doi: 10.4436/jass.89017

JASs Do it yourself
Journal of Anthropological Sciences
Vol. 89 (2011), pp. 201-219

Do it yourself

Tools which simulate the evolution of uni-parentally transmitted elements of the human genome

Sergio Tofanelli[1], Luca Taglioli[1], Davide Merlitti[2] & Giorgio Paoli[1]

1) Dipartimento di Biologia, Università di Pisa, via Derna 1, 56126, Pisa, Italy
e-mail: stofanelli@biologia.unipi.it

2) Scuola Normale Superiore, Piazza dei Cavalieri 1, 56126, Pisa, Italy

Keywords - *Computer simulations, mtDNA, Y chromosome, Haplotypes.*

Introduction

Computer simulations in the post-genomic era

High-throughput genotyping and next-generation sequencing technologies are providing unprecedented opportunities to detect signals of our evolutionary history in the human genome, opening new avenues for molecular anthropologists (Destro Bisol *et al.,* 2010).

Dense panels of SNP (Single Nucleotide Polymorphism) and STR (Short Tandem Repeats) markers as well as high coverage DNA sequences are now available for multiple individuals. The entire genome has started to be sequenced at population level (The 1000 Genomes Project Consortium, 2010) and the data are being assembled to reconstruct the finest-grained genetic picture of our demographic and evolutionary history.

The design of efficient tools that can manage the size of next-generation data is an urgent matter. Nevertheless, when tracing the causes of modern genetic variation, several sources of error are likely to intervene, which the increasing amount of information units can hardly prevent. For example, the most recent events might obscure ancient signatures. Secondly, when many factors are concurrently acting, it is extremely difficult to disentangle causes and effects. Thirdly, different processes can produce similar patterns.

In order to lower the risk of either a speculative or erroneous interpretation of the results, the whole available array of evidence regarding populations should be analyzed using statistical tools that can manage the complexity of the interactions. Hence, only systemic approaches networking a wide spectrum of disciplines other than population genetics (i.e. anthropology, history, paleo-ecology, linguistics, archeology among others) have the potential to provide a valid clue. This introduces further complexity in the hard task of selecting a hypothesis among a spectrum of possible hypotheses that could be tested.

A basic contribution in managing all this plus-value complexity is given by "*in silico*" experiments, such as computer simulations (Kitano, 2002). Properly designed simulations could be exploratory, checking the adequacy of the experimental design to research aims before performing new experiments or analyses. For instance, simulating datasets with a variable number of loci might help determine the minimum set of markers that must be genotyped in

order to discriminate between scenarios with differing time depths.

Alternatively, simulations could be explanatory. Simulated distributions under a given set of computable evolutionary parameters could form the theoretical framework upon which to test the consistency of either the experimental evidence or the method used to interpret it. For example, one could simulate the performance of different estimators under a given model (parameter-varying approach, i.e: Takezaki & Nei 2008) or among competing models establish the one that best explains a given dataset (model-varying approach, i.e. Fagundes *et al.,* 2007).

A third target of simulation computing is predictive, giving possible insights into future evolutionary processes. This aspect has been extensively applied to ecological and climatic contexts at a global scale but rarely to the genetic evolution of the human species.

An additional and valuable contribution of simulations is in learning approaches, whenever there is the need to explore the relationships between variables of a mathematical function, or for educational tasks.

The role of NRUP markers

Haploid non recombinant uni-parentally inherited (NRUP) markers, as variants of the plastid genomes (mitochondria and chloroplasts) and large segments of vertebrate sex heterochromosomes (Y and W), are widespread validated tools for phylogenetic/historical reconstructions, life management and forensic analyses (Moritz *et al.,* 1987; Jobling & Tyler-Smith, 2003; Lowe *et al.,* 2004; Sala *et al.,* 2004; Avise, 2004) whose data are treated and stored in a multilocus haplotype format. The reason behind the success in using haplotypes with uni-parental inheritance is either historical, due to the ease in extracting and sequencing plastid DNAs, and theoretical, as lineages branching along gene genealogies can be unambiguously traced. The fact that their effective size is one fourth that of markers with diploid bi-parental inheritance is a double-faced issue. It makes NRUPs a highly sensitive tool in detecting genetic introgression, in estimating divergence times or evaluating stochastic events such as bottlenecks, founder effects and population divergence due to genetic drift. However, it can enhance the role of drift compared to other demographic events.

