STUDY DESIGNS

Computer simulations: tools for population and evolutionary genetics

Sean Hoban[1,2], Giorgio Bertorelle[2] and Oscar E. Gaggiotti[1]

Abstract | Computer simulations are excellent tools for understanding the evolutionary and genetic consequences of complex processes whose interactions cannot be analytically predicted. Simulations have traditionally been used in population genetics by a fairly small community with programming expertise, but the recent availability of dozens of sophisticated, customizable software packages for simulation now makes simulation an accessible option for researchers in many fields. The *in silico* genetic data produced by simulations, along with greater availability of population-genomics data, are transforming genetic epidemiology, anthropology, evolutionary and population genetics and conservation. In this Review of the state-of-the-art of simulation software, we identify applications of simulations, evaluate simulator capabilities, provide a guide for their use and summarize future directions.

Stocking
Human-mediated supplementation of a native population with translocated or captive-bred individuals to increase population size or growth rates.

[1]*Laboratoire d'Ecologie Alpine, UMR CNRS 5553, Université Joseph Fourier, BP 53, 38041 Grenoble, France.*
[2]*Dipartimento di Biologia ed Evoluzione, Università di Ferrara, via L. Borsari, 46, 44100 Ferrara, Italy.*
Correspondence to O.E.G.
e-mail:
oscar.gaggiotti@ujf-grenoble.fr
doi:10.1038/nrg3130
Published online
10 January 2012

Computer simulations are computer programs that have been developed to simulate abstract models of particular systems. In evolutionary applications, the systems — that is, the individuals, populations or species — undergo multiple processes with complex interactions that are difficult to predict. Natural processes include mutation, disease, natural selection and spatial expansions, and human-mediated processes include stocking, fragmentation and artificial selection. Computer simulations are excellent tools for understanding the functioning of these complex systems. The general principle is to generate *in silico* data sets (known as pseudo data) of genetic polymorphism under specified scenarios describing the evolutionary history and genetic architecture of a species. For example, in understanding human evolution[1], one scenario might consider population expansion with a specific geographic origin, ancestral and descendant population sizes, and dispersal rates within and between continents. It would also include a specific type of genetic marker, a mutation rate and assumptions about linkage between markers. Repeated generation of pseudo data incorporates the inherent stochasticity of demographic and genetic mechanisms and sampling. These data sets are used to predict the effects of interacting forces, to infer historical processes (when compared to real data) or to understand the properties of newly developed methods.

Simulations have long been used to explore analytically intractable genetic models. An early example is the study[2] that introduced the stepwise mutation model and compared simulated allelic distributions to expectations under the infinite alleles model. They have also been used as a complement to analytical models based on approximations that needed verification[3]. As the use of molecular data and interest in statistical methods increased, so did the need to evaluate the performance of these methods with synthetic data (for example, REF. 4). The emergence of coalescent theory[5,6] provided new impetus to the use of simulations to estimate parameters[7], to compute parametric bootstrap confidence intervals[8] or to compare models[9]. However, simulation use was circumscribed to a small community that had the skills to design and code software. In the past decade, standalone simulation programs have become available, allowing non-specialists to address an ever-widening range of evolutionary, epidemiological and conservation problems.

Current simulation programs have achieved a sophistication level that matches modern genetic, historical and ecological data sets. For example, some generate pseudo data collected at multiple time points that are similar to data from a collection of museum specimens or ancient DNA[10]. Others generate hundreds of DNA markers or chromosome-length stretches of sequence under particular recombination and mutation models[11,12]; these are useful in the genomic age. Some programs model life cycle, mating system and even phenotypes, allowing the incorporation of phenotypic[13], geographic[14], ecological[15] and life history[16] data. Thus, simulation programs are truly multi-disciplinary. However, they have a wide array

© 2012 Macmillan Publishers Limited. All rights reserved

of options, capabilities, limitations, input formats and assumptions. Without a coherent guide, they are liable to misuse and are unlikely to achieve a wider use. Unlike previous reviews[17–19] that addressed technical issues, our goal is to connect particular simulators to particular research topics. We emphasize that simulations powerfully complement theoretical and statistical approaches for understanding genetic variation patterns.

First, we present an overview of the research problems that can be addressed by simulation and then provide a step-by-step guide to designing a simulation-based study. We describe the capabilities of 42 simulation packages (TABLE 1; Supplementary information S1,S2 (tables)). No simulator is an 'all-around' solution; this section and the accompanying tables will help the readers to match a simulator to their needs. We also explain the distinction between forwards-in-time simulations (also known as individual-based simulations) and backwards-in-time simulations (also known as coalescent simulations) (BOX 1) and explain which investigations each approach is suited for. We close with a discussion of practical issues, common problems and future improvements. Note that the links to the Web pages and references for all of the software packages described in this paper are listed in Supplementary information S1 (table).

Potential applications

We distinguish three main applications of population genetic simulations: predictive, statistical inference and evaluation of statistical genetics methods.

Predictive uses. Most scientific theories are built on mathematical models that predict outcomes under specific, pre-defined assumptions. Although they provide transparent analytical results, such models become mathematically intractable under realistic ecological and genetic scenarios, or they may use approximations that have an uncertain impact. Simulations avoid this problem and are an ideal complement to analytical approaches. Thus, they have become an established predictive tool in theoretical evolutionary biology for addressing problems such as the evolution of sexual reproduction[20] and speciation theory[21,22]. However, they are equally applicable in conservation genetics and genetic epidemiology, so here we explain their use in these settings.

Conservation, management and restoration of threatened or economically important species is complicated by many factors, including climate change, disease and exploitation. Typically, the goal is to predict the probable genetic impact of future ecological change or possible human interventions. A common approach is to simulate populations under various scenarios, to monitor them for genetic variation statistics that are considered relevant and to identify parameter values (for example, migration rates) under each scenario that allow a goal to be achieved. Examples include: evaluating the genetic consequences (for example, inbreeding, fecundity and genetic diversity) of maintaining a barrier between populations of European bison[23]; determining the minimum population size and immigration rates required to maintain heterozygosity in Yellowstone wolves[24]; determining the number of colonies required to maintain sex allele variation in managed bee populations[25]; and predicting the genetic impact of population reductions (for example, harvesting[26] and poaching[27]). These problems have been addressed using forward simulators (BOX 1), such as Vortex and BottleSim. Genetic viability analyses complement more traditional demographic population viability analysis (PVA), reflecting an increasing recognition that both genetic and demographic factors contribute to population extinction. Simulations can be parameterized with information collected from field studies of endangered species, including demography and environmental variables[28,29].

Predictive simulations are also useful in genetic epidemiology. An early application investigated how intergenerational differences in tobacco consumption could bias segregation analysis that is aimed at estimating the underlying genetic basis of lung cancer. It used an ad hoc program that simulated pedigrees and assumed an autosomal dominant mode of susceptibility to the disease[30]. More recent applications evaluate the plausibility of more general disease models. For example, the common disease–common variant (CDCV) hypothesis states that common diseases should have a few common alleles at each disease-susceptibility locus, but the necessary demographic and genetic conditions for the CDCV are debated. Forwards-in-time simulations have been used to test this hypothesis under two models — equilibrium and expansion — with a variety of parameters for mutation, migration, selection and genetic architectures[31]. SimuPop[12] was used to consider complex selection models and to obtain samples of the allele frequency spectra at multiple time points. Simulations under a model matching our current knowledge of human evolutionary history (recent expansion leading to non-equilibrium conditions) explains the high diversity of rare diseases well and gives support to the CDCV hypothesis. However, the >2,000 common variants linked to common diseases that have been uncovered by genome-wide association studies (GWASs) explain only a very small fraction of their estimated heritability[32]. Simulation studies are helping to explain this so-called 'missing heritability' phenomenon. A recent study[33] used Genome to simulate the genealogy of a population in which some individuals had rare variants associated with a disease. This study shows that weak signals detected for common variants could come from the effect of rare ones that could have large effects but that remain undetected by GWASs.

Simulations have also been used to explain several features of major histocompatibility complex (MHC) diversity, including the close associations between particular human leucocyte antigen (HLA) haplotypes and a large number of human pathologies that are due to partially recessive mutations[34]. A more general epidemiological study used forward simulations to show that historical bottlenecks can produce an excess of segregating damaging alleles, possibly explaining why human populations have a higher proportion of deleterious mutations in Europe than in Africa[35].

Stepwise mutation model
A mutation model in which the allelic states produced by mutation depend on the initial state of an allele. The basic version assumes mutations between adjacent states, but other versions allow larger mutational changes. This model is commonly used to model the microsatellite mutation process.

Infinite alleles model
A model in which each mutational event creates a new allele that is unlike any other that is currently present in the population.

Coalescent theory
A theory describing the genealogy of chromosomes or genes. The genealogy is constructed backwards-in-time, starting with the present-day sample. Lineages coalesce until the most recent common ancestor of the sample is reached.

Parametric bootstrap confidence intervals
These measure the accuracy of sample estimates using a bootstrapping approach where a parametric model is fitted to the data, and samples of parameter values are drawn from this fitted model.

Population viability analysis
(PVA). A probability-based modelling approach for assessing the future potential (such as reproduction and extinction) of populations or species.

