Infection, Genetics and Evolution 10 (2010) 826–833

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

ABC: A useful Bayesian tool for the analysis of population data

J.S. Lopes *, M.A. Beaumont

School of Biological Sciences, University of Reading, Reading RG6 6AJ, UK

ARTICLE INFO

Article history:
Received 24 June 2009
Received in revised form 20 October 2009
Accepted 21 October 2009
Available online 30 October 2009

Keywords:
Approximate Bayesian computation
Population genetics
Epidemiology
Phylogenetics
Population history
Coalescence models
Likelihood-free inference
Biogeography
Population models

ABSTRACT

Approximate Bayesian computation (ABC) is a recently developed technique for solving problems in Bayesian inference. Although typically less accurate than, for example, the frequently used Markov Chain Monte Carlo (MCMC) methods, they have greater flexibility because they do not require the specification of a likelihood function. For this reason considerable amounts of data can be analysed and more complex models can be used providing, thereby, a potential better fit of the model to the data. Since its first applications in the late 1990s its usage has been steadily increasing. The framework was originally developed to solve problems in population genetics. However, as its efficiency was recognized its popularity increased and, consequently, it started to be used in fields as diverse as phylogenetics, ecology, conservation, molecular evolution and epidemiology. While the ABC algorithm is still being greatly studied and alterations to it are being proposed, the statistical approach has already reached a level of maturity well demonstrated by the number of related computer packages that are being developed. As improved ABC algorithms are proposed, the expansion of the use of this method can only increase. In this paper we are going to depict the context that led to the development of ABC focusing on the field of infectious disease epidemiology. We are then going to describe its current usage in such field and present its most recent developments.

1. Introduction

1.1. Early use of Bayesian statistics in population biology and genetics

In the past decade, there has been a great increase on the use of Bayesian statistics in very many different fields. The practical advantages of such methods are well-known: they can easily accommodate unobserved data and latent variables; they provide an explicit mechanism for incorporating previous information via prior probability distributions; the resulting posterior probabilities are intuitively interpreted in contrast to the results from classic statistics; and they allow for analyses of complex data sets while assuming complicated models. The intention of this article is to provide, especially for empiricists, a brief overview of the application of Bayesian analysis in evolutionary biology, with special reference to infectious disease epidemiology, and then to give an overview of approximate Bayesian computation (ABC) methods. We have attempted to avoid technical details in our exposition, while at the same time conveying the salient facts.

The aim of Bayesian inference is to calculate the probability distribution of the (typically vector valued) parameter, Φ, given some empirical data, D, i.e. the posterior distribution $Pr(\Phi|D)$. This distribution can be obtained as the product of the prior distribution $Pr(\Phi)$, representing prior knowledge of the parameters, with the probability of observing particular values of the empirical data given particular values of the parameters, the likelihood $Pr(D|\Phi)$.

In the context of population genetics, the coalescent theory (Kingman, 1982; Hudson, 1983; Tajima, 1983) provided a useful modelling framework for describing genetic data. This theory has revolutionized that field by forming the basis for likelihood calculations in genealogical models (Felsenstein, 1992). Fields such as molecular ecology, phylogeography, the study of human origins and evolutionary genetics also underwent considerable advances as a consequence of these developments (Takahata et al., 2001; Pritchard et al., 1999; Ford, 1998; Edwards and Beerli, 2000). A particular field that also benefitted from coalescent theory was epidemiology (Crandall, 1999; Thompson, 2000), in particular, due to advances in the areas of linkage disequilibrium mapping (Nordborg and Tavare, 2002), assignment methods and demographic parameters estimation (Beaumont and Rannala, 2004). Rosenberg and Nordborg (2002) also pointed out that coalescent based analysis can be useful to estimate strictly epidemiological related parameters, such as, the timing of introduction of a pathogen into populations, selection coefficients favouring certain virulent lineages and the rate of evolution of pathogens occurring in hosts. In fact, host–pathogen systems have been an important source of problems for coalescent methods, especially due to their typically serially sampled data (e.g. Rodrigo et al., 1999; Fu, 2001).

* Corresponding author at: School of Biological Sciences, University of Reading, Philip Lyle Building, Whiteknights, PO Box 228, Reading RG6 6AJ, UK.
Tel.: +44 118 3785049; fax: +44 118 9310180.
E-mail address: joao.lopes@reading.ac.uk (J.S. Lopes).

1567-1348/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2009.10.010

1.2. Early influence of Bayesian statistics in the field of epidemiology

The amount of human genetic data has increased immensely in the last decade. Projects such as The Human Genome Project (Human Genome Sequencing Consortium, 2001; Venter, 2001) and HapMap (The International HapMap Consortium, 2003) produced considerable data describing polymorphisms across human populations. These data led to the development of new studies to identify genes involved in human diseases and to map them in the human genome (Rannala, 2001). The analyses of these data required rigorous statistical methods, many of which applied Bayesian approaches (Beaumont and Rannala, 2004). Disease mutations can be located using association between the presence or absence of alleles at genetically mapped loci and the presence or absence of the disease phenotype. However the presence of population stratification can lead to wrong conclusions. In an attempt to correct for that, methods that examine unlinked genetic markers have been proposed (Pritchard and Donnelly, 2001). Further improvements in this field led to the development of Bayesian approaches that could consider complex interactions between genes and interactions with the environment and account for the statistical uncertainty of genomic ancestries and admixture proportions (Sillanpaa et al., 2001; Hoggart et al., 2003).

The use of Bayesian assignment methods to cluster individuals based on their multilocus genotype or to assign them to populations has also been spreading greatly. The fundamental approach used in these methods is to calculate the probability of sampling an individual's multilocus genotype, given the allele frequencies at the loci in a population (Pritchard et al., 2000). These allele frequencies and the population to which an individual is assigned are treated as random variables, and inferred. These applications were originally developed to detect cryptic population admixture in association studies (Pritchard et al., 2000; Dawson and Belkhir, 2001). Since then the range of its application has broadened. In infectious disease epidemiology they can be especially useful to detect population sources of sporadic outbreaks or emerging epidemics (Davies et al., 1999; Bonizzoni et al., 2001).