Since the pioneering studies of Cann *et al.* (1987) on the genealogy of the mitochondrial hypervariable region (HVSI) and of Underhill *et al.* (2000) on the genealogy of unique mutational events at the non recombining region of the Y chromosome (NRY), NRUP haplotypes have been shown to be highly informative polymorphisms which can trace back the pathway of human evolution at different temporal and spatial scales. Moreover, NRUP sequences from animal and plant plastid genomes have been the basic, and sometimes unique, reference data for ancient DNA studies. Despite their limitations due to the large variance associated with inferences based on one locus only, the advent of the genomic era could have difficulty in replacing NRUP data when reconstructing the backbone processes on the male and female side of human evolution. Neither, could the overall neutrality and high sensitivity to drift and migration of NRUPs be ignored in any attempt to reconstruct recent demographic scenarios.

Publicly accessible databases of human mtDNA sequences currently host 12,247 HVSI-II haplotypes with 1,013 variants (EMPOP v2.1.3, MITOMAP r13) and a total of 5,344 complete sequences with 3,049 haplotype variants and 9,280 variant sites (NCIB , MITOMAP r13, HmtDB). Whole-genome sequencing has recently discovered 2,870 variable Y-SNPs in just 77 males from 3 continents (The 1000 Genomes Project Consortium, 2010), with a reasonable final score in the order of tens thousands YSNPs, and forensic databases host up to 97,575 Y-STR haplotypes (YHRD r37). The high rate at which novel variants are discovered ensure that all existing branches of NRUP genealogies will soon be discovered, thus generating new fascinating hypotheses on the historical relationships among human populations. Despite what have been mentioned above, few dedicated programs have been designed to model the evolution of NRUP haplotype data.

Aims

The present paper is a brief comparative and annotated overview on the simulators that have been recently developed in the field of population genetics which can be confidently applied to large NRUP data in order to test hypotheses on the forces that shaped human genetic variation.

Drawbacks and advantages of backward and forward strategies are discussed and the performance of selected applications from each class of programs are evaluated according to criteria of time efficiency and ease of use.

We aim to provide a basic guideline which can introduce non-programmer users interested in human genetic history to computer simulation tools.

Backward- and forward-in-time algorithms

Several applications which make it possible to simulate the virtual evolution of haploid DNA sequences under different scenarios are currently available (for a review see Carvajal-Rodríguez, 2008a, 2010). A first-order classification subdivides them into programs implementing backward-in-time (BiT) and forward-in-time (FiT) algorithms.

Most of the available applications are BiT, based on the standard coalescent model formerly developed by Kingman (1982) or modified versions. Their popularity is largely due to their computational efficiency, which makes it possible to simulate large haplotype data sets (e.g. many loci and individuals) very quickly. Simulating according to the coalescent model means following back the genealogy of a sample of unknown genotype "coalescing" individuals according to a stochastic process that depends on evolutionary parameters such as effective size and migration.

In its simplest version, the coalescent theory assigns the same probability to yield descendants to all the members of the genealogy (selective neutrality). The genealogical and the mutational processes are separable: after the most recent common ancestor (MRCA) of all the sampled individuals is found, the process runs forward in time and randomly assigns genetic information to individuals and lineages on the coalescent tree.

Mathematically, the Kingman coalescent, or n-coalescent, is a stochastic Markov process composed of n-1 independent, random, Poisson-distributed collisions of lineages, given an initial sample size n, a population size N tending to infinity and a time in generations rescaled by θ (for haploid data $\theta=2\mu N_e$, where N_e=effective size and μ=mutation rate). It has been shown to hold for a surprisingly wide variety of population models (Kingman, 1982; Fu 2006), including more realistic reproductive models such as the Moran and the Wright-Fisher (W-F) models, although it can be considered a good approximation of the latter only when the sample size is much smaller than the population effective size ($n \ll N_e$).

It is reasonable to presume that, a low n/N_e ratio will not occur in a near future, when applications face genome-wide panels from thousands of individuals. However, empirical cases with n dominating over N_e can already be found when genotyping the mtDNA control region and the NRY of the human genome. As reported above, there are respectively 12,247 and 97,575 currently available NRUP haplotypes, while long-term human N_e is estimated in the order of 2.5-8 thousand (Takahata *et al.,* 1995; Ingman *et al.,* 2000; Thomson *et al.,* 2000).

Nonetheless, the coalescent theory provides a well-suited approach in analytically inferring gene genealogies from the data using simple evolutionary models. It makes it possible to compute fundamental parameters in population genetics such as the time to coalescence of a number of alleles/lineages (with both, expected value and standard deviation, equal to $2N_e$) or the amount of variation expected from genetic drift alone (i.e. the mean heterozygosity that depends on θ).

In cases of more complex demographic scenarios, when analytical results are not accessible, applications implementing the coalescent model offer the opportunity to simulate genetic data to explore the effects of demographic parameters such as subdivision, size fluctuation, and migration.