© 2012 Macmillan Publishers Limited. All rights reserved

Table 1 | **Capabilities of simulators**

Program	SECs[a]	M/D[b]	MS[c]	Fec[d]	LC[e]	PG[f]	Events[g]	Selection[h]	Mutation[i]	Rec[j]
Backward simulators										
AquaSplatche	P, 1D	SS, E, DD	RM	RD	NA	L(Poi)	‡	NS	M(G), RF, Seq(K2, J), SNP, TB, H	N
Splatche	P, 2D	SS, E	RM	RD	NA	K, L	‡	NS	M(SMM), RF, Seq(K2, J), SNP, TB, H	N
CoaSim		UD	RM	RD	NA	Ex, PSp	PopS, GR, MM, FF, RP	NS	M(SMM, KA), SNP, UD	Y
cosi		UD	RM	RD	NA	Ex	PopS, CE, MM	SBS	Seq	Y, V
CoalFace		NA(1P)	RM	RD	NA	Ex	GR	NS	M(SMM), Seq(J, F81, K2, HK)	N
FastCoal		NA(1P)	RM	RD	NA	Ex	GR	NS	SNP	Y
GeneArtisan		NA(1P)	RM	RD	NA	Ex	NE	SBS	Seq(J), M(SMM)	Y
Genome		IM	RM	RD	NA	C, UD	PopS, FF	NS, MPG	SNP, Seq	Y, V
IBDsim	P, I, 2D	SS, DK, IM, IBD	RM	RD	NA	C	D, LS, DK	NS	M(KA, SMM, G, T), H, *Seq(K2, J, HK)*, *SNP*	N
MaCS		IM	RM	RD	NA	Ex	GR, MM, RP, FF, CE, PopS	NS	SNP	Y
ms		IM, UD	RM	RD	NA	Ex	PopS, GR, MM, FF	NS	Seq	Y
mbs		IM	RM	RD	NA	Ex	PopS, GR, MM, FF	SBS, DS, B	Seq, MH	Y, V
mshot		IM, UD	RM	RD	NA	Ex, PSp	PopS, GR, MM, FF	NS	Seq	Y, V
msms		IM, UD	RM	RD	NA	Ex	PopS, GR, MM, FF, RP, Sel	SBS, Ep, G, TV	S, *K2*	Y, *V*
mlcoalsim		IM, *UD*	RM	RD	NA	L	PopS, GR, FR	SBS, DS	Seq, MH, H	Y, V
Recodon		UD	RM	RD	NA	Ex	FF	NS	Seq, TB, Co	Y
SelSim		NA(1P)	RM	RD	NA	C	NE	SBS, DS, B	M(SMM), SNP	Y, V
Serial Netevolve		IM	RM	RD	NA	Ex	PopS, FF	NS	Seq(F81, HK, GT, K2, J)	Y, V
Bayesian Serial SimCoal (BayeSSC)	P, I, 2D	UD	RM	RD	NA	UD, Ex, PSp	PopS, GR, MM, FF, CE	NS	M(KA, T), RF, SNP, Seq, TB, H, TV	N
SimCoal2		UD	RM	RD	NA	UD, Ex, PSp	PopS, GR, MM, FF, CE	NS	M(KA, T, G, SMM), RF, SNP, Seq, TB, H	Y, V
fastsimcoal		UD	RM	RD	NA	Ex	PopS, GR, MM, FF, CE	NS	M(KA, T, G, SMM), SNP, Seq, MH, TB, H	Y, V
SNPsim		NA(1P)	RM	RD	NA	Ex, UD	GR, RP	NS	SNP, Seq(J)	Y, V
Forward simulators										
BottleSim		NA(1P)	RM, PS, CS, SM, SP	RD	O, T, AR, AM	UD	PopS	NS	NM	N
cdpop	I, P	SS, IBD, Sx, E	RM, SP, Pg, Pa, M, CS, AS	RD, Poi, CN	O, AM	C, PSp	D, LS, Sel‡	SBS, MBS, EB, *MMS*, *FS*, *ID*	M(KA, SMM)	N, Y
Easypop	P, 2D	SS(H), IM(H), IB	RM, Pg(PE), M(PE), H	RD	NA	C	MM	NS	M(KA, SMM, T)	Y
ForSim	I, 2D	UD(H, SF)	RM, AP	Poi	NA	K, L	Sel, CE, MM, GR	MBS, DS, TS, MPG	SNP	Y
ForwSim		NA(1P)	RM	RD	NA	C	NE	MPG, DS, B	(Y)	Y
FPG		IM	RM	RD	NA	C	NE	MBS, DS, ME, Ep	SNP	Un
FreGene		UD	RM	UDP	NA	K, Ex	GR, MM, RP	MBS, DS, B	SNP, MH	Y, V

© 2012 Macmillan Publishers Limited. All rights reserved

Table 1 (cont.) | **Capabilities of simulators**

Program	SECs[a]	M/D[b]	MS[c]	Fec[d]	LC[e]	PG[f]	Events[g]	Selection[h]	Mutation[i]	Rec[j]
GenomePop/ GenomePop2	P, 1D	SS, IM, UD	RM	RD	NA	C, K	PopS, *FF*	MBS, DS, ME, CB, EB	Seq(J, GT), SNP, Co, TB	Y, V
GenomeSimla		NA(1P)	RM	Un	NA	K, L, Ex	NE	Disease	SNP	Y, V
Mendels Accountant	P, 2D	IM, SS	RM, PS	CN(FI)	NA	K, Ex	PopS, GR	MBS, DS, TS, ME	Seq, H	Y, V
Nemo 2.2	P, 2D	IM, Pr, SS, UD, Sx, H, IBD	RM, SM, M, Pg, PS, CS, PE	Poi, ND, CN	NA	K, UD, PSp	MM, PopS, FF, EH(St), CE(St), Sel	SP, EB, IF, ID, SBS, MBS, MMS, MPG, B, ME, TV, Disr	M(KA, SMM), SNP	Y, V
QuantiNemo	P, 2D	IM, Pr, SS, UD, DD, Sx	RM, SM, M, PG, PS, CS	Poi, CN	NA	K, AS	MM, PopS, Sel, E(St)	MPG, DS, BM, TV, EB	M(SMM), RIQ, H	Y, V
Pedagog		UD, DD, Sx, AS	RM, M, Pg, Pa, Sx, AS	CN, UDD, AI, Sx	O, TM, AM	K, L, Ex, AS	PopS(St), EH	MPS, DS, BM, Disr, B	M(SMM), H	N
QMSim		UD	RM, AP, MMI, SM, SP	CN, UDP	O, AR, PM	C, UD	PopS, FF, GR, Sel, F, DE	MPG, TS	SNP, M	Y, V
RmetaSim/ metasim	P, 2D	UD, AS	RM, Sx, AS	Poi, DI	O, TM	K	F, MM, MP(St)	NS	M(SMM), Seq, H	N
KernelPop	I, 2D	DK	RM, Sx, AS	Poi, DI	O, TM	K	F, MM, MP(St)	NS	M(SMM), Seq, H	N
SFS_Code		IM, UD, SS, Sx	RM, PS, Sx, SM, CS, H	RD, B, FI	NA	K, L, Ex, UD	GR, FF, CE, PopS, Sel, MM, RP, DE, EH	MBS, MMS, ME, CB, DS, *B*	G, Seq(J, K2, GT, HK), MH, SNP, In, H, TB, Co	Y, V
SimuPop	P, I, 2D	IM(H), SS, UD, Sx	RM, M, Pg, PS, H, PP, SM, Pa, AS*	B, Poi, CN, Sx, UDD, FI	O, AR, AM, PM	UD	MM, PopS, FF, RH, Sel, MP	MBS, MMS, EB, ME, FS, CB, SP, MP, DS, B, TS, TV	M(SMM, G, T), Seq(J, GT, K2, F81, H, T92), SNP, H, In	Y, V
Spip/ Spip_m		UD, Sx	RM, PE, M, SM	CN, AI, Sx, B	AM, O, PM	UD	MM, PopS	NS	NM	N
Vortex		UD, Sx, AS	Pg, M, Sx, AS*	CN, UDD, AI, Sx, DI	O, TM, AR, AM, PM	K, PSp	GR, PopS, EH, S(St), MM, CE	IF, MMS	Mitochondria only, *K2*	N

Additional information is included in Supplementary information S1,S2 (tables). Features in normal type are currently available; features in italics are planned to be released by approximately February 2012. | [a]Spatially explicit considerations (spatial coordinates). Populations (P) or individuals (I) can be modelled on a lattice. 1D, 1-dimensional; 2D, 2-dimensional. | [b]Migration or dispersal. AS, age- or stage-specific; DD, density dependence; DK, dispersal kernel (for example, Weibull distribution); E, influenced by environmental factors; H, hierarchical structure possible; IM, Wright's Island model; IBD, isolation by distance (exponential decay function); NA(1P), not applicable (single population modelled); Pr, propagule pool; SF, influenced by selective forces; SS, stepping stone; Sx, sex-specific migration rates; UD, user-defined matrix. | [c]Mating system. AP, assortative based on phenotype; AS, age- or stage-specific; CS, complete self-fertilizing (selfing); H, haplodiploid; M, monogamous; Pa, polyandrous; PE, proportion extramarital; Pg, polygamous; PP, defined per population; PS, partial selfing; RM, random mating; Sx, can define specific sex ratios; SM, single male; SP, single pair. *Also allowed are definitions of maximum breeding age and percent male and females able to breed. | [d]Fecundity. AI, age-influenced; B, binomial distribution; CN, constant number; DI, density-influenced; FI, influenced by fitness; ND, normal distribution; Poi, Poisson distribution; RD, random distribution; Sx, sex-influenced; UDD, user-defined distribution; UDP, user-defined probability; Un, unknown. | [e]Life cycle. AR, user-defined age of reproduction; AM, user-defined age of sexual maturity; NA, not applicable; O, overlapping; PM, user-defined percentage mortality at each stage; TM, user-defined transition matrices. | [f]Population growth. AS, age- or stage-specific carrying capacities; C, only allows constant size — size does not change (except at event times); Ex, exponential growth; K, carrying capacity; L, carrying capacity defined by logistic growth; PSp, population-specific; Poi, carrying capacity is determined each generation by a Poisson distribution; UD, user-defined population size every time step. | [g]Events allowed. CE, colonization or extinction (change in number of populations); D, density; DE, domestication event; DK, dispersal kernel; EH, extinction or harvesting; F, fecundity; FF, population fission (splitting) or fusion (admixture); FR, colonization from a refuge; GR, population growth rate; LS, lattice size; MM, migration matrix; MP, mating probabilities (between sexes and stages); NE, no events; PopS, population size; RP, recombination parameters; S, stocking from a captive source; Sel, selection strength; St, stochastic option (that is, events can be modelled to occur with a given probability). ‡The environment (friction and carrying capacity) can change at time points, which subsequently affects migration rates and population size. | [h]Selection. B, balancing selection; BM, built-in trait models (such as size, age at maturity, movement and survival); CB, codon-based selection; Disr, disruptive; DS, directional selection; Ep, epistasis; EB, environment-based or population specificity instead of global fitness; FS, frequency-dependent selection; ID, infection dynamics; IF, inbreeding affects fitness; ME, multiplicative effects; MBS, multiple biallelic sites; MMS, multiple multilocus selection; MPG, selection on multiple phenotypes or genetic values (which are determined by quantitative trait loci (QTLs)); NS, no selection — only neutral markers; SBS, single biallelic site; SP, selection on single phenotype; TS, threshold selection removes all individuals with a phenotype above a threshold — stochastic selection removes individuals above threshold with set probabilities; TV, time variable. | [i]Mutation. Co, codons; F81, Felsenstein 81; G, general stepwise mutation model; GT, general time reversible model; H, heterogeneity in mutation among sites; HK, HKY85; In, insertions or deletions (indels) allowed; K2, Kimura 2; J, Jukes Cantor; KA, *k* allele model; M, microsatellites; MH, multiple hits allowed; NM, no mutation; RF, restriction fragment length polymorphisms (RFLPs); RIQ, mutations at a QTL can have a random or incremental effect; Seq, sequence; SMM, strict stepwise mutation model; SNP, single-nucleotide polymorphism; T, two-phase model; TB, transition bias; TV, mutation rate can vary in time; UD, user-defined — user writes their own script; (Y), mutations allowed but parameters unclear. | [j]Recombination. N, no; Un, unknown; V, variation allowed by defining hot spots or a genetic map; Y, yes.