Models used in epidemiology have often many parameters to consider. With the introduction of Bayesian methods it became possible to infer their values separately through the use of marginal posterior distributions. Since the first fully Bayesian genealogical analysis (Wilson and Balding, 1998), these methods have been developed significantly allowing for analysis of varied and complex population structures (Beaumont, 1999; Nielsen and Wakeley, 2001; Beaumont, 2003; Rannala and Yang, 2003; Wilson et al., 2003). Bayesian methods have also been developed to deal with genetic data taken at different times (Drummond et al., 2002; Beaumont, 2003). This last method has been particularly useful to deal with viral epidemiology (Pybus et al., 2003; Drummond et al., 2005).

1.3. Current and future developments of Bayesian methods

Due to the increase of available data and the desire to use models with increasing complexity, modifications to standard MCMC methods have been explored (O'Neill et al., 2000; Beaumont, 2003). Furthermore, MCMC methods often have problems in obtaining a good sample from the posterior distribution within a reasonable computation time (convergence problems). This has led to the development of alternatives to MCMC (e.g. Del Moral et al., 2006). A group of these alternative methods has come to be known as "approximate Bayesian computation" (ABC, Beaumont et al., 2002; Marjoram et al., 2003). ABC is mostly useful when the likelihood function is intractable or for cases in which there is a large number of parameters that are not of interest, i.e. nuisance parameters. They are also particularly valuable for choosing between different models applied to the same data set (e.g. Fagundes et al., 2007). Since the first application of ABC by Pritchard et al. (1999) on Y-chromosome data different studies have been performed that demonstrated its competitiveness with correspondent full-likelihood methods (Beaumont et al., 2002; Hickerson et al., 2006; Sousa et al., 2009).

1.4. Approximate Bayesian computation

Although ABC is a statistical tool developed originally to perform population genetics analysis (Pritchard et al., 1999), the technique actually provides a framework within which to perform Bayesian analysis in a wide range of fields (e.g. Ratmann et al., 2007; Bortot et al., 2007; Grelaud et al., 2009. Also see a short review in Lopes and Boessenkool, 2009).

The approach is characterized by two main features: the use of summary statistics to summarise data, which significantly reduces the size of data to handle; and the use of Monte Carlo simulations that avoid the need to use explicit likelihood functions. These characteristics make ABC a versatile method: it can assume complex models; allows the use of large data sets without being too computational demanding; and enables the possibility to estimate parameters otherwise intractable. As its flexibility and robustness were recognized its use has widespread both in population genetics and in such fields as phylogenetics, conservation genetics, molecular evolution and infectious disease epidemiology (Fig. 1).

1.5. Use of ABC in epidemiology

Currently, ABC is seldom used in the field of epidemiology. There are now, however, a number of recent examples of its application (Tanaka et al., 2006; Shriner et al., 2006; Luciani et al., 2009; McKinley et al., 2009; Wilson et al., 2009; Toni et al., 2009). These give a good insight on the possible range of epidemiological problems that can be tackled with this statistical method. An ABC algorithm was used by Tanaka et al. (2006) to estimate typical epidemic parameters of tuberculosis transmission using genotype data from *Mycobacterium tuberculosis* isolated from patients in San Francisco. They used a stochastic model of tuberculosis transmission with parameters such as net transmission rate, doubling time and the reproductive value of the pathogen. Shriner et al. (2006) studied diverse hypothesis of evolution of HIV-1 genetic diversity within a single individual. For the study they have used a large data set of RNA sequence from the HIV-1 *env* locus. The proposed models varied in terms of population structure and presence or absence of negative and positive selective pressures. Wilson et al. (2009) presented results on the evolutionary history of *Campylobacter jejuni* using an ABC method to test a codon model firstly described by Nielsen and Yang (1998). In this work they modelled selection pressures as a mutational bias using such parameters as synonymous mutation rate, transition–transversion ratio and d*N*/d*S* ratio. They used about 900 isolates from patients diagnosed with campylobacteriosis sequenced at seven loci (*aspA*, *glnA*, *gltA*, *glyA*, *pgm*, *tkt* and *uncA*). As a final example, Toni et al. (2009) applied a sophisticated ABC method to SIR models, models using values of susceptible (S), infected (I) and recovered (R) individuals, to describe common cold outbreaks. The data studied was from an isolated island from the Atlantic Ocean called Tristao da Cunha. In this work the authors first selected the model that best described the data from four SIR models with different complexity levels. Later they estimated the parameters of the most supported model: infection rate, recovery rate, the transition rate from the latent to the infective stage and the initial susceptible population.

