forests ([55], Box 2).

Although the use of ABC is widespread in many areas of evolutionary biology, it has become contentious in the field of statistical phylogeography [56,57] (Box 3). The main objections to ABC are that inference is limited to a finite set of phylogeographical models, and that models hypothesized in ABC studies are complex. As a consequence, conclusions from ABC could be influenced by subjective inclusion of evolutionary scenarios and implicit model assumptions that are not foreseen by the modeler. A recent article addresses these concerns in great detail, and points out that they are general criticisms of model-based approaches and are not specific to ABC [58]. We acknowledge that difficulties can arise at several levels if using a simulation-based method such as ABC, thus the principal aim of this review is to encourage good practice in the use of the method. Here, we review some elementary Bayesian principles and emphasize some facets of Bayesian analysis that are often neglected in ABC studies. Then we highlight ways to improve inference with ABC via application of these principles.

**Bayesian data analysis: building, fitting, and improving the model**

The three main steps of Bayesian analysis are formulating the model, fitting the model to data, and improving the

## Box 2. Inferring the parameters of the neutral theory of biodiversity

### Hubbell's theory

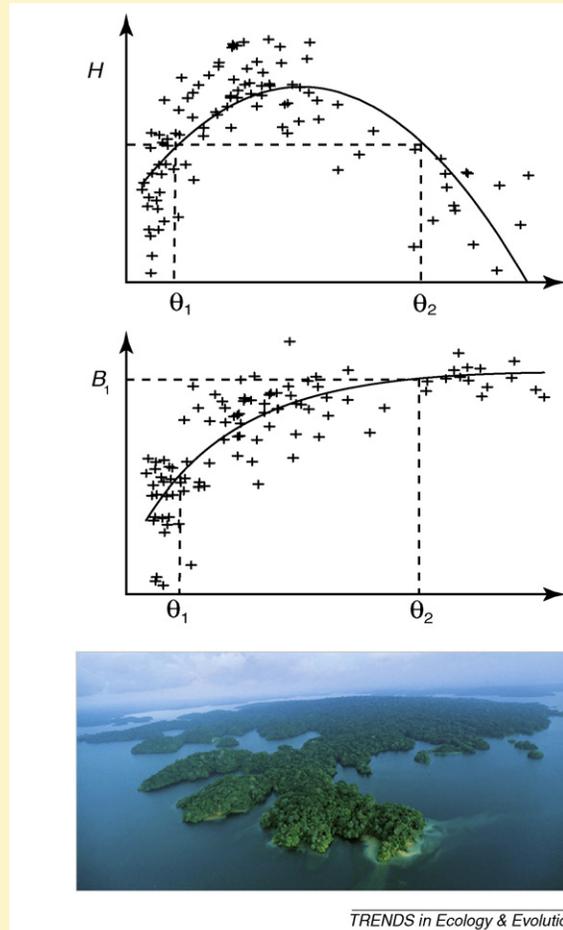
In community ecology, the unified neutral theory of biodiversity (UNTB) [102] states that all individuals in a regional community are governed by the same rates of birth, death and dispersal. The UNTB is a stochastic, individual-based model that gives quantitative predictions for species abundance data. The theory has a practical merit for the evaluation of biodiversity: it provides a fundamental biodiversity number. The biodiversity number can be interpreted as the size of an ideal community that best approximates the real, sampled community [103].

### Inference of the fundamental biodiversity number

Inference under the UNTB encounters a serious theoretical issue: species abundance data generally leads to two distinct and equally likely estimates of the biodiversity number. One solution is to use *a-priori* information to discard the unrealistic estimate of the biodiversity number. Jabot and Chave [55] provide an elegant alternative solution to the estimation of the biodiversity number by combining phylogenetic information with ABC. They find that the phylogenetic relationships between the species living in the local community convey useful information for parameter inference. Exploiting the information contained in 'tree imbalance' statistics leads to a unique, most-likely estimate of the biodiversity number (Figure 1).

### Application of ABC to tropical rainforests

Using ABC, Jabot and Chave [55] give an estimate of the biodiversity number of the regional pool for two tropical-forest tree plots in Panama and Columbia. The biodiversity numbers are one order of magnitude greater than those computed by maximum-likelihood methods that do not use phylogenetic information. One explanation for those larger values is that the Central American plots extend over areas of the size of the entire Neotropical ecozone. The two forest plots in Central America would then be connected *via* dispersal to a large ecozone including Central and South America.