© 2012 Macmillan Publishers Limited. All rights reserved

F_{IS}
Wright's inbreeding coefficient, measuring the level of correlation between two genes drawn from an individual relative to two genes drawn from the population. Also defined as the probability that two alleles in an individual are both descended from a single allele in an ancestor.

Summary statistics
Numerical values for summarizing the characteristics of a genetic data set; these often summarize features such as variability (number of alleles) or population differentiation (F_{ST}).

Statistical inference. The past decade has witnessed an increased interest in unravelling the evolutionary history of species, and simulations are crucial for inferring such histories. In particular, simulations have been embraced for inferring demographic expansion and migration in humans[1,36,37]. A common problem is to evaluate the plausibility of alternative hypotheses and to estimate demographic and genetic parameters under the best-supported model. Typically, data are generated under alternative models of evolutionary history, and relevant population-genetics statistics (such as F_{ST}, number of alleles and F_{IS}) are used to summarize each data set, creating a distribution of possible values under each scenario. Summary statistics obtained from observed data are matched to these distributions using a variety of ad hoc methods or, more recently, using approximate Bayesian computation (ABC), which is a general statistical approach that has revolutionized the use of simulations for statistical inference (BOX 2).

Box 1 | **Forward and backward approaches**

The two categories of simulation algorithms, forward and backward, differ in approach, capabilities and computation times, so they are suitable for addressing different questions.

Forwards-in-time simulators are centred on individuals: each individual in the simulated population (or populations) follows a life cycle (that is, birth, selection, mating, reproduction, mutation, migration and death). Looking forward, the demographic and genetic makeup of subsequent generations is determined by the current generation and a series of transition matrices[16]. This approach allows researchers to monitor changes in the genetic composition of a population analysing samples at specific time intervals.

Backwards-in-time simulators take a lineage approach. For each gene, a sample of copies is followed back in time to the most recent common ancestor. Backwards in time, lineages coalesce progressively with a probability that is influenced by features of the sample and the evolutionary history[69,94,95]. After the gene tree has been generated, mutations are added to the branches (following the mutation model chosen by the user). There are two strategies for this[96] — to place mutations at a constant rate of $\theta/2$ along each of the branches (where θ is the population mutation parameter) or to place S mutations on the tree (where S is the observed number of segregating sites).

Because the coalescent approach used in backward simulators only considers the genealogy of the sampled DNA fragments and not each single individual in the population, these simulators are generally faster (much faster when population sizes are large relative to sample size). However, this approach precludes life history modelling, so backward simulators are suited to inferential questions on an evolutionary timescale, and situations in which deviations from the reproductive scheme assumed by the Wright–Fisher model are minor[47]. Forward simulators are slower because they follow each individual, but they can model more complexity, making them more suited to predictive questions at a short timescale, particularly with well-characterized species[28,29]. Additionally, backward simulators are more limited than forward simulations for modelling natural or artificial selection. Currently, the backward simulators ms and SimCoal2 are the most widely used, probably owing to their flexibility and ease of use, as well as their speed.

Another important difference is that forward simulations require defining initial conditions of genetic variation (each individual needs an initial genotype). This is a disadvantage when the starting point is relevant and no real data are available. In this case, the standard practice is to use fairly arbitrary initial conditions (for example, uniform allele frequencies) and to include a 'burn-in' period in each run during which no output is generated (sometimes this period can be thousands of generations). This allows the system to reach a stochastic equilibrium — that is, in which the final state of the system is not influenced by the initial conditions. Alternatively, the user can 'seed' the forward simulator with samples generated using backward simulators (for example, RmetaSim can be seeded with output from SimCoal2); this strategy is termed the 'sideways approach'[97]. Any forward simulator that allows specifying initial allele frequencies as starting conditions could use this approach.

Simulations have been used for statistical inference in evolutionary biology, ecology, conservation and epidemiology. An economically and ecologically important use is in estimating demographic parameters of species invasions. Examples include estimation of the number and source of founder individuals of the highly invasive American bullfrog (using RmetaSim)[38] and inferring connectivity between recently established populations of an invasive, ecosystem-altering sea urchin (*Centrostephanus rodgersii*) and putative Australian source populations (using Splatche)[39]. Evolutionary uses include estimating the size and duration of bottlenecks[40,41], migration patterns in fragmented seascapes or landscapes[42] and the strength of selection at particular loci[43]. Demographic inferences are also possible: for example, to distinguish a metapopulation signal from that of population expansion[44] (a goal that was later revisited using ABC[45]). Simulations have also been used to investigate species or population divergence[46,47] and the signal of recent versus ancient fragmentation events[48]. The most popular simulator for these efforts has been SimCoal2.

Epidemiological uses include exploring selective and demographic processes to explain the low effective population size of HIV-1 (REF. 49) and modelling tuberculosis transmission to estimate epidemiological parameters[50]. Simulations have also been used to determine the strength of selection on particular loci[43] and in plant and animal breeding for estimation of kinship coefficients[51].

Validating statistical methods. The expanding wealth of genetic data demands new statistical tools and methods for summarizing data, estimating parameters and testing hypotheses. As these emerge, simulations are used for validation. Examples include methods for detecting bottlenecks[52] or for estimating migration rates[53]. In testing methods, authors explore the effect of varying parameters, test the consequences of relaxing the assumptions of the underlying model and estimate mean square errors (MSEs) or other quality indices (for example, REFS 54–57). Applications have included testing the power of several assignment tests under different levels of F_{ST} and marker variability[58], testing the efficiency and accuracy of Bayesian methods for identifying hybrids or distinct population clusters[59,60] and testing the effects of assuming panmixia in populations that actually exhibit limited dispersal and local structure[61]. Simulations are also used to infer the power of genome scan methods[62], of genomic selection methods for domesticated plants and animals[63] and of association studies[64,65], as well as being used to evaluate the accuracy of methods for predicting the individual genetic risk to disease[66,67]. It is also possible to analyse the sensitivity of summary statistics to changes in model parameters to identify the most informative statistics. For example, simulations were used to identify F_{IS} as a key summary statistic in distinguishing the genetic signal of population subdivision from that of bottlenecks[45].

© 2012 Macmillan Publishers Limited. All rights reserved

Box 2 | **Approximate Bayesian computation: a formal way of comparing models**

Simulations are still used with a variety of ad hoc methods to evaluate the fit between observed and simulated data. Approximate Bayesian computation (ABC) is replacing these approaches and is different from the ad hoc pre-ABC methods in several important ways. First, ABC generates simulated data sets by drawing parameter values from prior distributions rather than using a fixed set of values, as is done in most pre-ABC work. For example, rather than running simulations with divergence times of 10^4, 5×10^4 and 10^5 years, the value for each ABC simulation is drawn at random from a prior distribution between 10^4 and 10^5, allowing better exploration of parameter space. Second, ABC rigorously estimates parameter uncertainty from the posterior distributions of parameters using specific algorithms. Additionally, ABC provides a formal framework for model comparison using Bayes factors or deviance information criteria (DIC)[98,99], which was not possible with pre-ABC approaches. ABC is also suited to post-analysis estimation of power and precision using pseudo data sets[100,101]. Its sophisticated nature presents an obstacle to new users, but several user-friendly packages are now available (for example, DIYABC and ABCToolbox). Technical details of ABC continue to be debated, including how to choose summary statistics[100], compute indices for model comparison[98] and combine *P* values across summary statistics[100], but its use in addressing different questions is well-established[98–101].

Bayesian
A scientific paradigm that uses probability as a means of quantifying the analyst's knowledge or uncertainty concerning the model and/or its parameters, given the data observed. Given a particular model described by a likelihood function, the approach involves choosing a prior distribution and then updating this with the information provided by the observed data.

Most recent common ancestor
In the case of a sample of genes, this is the most recent gene from which all alleles in the sample are directly descended.

Number of segregating sites
The number of polymorphic sites in a sample of homologous DNA sequences. It measures the degree of DNA sequence variation that is present in the sample.

Assignment tests
A broad category of methods whose goal is to determine with a degree of confidence the population of origin of individuals using genetic data.

Panmixia
The random mating of individuals within a breeding population.

Genome scan
Large-scale genotyping (thousands of markers) that is usually used to detect outliers such as regions of the genome under selection.

Prior distributions
The probability distributions of parameter values before observing the data. They reflect the observer's knowledge about what values the model parameters might take before having seen the data.

Posterior distributions
The conditional distributions of the parameter given the observed data. They reflect both the likelihood of the data and the prior distribution. They represent what we know about the model parameters, having observed the data.