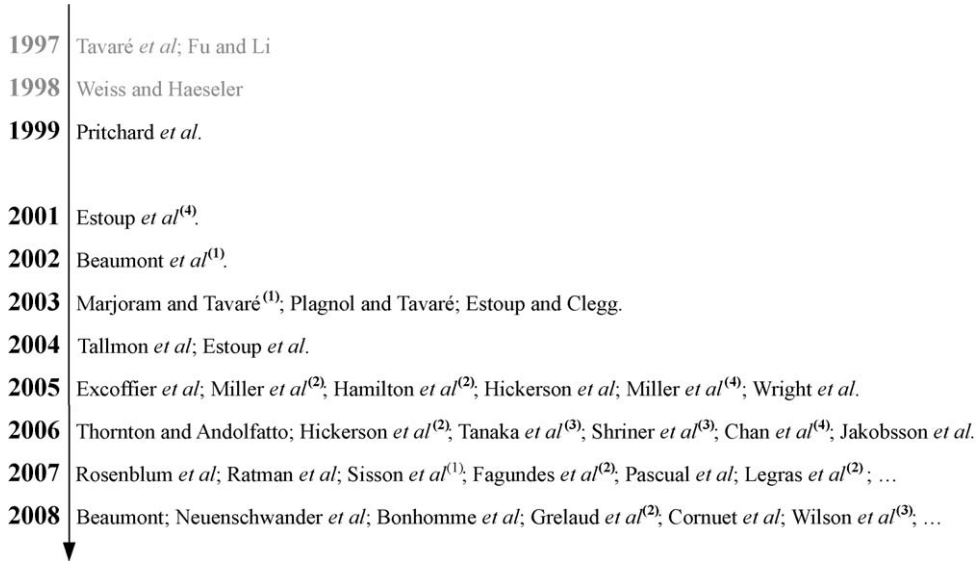

Fig. 1. Use of ABC since its appearance until 2008. The use of ABC in particular fields is emphasized: (1) phylogenetic problems; (2) epidemiological problems; (3) theoretical papers; (4) conservation problems (See ref. Bonhomme et al., 2008; Chan et al., 2006; Estoup et al., 2001, 2004; Estoup and Clegg, 2003; Fu and Li, 1997; Jakobsson et al., 2006; Kruglyak, 1999; Legras et al., 2007; Liu et al., 2001; Miller et al., 2005; Morris et al., 2002; Pascual et al., 2007; Plagnol and Tavare, 2003; Pritchard and Przeworski, 2001; Rannala and Reeve, 2001; Rosenblum et al., 2007; Tavare et al., 1997; Weiss and von Haeseler, 1998 and Wright et al., 2005).

1.6. Available software

As stated above, ABC has been primarily developed for population genetics studies. For this reason, the existing packages to perform ABC were developed mostly to be applied in that field. In 2002, Richard Hudson published a very flexible coalescence simulator called ms. This software was later pipelined in the msBayes package (Hickerson et al., 2007), which comprises several scripts that allow the user to run a complete ABC analysis. In 2004 Laval and Excoffier released the coalescent simulator SIMCOAL2.0. This software simulates genetic trees using the discrete-generation coalescent approach which, although slowing down the simulations, allows a greater flexibility for defining demographic models. Later this simulator was too integrated in a package to perform ABC analysis (Serial SimCoal, Anderson et al., 2005). This package has been extensively developed to allow for very detailed model parameterisation. More recently two other packages have been developed: DIY ABC (Cornuet et al., 2008) and popABC (Lopes et al., 2009). These programs provide a fast and user-friendly way to run ABC studies, in particular for model-choice inferences. Another program that was recently developed and has been used fairly frequently is ONeSAMP (Tallmon et al., 2008). This software can only deal with a single Wright–Fisher population but has a very friendly web-based interface. Table 1 summarizes the characteristics of these packages.

2. Standard ABC methodology

2.1. Rejection

The base algorithm, as introduced by Pritchard et al. (1999) can be written as follows:

(1) Sample (vector valued) parameter, Φ, from the prior: $\Phi_i \sim p(\Phi)$;
(2) Simulate data, D, given Φ_i: $D_i \sim p(D|\Phi_i)$;
(3) Summarize D_i with a set of chosen summary statistics to obtain S_i; go to (1) until N sample points from the joint distribution $p(S,\Phi)$ have been created;
(4) Accept the points whose S_i is within a distance δ from s', the real data summarized by the same set of summary statistics, $|S_i - s'| < \delta$.

In steps (1)–(3) we simulate independent pairs (Φ_i, S_i), $i = 1, 2, \ldots, N$, where each Φ_i is an independent draw from the prior distribution Φ, and the S_i are values that summarize simulated values of D with $\Phi = \Phi_i$. The (Φ_i, S_i) are then random draws from the joint density. Step (4), the rejection-step, provides an estimate of the conditional density, when $S = s'$, that is, the posterior distribution $p(\Phi|S = s')$. The idea is that the Φ_i for which $|S_i - s'|$ is small form an approximation of a random sample from the desired posterior distribution $p(\Phi|D = d')$ (Beaumont et al., 2002).

2.2. Regression

In the regression method two innovations are proposed at step (4): smooth weighting and regression adjustment. The aim is to improve the sampling of the posterior density by weighting the Φ_i according to its distance from the real data by evaluating $|S_i - s'|$ and to adjust the Φ_i using local-linear regression to weaken the effect of that discrepancy (Beaumont et al., 2002).

The proposed algorithm (Beaumont et al., 2002) can be described as

(4) Apply a weighting scheme to the points according to their distance to the s'. And then perform a weighted linear multiple regression with points assigned non-zero weight. Adjust them according to $\Phi_i^* = \Phi_i + R_i$, where R_i is the residual of the ith point.

Table 1
Available software to perform ABC[a] computation analysis.

	Algorithm	Interface	Reference
ms	Simulator	–	Hudson (2002)
SIMCOAL 2.0	Simulator	–	Laval and Excoffier (2004)
Serial SimCoal	ABC[a] regression	–	Anderson et al. (2005)
msBayes	ABC[a] regression	–	Hickerson et al. (2007)
DIY ABC	ABC[a] regression	Yes	Cornuet et al. (2008)
ONeSAMP	ABC[a] rejection	Yes	Tallmon et al. (2004)
popABC	ABC[a] rejection	Yes	Lopes et al. (2009)

[a] Approximate Bayesian computation.

As noted by Beaumont and co-workers the rejection method can be viewed as a special case of the local-linear regression approach when using a uniform kernel and a local-constant regression. Furthermore, as δ tends to zero, the regression and rejection methods become equivalent.