Sophisticated strategies in computational statistics have been recently developed which, once coupled to the coalescent, offer a framework to maximize the extraction of the information contained in the data even under complex demographic models.

The Bayesian statistical paradigm, for instance, makes it possible to calculate the distribution of a parameter given the observed data (Posterior distribution) as the probability to obtain the data under a certain model (Likelihood) multiplied by the uncertainty of the parameter distribution based on the knowledge available before looking at the data (Prior distribution). In practice, it makes it possible to start with some previous beliefs (prior) about the population parameter we are investigating and modify those beliefs with empirical data to make new inferences on the history of the population (posterior). Three steps are intimately linked when implementing the Bayesian paradigm in a program. First, models are built using background information. Then, the models are fit to the data by using simulations to calculate the posterior probability distribution and credible intervals for the parameter of interest. Finally, the goodness-of-fit of the conditional distributions between alternative models is evaluated.

Because more solid conclusions can be drawn when empirical data-sets are much more informative to the point that the likelihood dominates over prior knowledge, the Bayesian inference will benefit from the current exponential rate of data growing. Independent of the amounts of empirical data, however, strong prior distributions (i.e. with small variance) have a large influence on the posterior distribution and, hence, on the final inference regarding population history. Thus, situations where there are large amounts of new data and weak priors will make the best use of the Bayesian framework in the near future.

Markov chain Monte Carlo computational models (MCMC, Nielsen & Wakeley, 2001) are a class of algorithms which simplify the calculation of posterior distribution. They search for the set of parameters which give a maximum a posteriori probability under a given model by iteratively generating samples from the posterior distribution until a convergence state is reached. Point estimates and credible intervals can be calculated for any deterministic function of the underlying parameters so that they can be used under a Bayesian framework to simulate gene genealogies under virtually any demographic model.

Unfortunately, the more realistic the model, the more computationally intensive it is to calculate the likelihood function analytically. This is why likelihood-free computational methods have been explored and applied in recent years. One of these families of algorithms is known as Approximate Bayesian Computation (ABC, Beaumont *et al.*, 2002; Marjoram *et al.*, 2003; Sisson *et al.*, 2007; Csillery *et al.,* 2010; Bertorelle *et al.,* 2010). ABC replaces the exact computation of the likelihood function with an approximate version obtained by using summary statistics and simulations under the model. In synthesis, among the millions of genealogies obtained by simulating under different models, those producing the variation which is closest to the observed data are selected to estimate the posterior distribution of the parameter/s. The criteria of selection of the simulated data are crucial. They are usually based on the distance between summary statistics computed in real and simulated data-sets and the cutoff of those data which are lower than an appropriate threshold ε. To date, there is no consensus on the best choice of ε or the best methods to explore the parameter space and the proper number of summary statistics. They are, in fact, case-dependent. Manual adjustments and careful controls of the various steps are needed to achieve good results. The accuracy of the inferences (model checking) should often be validated by the user with the help of statistical tools which are not implemented in the same application with the simulator.

An additional but not trivial drawback of BiT simulations is that evolutionary processes that need a population level to be fully understood (i.e. the fate of rare variants) or non random sampling, such as selection, cannot be treated unless confining them into a very specific context (i.e. strong positive selection, Kaplan *et al.,* 1988). In

case of selection, the assumption of the standard coalescent that individuals have the same probability of yielding descendants does not hold because the reproductive success depends on the alleles/haplotypes codified in each genome.

A strategy implemented in SELSIM (Spencer & Coop, 2004), for instance, bypasses the indetermination regarding the transmission of selected alleles from ancestors to descendants inherent to the coalescent process by admitting the conditional reversibility between the backward trajectory of an allele from the present state to a frequency *x* and the forward *x*-to-loss trajectory of the same allele. A similar approach is developed in MSMS (Ewing & Hermisson, 2010), which extends the performance of MS to the case of selection at a single-locus with two alleles using a three-step procedure that assumes forward and backward conditional processes.

Another limit might be the fact that the most popular applications that implement Bayesian computations, and in particular ABC, are not user-friendly. Efforts towards friendliness are being performed: non-standard ABC have become available for general users within the set of programs called ABCTOOLBOX (Wegmann *et al.,* 2010); recent graphical tools such as M4S2 (Antao *et al.,* 2007), REJECTOR (Jobin & Mountain, 2008) and DIYABC (Cornuet *et al.,* 2008) have been developed to provide a more friendly interface for existing programs.

Usually, ABC steps are not fully integrated with the simulator and it is necessary to pipeline two or more algorithms to complete the analysis. For particularly complex models, ABCs require millions of replications to reach convergence, if there is any, thus decreasing the time efficiency of BiT simulations, unless one resorts to big CPU clusters.