Other uses. Simulations can also be used to test the power of a particular population-genetics sampling scheme (such as the number of loci and samples needed) to carry out a specific test[68,69]. When planning an experiment, simulations allow the evaluation of molecular and field resources needed to achieve a desired statistical power[60,70]; after an experiment, post hoc power analyses can help to interpret nonsignificant results[71]. The project ConGRESS, which is underway, will feature a simulation engine to assist in planning the sample scheme of conservation genetic studies. Another use is in teaching population genetics and evolution to help students explore simple and complex population histories[72–74]. The graphical user interface of some simulators is especially amenable to classrooms[15,75].

Designing a simulation study

Simulation studies comprise several steps; a general guide is provided in FIG. 1. Although all three kinds of application involve defining the genetic and demographic scenarios that are under consideration and choosing an appropriate simulator based on these, each application differs in several important aspects (FIG. 2). Here we focus on using available software tools, and we do not provide a blueprint for validating methods, as we feel that most users who are interested in method development are already familiar with the steps involved (for examples, see REFS 53,55,76,77).

For predictive uses. Simulations have been used for predictive purposes in a range of disciplines, but the steps that need to be followed are the same. Here we use an example from conservation genetics to describe them. The first step is the definition of scenarios corresponding to different conservation strategies (or theories in the case of epidemiological or evolutionary applications). Bruford *et al.*[28] focus on a population of orangutans that inhabit a fragmented riparian forest with low population connectivity, resulting in inbreeding and population decline. The authors consider four management actions: translocation, migration corridors, both translocation and migration corridors, and no intervention. The second step is to define the range of parameter values to explore. Although the number of parameters can be very large, few need to be explored in detail; remaining parameters are assigned values estimated by previous studies. For example, although simulations in Bruford *et al.*[28] considered >20 parameters (such as age-specific vital rates and population-specific carrying capacities), the authors only varied the translocation rate under scenarios with or without a gradual increase in carrying capacity. The next step is to choose statistics to monitor the outcome. In this case[28], the authors used probability of extinction, final population size, expected heterozygosity, inbreeding coefficient and level of genetic differentiation.

After completing the first three steps, the simulator must be chosen. Bruford *et al.*[28] chose Vortex because they required a simulator that allows the consideration of life history information and modelling of inbreeding effects on fecundity (see the discussion below on the capabilities of simulators). Next, before running the simulations, it is necessary to choose the number of replicate runs and, for forward simulations, the length (number of generations) of each run. This may be hundreds for conservation applications or thousands or more for evolutionary applications. Forward simulators (BOX 1) that model many life history parameters are computationally demanding and can limit the number and length of runs. Also, some applications require starting allele frequencies to initialize the simulations.

After simulations have been completed, the next step is to analyse pseudo data with population genetic data analysis software to quantify the outcome using the chosen summary statistics. All replicate runs generated need to be processed, so software that allows batch analysis is required (for example, Arlequin[78] or Genepop[79]). However, some simulators (such as msms, mlcoalsim, SimuPop and a soon-to-be-released version of cdpop) integrate data analysis features that meet all the needs of a study. Results can be described using the mean and variance of the summary statistics (as discussed by Bruford *et al.*[28]) across replicate runs under each scenario considered, but there are many other ways of describing the distribution of summary statistics (for example, quantiles, median, mode and skew). With these, it is possible to compare the outcome of the different management scenarios (or, in general, the different hypotheses or models) considered and to draw conclusions. In their study, Bruford *et al.*[28] found that

© 2012 Macmillan Publishers Limited. All rights reserved

non-intervention led to rapid extinction, but the exclusive use of either translocation or corridor establishment was insufficient to prevent inbreeding and extinction. The authors concluded that a mixed strategy is essential.

For statistical inference. Simulations may be used for statistical inference at a wide range of temporal and spatial scales, but the procedure is the same. Here we use an example from evolutionary genetics — the divergence between gorilla subspecies[47] — as an illustration. The first step is to define alternative models of genetic and demographic history and associated parameters (such as divergence times and population sizes). The second step is to choose values for the parameter combinations that will be considered. Thalmann *et al.*[47] defined three models of connectivity following a split between eastern and western gorillas: equal migration, no migration and asymmetrical migration. Parameters included the timing of the split and the population size for each subspecies. The relevant parameter space is user-defined based on prior knowledge, such as field data. Users should explore a wide and realistic range of parameters and assumptions. Thalmann *et al.*[47] used historical information (specifically, climatic changes that result in contraction of forest habitat) to set upper and lower bounds on the timing of the split. A pre-ABC or ABC algorithm can be chosen to explore the parameter space. The traditional pre-ABC approach is to choose a default set of parameter values and, varying one parameter at a time, to create a grid of parameter combinations covering a portion of the parameter space deemed relevant. As with all simulations, some assumptions are necessary; the model used by Thalmann *et al.*[47] assumed constant growth rate over time.

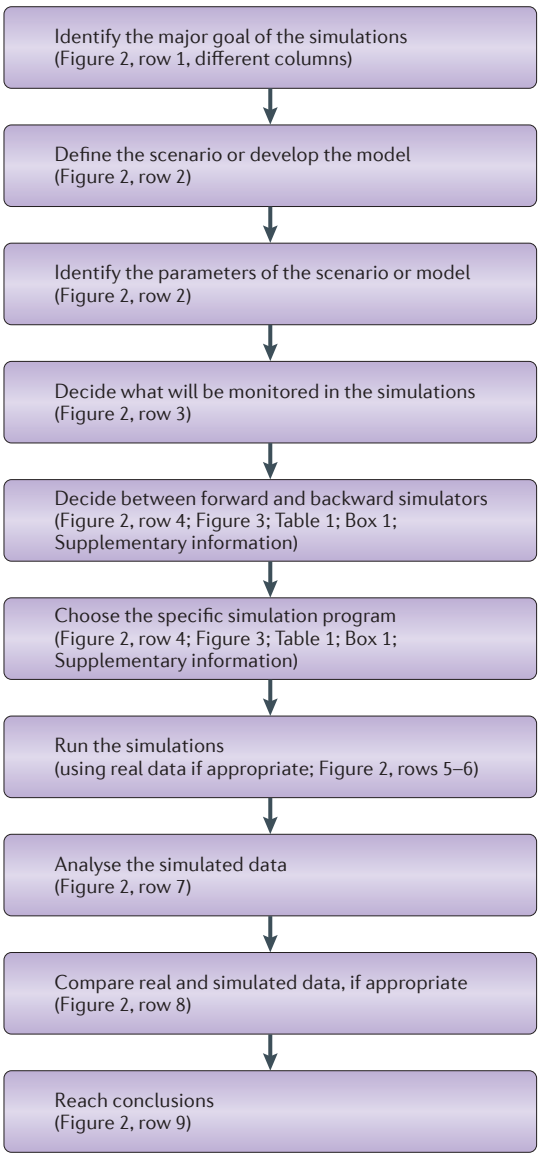

Figure 1 | **Overview of simulation studies.** A flow diagram of the steps in a simulation study. Further information about the various stages is provided in this Review, as indicated on the figure.

Next, the user chooses summary statistics for comparing pseudo data and real data. The chosen statistics should be informative for distinguishing among competing scenarios that are under consideration[45] and should be strongly influenced by the parameters to be estimated. For example, F_{ST} is influenced by migration rates and population size, the number of alleles is influenced by bottlenecks, and F_{IS} is influenced by admixture or inbreeding. Thalmann *et al.*[47] used nucleotide diversity (π) and F_{ST} and chose the backward simulator SimCoal2 owing to its speed and ability to simulate DNA sequences, multiple populations and changing migration matrices. The number of simulations to run depends on the number and complexity of alternative models and on the computing resources available; typically between 10^2 and 10^4 runs[38,46–48] are used with the pre-ABC method and ~10^6 with ABC.

As with predictive uses, analysis of pseudo data produces a distribution of summary statistics for each model. To compare these distributions to the statistics calculated from the real samples and to identify the model that best matches the observed data, several ad hoc methods exist. It is possible to determine where the observed values of the summary statistic fell on the distribution of simulated values in order to obtain a *P* value[38,39]. Thalmann *et al.*[47] put the observed values into a linear regression equation obtained from the pseudo data sets (for example, time of split on F_{ST}) to identify the particular values of parameters. These ad hoc procedures are in the process of being replaced by the ABC approach (BOX 2).

Choosing simulation software

Simulation programs differ greatly in the evolutionary and demographic scenarios that they consider. In this section, we classify the areas of interest that can be addressed using simulations into five categories: historical events, migration, life cycle and population growth, recombination and/or mutation and selection and/or phenotype. For each category, we describe available options (FIG. 3; TABLE 1). We suggest that the general categories in FIG. 3 should be used to choose several

Bayes factors
The relative odds that the hypothesis is true before and after examining the data. Calculated as the ratio of the prior probabilities of the null hypothesis versus the alternative hypothesis over the ratio of the posterior probabilities.

Deviance information criteria
(DIC). A method of model comparison or selection in which increased fit owing to addition of terms is balanced by a penalty for each additional term.

Carrying capacities
The maximum population size of a species that a habitat can sustain. It is determined by availability of space and resources.

Admixture
The interbreeding of individuals issued from two or more distinct populations or species.