**Figure 1.** Estimation of the biodiversity number,  $\theta$ , for communities of tropical trees in Barro Colorado Island (Panama) with ABC. While the variation of the species evenness (measured by Shannon diversity,  $H$ ) leads to two likely values of the biodiversity number ( $\theta_1$  and  $\theta_2$ ), the use of a measure of phylogenetic tree balance,  $B_1$ , leads to a unique most-likely biodiversity number ( $\theta_2$ ). Photograph reproduced with permission from Christian Ziegler (<http://www.naturphoto.de>), and figure reproduced with permission from Wiley-Blackwell [55].

model by checking its fit and comparing it with other models [59] (Box 4). When formulating a model, we use our experience and prior knowledge, and sometimes resort to established theories. This step is often mathematical in essence because it encompasses explicit definitions for the likelihood (or a generating mechanism) and the prior distribution. These quantities summarize background information about the data and the parameter values. Fitting the model is at the heart of Bayesian data analysis and, in modern approaches, it is often carried out with Monte Carlo algorithms. The aim of this second step is to calculate the posterior probability distribution and a credibility interval for the parameter of interest, starting with its prior distribution, and updating it based on the data [2]. Models are always approximations of reality, so evaluating their goodness-of-fit and improving them is the third step of a Bayesian analysis. In this step, we often need to confront two or more models that differ in their level of complexity. The three steps of a Bayesian analysis are strongly inter-dependent and should be considered as a unified approach, with a possibility of cycling through the three stages.

### Model building

Two intimately linked questions arise when formulating or improving models. First, why do we want models? Second, how complex should the models be? Models can be used towards two distinct ends: explanation and prediction [60]. Although we might be concerned with prediction when investigating the decrease of biodiversity or the consequences of global change [61], evolutionary models tend to be explanatory, i.e. they are used to help describe the evolutionary processes that have generated the data. There are often several potential explanatory models of a phenomenon, and model formulation is not restricted to hypothesizing a unique scenario. Often, many different explanatory models can be proposed, with the main objective of finding the most parsimonious explanation. As Einstein nicely stated, "models should be as simple as possible, but not more so" [62].

### Model fitting

ABC algorithms can be classified into three broad categories. The first class of algorithms relies on the basic rejection algorithm [8,9]. Technical improvements of this

### Box 3. Controversy surrounding ABC

The ABC approach has been vigorously criticized by Templeton [56,57] and forcefully defended by a group of statistical geneticists [58]. The purpose of this Box is, therefore, to acknowledge the existence of this controversy and to provide a brief comment on the criticisms that we deem most relevant.

The issue that is most thoroughly discussed is the testing of complex phylogeographic models *via* computer simulation. According to Templeton, the impossibility of including an exhaustive set of alternative hypotheses in the testing procedure forces us to choose only one subset, introducing a great deal of subjectivity in the process. Moreover, in this situation one runs the risk of not including the true model in the restricted set of models, in which case the results of the test would be meaningless. According to Beaumont

et al. [58], this situation is applicable to all model-based methods whether they rely on simulations or not. Although in principle this is a potentially important problem, in reality scientific arguments often revolve around a limited number of hypotheses or scenarios without the need to consider an infinite set of alternative models. The issue then becomes the possibility of model mis-specification, something that is also addressed by Templeton but again is not restricted to ABC approaches. This problem can in part be addressed by using the many statistical techniques for assessing the fit of a model [58]. Models can always be improved and refined by other authors, allowing an open discussion that can greatly increase our understanding of the problem being studied. This is the way scientific progress is made.

basic scheme correct for the discrepancy between the simulated and the observed statistics by using local linear or non-linear regression techniques (Figure 1) [5,10,63]. A second class of methods, ABC-MCMC algorithms, explore the parameter space iteratively using the distance between the simulated and the observed summary statistics to update the current parameter values [11,64]. In this way, parameter values that produce simulations close to the observed data are visited preferentially. As in standard

MCMC methods, ABC-MCMC requires that the Markov chain converges to its stationary state, a condition that can be difficult to verify [4]. A third class of algorithms is inspired by Sequential Monte Carlo methods (SMC) [65]. SMC-ABC algorithms approximate the posterior distribution by using a large set of randomly chosen parameter values called ‘particles’. These particles are propagated over time by simple sampling mechanisms or rejected if they generate data that match the observation poorly.

### Box 4. Tutorial on ABC: estimating the effective population size ( $N_e$ )

Estimating the effective population size from molecular data is of interest to many evolutionary biologists. Here, we propose a sample of 40 haploid individuals genotyped at 20 unlinked microsatellite loci.

#### Models

Before seeing the data, three candidate demographic scenarios are hypothesized: constant population size, bottleneck, and divergence (where an ancestral population splits into two sub-populations of equal size). Uniform prior distributions are assumed. Simulations are carried out with Hudson’s coalescent sampler, ms [88].

#### Inference and model choice

Model fitting is based on two classic summary statistics: genetic diversity and the Garza–Williamson statistic [104]. Both measures are known to be sensitive to historical variation in population size. Estimation of the present effective population size,  $N_e$ , is carried out according to the algorithm of Beaumont et al. [5]. Additional model parameters such as the divergence time or the duration and severity of the bottleneck are considered to be ‘nuisance parameters’ and estimates of  $N_e$  are averaged over these variables. Estimates obtained under the bottleneck and divergence models are close to the value we used for generating the example dataset ( $N_e = 600$ , Table I). Posterior model probabilities computed by multinomial logistic regression [66] reveal that the bottleneck model is the most supported by the data (61%), but the divergence model also receives considerable support (38%).