2.3. Comparison between traditional ABC methods

In order to compare previously described ABC-rejection and ABC-regression methods we provide a motivating example created by simulating a toy data set using a simple model of population divergence (often termed vicariance) of 2 populations; that is, two sister populations descending from a common ancestral population. We chose a point prior for the effective size of the ancestral population (1000 individuals), the time of divergence (100 years ago) and the effective size of one of the modern populations (100 individuals). The aim of the ABC method is then to infer the effective size of the second modern population. The data set comprises a sample of 50 gene copies taken from 10 loci of sequence data for each modern population. The mutation rate is set to 0.01 mutations per locus per generation and the true value to be estimated is 50 individuals. The prior for the parameter in study is set to a uniform distribution between 0 and 100,000 individuals. For this study the following summary statistics are used: average pairwise distance, number of segregating sites and number of different haplotypes. These three summary statistics are calculated within each population, and then recalculated from the pooled sample, resulting in a total of 9 summary statistics. All the remaining analyses presented in this work are based in this motivating example.

To compare both methods we use the relative mean integrated square error (RMISE):

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(Ne_i - Ne')^2}{(Ne')^2}\right),$$

where n is the number of sampled points from the posterior distribution, Ne_i is the ith sampled point from the posterior distribution and Ne' is the true value of the effective population size in study.

Table 2 presents three ABC runs on the same data set with different number of simulations (1000, 10,000 and 100,000). In all three runs the closest 100 simulated points were accepted. Using these accepted points we calculate the RMISE before and after performing a regression step. As we can observe from the results (Table 2), within the conditions of the study the regression step highly improves the ABC method, that is, when considering the regression step the values of RMISE are lower. As expected, as the number of simulated points increases the difference between both methods (with and without the regression step) starts to fade. The regression step is used to correct the accepted simulated points according to their distance to the true data set. By fixing the number of accepted points and increasing the pool from which these points are taken we expect to accept points less distant to the studied data set. In fact, when using enough simulated points we expect to obtain almost identical results for both methods.

Table 2
Comparison between the ABC[a] rejection and the ABC[a] regression methods using RMISE.[b] Each different line corresponds to a study using a different number of simulated points.

Simulated points	Accepted points	Rejection	Regression
1,000	100	24948.17	1.37
10,000	100	164.92	0.06
100,000	100	1.28	0.03

[a] Approximate Bayesian computation.
[b] Relative mean integrated square error.

From Table 2 we can also have some insight on a complicated problem: how many simulations should one use to perform an ABC analysis? Unfortunately there is no general rule to calculate the minimum needed simulations. Often such a number is chosen by performing some empirical studies, which can provide guidelines solely for the problem in analysis. As an example of such studies we can look at the column regarding the ABC-regression analyses in Table 2. The values of RMISE for 10,000 and 100,000 simulations are very close, and one may decide that 10,000 points are enough to perform analysis, at least in this example.

3. Future advances in ABC

Although the core algorithm of ABC methods (presented in Section 2) is fairly well established, modifications to improve it have been proposed. Both the capabilities and the limitations of the standard ABC algorithm have been extensively studied (Hamilton et al., 2005; Hickerson et al., 2005; Joyce and Marjoram, 2008; Wilkinson, 2008; Blum, 2009). Recently, there have been some attempts to overcome the latter by proposing modifications at different stages of this Bayesian framework (Sisson et al., 2007; Blum and Francois, 2009; Leuenberger et al., 2009). We present a graphic representation of the standard algorithm (Fig. 2) in order to provide a overview of the general ABC method and to better

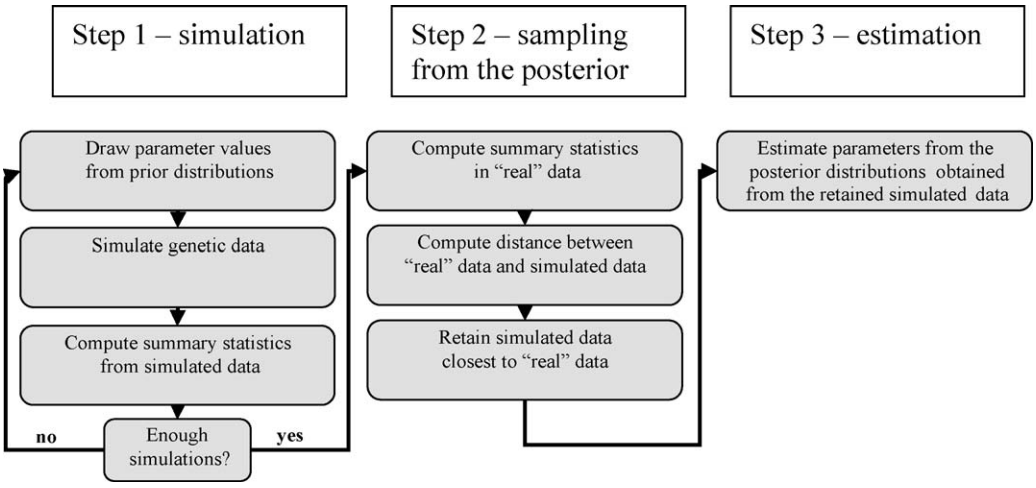

Fig. 2. Schematic representation of the ABC method as described in Beaumont et al. (2002). Adapted from Excoffier et al. (2005).

identify the different stages where improvements have been suggested.

3.1. Efficient sampling from the priors

As mentioned before, one advantage of using Bayesian approaches is to be able to use previously available information in the studies through prior distributions. In the absence of information, however, a broad, flat distribution should be used. Analysis using these vague priors can be rather inefficient when using a standard rejection- or regression-based ABC approach.

3.1.1. Sequential methods

One way to improve the efficiency of vague prior distributions is to use sequential algorithms (Sisson et al., 2007; Beaumont et al., 2009; Toni et al., 2009). A basic example of such algorithms, based on standard rejection, is

(1) Run the standard ABC algorithm once:
  (1.1) Sample parameters, Φ, from the prior: $\Phi \sim p(\Phi)_1$;
  (1.2) Obtain the 1st posterior, through rejection-based ABC procedures: $p(\Phi|D = D')_1$
(2) Run the ABC method n times ($i = 1, 2, 3, \ldots, n$), using a rough estimate of the ith posterior as the $i+1$th prior, but then weighting to correct for not using the true prior:
  (2.1.) Sample, parameters, Φ, from the $i+1$th prior: $\Phi \sim p(\Phi)_{i+1}$;
  (2.2.) Weight the sampled value Φ with: $p(\Phi)_1/p(\Phi)_{i+1}$;
  (2.3.) Obtain samples from the $i+1$th posterior, through standard ABC procedures: $p(\Phi|D = D')_i$.