Forward simulations model the evolution of all the sequences contained in a given ancestral population. The properties of the initial population are followed generation by generation under a certain set of genetic or demographic conditions and the final sample can be considered as being representative of either the current or the forthcoming population. As a general rule, every member of a generation can yield descendants following random sampling probabilities (Wright–Fisher model). Access to every individual which forms every generation through the history of the population makes it easy modify demographic (size, migrants) and evolutionary (mutation rates, probabilities of having descendants) variables. It is also easier to model processes determined by population sub-structures.

The higher performance in flexibility and accuracy makes them an ideal tool to approach the complexity of evolutionary pathways actually followed by real populations. In fact, unlike basic Bit approaches, FiT strategies allow for a wider range of demographic scenarios to be tested as null hypothesis. What is more, the uncertainty due to the number of possible trajectories of a backward induction process is reduced and the design of predictive models is allowed.

The main limitation when simulating FiT is the intensive computational effort required: the more complex the model or the larger the loci x chromosomes product (LC-grid), the longer the runtime. Some programs have incorporated strategies to shortcut simulation times. GENOMEPOP (Carvajal-Rodríguez, 2008b) and FREEGENE (Hoggart *et al.,* 2007) scale the population size N and the time t provided to keep the products Nμ, Nm constant. Using a 10-fold re-scaling implies that N and t should be divided by 10, introducing *de facto* a sampling error to the final estimates, while mutation and migration rates should be multiplied by the same factor. In ASHEs (Merlitti & Tofanelli, 2009) and EASYPOP (Balloux, 2001), the computer code has been designed for a more efficient use of the memory. Even using an optimized software, however, simulation designs with thousands of iterations and large LC-grids (>1 M) over hundreds of generations require too much time without access to distributed computing systems.

Comparing features and performance of available programs

The landscape of the simulation software available to population geneticists has been exhaustively

reviewed by Carvajal-Rodríguez (2008a, 2010). The ideal case for handling haploid genomes is a program that can manage NRUP haplotypes in order to mimic the complexity of human genetic history at a global as well as at a regional scale, over large genomic regions, in a time-efficient way, using an easy-to-use interface. Such a program is not yet available today.

Here, we have selected the applications that most closely approach that case: MS (Hudson, 2002), BAYESSC (Anderson *et al.,* 2005), SIMCOAL (Laval & Excoffier, 2004), FASTSIMCOAL (Excoffier & Foll, 2011), POPABC (Lopes *et al.,* 2009) simulating BiT; EASYPOP and ASHEs simulating FiT. Separately or concurrently, they can be confidently applied to large NRUP data in order to reliably test hypotheses on human genetic evolution. Each program can be preparatory either to test alternative demographic histories, by comparing randomized replicates of user-specified models with empirical data, or to model the evolution of specific parameters.

A synopsis of the features and performance of the selected applications are provided in Appendix. The most popular BiT simulator is MS, a powerful and fast sampler that simulates genealogies under any neutral demography (subdivision, divergence, constant size, growth, bottlenecks) once a user-defined population mutation rate (θ) is given. As in most coalescent simulators, the output is software-specific, but it can be piped into SEQGEN (Rambaut & Grassly, 1997) to generate sequences in a NEXUS format or into MLCOALSIM (Ramos-Onsins & Mitchell-Olds, 2007) to construct sample sequences in FASTA format and calculate several neutrality tests. All the above are command line programs that should be compiled on a UNIX system.

Only ASHEs, SIMCOAL, FASTSIMCOAL and BAYESSC are conceived to handle data directly in a NRUP format, namely binary or multistate multilocus haplotypes without recombination across loci. The other applications have options to arbitrarily set the recombination rate and the ploidy level but the outputs are often given in an unfriendly format. This further affects the computational time required to post-process the data, which sometimes can be so slow that it becomes untreatable with a common hardware platform.

BAYESSC, FASTSIMCOAL and SIMCOAL are largely inter-exchangeable applications under different variants of the coalescent. They differ from POPABC and MS in the array of usable markers (i.e. POPABC cannot simulate binary data such as SNPs or RFLPs, MS cannot simulate STR data), in the settable input parameters and in output options (see Appendix). BAYESSC, which is a Bayesian version of SERIALSIMCOAL, and FASTSIMCOAL are the only applications able to add ancient DNA sequences to simulation parameters. POPABC is a package of algorithms which implements an ABC framework under an Isolation with Migration model (Nielsen & Wakeley, 2001; Beaumont & Nichols, 1996).