#### Model checking and model averaging

Replicates of the data under the three models were simulated using the posterior estimates of the parameters. The distributions of the replicated data (posterior predictive distributions) were compared with the observed data in terms of the expected heterozygosity, a ‘test’ statistic that was not used during the model-fitting stage. We find that the observed value of heterozygosity lay within the tails of the posterior predictive distribution under the bottleneck and divergence models, but well outside under the constant population size model. We decided to discard the constant population size scenario, and estimated  $N_e$  by weighting the posterior values obtained from the bottleneck and divergence models (Figure 1).

Table I. Posterior estimates of  $N_e$

Model	Posterior median	95% credible interval
Constant population size	3274	3974, 4746
Bottleneck	588	238, 1065
Divergence	550	236, 1310

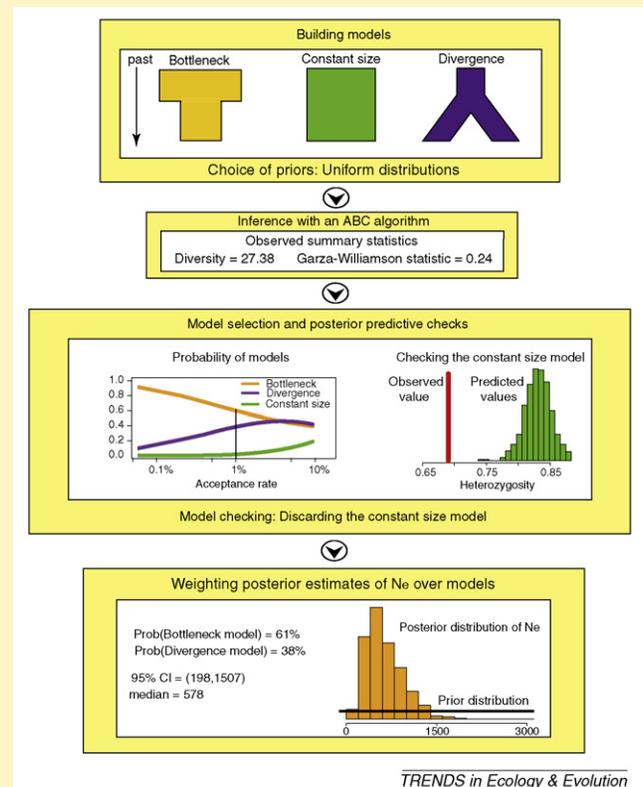
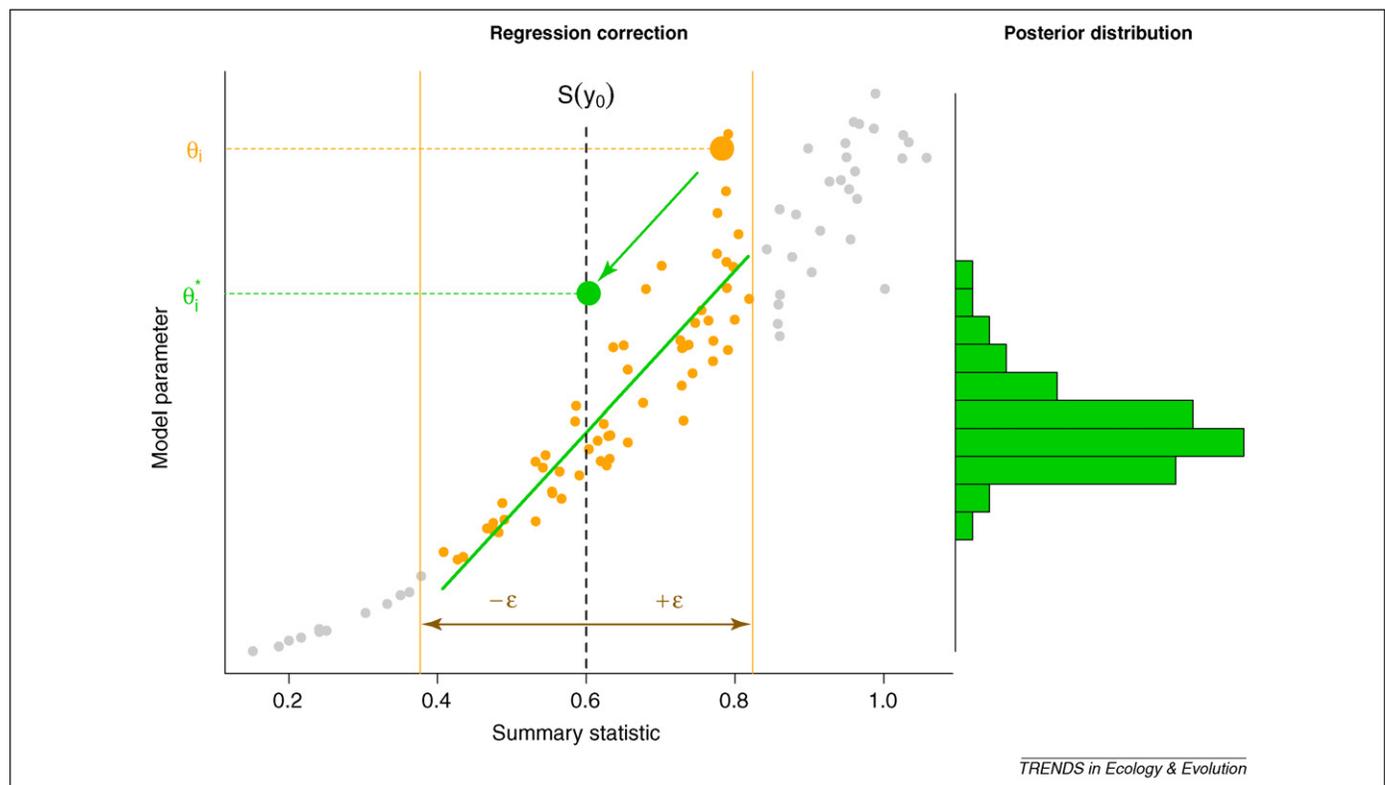