Note that in step (2.2) of this algorithm, we require an analytical expression for $p(\Phi)_{i+1}$, which is provided by fitting a kernel density. Note that this does not need to be an accurate density. This description of the sequential procedure in terms of importance sampling (step (2.2)) is quite intuitive. A more general description, based on theory in Del Moral et al. (2006), which is rather less intuitive, is used in Sisson et al. (2007), but this can lead to some problems, highlighted in Beaumont et al. (2009). When using a sequential algorithm the ABC method can be rerun as much as it is necessary in order to sharpen the posterior distribution (Fig. 3). This plot was produced using the same conditions as the previously given example (Section 2) but instead of using a standard ABC approach we used a sequential ABC approach based on the algorithm above. In Fig. 3 we present several estimates of the posterior distribution using an increased number of iterations of the sequential ABC analysis. As we can observe, as the number of iterations increases the posterior distribution becomes sharper and located closer to the true value.

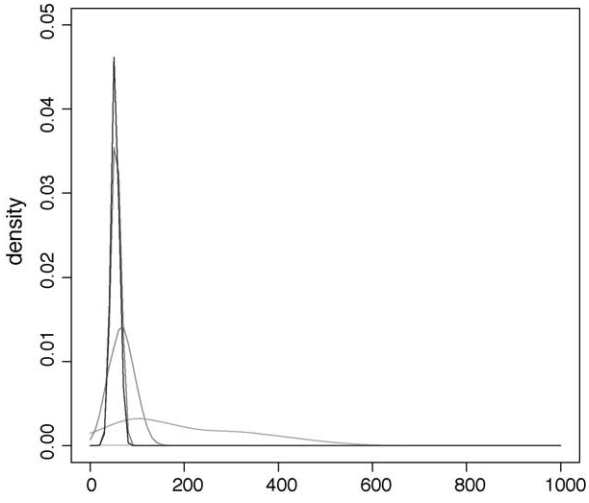

Fig. 3. Posterior distributions obtained using increasing number of populations of points in a sequential ABC method without a final regression step for the parameter Ne1 (true value is 50 individuals).

3.1.2. Hierarchical Bayesian models

Hierarchical Bayesian models are based on the use of a hierarchical scheme for the prior distribution (Storz and Beaumont, 2002; Storz et al., 2002). In these models the values that characterise the prior distribution of its parameters, rather than being fixed and chosen by the modeller, may themselves be drawn from another prior distribution (the hyperprior). This hierarchy can be extended indefinitely. As an example we can consider two marginal prior distributions defining the mean, Π_μ, and the standard deviation, Π_σ, of a third distribution, $\Pi(\mu,\sigma)$, the former two are hyper priors of the latter. These approaches can be especially useful when analysing parameter-rich models that consist of groups of parameters related to each other. For example, rather than modelling mutation rate independently for each locus, or assuming that mutation rate is the same, we can model an intermediate stage in which the standard deviation and mean mutation rate across loci can also be inferred. Examples of ABC applications in hierarchical Bayesian models are given in Excoffier et al. (2005), Hickerson et al. (2006), Beaumont (2008) and Cornuet et al. (2008).

3.2. Choice of the summary statistics

The choice of the most suitable summary statistics is still a problematic area for ABC researchers (Beaumont, 2008). Several attempts to provide a methodology for such choice have been proposed (Hamilton et al., 2005; Joyce and Marjoram, 2008), but there is little general consensus on the approach to take. The choice of the summary statistics is however of great importance in ABC. The general rule is to choose summary statistics close to sufficient for a given model while taking in account the dimensionality problems that arise with the increasing number of summary statistics (Beaumont et al., 2002).

In order to choose summary statistics to estimate parameters using ABC one should be aware of their sensitivity to variation in such parameters. Ideally, the summary statistics should be highly correlated with the parameters in study. A simple linear correlation can provide a first insight into the relationship between one parameter and one particular summary statistic. This single value, however, fails to provide insights into the nature of the relationship. The simplest way to attain a better knowledge is from visual plots as in Hickerson et al. (2005).

However, the correlation between a parameter and a summary statistic can lead to an erroneous choice of the summary statistics because it does not take into account the correlation between different summary statistics. Ideally, the summary statistics should be weakly correlated with each other so that the extraction of information from the data is optimized. One way to deal with this problem is to calculate a multiple correlation coefficient between a parameter and different sets of summary statistics. One should be aware, however, that most of the available statistical tools for such calculations can only deal with linear relations. Another problem is that this approach does not consider all the parameters jointly. There is a need, then, to find the union between each "best" set of summary statistics for each considered parameter. So far, the most useful method described in the literature to take all the information in consideration is to run ABC analysis on simulated data, in which the true parameter values are known, and compare the errors obtained from the use of different sets of summary statistics (Hickerson et al., 2006; Neuenschwander et al., 2008). A common error measure is the already mentioned RMISE.

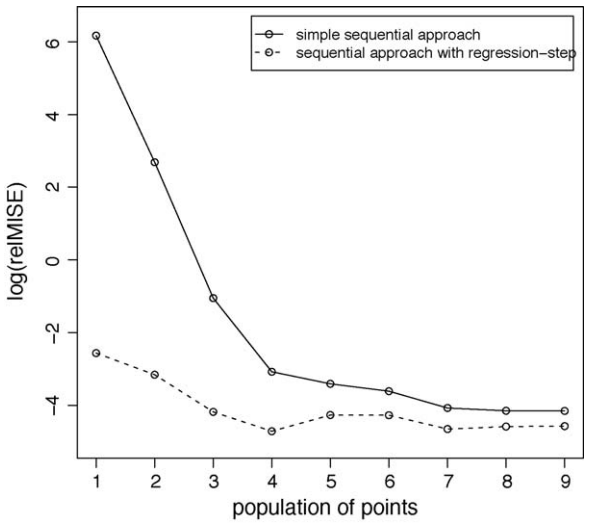

Fig. 4. Performance of the sequential approach with a final regression step (dashed line) and without a final regression step (full line) measured by the RMISE as the number of population of points used increases.