© 2012 Macmillan Publishers Limited. All rights reserved

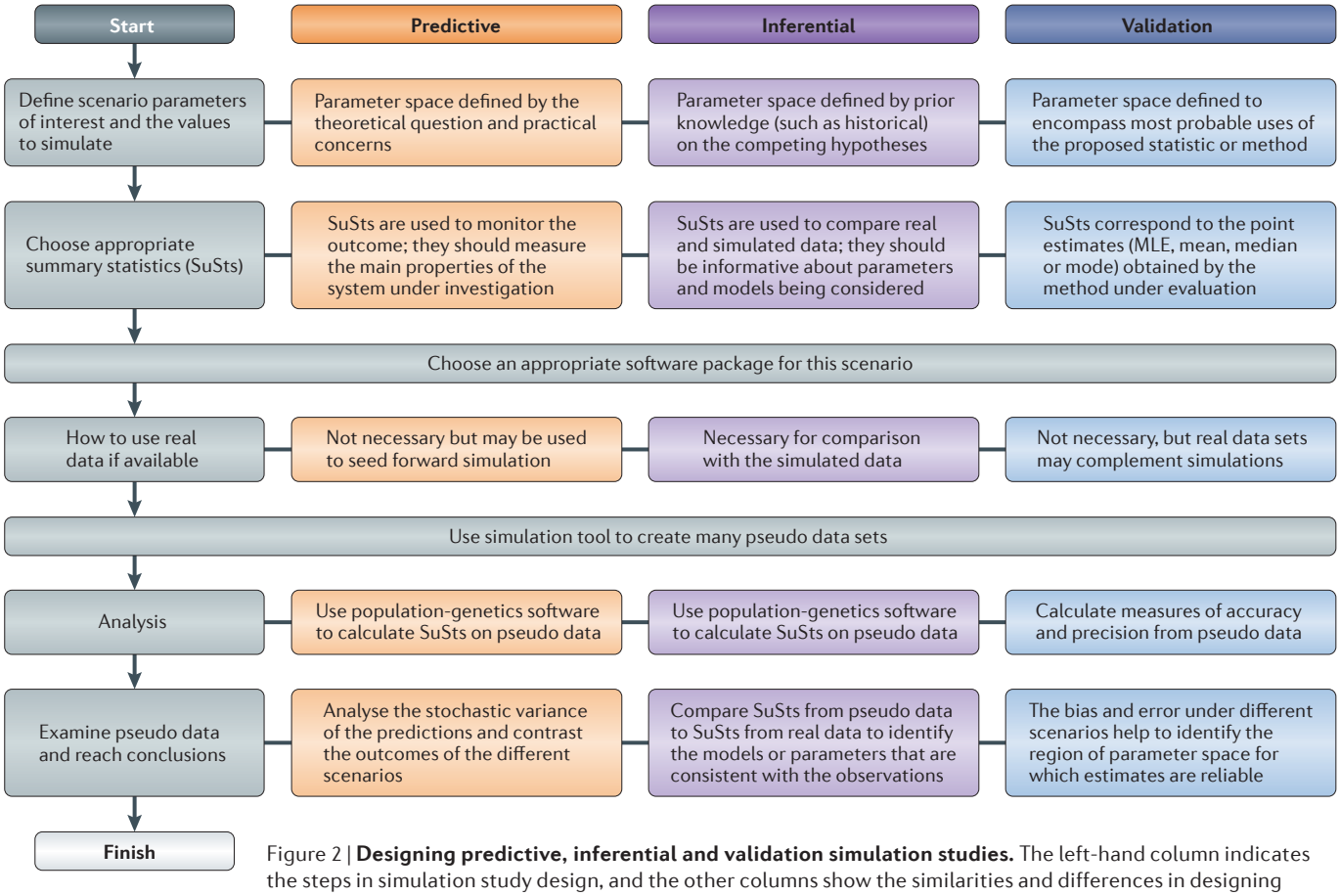

Figure 2 | **Designing predictive, inferential and validation simulation studies.** The left-hand column indicates the steps in simulation study design, and the other columns show the similarities and differences in designing predictive, inferential and validation studies. Grey boxes designate actions that are applicable to all; coloured boxes denote differences. MLE, maximum likelihood estimation.

possible simulators, and then the details in TABLE 1 should be used to decide on the most appropriate. Trade-offs are apparent, as most simulators specialize in detailed modelling of demography and/or ecology or genetic architecture. When applicable, we mention 'speciality' software programs that address a particular challenging or emerging topic. Supplementary information S1 (table) provides a list of original citations and (if available) a case study that incorporates the simulator, whereas Supplementary information S2 (table) provides a short description of all of the simulators we review. Readers should bear in mind the major methodological distinction between forward and backward simulators (BOX 1), as this difference strongly affects the choice of simulator for a specific study.

Historical events. One reason for the development of the simulation approach has been to improve understanding of the dynamics of genetic variation when the effect of multiple events over time is difficult to predict with classical population-genetics models. Most programs allow several historical events — including population size changes, population fusion or fission and speciation (TABLE 1). Typically, users need to consider the timing, length, intensity and sometimes the spatial location of events. Often they need to compare the effect of a particular event or sequence of events with the expected outcome under a null model, such as constant population size over time. Four simulation programs are of note. BottleSim simulates bottleneck dynamics in a single population of long-lived organisms with multiple overlapping generations (for example, forest trees). Splatche and AquaSplatche simulate colonization and spatial expansions, and they model changes in the environment (for example, fragmentation), rather than direct changes to population sizes and migration rates, allowing increased ecological realism. Last, IBDsim allows changes in both the population density and size of the 'world', which are fixed quantities in most simulators.

Historical events can be deterministic or stochastic. Most programs allow for deterministic events, but several forward simulators explicitly address stochasticity, in that events do not necessarily occur at specific time points; this is useful in simulating fragile environments, disease outbreaks or climatic instability. RmetaSim, KernelPop and Vortex allow for stochasticity in timing and order of historical events, Nemo and QuantiNemo allow for stochasticity in extinction and harvest events, and Pedagog allows events to occur at any generation with a set probability, which is useful for recurring disturbances (such as fire and hurricanes).

© 2012 Macmillan Publishers Limited. All rights reserved

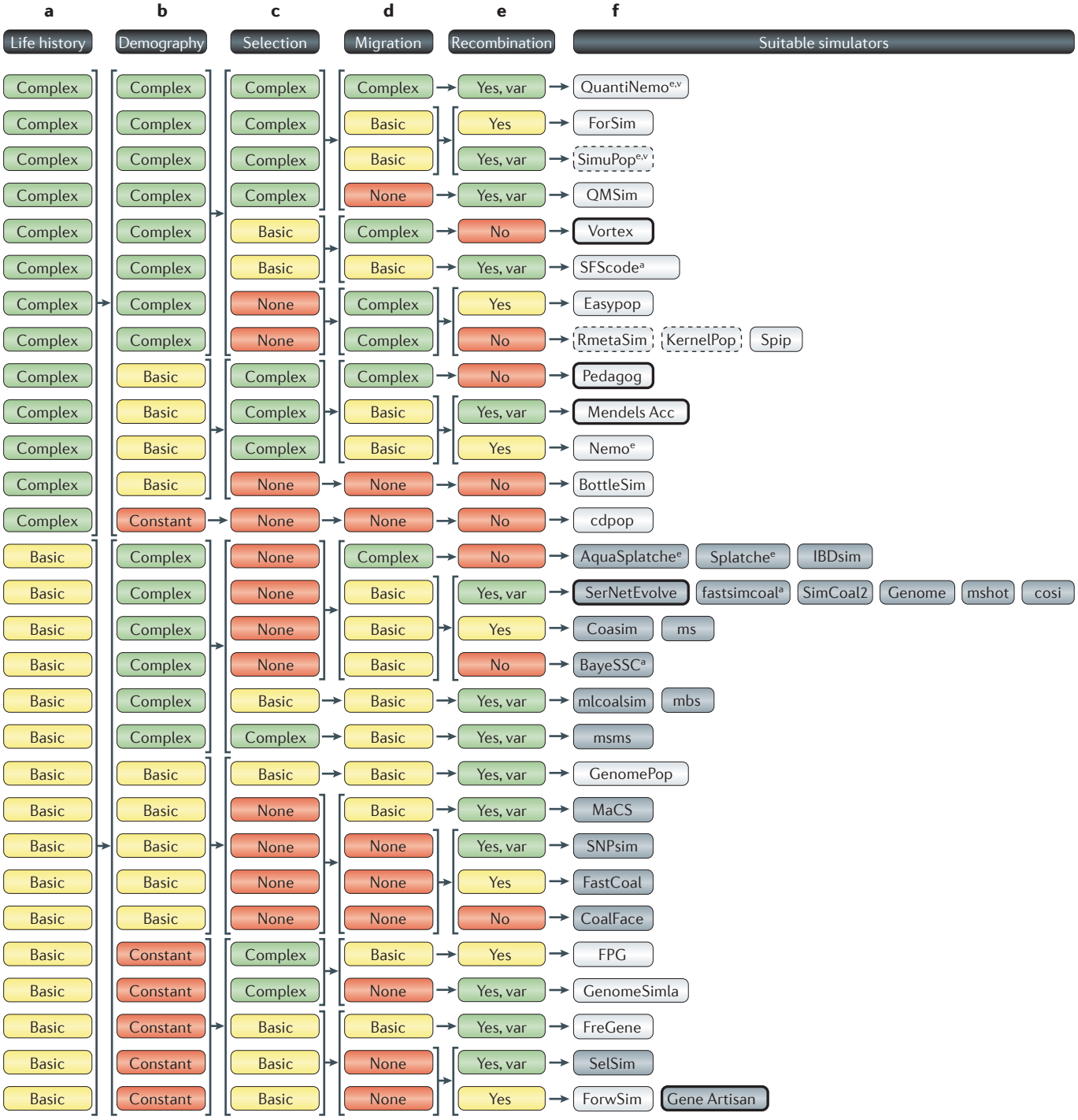

Figure 3 | **Decision matrix for choosing a simulator.** The figure presents a decision matrix with four categories to help the user match a simulator to their objective and study system. Details of the simulators are in TABLE 1 and Supplementary information S1,S2 (tables). The user can choose an option in each column, starting on the left, and appropriate simulators are given on the right. **a** | Life history options. Basic, discrete generations and random mating within demes; complex, overlapping generations, complex mating or age-, stage- or population-specific fecundity. **b** | Demography options. Complex, changes to environment, migration or density or population fusion or fission; constant, constant population size; basic, changes in population size or growth rate. **c** | Selection options. Basic, directional and/or balancing selection; complex, selection that is environment-based, time variable, epistatic, frequency dependent, and so on; none, no selection. **d** | Migration options. Basic, user-defined matrix, stepping stone and/or Wright's island model; complex, spatially explicit, density dependence, selective, age, sex or environmental effects; none, no migration. **e** | Recombination options. No, there is no recombination; yes, recombination rate is constant across sites; yes, var, variation in recombination rates (for example, hotspots) among sites possible. **f** | Simulators in grey boxes are backward simulators, and those in white boxes are forward simulators. Bold-bordered boxes have graphical user interfaces (GUIs), dashed lines require writing scripts in R or Python, and single grey lines use command lines. Several special features may be key to decision making; these are indicated by superscript letters: a, ancient DNA; e, environmental effects on selection or migration; v, time- or frequency-variable selection.