Also EASYPOP and ASHEs have complementary functions. EASYPOP does not manage DNA sequence formats and returns less output variables but it implements a larger number of migration and mutation models. ASHEs is limited to models involving only one or two populations/demes but is able to manage all data types and a wider number of variables (growth rates, non random sampling probabilities, distance measures). Thanks to its interactive graphical user interface, which makes it possible to visually monitor the output of each variable in real time, ASHEs is the program having the easiest-to-use interface and, hence, is the most suitable for educational tasks.

In order to test the performance of the algorithms implemented in the available simulation programs under different conditions, the outputs of simulations performed by applications under the same model were compared. Among the various measures and models we referred to, the most congruent between programs were chosen (Fig. 1), even if it meant accepting simplicity. All populations were assumed to evolve under neutrality and constant size with no recombination among loci. Accuracy was evaluated in terms of standardized deviation from theoretical expectation (delta between observed and expected

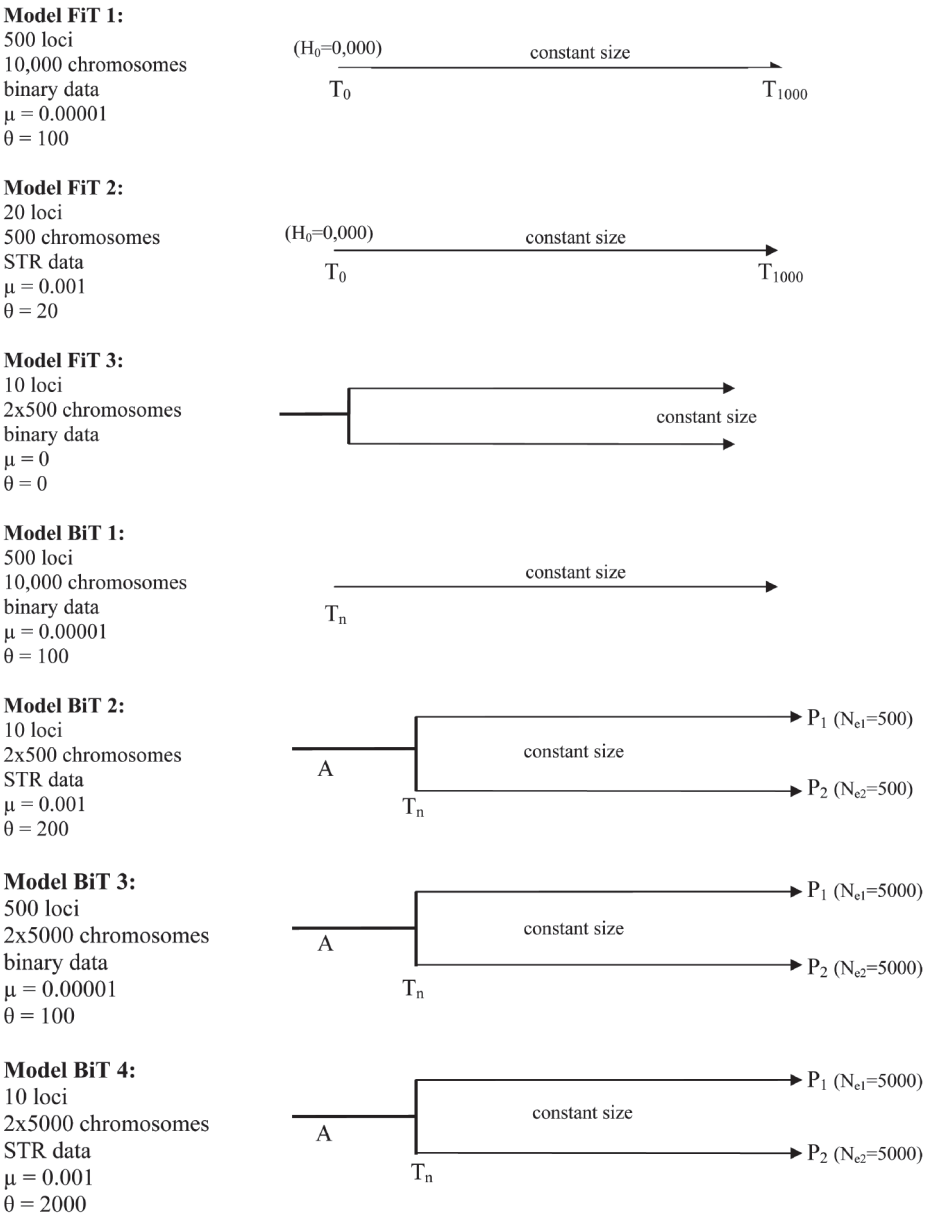

Fig. 1 - Graphic and parameter specifications of the simulated models.

value/expected value) and extent of variation (Coefficient of Variation, CV). Time efficiency was calculated as runtime in seconds. For the sake of uniformity, we performed 100 iterations for each simulation but the reader must bear in mind that runs in the order of thousands should be carried out to obtain satisfactory support to estimates.