3.3. Conditional density estimation

The main approximation of the ABC algorithm concerns the final step, i.e. when attempting to sample from the posterior distribution. It is essentially a problem in conditional density estimation. In the standard algorithm instead of recovering samples corresponding to the exact posterior distribution, $p(\Phi|D = d')$, samples are taken from an approximation of that distribution using summary statistics, $p(\Phi|S = s')$. Additionally, in order to make the sampling more efficient, that approximation is further relaxed with the use of the rejection-step (Pritchard et al., 1999), $p(\Phi|S - s'| < \delta)$. As seen before, the regression step (Beaumont et al., 2002) allows the use of bigger values of δ by correcting the sampled values according to their distance from the real data set. We can interpret this as an improvement in conditional density estimation. The regression step proposed by Beaumont and co-authors uses a linear regression to study the variation of the expected value of the summary statistics within the local interval of points with a non-zero weight. This method makes two assumptions: a linear relation between the variables in study; and that only the first moment of the distribution, the expected value, is varying within the accepted interval. Both assumptions should hold when dealing with small values of δ. However, methods that use non-linear regression and relax the assumption of constancy of all but the first moment of the sampled points have been explored, and found to produce some improvements on the ABC analyses (Blum and Francois, 2009).

The incorporation of several ABC improvements together should favour further optimization of the method. Fig. 4 presents two sequential approaches consisting of 10 iterations: a standard sequential approach and a sequential approach with a final regression step added in. Both approaches were performed under the same conditions as the previously described results. The inclusion of the final regression step further improves the standard sequential approach. In fact, by the end of the 3rd iteration the sequential approach with the regression step gives results that are as good as or better than the simple sequential approach using 10 iterations.

4. Conclusions

The usage of approximate Bayesian computation methods is spreading fast across many fields. Originally designed to tackle population genetics problems, ABC is now used in vary many different areas. Its only requirement is to be able to produce simulations of the data that one wants to analyse. It is then very simple to export it to different scientific areas. Its main advantage is to allow investigators to assume models with a complexity level impossible to handle by traditional statistic approaches. ABC is also particularly suitable to choose the model that best suits the data from a set of previously chosen models. For these reasons, this Bayesian algorithm has been helping researchers to increase the complexity of the models assumed and the amount of data analysed in such fields as phylogenetics, ecology, conservation, molecular evolution and epidemiology of infectious diseases. Furthermore, due to their close relationship with population genetics, phylogeographic and epidemiological analysis seem ideal for applying the already available ABC tools. These algorithms are still in the process of being optimized and there is considerable scope for improvement. With the increasing development of these methods their range of use can only expand.

Acknowledgements

We would like to acknowledge David Balding for holding stimulating meetings on cutting-edge topics of ABC in Imperial College, which greatly contributed to a better insight on this novel method. We also want to thank two anonymous reviewers for help raising the quality of the original manuscript. This work was funded by an EPSRC grant EP/C533550/1 and a FCT grant SFRH/BD/43588/2008.

References

Anderson, C.N.K., Ramakrishnan, U., Chan, Y.L., Hadly, E.A., 2005. Serial SimCoal: A Population Genetics Model for Data from Multiple Populations and Points in Time. Oxford Univ Press, pp. 1733–1734.
Beaumont, M., 2008. Joint determination of topology, divergence time, and immigration in population trees. In: Matsumura, S., Forster, P., Renfrew, C. (Eds.), Simulations, Genetics, and Human Prehistory. McDonald Institute for Archaeological Research, Cambridge, pp. 135–154.
Beaumont, M., Cornuet, J.M., Marin, J.M., Robert, C.P., 2009. Adaptive approximate Bayesian computation: the ABC-PMC scheme. Biometrika Arxiv preprint arXiv:0805.2256v9.
Beaumont, M.A., 1999. Detecting population expansion and decline using microsatellites. Genetics 153, 2013–2029.
Beaumont, M.A., 2003. Estimation of population growth or decline in genetically monitored populations. Genetics 164, 1139–1160.
Beaumont, M.A., Rannala, B., 2004. The Bayesian revolution in genetics. Nat. Rev. Genet. 5, 251–261.
Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian computation in population genetics. Genetics 162, 2025–2035.
Blum, M.G.B., Francois, O., 2009. Non-linear regression models for approximate Bayesian computation. Stat. Comput., doi:10.1007/s11222-009-9116-0.
Blum, M.G.B., 2009. Approximate Bayesian Computation: A Non-parametric Perspective. , Arxiv preprint arXiv:0904.0635.
Bonhomme, M., Blancher, A., Cuartero, S., Chikhi, L., Crouau-Roy, B., 2008. Origin and number of founders in an introduced insular primate: estimation from nuclear genetic data. Mol. Ecol. 17, 1009–1019.
Bonizzoni, M., Zheng, L., Guglielmino, C.R., Haymer, D.S., Gasperi, G., Gomulski, L.M., Malacrida, A.R., 2001. Microsatellite analysis of medfly bioinfestations in California. Mol. Ecol. 10, 2515–2524.
Bortot, P., Coles, S.G., Sisson, S.A., 2007. Inference for stereological extremes. J. Am. Stat. Assoc. 102, 84–92.
Chan, Y.L., Anderson, C.N.K., Hadly, E.A., 2006. Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA. PLoS Genet. 2, 59.
Cornuet, J.M., Santos, F., Beaumont, M.A., Robert, C.P., Marin, J.M., Balding, D.J., Guillemaud, T., Estoup, A., 2008. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics 24, 2713.
Crandall, K.A., 1999. The Evolution of HIV. Johns Hopkins University Press.
Davies, N., Villablanca, F.X., Roderick, G.K., 1999. Bioinvasions of the medfly Ceratitis capitata source estimation using DNA sequences at multiple intron loci. Genetics 153, 351–360.
Dawson, K.J., Belkhir, K., 2001. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genet. Res. 78, 59–77.
Del Moral, P., Doucet, A., Jasra, A., 2006. Sequential Monte Carlo samplers. J. R. Stat. Soc. B 68, 411–436.