© 2012 Macmillan Publishers Limited. All rights reserved

Life histories. Organisms differ widely in life cycle and reproductive biology, including in lifespan, mating system, sex ratios and age structure. Backward programs focus on genes rather than on individuals and only allow very simple customization of life history (BOX 1); thus, forwards-in-time simulations are mainly considered here. In FIG. 3, we broadly classify simulators into 'basic' and 'complex' for life history, and TABLE 1 provides more detailed information about their capabilities.

Several programs (including BottleSim, cdpop, RmetaSim, KernelPop, Pedagog and Vortex) allow an arbitrarily complex life cycle with multiple overlapping stages that differ in survival, fecundity and dispersal capabilities; these are useful for simulating long-lived plant and animal species. The other programs assume two-stage lifecycles (non-reproductive juveniles and reproductive adults). The simplest mating scheme is a hermaphroditic species with random mating and a set proportion of self-fertilizing ('selfing'), but some simulators consider polygamy, polyandry, monogamy and other complex mating behaviours (TABLE 1), enabling species-specific customization. SimuPop is unique in allowing the mating system to vary across populations; for example, it allows different selfing rates in different populations, which is the case for many partial-selfing plants (such as *Arabidopsis* spp.). ForSim and QMSim allow phenotype-based assortative mating.

Life history is also modelled through fecundity and population growth parameters. Generally, forward simulators model growth based on individual density-independent fecundity up to a carrying capacity, but Vortex, RmetaSim and KernelPop allow fecundity to depend on density and/or selective pressures. Pedagog is the only simulator to offer carrying capacities for each age or stage group, which is useful for strongly age-structured species. Splatche and AquaSplatche simulate spatial expansions by allowing populations over their carrying capacity to 'overflow' into empty adjacent populations. Fecundity is typically Poisson-distributed, but some simulators (namely, Pedagog, Vortex and SimuPop) allow the use of distributions that are more appropriate for organisms with a high potential reproductive output and a wide variance in reproductive success (for example, fish and plants). Backward simulators cannot model fecundity, because they are not individual-based, and so they generally consider constant population size or logistic or exponential growth.

Migration. The simplest migration model that is implemented by simulators is Wright's island model[80], which can be considered to be a null model against which other more complex models are compared. Most simulators allow a user-defined matrix of migration rates between each pair of populations, or they allow the stepping stone model, in which migration only occurs between adjacent populations (TABLE 1). There are numerous other possibilities (TABLE 1) — for example, a hierarchical island model (using Easypop or SimuPop), a propagule-pool migration model, in which immigrants to a given patch all originate in the same source patch, rather than in a migrant pool (using Nemo) and others. Usually, the migration matrix can be changed as a historical event.

Several programs have been tailored to interesting, challenging migration and dispersal scenarios. In Splatche and AquaSplatche, the environment quality (defined by a user-provided geographic information system (GIS) format file) is described by a 'resistance', which reduces optimal migration rates by reducing the survival of migrants; this is useful for a heterogeneous environment (for example, fragmentation). In KernelPop and IBDsim, every individual has a particular spatial coordinate, rather than being a member of a population, and offspring move away from parents according to a dispersal function, as in continuously distributed species or in those with long-distance dispersal. Easypop offers an isolation-by-distance model, in which migration rates exponentially decay with increasing distance, and Nemo and QuantiNemo allow different migration rates for low and high population density.

Recombination and mutation. Nearly all simulators allow mutation, but they differ in the mutation models considered. Most implement the infinite alleles model and/or the *k* allele model. A number of simulators (TABLE 1) also model microsatellites (using various stepwise mutation models), SNPs and/or sequence data (including transition bias and heterogeneity in mutation rates among sites). In general, SNP data are simulated as simple biallelic loci; FreGene, and to a lesser extent SimCoal2 and CoaSim, can simulate ascertainment bias. Few programs simulate amplified fragment length polymorphisms (AFLPs) or restriction fragment length polymorphisms (RFLPs). Mutation rate may be constant across loci or may be drawn from a distribution (for example, gamma distribution for microsatellites). Only SFScode models insertions and deletions.

A common feature of simulators is to consider recombination; this is done by including either a constant rate across loci or hotspots of recombination. Some simulators allow definition of the location and distance between each marker (such as SelSim, SimCoal2, QuantiNemo, SimuPop and FreGene). Until recently, backward simulation of entire genomes was impractical because, when going back in the past, recombining DNA sequences could join (owing to coalescence) but could also split (owing to recombination); thus, forward simulations were preferred[63,81]. However, the sequential Markov coalescent[82,83] process, which is an approximation of the coalescent model (which is implemented in MaCS and fastsimcoal), has solved this problem of recombination increasing the speed of backward genomic simulators.

Users with basic bioinformatic skills can post-process simulation output to implement mutation or recombination models that are not directly supported by the simulator (for example, see REF. 84, or see Supplementary information S2 (table) for further information about its implementation in ms2ms) or to modify the simulated data sets to introduce biases that are specific to certain markers or sequencing techniques (for example, SNP ascertainment bias). A modular approach — linking a simulator to another piece of software — can completely cover an individual researcher's needs.

Wright's island model
A population-genetics model in which all populations are of equal size and contribute equally to a global migrant pool, from which each population draws an equal proportion of immigrants each generation.

Hierarchical island model
A variation on Wright's island model in which local sets of populations are connected to each other by a relatively high migration rate and to other local sets of populations by a relatively low rate. They are well-suited to modelling species that are distributed over several continents.

Geographic information system
(GIS). A collection of spatially referenced data, such as geographical and altitudinal coordinates of individuals.

k allele model
A mutation model in which each allele can mutate to any of the other $k - 1$ possible alleles with equal probability.

Sequential Markov coalescent
A simplified genealogical process that aims to capture the essential features of the full coalescent model with recombination while being scalable in the number of loci. Computation time is saved by only accounting for coalescence between lineages without overlapping ancestral material.

© 2012 Macmillan Publishers Limited. All rights reserved

Selection and phenotypes. Modelling selection has various uses, including association studies[31], quantitative trait locus (QTL) mapping and removing outlier loci from studies of neutral processes. The coalescent limits backward simulators to simple selection on single biallelic sites, but more than half of the forward simulators incorporate selection (TABLE 1) to varying degrees, some being quite complex. Because these simulators focus on individuals, selection is modelled at the phenotypic or genotypic level, on multiple sites for more than two alleles per site and on traits controlled by QTLs. Selection can be modelled with either multiplicative or epistatic effects (such as in FPG), can vary through time (such as in QuantiNemo) or can interact with the environment at the population level (such as in Nemo). Additionally, because pedigrees are known, selection that is due to inbreeding depression or deleterious mutations can be modelled (such as in Vortex and Nemo). There are many specialized approaches, including codon-based selection (implemented in GenomePop), species interaction between a parasite and its host (using Nemo), population-specific selection (using FreGene), overdominance (using FreGene), frequency-dependent selection (using SimuPop) and variance in effect sizes (using Mendels Accountant). For users with little expertise in genetics, Pedagog has built-in models for evolution of selected phenotypic traits (such as size, age at maturity, movement and survival). One specialized piece of software[17] (GenomeSimla) simulates large-scale genomic data in population based case–control samples.

Practical considerations

Interface. Simulation packages differ in input and output (Supplementary information S1 (table)). Most packages run from the command line; model parameters are provided in a text input file or as command line arguments. Others have a graphical user interface (GUI) for the input of parameters using text boxes, radio buttons, and so on. Command line interfaces are more amenable to batch analyses and incorporation in a software pipeline and are usually available on multiple operating systems. Parameter input files are also useful when creating many scenarios that differ at only one or two parameters. GUIs may be easier for novice computer users. Graphical output also facilitates quicker exploration, presentation and communication of results but provides less access to raw data. Several simulators are coded in high-level programming languages. For example, RmetaSim runs within R, enabling access to the analysis and plotting functions of R and genetic packages such as ade4. SimuPop and CoaSim are coded in Python and use R graphical functions.

All programs output genotypes, haplotypes or allele frequencies, which are usually formatted for further analysis with specific software packages. Additional output may include data on genetic diversity and differentiation, demography, selection, phylogenetics, mutation and pedigrees. A few programs offer an interactive graphical tool for visualizing data trends (for example, Splatche, AquaSplatche, Vortex and Mendels Accountant). Mlcoalsim implements a large number of statistical tests (for example, neutrality tests). Some forward simulators collect output information at user-defined time points rather than only at the simulation end (TABLE 1; Supplementary information S1 (table)), which is useful when dynamics over time are of interest or if ancient DNA or museum specimens are available.

Avoiding common problems. Many simulators provide manuals with tips, warnings, assumptions of the model and explanations of error messages; some provide complete tutorials, example scripts and troubleshooting instructions. Most are also published in peer-reviewed journal articles that outline the methodology and capabilities of the simulator and provide an empirical example (Supplementary information S1 (table)). Documentation should be used when choosing a simulator to understand how closely model assumptions match the investigator's species and/or scenario or to identify features of speciality programs; it should also frequently be used during program use to run simulations and to analyse and interpret results successfully. Also, simulators do not model missing data or genotyping errors (the exception being Pedagog, which does), which may complicate comparison of real and simulated data. However, this problem can be overcome by using the modular approach described above.

A common problem is errors during simulation computation that are due to typographical errors in input or end-of-line characters that are specific to the operating system used to generate the input files. These can be avoided by using the functional example input files provided by some programs as a starting template. Also, two command line programs — ms and SimCoal2 — have GUIs to assist in creating input files. Simulation studies require several trial runs for familiarization and proofreading of input files. Preliminary checking under simple scenarios for which analytical results are available is recommended.