Figure 1. The main steps of an ABC analysis.



**Figure 1.** Linear regression adjustment in the ABC algorithm. In ABC, we repeatedly sample a parameter value,  $\theta_i$ , from its prior distribution to simulate a dataset,  $y_i$ , under a model. Then, from the simulated data, we compute the value of a summary statistic,  $S(y_i)$ , and compare it with the value of the summary statistic in the observed data,  $S(y_0)$ , using a distance measure. If the distance between  $S(y_0)$  and  $S(y_i)$  is less than  $\varepsilon$  (the so-called ‘tolerance’), the parameter value,  $\theta_i$ , is accepted. The plot shows how the accepted values of  $\theta_i$  (points in orange) are adjusted according to a linear transform,  $\theta_i^* = \theta_i - b(S(y_i) - S(y_0))$  (green arrow), where  $b$  is the slope of the regression line. After adjustment, the new parameter values (green histogram) form a sample from the posterior distribution.

Ongoing work is seeking to improve the parameter space exploration and develop efficient sampling strategies that drive particles toward regions of high posterior probability mass [12,13,66].

#### Model improvement

Comparing models and evaluating their goodness-of-fit are fundamental steps in the modeling and inference process (Box 4). Model comparison is often based on a decision theoretic framework, the objective of which is to choose models receiving high posterior support. In ABC studies, the posterior probability of a given model can be approximated by the proportion of accepted simulations given the model [9,28], by logistic regression estimates [25,67], or updated through a sequential Monte Carlo algorithm [66,68]. When comparing two models, often the Bayes factor is reported [9,18,69]. In this process, model choice does not imply the selection of a single ‘best’ model [60]. Different mechanisms could lead to the same data patterns, so several models could explain the data equally well. For example, both a weak population bottleneck and population subdivision with migration produce gene genealogies having long internal branches [70]. Thus, both demographic models can provide equally good explanations for an observed pattern of genetic variation [71]. Instead of focusing on a single model, we should consider the plausibility of each alternative model, and eventually weight parameter estimates over several models [72] (see Box 4).

The objective of model checking is to understand the ways in which models do not fit the data. The current

practice in Bayesian data analysis is oriented towards graphical checks whereby data simulated from the fitted model are compared with the observation through ‘test’ statistics (so-called ‘posterior predictive checks’, see Box 4). Specifying the test statistics amounts to deciding which aspects of the model are relevant to criticism. Posterior predictive checks were applied recently to coalescent models in ABC [16,20,36]. Another example of model checking is given by Schaffner and colleagues [73], who calibrated a model of human evolutionary history to generate data that closely resembled empirical data in terms of allele frequencies, linkage disequilibrium, and population differentiation.

#### Choice and dimension of summary statistics

Carrying out inference based on summary statistics instead of the full dataset inevitably implies discarding potentially useful information. More specifically, if a summary statistic is not sufficient for the parameter of interest, the posterior distribution computed with this statistic would not be equal to the posterior distribution computed with the full dataset [4]. Many areas of evolutionary biology focus on developing informative statistics. An example of a recently developed statistic is the extended haplotype homozygosity that aims to detect ongoing positive selection using genomic data [74]. The choice of summary statistics is crucial, and is closely linked to the particular inference questions addressed. In fact, ABC can be limited by the availability of informative statistics for any particular model parameter [75]. For example, the number

of polymorphic sites is useful in population genetics for the estimation of scaled mutation rates, but it says little about the demographic history [75]. As another example,  $F_{ST}$ , a well-known measure of population differentiation, is informative for estimating the migration rate between two populations under a symmetric island model [6], but less informative when estimating the migration rate in a model of divergence with migration [76].