Tab. 1 - Simulation parameters and outputs obtained modeling NRUP haplotypes by BiT and FiT programs (n/ Ne ratio, 1:1).

	EXP. VALUES (VAR)	ASHES	EASYPOP
Model FiT1		**mean**	**mean**
Heq	0.9804 (0.0032)	0.9862 (<0.0001)	0.9903 (<0.0001)
Computation time (sec)		50,148	52,950
Post processing program		EXCEL	ARLEQUIN 3.5
Post processing time (sec)		50	13,850
Model FiT2		**mean**	**mean**
Heq	0.9524 (0.0002)	0.9510 (0.0012)	0.9537 (0.0117)
Computation time (sec)		732	135
Post processing program		EXCEL	ARLEQUIN 3.5
Post processing time (sec)		50	3,045
Model FiT3		**mean**	**mean**
FST	0.1993	0.1845 (0.0062)	0.1742 (0.0053)
H	0.8179	0.8173 (0.0032)	0.8224 (0.0026)
Computation time (sec)		400	145
Post processing program		EXCEL	ARLEQUIN 3.5
Post processing time (sec)		100	160

	EXP. VALUES	BAYESSC	SIMCOAL 1.0	MS	FASTSIMCOAL
Model BiT1					
Tn	19,998	19,669	19,514	19,647	19,737
(95%CI)		(1,154-38,184)	(10,567-28,461)	(0-39,625)	(954-38,519)
Computation time (sec)		29	228	35	298
Post processing program		EXCEL	-	-	-
Post processing time (sec)		190	-	-	-
Model BiT2					
Tn	1,998	2,258	1,924		2,229
(95%CI)		(287-4,229)	(362-3,486)	ND	(0-4,477)
Computation time (sec)		12	4		3
Post processing program		EXCEL	-		-
Post processing time (sec)		190	-		-
Model BiT3					
Tn	19,998	22,906	19,471	21,157	20,403
(95%CI)		(1,966-43,845)	(9,043-29,899)	(0-47,919)	(0-43,930)
Computation time (sec)		32	245	40	280
Post processing program		EXCEL	-	-	-
Post processing time (sec)		190	-	-	-
Model BiT4					
Tn	19,998	20,642	18,479		20,302
(95%CI)		(0-42,406)	(9,622-27,336)	ND	(1,568-39,035)
Computation time (sec)		11	1363		9
Post processing program		EXCEL	-		-
Post processing time (sec)		190	-		-

FiT simulation tests

FiT1 models the evolution of large binary data (5M LC-grid) in one population under a mutation/drift equilibrium. The expected probability at equilibrium under the Infinite-Allele Model (IAM) that two haplotypes chosen at random are different was estimated according to Kimura & Crow (1964) and Watterson (1974) and then compared with the mean haplotype diversity obtained with the stochastic algorithms implemented in EASYPOP and ASHEs after a time span largely sufficient to reach the stabilization of diversity values (1,000 generations). FiT2 mirrors the previous model in the case of few STR data (10K LC-grid). The expected diversity at equilibrium under a Stepwise Mutation Model (SMM) was estimated according to Ohta & Kimura (1973) and Moran (1976). FiT3 models the evolution of two equally-sized Wright-Fisher populations (5K LC binary grid) split from an ancestral population in absence of mutation and migration. Expected values of F_{ST} and H (haplotype diversity) after 100 generations were calculated according to Wright (1951) and Nei (1987).

BiT simulation tests

We performed a comparative test of the accuracy of the coalescent algorithms in the estimation of a key parameter of anthropological research: the time since the most recent common ancestor (TMRCA). Under the standard coalescent model, where only one collision per generation is assumed, the expected TMRCA is given by $2N_e(1-1/n)$. When the sample size is much larger than the effective size, multiple simultaneous collisions need to be implemented to avoid an overestimation of the TMRCA. We tested the deviation of simulated values from the theoretical expectation under four simple models: BiT1, which reproduces basically the same conditions of FiT 1 (5M binary LC-grid); BiT2, which simulates the effect of drift in two pools of 10-locus Y-STR haplotypes of equal size 1,000 (10K LC-grid) over 100 generations; BiT3 follows the same divergence scheme as BiT2 but with BiT1 parameters; BiT4 mirrors BiT2 with a larger LC grid. The four models were simulated under the n-coalescent (MS, SIMCOAL 1.0), under a coalescence adjusted for multiple collisions per generation (BAYESSC) and a sequential Markov coalescence (FASTSIMCOAL), each time assuming the n/N_e ratio varying from 0.001 to 8 (Fig. 2).