Closing remarks and future directions

There are several areas of improvement for simulators. The increasing need to model genome-level data requires substantial improvements in efficiency. A new forward-simulation method[85], ForwSim, uses the simulated genealogy over several generations to identify chromosomes that are destined to disappear, which do not need to be included in ensuing generations. This results in marked gains in efficiency, but it currently assumes the standard Wright–Fisher model. Efficiency gains have been achieved for backward simulators using the sequential Markov coalescent[82,86], as noted above. Despite these advances, substantial improvements are still required to efficiently simulate data sets that consist of polymorphisms throughout the genome on the scale that new high-throughput sequencing technology allows[87]. It is important to note that advances in efficiency do not suffice to simulate next-generation sequencing data sets properly; simulators should either model the genotyping errors and biases that are specific to this technology[88,89], or they should at least produce raw data that can easily be modified by additional software modules.

An emerging trend is to allow samples from multiple time points in the coalescent (for example, Bayesian Serial

© 2012 Macmillan Publishers Limited. All rights reserved

SimCoal). The interest in more complex demographic considerations is also increasing[90]; in particular, spatially realistic scenarios with cost–distance matrices, metapopulation dynamics and environmental effects are of special interest to landscape genetics[91]. Simulators with these capabilities exist (such as cdpop, KernelPop and Splatche), but they all assume very simplistic demographic histories. Consideration of stochasticity in timing and severity of 'events' is another current need that has been addressed by few simulators. Clever use of programs may circumvent this and other limitations; for example, Mardulyn and Milinkovitch[44] used many repeated events to simulate frequent extinction and colonization of habitat patches. Using distributions rather than point values for parameters will also help to simulate year-to-year stochasticity.

Another consideration is designing programs that can easily be integrated in bioinformatic pipelines. Output in a commonly used format (for example, genepop, FSTAT or Arlequin) and/or summaries across replicates will make programs more useful. Furthermore, the usability and accessibility of many programs could be markedly improved; some require at least moderate programming abilities, have difficult-to-construct input files or provide no documentation. Simultaneously, users must become increasingly agile in command line and R environments[92]; benefits of mastering basic bioinformatic skills include the ability to post-process simulated genotypes in order to compare them meaningfully to real data (for example, introducing genotyping errors or missing data), to manage large data sets, to carry out ABC or to create a sideways simulation pipeline (BOX 1).

Users and designers of software packages will need to understand further which aspects of real population demography and genetics are the most important under the different evolutionary scenarios of interest to avoid overparameterizing studies. Also, as simulation programs rely on summary statistics, their usefulness will increase with the development of new statistics designed to provide information about particular evolutionary histories. A final direction is bridging ecological, population genetic and evolutionary timescales and developing processes to further study range-shift dynamics and species extinction related to climate change. Substantial further work is also required to achieve multiple taxon simulations[93].

Computer simulations are essential for explaining the origin and maintenance of genetic variation and, as this Review shows, they have applications in many disciplines. It is clear that simulators need further improvements to mimic modern data sets. However, the past decade has witnessed much progress, and we expect simulations to become a standard tool for the study of genetic variation.

1. Ray, N. & Excoffier, L. Inferring past demography using spatially explicit population genetic models. *Hum. Biol.* **81**, 141–157 (2009).
2. Ohta, T. & Kimura, M. Simulation studies on electrophoretically detectable genetic variability in a finite population. *Genetics* **76**, 615–624 (1974).
3. Kimura, M. Average time until fixation of a mutant allele in a finite population under continued mutation pressure: studies by analytical, numerical, and pseudo-sampling methods. *Proc. Natl Acad. Sci. USA* **77**, 522–526 (1980).
4. Sokal, R. R. & Wartenberg, D. E. A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics* **105**, 219–237 (1983).
5. Kingman, J. F. C. The coalescent. *Stoch. Proc. Appl.* **13**, 235–248 (1982).
6. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. & Antonovics, J.) 1–44 (Oxford Univ. Press, UK, 1990).
This book is the most easy-to-follow and authoritative review on the coalescent.
7. Fu, Y. X. & Li, W. H. Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.* **14**, 195–199 (1997).
8. Schneider, S. & Excoffier, L. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* **152**, 1079–1089 (1999).
9. Weiss, G. & von Haeseler, A. Inference of population history using a likelihood approach. *Genetics* **149**, 1539–1546 (1998).
10. Anderson, C. N. K., Ramakrishnan, U., Chan, Y. L. & Hadly, E. A. Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics* **21**, 1733–1734 (2004).
11. Chadeau-Hyam, M. *et al.* Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics* **9**, 364 (2008).
12. Peng, B. & Kimmel, M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **21**, 3686–3687 (2005).
13. Coombs, J. A., Letcher, B. H. & Nislow, K. H. Pedagog: software for simulating eco-evolutionary population dynamics. *Mol. Ecol. Resour.* **10**, 558–563 (2010).
14. Estoup, A. *et al.* Combining genetic, historical and geographical data to reconstruct the dynamics of bioinvasions: application to the cane toad *Bufo marinus*. *Mol. Ecol. Resour.* **10**, 886–901 (2010).
15. Currat, M., Ray, N. & Excoffier, L. Splatche: a program to simulate genetic diversity taking into account environmental heterogeneity. *Mol. Ecol. Notes* **4**, 139–142 (2004).
This paper introduces Splatche, one of the first simulators to account for environmental variation, which was a major step in use of simulations in landscape genetics.
16. Strand, A. E. metasim 1.0: an individual-based environment for simulating population genetics of complex population dynamics. *Mol. Ecol. Notes* **2**, 373–376 (2002).
This paper introduces metasim (later RmetaSim), one of the earlier user-friendly software packages for simulation studies; the manuscript provides a coherent description of how to construct migration and transition state matrices, as well as a review of uses of simulations.
17. Ritchie, D. Genome simulation approaches for synthesizing *in silico* datasets for human genomics. *Adv. Genet.* **72**, 1–24 (2010).
18. Carvajal-Rodriguez, A. Simulation of genomes: a review. *Curr. Genomics* **9**, 155–159 (2008).
19. Carvajal-Rodriguez, A. Simulation of genes and genomes forward in time. *Curr. Genomics* **11**, 58–61 (2010).
20. Lively, C. A review of red queen models for the persistence of obligate sexual reproduction. *J. Hered.* **101**, S13–S20 (2010).
21. Church, S. & Taylor, D. The evolution of reproductive isolation in spatially structured populations. *Evolution* **56**, 1859–1862 (2002).
22. Servedio, M. The evolution of premating isolation: local adaptation and natural and sexual selection against hybrids. *Evolution* **58**, 913–924 (2004).
23. Daleszczyk, K. & Bunevich, A. N. Population viability analysis of European bison populations in Polish and Belarusian parts of Bialowieza Forest with and without gene exchange. *Biol. Conserv.* **142**, 3068–3075 (2009).
24. Vonholdt, B. M. *et al.* The genealogy and genetic viability of reintroduced Yellowstone grey wolves. *Mol. Ecol.***17**, 252–274 (2008).
25. Alves, D. A. *et al.* Successful maintenance of a stingless bee population despite a severe genetic bottleneck. *Conserv. Genet.* **12**, 647–658 (2011).
26. Ng, K. K. S., Lee, S. L. & Ueno, S. Impact of selective logging on genetic diversity of two tropical tree species with contrasting breeding systems using direct comparison and simulation methods. *For. Ecol. Manage.* **257**, 107–116 (2009).
27. Kenney, J. S., Smith, J. L. D., Starfield, A. M. & McDougal, C. W. The long-term effects of tiger poaching on population viability. *Conserv. Biol.* **9**, 1127–1133 (1995).
28. Bruford, M. *et al.* Projecting genetic diversity and population viability for the fragmented orang-utan population in the Kinabatangan floodplain, Sabah, Malaysia. *Endanger. Species Res.* **12**, 249–261 (2010).
29. Yang, Z., Hu, J. & Liu, N. The influence of dispersal on the metapopulation viability of Giant Panda (*Aliuropoda melanoleuca*) in the Minshan Mountains. *Acta Zool. Academ. Sci. Hung.* **53**, 169–184 (2007).
30. Sellers, T. A., Weaver, T. W., Phillips, B., Altmann, M. & Rich, S. S. Environmental factors can confound identification of a major gene effect: results from a segregation analysis of a simulated population of lung cancer families. *Genet. Epidemiol.* **15**, 251–262 (1998).
31. Peng, B. & Kimmel, M. Simulations provide support for the common disease–common variant hypothesis. *Genetics* **175**, 763–776 (2007).
32. Maher, B. The case of the missing heritability. *Nature* **456**, 18–21 (2008).
33. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
34. van Oosterhout, C. A new theory of MHC evolution: beyond selection on the immune genes. *Proc. R. Soc. B* **276**, 657–665 (2009).
35. Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994–997 (2008).
36. Fagundes, N. J. R. *et al.* Statistical evaluation of alternative models of human evolution. *Proc. Natl Acad. Sci. USA* **104**, 17614–17619 (2007).
37. Hellenthal, G., Auton, A. & Falush, D. Inferring human colonization history using a copying model. *PLoS Genet.* **4** e1000078 (2008).
38. Ficetola, G. F., Bonin, A. & Miaud, C. Population genetics reveals origin and number of founders in a biological invasion. *Mol. Ecol.* **17**, 773–782 (2008).
This is a simple but striking example of the use of simulations to infer parameters of a historical process (in this case, introduction of non-native species).
39. Banks, S. C. *et al.* Genetic structure of a recent climate change-driven range extension. *Mol. Ecol.* **19**, 2011–2024 (2010).