An intuitive approach to the lack of sufficient summary statistics in most problems is to increase the number of summary statistics, thereby increasing the amount of information available to the ABC algorithm [38]. However, increasing the number of statistics can reduce the accuracy of the inference [5]. The need to manipulate large numbers of summary statistics is amplified by the ongoing increase in the amount of available data. With hundreds of loci, one approach is to average summary statistics over loci. Another approach is to use the quantiles and the moments of the summary statistic distribution across the loci [18,33,48] or the allele frequencies themselves [38], thereby increasing the dimensionality of the set of summary statistics. This might be a concern because ABC algorithms attempt to sample from a small, multidimensional sphere around the observed summary statistics, and the probability of accepting a simulation decreases exponentially as dimensionality increases. As a result, the number of summary statistics that can be handled with a reasonable number of simulations is limited in any ABC algorithm [5]. To circumvent this problem, a method for selecting summary statistics has been proposed that scores summary statistics according to whether their inclusion in the analysis substantially improves the quality of inference [77]. Alternative methods use dimension reduction techniques. Dimension reduction approaches are more robust to an increase in the number of summary statistics than the standard ABC rejection algorithm, and they lead to more accurate estimates. These methods encompass non-linear feed forward neural networks [10] and partial least square regression [64]. Neural networks and similar algorithms can also find combinations of summary statistics that contain the maximum information about the parameters of interest, and so are also useful in the selection of summary statistics [10].

## The future of ABC

### *Inference under complex models*

ABC is extremely flexible and relatively easy to implement, so inference can be carried out for many complex models in evolution and ecology as long as informative summary statistics are available and simulating data under the model is possible. Templeton [56,57] argues that the interpretation of why a complex model is preferred is highly subjective when using computer simulations. With the proliferation of highly complex models, Bayesian statisticians and evolutionary biologists have made efforts to reduce subjectivity by using hierarchical modeling techniques [48,59] and information theoretic measures of model selection [78,79]. Hierarchical models help in the understanding of parameter dependencies and remove implicit assumptions that are not foreseen by the modeler. With the development of highly structured hierarchical models, new model selection

methods have come to the fore. These methods provide explicit evaluation of model complexity based on information theoretic measures such as Akaike's Information Criterion (AIC)[72,79,80] or the Deviance Information Criterion (DIC)[78], which can be readily approximated in ABC approaches. Several aspects of Bayesian thinking have yet to be explored in ABC, but subjectivity can be reduced via careful application of all steps of Bayesian data analysis.

### *Future applications*

The most widespread application of ABC is in making inferences about demographic history and local adaptation [81,82], but applications outside evolutionary biology have already appeared for example in epidemiology [12,23,83] and in systems biology [13,68,84]. With the advent of high-throughput genotyping technologies to address evolutionary questions [85,86], ABC applications to genome-wide data will arrive very soon. Genome-wide approaches present an increased level of complexity, and fast genome samplers will become increasingly important when dealing with haplotypic diversity and patterns of linkage disequilibrium shaped by meiotic processes and natural selection [87]. In this context, it will become more and more important to develop simulation programs that capture some essential features of the biological problem, but which can also sample whole genomes. An example of 'trading-off high accuracy for decreased run-time is given in the estimation of recombination rates from genome-wide haplotypic data based on the coalescent and its extensions [88]. Various approximations to the coalescent allow for fast simulation of recombining haplotypes [89,90]. Although the mathematical 'tricks' used for improving speed might be unrealistic biologically, these models continue to accurately relate patterns of linkage disequilibrium to the underlying recombination process, which make them appropriate in an ABC inference framework.

### *Software for ABC inferences*

ABC users can base their analysis on simulation programs such as SIMCOAL [92], ms [93] or MaCS [91] and use statistical software to carry out the model-fitting step. More recently, specific ABC software have been developed that carry out the data simulation as well as the rejection and regression steps such as msBayes (<http://msbayes.sourceforge.net/>) [94], DIYABC (<http://www1.montpellier.inra.fr/CBGP/diyabc/>) [95], ONeSAMP (<http://genomics.jun.alaska.edu/asp/>) [96], ABC4F (<http://www-leca.ujf-grenoble.fr/logiciels.htm>) [97], PopABC (<http://code.google.com/p/popabc/>) [98] and 2BAD [99]. Even though these ABC software packages greatly facilitate the inference step of the algorithm, users must check their models. As with the direction followed by MCMC programs, we predict that efficient ABC programs will be developed to address specific questions. For these programs, the accuracy of inferences could also be extensively validated by model-checking techniques [100].

## Conclusions

Biology is a complex science, so it is inevitable that the observation of biological systems leads us to build complex models. However, the apparent ease of using an inference

algorithm such as ABC should never hide the general difficulties of making inferences under complex models. The automatic process of inference is hampered by the model definition and model checking steps, which are case-dependent and highly user-interactive. ABC is far from being as 'easy as 123'. Important *caveats* when using ABC algorithms can be summarized in the following points. First, credibility intervals obtained from ABC algorithms are potentially inflated due to the loss of information implied by the partial use of the data. This source of error should not be ignored, and users must be cautious about interpretations of their parameter estimates. Second, all models are wrong, thus model checking is a way to explore and understand differences between model and data, and to improve the fit between them [56]. Third, there can be several models that explain the data equally well, and support from the data for one model does not imply that the model is true. If models have common parameters, weighting parameter estimates over several well-supported models can produce more robust estimates than estimates from a single model (Box 4). With these *caveats* in mind, we argue that ABC is an extremely useful tool to make inferences with complex models. We envisage that many aspects of the ABC algorithms will be further improved in the near future, such as dealing with high-dimensional sets of summary statistics, evaluating model complexity, and devising efficient ways for model checking in ABC.