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., Solomon, W., 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161, 1307–1320.
Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22, 1185–1192.
Edwards, S.V., Beerli, P., 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54, 1839–1854.
Estoup, A., Beaumont, M., Sennedot, F., Moritz, C., Cornuet, J.M., 2004. Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, Bufo marinus. Evolution 58, 2021–2036.
Estoup, A., Clegg, S.M., 2003. Bayesian inferences on the recent island colonization history by the bird Zosterops lateralis lateralis. Mol. Ecol. 12, 657–674.
Estoup, A., Wilson, I.J., Sullivan, C., Cornuet, J.M., Moritz, C., 2001. Inferring population history from microsatellite and enzyme data in serially introduced cane toads, Bufo marinus. Genetics 159, 1671–1687.
Excoffier, L., Estoup, A., Cornuet, J.M., 2005. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics 169, 1727–1738.
Fagundes, N.J.R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F.M., Bonatto, S.L., Excoffier, L., 2007. Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. U.S.A. 104, 17614.
Felsenstein, J., 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. 59, 139.
Ford, M.J., 1998. Testing models of migration and isolation among populations of chinook salmon (Oncorhynchus tschawytscha). Evolution 52, 539–557.
Fu, Y.-X., 2001. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. Mol. Biol. Evol. 18, 620–626.
Fu, Y.X., Li, W.H., 1997. Estimating the age of the common ancestor of a sample of DNA sequences. Mol. Biol. Evol. 14, 195–199.
Grelaud, A., Robert, C.P., Marin, J.M., 2009. ABC methods for model choice in Gibbs random fields. Comptes rendus-Mathématique, doi:10.1016/j.crma.2008.12.009.
Hamilton, G., Currat, M., Ray, N., Heckel, G., Beaumont, M., Excoffier, L., 2005. Bayesian estimation of recent migration rates after a spatial expansion. Genetics 170, 409–417.
Hickerson, M.J., Dolman, G., Moritz, C., 2005. Comparative phylogeographic summary statistics for testing simultaneous vicariance. Mol. Ecol. 15, 209–223.
Hickerson, M.J., Stahl, E.A., Lessios, H.A., 2006. Test for simultaneous divergence using approximate Bayesian computation. Evolution 60, 2435–2453.
Hickerson, M.J., Stahl, E., Takebayashi, N., 2007. msBayes: pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. BMC Bioinformatics 8, 268.
Hoggart, C.J., Parra, E.J., Shriver, M.D., Bonilla, C., Kittles, R.A., Clayton, D.G., McKeigue, P.M., 2003. Control of confounding of genetic associations in stratified populations. Am. J. Hum. Genet. 72, 1492–1504.
Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23, 183–201.
Hudson, R.R., 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18, 337–338.
International HapMap C, 2003. The international HapMap project. Nature 426, 789–796.
International Human Genome Sequencing C, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
Jakobsson, M., Hagenblad, J., Hagenblad, J., Tavare, S., Sall, T., Hallden, C., Lind-Hallden, C., Nordborg, M., 2006. A unique recent origin of the allotetraploid species arabidopsis suecica: evidence from nuclear DNA markers. Mol. Biol. Evol. 23, 1217–1231.
Joyce, P., Marjoram, P., 2008. Approximately sufficient statistics and Bayesian computation. Stat. Appl. Genet. Mol. Biol. 7, 26.
Kingman, J.F., 1982. The coalescent. Stoch. Proc. Appl. 13, 235–248.
Kruglyak, L., 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. 22, 139–144.
Laval, G., Excoffier, L., 2004. SIMCOAL 2.0: A Program to Simulate Genomic Diversity Over Large Recombining Regions in a Subdivided Population with a Complex History. Oxford Univ. Press, pp. 2485–2487.
Legras, J., Merdinoglu, D., Cornuet, J.M., Karst, F., 2007. Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. Mol. Ecol. 16, 2091–2102.
Leuenberger, C., Wegmann, D., Excoffier, L., 2009. Bayesian Computation and Model Selection in Population Genetics. , Arxiv preprint arXiv:0901.2231v1.
Liu, J.S., Sabatti, C., Teng, J., Keats, B.J.B., Risch, N., 2001. Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res. 11, 1716–1724.
Lopes, J.S., Balding, D., Beaumont, M.A., 2009. PopABC: a program to infer historical demographic parameters. Bioinformatics, doi:10.1093/bioinformatics/btp487.
Lopes, J.S., Boessenkool, S., 2009. The use of approximate Bayesian computation in conservation genetics and its application in a case study on yellow-eyed penguins. Conserv. Genet..
Luciani, F., Sisson, S.A., Jiang, H., Francis, A.R., Tanaka, M.M., 2009. The epidemiological fitness cost of drug resistance in Mycobacterium tuberculosis. Proc. Natl. Acad. Sci. U.S.A. 106 (34), 14711–14715.
Marjoram, P., Molitor, J., Plagnol, V., Tavare, S., 2003. Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. U.S.A. 100, 15324–15328.
McKinley, T., Cook, A.R., Deardon, R., 2009. Inference in epidemic models without likelihoods. Int. J. Biostat. 5 (1), 24.
Miller, N., Estoup, A., Toepfer, S., Bourguet, D., Lapchin, L., Derridj, S., Kim, K.S., Reynaud, P., Furlan, L., Guillemaud, T., 2005. Multiple transatlantic introductions of the western corn rootworm. Science 310, 992.
Morris, A.P., Whittaker, J.C., Balding, D.J., 2002. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am. J. Hum. Genet. 70, 686–707.
Neuenschwander, S., Largiader, C.R., Ray, N., Currat, M., Vonlanthen, P., Excoffier, L., 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.
Nielsen, R., Wakeley, J., 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158, 885–896.
Nielsen, R., Yang, Z., 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148, 929–936.
Nordborg, M., Tavare, S., 2002. Linkage disequilibrium: what history has to tell us. Trends Genet. 18, 83–90.
O'Neill, P.D., Balding, D.J., Becker, N.G., Eerola, M., Mollison, D., 2000. Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. J. R. Stat. Soc. C: Appl. Stat. 49, 517–542.
Pascual, M., Chapuis, M.P., Mestres, F., Balanya, J., Huey, R.B., Gilchrist, G.W., Serra, L., Estoup, A., 2007. Introduction history of Drosophila subobscura in the New World: a microsatellite-based survey using ABC methods. Mol. Ecol. 16, 3069–3083.
Plagnol, V., Tavare, S., 2003. Approximate Bayesian computation and MCMC. In: Niederreiter, H. (Ed.), Monte Carlo and Quasi-Monte Carlo Methods 2002. Springer-Verlag, Heidelberg.
Pritchard, J.K., Donnelly, P., 2001. Case-control studies of association in structured or admixed populations. Theor. Popul. Biol. 60, 227–238.
Pritchard, J.K., Przeworski, M., 2001. Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. 69, 1–14.
Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W., 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. 16, 1791–1798.
Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.
Pybus, O.G., Drummond, A.J., Nakano, T., Robertson, B.H., Rambaut, A., 2003. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. Mol. Biol. Evol. 20, 381–387.
Rannala, B., 2001. Finding genes influencing susceptibility to complex diseases in the post-genome era. Am. J. Pharmacogenomics 1, 203.
Rannala, B., Reeve, J.P., 2001. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am. J. Hum. Genet. 69, 159–178.
Rannala, B., Yang, Z., 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656.
Ratmann, O., Jørgensen, O., Hinkley, T., Stumpf, M., Richardson, S., Wiuf, C., 2007. Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. PLoS Comput. Biol. 3, e230.
Rodrigo, A.G., Shpaer, E.G., Delwart, E.L., Iversen, A.K.N., Gallo, M.V., Brojatsch, J., Hirsch, M.S., Walker, B.D., Mullins, J.I., 1999. Coalescent estimates of HIV-1 generation time in vivo. Proc. Natl. Acad. Sci. U.S.A. 96, 2187–2191.
Rosenberg, N.A., Nordborg, M., 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3, 380–390.
Rosenblum, E.B., Hickerson, M.J., Moritz, C., 2007. A multilocus perspective on colonization accompanied by selection and gene flow. Int. J. Organ. Evol. 61, 2971–2985.
Shriner, D., Liu, Y., Nickle, D.C., Mullins, J.I., 2006. Evolution of intrahost HIV-1 genetic diversity during chronic infection. Evolution 60, 1165–1176.
Sillanpaa, M.J., Kilpikari, R., Ripatti, S., Onkamo, P., Uimari, P., 2001. Bayesian association mapping for quantitative traits in a mixture of two populations. Genet. Epidemiol. 21, S692–S699.
Sisson, S.A., Fan, Y., Tanaka, M.M., 2007. Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. U.S.A. 104, 1760.
Sousa, V.M., Fritz, M., Beaumont, M.A., Chikhi, L., 2009. Approximate Bayesian computation (ABC) without summary statistics: the case of admixture. Genetics, doi:10.1534/genetics.108.098129.
Storz, J.F., Beaumont, M.A., 2002. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. Evolution 56, 154–166.
Storz, J.F., Beaumont, M.A., Alberts, S.C., 2002. Genetic evidence for long-term population decline in a savannah-dwelling primate: inferences from a hierarchical Bayesian model. Mol. Biol. Evol. 19, 1981–1990.
Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105, 437–460.
Takahata, N., Lee, S.H., Satta, Y., 2001. Testing multiregionality of modern human origins. Mol. Biol. Evol. 18, 172–183.
Tallmon, D.A., Luikart, G., Beaumont, M.A., 2004. Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. Genetics 167, 977–988.
Tallmon, D.A., Koyuk, A., Luikart, G., Beaumont, M.A., 2008. ONeSAMP: a program to estimate effective population size using approximate Bayesian computation. Mol. Ecol. Resour. 8, 299–301.
Tanaka, M.M., Francis, A.R., Luciani, F., Sisson, S.A., 2006. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics 173, 1511–1520.