© 2012 Macmillan Publishers Limited. All rights reserved

40. Martínez-Cruz, B., Godoy, J. A. & Negro, J. J. Population genetics after fragmentation: the case of the endangered Spanish imperial eagle (*Aquila adalberti*). *Mol. Ecol.* **13**, 2243–2255 (2004).
41. Fabbri, E. *et al.* From the Apennines to the Alps: colonization genetics of the naturally expanding Italian wolf (*Canis lupus*) population. *Mol. Ecol.* **16**, 1661–1671 (2007).
42. Alberto, F. *et al.* Habitat continuity and geographic distance predict population genetic differentiation in giant kelp. *Ecology* **91**, 49–56 (2010).
43. Enard, D., Depaulis, F. & Crollius, H. R. Human and non-human primate genomes share hotspots of positive selection. *PLoS Genet.* **6** e1000840 (2010).
44. Mardulyn, P. & Milinkovitch, M. C. Inferring contemporary levels of gene flow and demographic history in a local population of the leaf beetle *Gonioctena olivacea* from mitochondrial DNA sequence variation. *Mol. Ecol.* **14**, 1641–1653 (2005).
45. Peter, B. M., Wegmann, D. & Excoffier, L. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Mol. Ecol.* **19**, 4648–4660 (2010).
This paper provides a coherent contemporary example of use of ABC in model choice.
46. Mardulyn, P., Mikhailov, Y. E. & Pasteels, J. M. Testing phylogeographic hypotheses in a Euro–Siberian cold-adapted leaf beetle with coalescent simulations. *Evolution* **63**, 2717–2729 (2009).
47. Thalmann, O., Fischer, A., Lankester, F., Paabo, S. & Vigilant, L. The complex evolutionary history of gorillas: insights from genomic data. *Mol. Biol. Evol.* **24**, 146–158 (2007).
48. Haanes, H., Roed, K. H., Flagstad, O. & Rosef, O. Genetic structure in an expanding cervid population after population reduction. *Conserv. Genet.* **11**, 11–20 (2010).
49. Shriner, D., Liu, Y., Nickle, D. C. & Mullins, J. I. Evolution of intrahost HIV-1 genetic diversity during chronic infection. *Evolution* **60**, 1165–1176 (2006).
50. Tanaka, M. M., Francis, A. R., Luciani, F. & Sisson, S. A. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics* **173**, 1511–1520 (2006).
51. Jorjani, H. A general genomics simulation program. *Interbull Bull.* **40**, 202–206 (2010).
52. Luikart, G., Allendorf, F. W., Cornuet, J.-M. & Sherwin, W. B. Distortion of allele frequency distributions provides a test for recent population bottlenecks. *J. Hered.* **89**, 238–247 (1998).
53. Beerli, P. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Mol. Ecol.* **13**, 827–836 (2004).
54. Waples, R. S. Temporal variation in allele frequencies: testing the right hypothesis. *Evolution* **43**, 1236–12351 (1989).
55. Faubet, P., Waples, R. S. & Gaggiotti, O. E. Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. *Mol. Ecol.* **16**, 1149–1166 (2007).
56. Gaggiotti, O. E., Lange, O., Rassmann, K. & Gliddon, C. A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data. *Mol. Ecol.* **8**, 1513–1520 (1999).
57. Hardy, O. J., Charbonnel, N., Freville, H. & Heuertz, M. Microsatellite allele sizes: a simple test to assess their significance on genetic differentiation. *Genetics* **163**, 1467–1482 (2003).
58. Manel, S., Berther, P. & Luikart, G. Detecting wildlife poaching: identifying the origin of individuals with Bayesian assignment tests and multilocus genotypes. *Conserv. Biol.* **16**, 650–659 (2002).
59. Francois, O. & Eric, D. Spatially explicit Bayesian clustering models in population genetics. *Mol. Ecol. Res.* **10**, 773–784 (2010).
60. Vaha, J.-P. & Primmer, C. R. Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Mol. Ecol.* **15**, 63–72 (2006).
61. Landguth, E. L., Cushman, S. A., Murphy, M. A. & Luikart, G. Relationships between migration rates and landscape resistance assessed using individual-based simulations. *Mol. Ecol. Res.* **10**, 854–862 (2010).
62. Vasemagi, A. & Primmer, C. R. Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Mol. Ecol.* **14**, 3623–3642 (2005).
63. Meuwissen, T. & Goddard, M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* **185**, 623–631 (2010).
64. Huang, J., Johnson, A. & O'Donnell, C. PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. *Bioinformatics* **27**, 1201–1206 (2011).
65. Spencer, C. C. A., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* **5**, e1000477 (2009).
66. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).
67. Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* **3**, e3395 (2008).
68. Ryman, N. *et al.* Power for detecting genetic divergence: differences between statistical methods and marker loci. *Mol. Ecol.* **15**, 2031–2045 (2006).
69. Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Rev. Genet.* **3**, 380–390 (2002).
70. Crandall, K. A., Bininda-Emonds, O. R. P., Mace, G. M. & Wayne, R. K. Considering evolutionary processes in conservation biology. *Trends Ecol. Evol.* **15**, 290–295 (2000).
71. Ryman, N. & Palm, S. POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Mol. Ecol.* **6**, 600–602 (2006).
72. Hamilton, M. *Population Genetics* (Wiley–Blackwell, Chichester, 2009).
73. Jones, T. C. & Laughlin, T. F. Popgen fishbowl: a free online simulation model of microevolutionary processes. *Am. Biol. Teach.* **72**, 100–103 (2010).
74. Delport, W. Coalface: a graphical user interface program for the simulation of coalescence. *Mol. Ecol. Notes* **6**, 281–284 (2006).
75. Neuenschwander, S. AquaSplatche: a program to simulate genetic diversity in populations living in linear habitats. *Mol. Ecol. Notes* **6**, 583–585 (2006).
76. Gaggiotti, O. E. & Excoffier, L. A simple method of removing the effect of a bottleneck and unequal population sizes on pairwise genetic distances. *Proc. R. Soc. B* **267**, 81–87 (2000).
77. Garza, J. C. & Williamson, E. G. Detection of reduction in population size using data from microsatellite loci. *Mol. Ecol.* **10**, 305–318 (2001).
78. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Res.* **10**, 564–567 (2010).
79. Rousset, F. Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Mol. Ecol. Res.* **8**, 103–106 (2008).
80. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).
81. Hoggart, C. J. *et al.* Sequence-level population simulations over large genomic regions. *Genetics* **177**, 1725–1731 (2007).
82. Marjoram, P. & Wall, J. Fast "coalescent" simulation. *BMC Genetics* **7**, 16 (2006).
This was a description of one of the first implementations of the sequentially Markovian coalescent algorithm to improve the efficiency of the coalescent simulators.
83. McVean, G. A. T. & Cardin, N. J. Approximating the coalescent with recombination. *Phil. Trans. R. Soc. B* **360**, 1387–1393 (2005).
This paper describes the introduction of the approximation to the coalescent with recombination that allows fast simulation of genealogies sequentially along a sequence. This development paved the way for coalescent simulators able to generate genomic data.
84. Wall, J. D. & Pritchard, J. K. Assessing the performance of the haplotype block model of linkage disequilibrium. *Am. J. Hum. Genet.* **73**, 502–515 (2003).
85. Padhukasahasram, B., Marjoram, P., Wall, J. D., Bustamante, C. D. & Nordborg, M. Exploring population genetic models with recombination using efficient forward-time simulations. *Genetics* **178**, 2417–2427 (2008).
This paper describes a major innovation in the forward-in-time simulations, which determine the fate of individuals several generations in the future; the program saves efficiency by not following those that will not leave descendants.
86. Chen, G. K., Marjoram, P. & Wall, J. D. Fast and flexible simulation of DNA sequence data. *Genome Res.* **19**, 136–142 (2009).
87. Dudek, S., Motsinger, A., Velez, D., Williams, S. & Ritchie, M. Data simulation software for whole-genome association and other studies in human genetics. *Pac. Symp. Biocomput.* **11**, 499–510 (2006).
88. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10**, R32 (2009).
89. Schwartz, S., Oren, R. & Ast, G. Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS ONE* **6**, e16685 (2011).
90. Wernsdörfer, H. *et al.* Relationships between demography and gene flow and their importance for the conservation of tree populations in tropical forests under selective felling regimes. *Conserv. Genet.* **12**, 15–29 (2010).
91. Epperson, B. K. *et al.* Utility of computer simulations in landscape genetics. *Mol. Ecol.* **19**, 3549–3564 (2010).
92. Haddock, S. & Dunn, C. *Practical Computing for Biologists* (Sinauer Associates, 2010).
This book is a useful, practical and brand new guide to basic bioinformatics including scripting, data management and command line; especially recommended for graduate students in all areas of biology.
93. Ilves, K., Huang, W., Wares, J. & Hickerson, M. Colonization and/or mitochondrial selective sweeps across the North Atlantic intertidal assemblage revealed by multi-taxa approximate Bayesian computation. *Mol. Ecol.* **19**, 4505–4519 (2010).
94. Hudson, R. R. Island models and the coalescent process. *Mol. Ecol.* **7**, 413–418 (1998).
95. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
This paper introduces ms, one of the earliest and still one of the most popular software packages for simulation studies.
96. Wall, J. & Hudson, R. Coalescent simulations and statistical tests of neutrality. *Mol. Biol. Evol.* **18**, 1134–1135 (2001).
97. Liu, Y., Athanasiadis, G. & Weale, M. A survey of genetic simulation software for population and epidemiological studies. *Hum. Genomics* **3**, 79–86 (2008).
98. Csilléry, K., Blum, M. G. B., Gaggiotti, O. E. & François, O. Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.* **25**, 410–418 (2010).
99. Beaumont, M. A. Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* **41**, 379–406 (2010).
This provides overview of the theory, methodology, major advances and current and future uses of ABC.
100. Bertorelle, G., Benazzo, A. & Mona, S. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol. Ecol.* **19**, 2609–2625 (2010).
101. Lopes, J. & Boessenkool, S. 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 (2009).

Acknowledgements
This work was supported by the European Project CONGRESS funded by the European Union under FP7. We also thank E. Anderson and two anonymous referees for very helpful suggestions, as well as all of the software developers (see the full list in the Supplementary information) for their assistance in checking the information presented in TABLE 1 and in Supplementary information S1,S2 (tables).

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Sean Hoban's homepage: http://sites.google.com/site/hoban3
Giorgio Bertorelle's homepage: http://docente.unife.it/giorgio.bertorelle
Oscar E. Gaggiotti's homepage: http://www-leca.ujf-grenoble.fr/membres/gaggiotti.htm
ConGRESS: http://www.congressgenetics.eu page.aspx?SP=About#about2
***Nature Reviews Genetics* Series on Study Designs:** http://www.nature.com/nrg/series/studydesigns/index.html

SUPPLEMENTARY INFORMATION
See online article: S1 (table) | S2 (table)
ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2012 Macmillan Publishers Limited. All rights reserved