### Acknowledgements

KC is funded by a postdoctoral fellowship from the Université Joseph Fourier (ABC MSTIC). OF was partially supported by a grant from the Agence Nationale de la Recherche (BLAN06-3146282 MAEV). MB and OF acknowledge the support of the Complex Systems Institute (IXXI). OEG acknowledges support from the EcoChange Project.

### References

- Avice, J.C. (2004) *Molecular Markers, Natural History and Evolution*, (2nd edn), Sinauer Associates
- Beaumont, B.A. and Rannala, B. (2004) The Bayesian revolution in genetics. *Nat. Rev. Genet.* 5, 251–261
- Kuhner, M.K. (2009) Coalescent genealogy samplers: windows into population history. *Trends Ecol. Evol.* 24, 86–93
- Marjoram, P. and Tavaré, S. (2006) Modern computational approaches for analysing molecular genetic variation data. *Nat. Rev. Genet.* 7, 759–770
- Beaumont, M.A. *et al.* (2002) Approximate Bayesian Computation in population genetics. *Genetics* 162, 2025–2035
- Wright, S. (1951) The genetical structure of populations. *Ann. Eugen.* 15, 323–354
- Cavalli-Sforza, L.L. and Zei, G. (1967) Experiments with an artificial population. In *Proceedings of the Third International Congress of Human Genetics* (Crow, J.F. and Neel, J.V., eds), pp. 473–478, John Hopkins Press
- Tavaré, S. *et al.* (1997) Inferring coalescence times from DNA sequence data. *Genetics* 145, 505–518
- Pritchard, J.K. *et al.* (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* 16, 1791–1798
- Blum, M.G.B. and François, O. (2010) Non-linear regression models for Approximate Bayesian Computation. *Stat. Comput.* 20, 63–73
- Marjoram, P. *et al.* (2003) Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15324–15328
- Sisson, S.A. *et al.* (2007) Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. U. S. A.* 104, 1760–1765
- Toni, T. *et al.* (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* 6, 187–202
- Tallmon, D.A. *et al.* (2004) Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. *Genetics* 167, 977–988
- Chan, Y.L. *et al.* (2006) Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA. *PLoS Genet.* 2, e59
- Thornton, K.R. and Andolfatto, P. (2006) Approximate Bayesian inference reveals evidence for a recent, severe, bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* 172, 1607–1619
- Pascual, M. *et al.* (2007) Introduction history of *Drosophila subobscura* in the New World: a microsatellite based survey using ABC methods. *Mol. Ecol.* 16, 3069–3083
- François, O. *et al.* (2008) Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genet.* 4, e1000075
- Ross-Ibarra, J. *et al.* (2008) Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE* 3, e2411
- Ingvarsson, P.K. (2008) Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics* 180, 329–340
- Gao, L.Z. and Innan, H. (2008) Non-independent domestication of the two rice subspecies, *Oryza sativa* subsp. *indica* and subsp. *japonica*, demonstrated by multilocus microsatellites. *Genetics* 179, 965–976
- Guillemaud, T. *et al.* (2010) Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity* 104, 88–99
- Tanaka, M. *et al.* (2006) Estimating tuberculosis transmission parameters from genotype data using approximate Bayesian computation. *Genetics* 173, 1511–1520
- Shriner, D. *et al.* (2006) Evolution of intrahost HIV-1 genetic diversity during chronic infection. *Evolution* 60, 1165–1176
- Fagundes, N.J.R. *et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. U. S. A.* 104, 17614–17619
- Cox, M.P. *et al.* (2008) Testing for archaic hominin admixture on the X chromosome: model likelihoods for the modern human RRM2P4 region from summaries of genealogical topology under the structured coalescent. *Genetics* 178, 427–437
- Gerbault, P. *et al.* (2009) Impact of selection and demography on the diffusion of lactase persistence. *PLoS ONE* 4, e3669
- Patin, E. *et al.* (2009) Inferring the demographic history of African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet.* 5, e1000448
- Verdu, P. *et al.* (2009) Origins and genetic diversity of Pygmy hunter-gatherers from western Central Africa. *Curr. Biol.* 19, 312–318
- Bonhomme, M. *et al.* (2008) Origin and number of founders in an introduced insular primate: estimation from nuclear genetic data. *Mol. Ecol.* 17, 1009–1019
- Estoup, A. *et al.* (2004) Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad. *Bufo marinus*. *Evolution* 58, 2021–2036
- Miller, N. *et al.* (2005) Multiple transatlantic introductions of the Western corn rootworm. *Science* 310, 992
- Rosenblum, E.B. *et al.* (2007) A multilocus perspective on colonization accompanied by selection and gene flow. *Evolution* 61, 2971–2985
- Neuenschwander, S. *et al.* (2008) Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Mol. Ecol.* 17, 757–772
- Ray, N. *et al.* (2010) A statistical evaluation of models for the initial settlement of the American continent emphasizes the importance of gene flow with Asia. *Mol. Biol. Evol.* 27, 337–345
- Ghirotto, S. *et al.* (2010) Inferring genealogical processes from patterns of bronze-age and modern DNA variation in Sardinia. *Mol. Biol. Evol.* 27, 875–886
- Excoffier, L. *et al.* (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* 169, 1727–1738
- Sousa, V.C. *et al.* (2009) Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics* 181, 1507–1519
- Cornuet, J.M. *et al.* (2009) Bayesian inference under complex evolutionary scenarios using microsatellite markers: multiple divergence and genetic admixture events in the honey bee *Apis*