Tavare, S., Balding, D.J., Griffiths, R.C., Donnelly, P., 1997. Inferring coalescence times from DNA sequence data. Genetics 145, 505–518.
Thompson, R.C.A., 2000. Molecular Epidemiology of Infectious Diseases. Arnold.
Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.H., 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface 6, 187–202.
Venter, J.C., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
Weiss, G., von Haeseler, A., 1998. Inference of population history using a likelihood approach. Genetics 149, 1539–1546.
Wilkinson, R.D., 2008. Approximate Bayesian Computation (ABC) Gives Exact Results under the Assumption of Model Error. , Arvix preprint arXiv:0811.3355.
Wilson, D.J., Gabriel, E., Leatherbarrow, A.J.H., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Hart, C.A., Diggle, P.J., Fearnhead, P., 2009. Rapid evolution and the importance of recombination to the gastroenteric pathogen Campylobacter jejuni. Mol. Biol. Evol. 26, 385.
Wilson, I.J., Balding, D.J., 1998. Genealogical inference from microsatellite data. Genetics 150, 499–510.
Wilson, I.J., Weale, M.E., Balding, D.J., 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities (with discussion). J. R. Stat. Soc. Stat. Soc. 166, 155–202.
Wright, S.I., Bi, I.V., Schroeder, S.G., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., 2005. The effects of artificial selection on the maize genome. Science 308, 1310.