Simulation results

As a general rule, the mean values of the demographic parameters obtained both by FiT and BiT algorithms showed a fairly good adhesion to the theoretical expectations when the n/Ne ratio approaches one (see Tab. 1 and Fig. 2a). Estimates fluctuated between +25% and -10% of the expected value with the exception of TMRCAs obtained simulating by SIMCOAL 1.0. In the latter case, clear deviations towards underestimation (up to -50% than expected) were observed for $n<N_e$ and towards overestimation (up to +150% than expected) for $n>N_e$. As a consequence, the new released versions of SIMCOAL 1.0 are to be preferred (SIMCOAL 2, FASTSIMCOAL).

The pros and cons of the two kinds of approach were here confirmed: simulating FiT makes it possible to obtain higher accuracy but with low time efficiency (Tab. 1). Time consumption depends on data size and on post-processing stages but one should expect, on average, to spend >100 times longer to simulate forward-in-time.

The stochastic process inherent to the coalescent induces high variation in TMRCA estimates from one simulation to the next (average CV around 0.50) but computation times are only a few seconds as long as the sample size remains low (Fig. 2b).

It's worth noting that advanced versions (FASTSIMCOAL) of the same basic script (SIMCOAL 1.0) scarcely affect time-efficiency for large LC-grids of NRUPs and that large data (n/Ne 8:1, LC>20M) drive most applications to run failure, as would be expected from coalescent assumptions.

It still remains to be demonstrated whether more biased trees are produced when much more

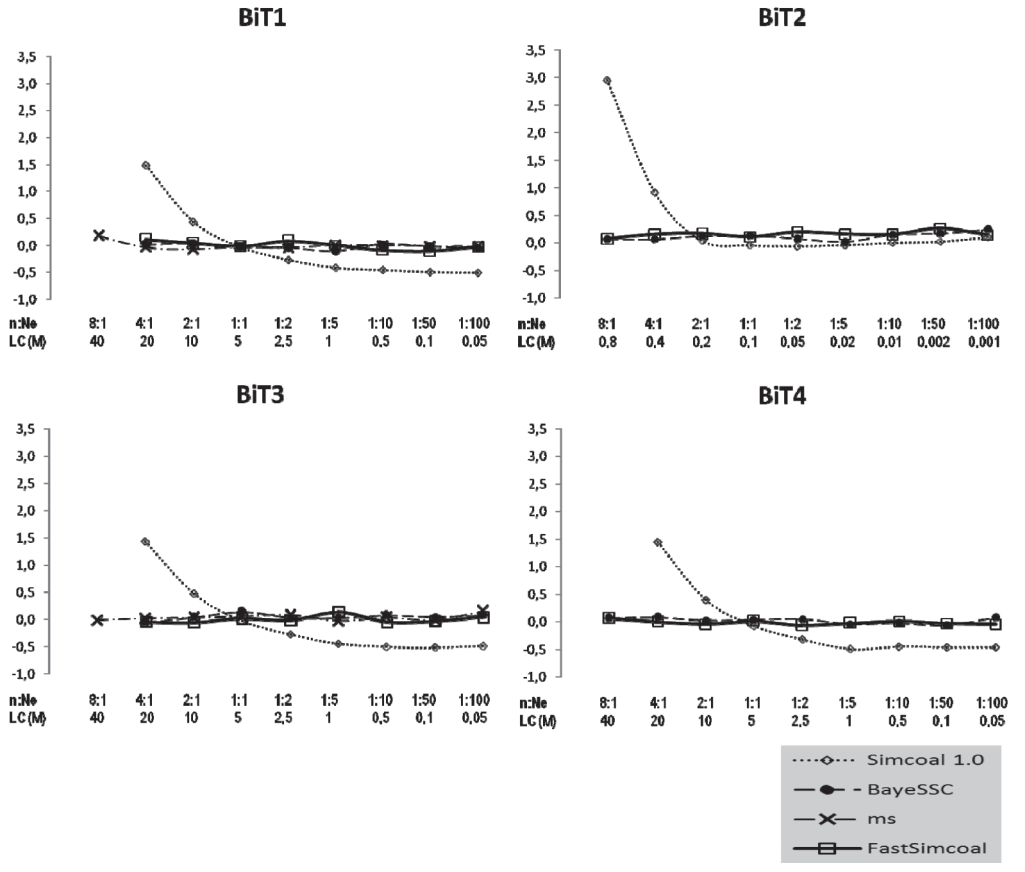

Fig. 2 – a) Deviations from expected TMRCA values [(Tobs-Texp)/Texp] after simulating 100 iterations under Bit1-4 using applications implementing coalescent models with varying n:Ne ratios, loci x chromosomes (LC) grids and theta values (θ).

complex scenarios or population parameters other than TMRCA are modeled.