## Review

- mellifera. In *Genetic diversity* (Mahoney, C.L. and Springer, D.A., eds), pp. 229–246, Nova Science Publishers Inc
- 40 Hamilton, G. *et al.* (2005) Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* 170, 409–417
- 41 Lopes, J.S. and Boessenkool, S. (2010) The use of approximate Bayesian computation in conservation genetics and its application in a case study on yellow-eyed penguins. *Conserv. Genet* 11, 421–433
- 42 Tiemann-Boege, I. *et al.* (2006) High-resolution recombination patterns in a region of human chromosome 21 measured by sperm typing. *PLoS Genet.* 2, e70
- 43 Padhukasahasram, B. *et al.* (2006) Estimating recombination rates from single-nucleotide polymorphisms using summary statistics. *Genetics* 174, 1517–1528
- 44 Touchon, M. *et al.* (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* 5, e1000344
- 45 Jensen, J.D. *et al.* (2008) An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. *PLoS Genet.* 4, e1000198
- 46 Quach, H. *et al.* (2009) Signatures of purifying and local positive selection in human miRNAs. *Am. J. Hum. Genet.* 84, 316–327
- 47 Itan, Y. *et al.* (2009) The origins of lactase persistence in Europe. *PLoS Comput. Biol.* 5, e1000491
- 48 Hickerson, M.J. *et al.* (2006) Test for simultaneous divergence using approximate Bayesian computation. *Evolution* 60, 2435–2453
- 49 Becquet, C. and Przeworski, M. (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* 17, 1505–1519
- 50 Leaché, A. *et al.* (2007) Two waves of diversification in mammals and reptiles of Baja California revealed by hierarchical Bayesian analysis. *Biol. Lett.* 3, 646–650
- 51 Putnam, A.S. *et al.* (2007) Discordant divergence times among Z chromosome regions between two ecologically distinct swallowtail butterfly species. *Evolution* 61, 912–927
- 52 Wilkinson, R.D. and Tavaré, S. (2009) Estimating primate divergence times by using conditioned birth-and-death processes. *Theor. Popul. Biol.* 75, 278–285
- 53 Jakobsson, M. *et al.* (2006) A recent unique origin of the allotetraploid species *Arabidopsis suecica*: evidence from nuclear DNA markers. *Mol. Biol. Evol.* 23, 1217–1231
- 54 Carnaval, A.C. *et al.* (2009) Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science* 323, 785–789
- 55 Jabot, F. and Chave, J. (2009) Inferring the parameters of the neutral theory of biodiversity using phylogenetic information, and implications for tropical forests. *Ecol. Lett.* 12, 239–248
- 56 Templeton, A.R. (2009) Statistical hypothesis testing in intraspecific phylogeography: NCPA versus ABC. *Mol. Ecol.* 18, 319–331
- 57 Templeton, A.R. (2010) Coalescent-based, maximum likelihood inference in phylogeography. *Mol. Ecol.* 19, 431–435
- 58 Beaumont, M.A. *et al.* (2010) In defence of model-based inference in phylogeography. *Mol. Ecol.* 19, 436–446
- 59 Gelman, A. *et al.* (2004) *Bayesian Data Analysis*, (2nd edn), Chapman and Hall
- 60 Ripley, B.D. (2004) Selecting amongst large classes of models. In *Methods and Models in Statistics: In Honour of Professor John Nelder, FRS* (Adams, N. *et al.*, eds), pp. 155–170, Imperial College Press
- 61 Purvis, A. *et al.* (2000) Predicting extinction risk in declining species. *Proc. Biol. Sci.* 267, 1947–1952
- 62 May, R.M. (2004) Uses and abuses of mathematics in biology. *Science* 303, 790–793
- 63 Leuenberger, C. and Wegmann, D. (2010) Bayesian computation and model selection without likelihoods. *Genetics* 184, 243–252
- 64 Wegmann, D. *et al.* (2009) Efficient Approximate Bayesian Computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* 182, 1207–1218
- 65 Liu, J.S. (2001) *Monte Carlo Strategies in Scientific Computing*, Springer
- 66 Beaumont, M.A. *et al.* (2009) Adaptive approximate Bayesian computation. *Biometrika* 96, 983–990
- 67 Beaumont, M.A. (2008) Joint determination of topology, divergence time, and immigration in population trees. In *Simulation, Genetics, and Human Prehistory* (Matsumura, S. *et al.*, eds), pp. 135–154, McDonald Institute for Archaeological Research
- 68 Toni, T. and Stumpf, M.P.H. (2010) Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* 26, 104–110
- 69 Grelaud, A. *et al.* (2009) ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Anal.* 4, 317–336
- 70 Hein, J. *et al.* (2005) *Gene Genealogies, Variation, and Evolution*, Oxford University Press
- 71 Nielsen, R. and Beaumont, M.A. (2009) Statistical inferences in phylogeography. *Mol. Ecol.* 18, 1034–1047
- 72 Johnson, J.B. and Omland, K.S. (2004) Model selection in ecology and evolution. *Trends Ecol. Evol.* 19, 101–108
- 73 Schaffner, S.F. *et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15, 1576–1583
- 74 Sabeti, P.C. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837
- 75 Hey, J. and Machado, C.A. (2003) The study of structured populations - new hope for a difficult and divided science. *Nat. Rev. Genet.* 4, 535–543
- 76 Nielsen, R. and Wakeley, J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158, 885–896
- 77 Joyce, P. and Marjoram, P. (2008) Approximately sufficient statistics and Bayesian computation. *Stat. Appl. Genet. Mol. Biol.* 7, 26
- 78 Spiegelhalter, D.J. *et al.* (2002) Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Series B Stat. Methodol.* 64, 583–639
- 79 Carstens, B.C. *et al.* (2009) An information-theoretical approach to phylogeography. *Mol. Ecol.* 18, 4270–4282
- 80 Akaike, H. (1974) New look at statistical model identification. *IEEE Trans. Automat. Control* 19, 716–723
- 81 Hickerson, M.J. *et al.* (2010) Phylogeography's past, present, and future: 10 years after Avise, 2000. *Mol. Phylogenet. Evol.* 54, 291–301
- 82 Segelbacher, G. *et al.* (2001) Applications of landscape genetics in conservation biology: concepts and challenges. *Conserv. Genet.* 11, 375–385
- 83 Lopes, J.S. and Beaumont, M.A. (2009) ABC: A useful Bayesian tool for the analysis of population data. *Infect. Genet. Evol.* DOI: 10.1016/j.meegid.2009.10.010 (<http://ees.elsevier.com/meegid/>)
- 84 Ratmann, O. *et al.* (2009) Model criticism based on likelihood-free inference. *Proc. Natl. Acad. Sci. U. S. A.* 106, 10576–10581
- 85 Jakobsson, M. *et al.* (2008) Genotype, haplotype, and copy number variation in worldwide human populations. *Nature* 451, 998–1003
- 86 Li, J. *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100–1104
- 87 Hoggart, C.J. *et al.* (2007) Sequence-level population simulations over large genomic regions. *Genetics* 177, 1725–1731
- 88 Hudson, R.R. (2002) Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 18, 337–338
- 89 Wiuf, C. and Hein, J. (1999) Recombination as a point process along sequences. *Theor. Popul. Biol.* 55, 248–259
- 90 Marjoram, P. and Wall, J. (2006) Fast “coalescent” simulation. *BMC Genet.* 7, 16
- 91 Chen, G.K. *et al.* (2009) Fast and flexible simulation of DNA sequence data. *Genome Res.* 19, 136–142
- 92 Laval, G. and Excoffier, L. (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* 20, 2485–2488
- 93 Hudson, R.R. (1983) Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* 23, 183–201
- 94 Hickerson, M.J. *et al.* (2007) msBayes: pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. *BMC Bioinformatics* 8, 268
- 95 Cornuet, J.M. *et al.* (2008) Inferring population history with DIY ABC: a user-friendly approach to Approximate Bayesian Computation. *Bioinformatics* 24, 2713–2719
- 96 Tallmon, D.A. *et al.* (2008) ONeSAMP: a program to estimate effective population size using approximate Bayesian computation. *Mol. Ecol. Resour.* 8, 299–301
- 97 Foll, M. *et al.* (2008) An approximate Bayesian computation approach to overcome biases that arise when using amplified fragment length

- polymorphism markers to study population structure. *Genetics* 179, 927–939
- 98 Lopes, J. *et al.* (2009) *PopABC*: a program to infer historical demographic parameters. *Bioinformatics* 25, 2747–2749
- 99 Bray, T.C. *et al.* (2009) 2BAD: an application to estimate the parental contributions during two independent admixture events. *Mol. Ecol. Resour.* 10, 538–541
- 100 Cook, S. *et al.* (2006) Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Stat.* 15, 675–692
- 101 Lachaise, D. *et al.* (1988) Historical biogeography of the *Drosophila-melanogaster* species subgroup. *Evol. Biol.* 22, 159–225
- 102 Hubbell, S.P. (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*, Princeton University Press
- 103 Alonso, D. *et al.* (2006) The merits of neutral theory. *Trends Ecol. Evol.* 21, 451–457
- 104 Garza, J.C. and Williamson, E.G. (2001) Detection of reduction in population size using data from microsatellite loci. *Mol. Ecol.* 10, 305–318