Wakeley & Takahashi (2003) have measured the side-effects of modeling genealogies under the n-coalescent with sample sizes which are as large as or larger than the effective size (i.e. structured populations). They showed that inflated rates of singletons (namely, the polymorphisms found in only one sequence) are generated during the first backwards generations relative to model predictions that may erroneously mimic population growth or positive selection. This makes it harder to distinguish between adaptive and demographic effects by means of neutrality tests based on the shape of the allele frequency spectrum, or AFS (i.e. Tajima's D, Fay & Wu's H statistic, Fu & Li parameters). An AFS which is more skewed than expected under neutrality might yield misleading demographic inferences such as false positive results of bottleneck-shaped or admixed populations (Gabor *et al.,* 2004; Fu, 2006; Lohmueller *et al.,* 2010).

Conclusions

The post-genomic era makes it paramount complement the increasing amounts of genetic data with *in-silico* tools able to model the

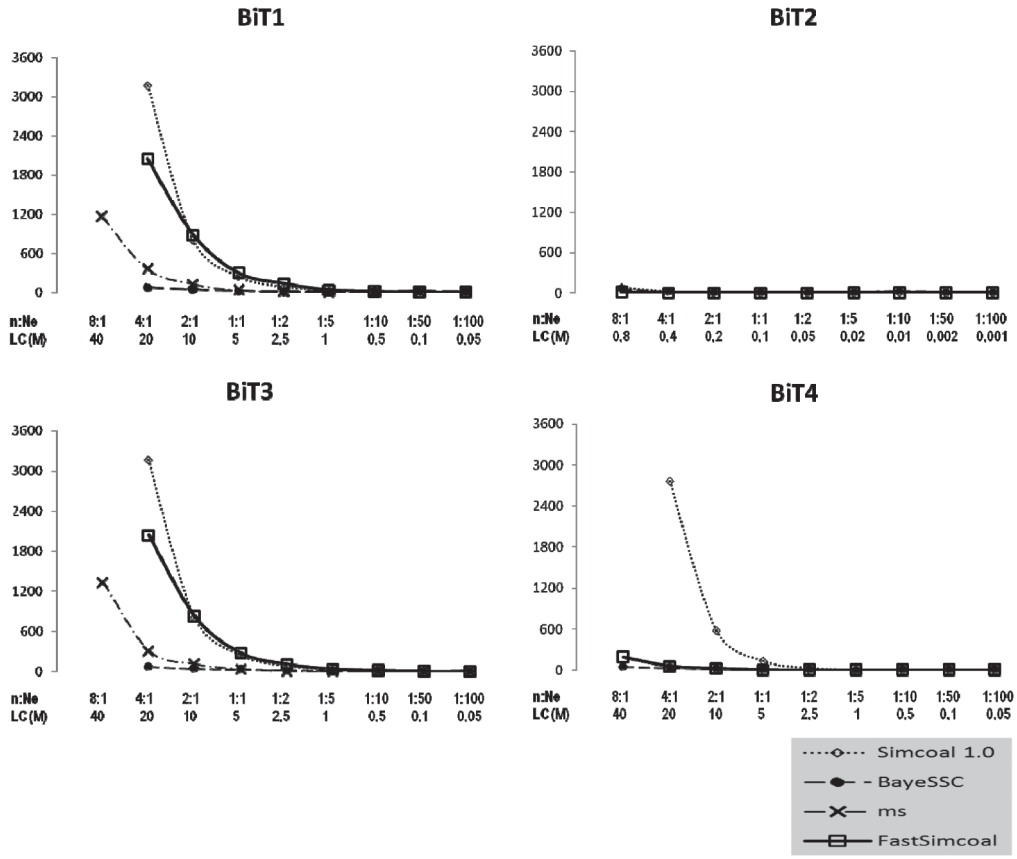

Fig. 2 - b) Runtime (in seconds with AMD PhenomII X4 925 Processor 2.8 GHz and 4 GB RAM + 2 Gb graphic memory under the WINDOWS7 64 bit operating system and under LINUX UBUNTU 10.10 64 bit VirtualBox emulator for MS simulations) after simulating 100 iterations under Bit1-4 using applications implementing coalescent models with varying n:Ne ratios, loci x chromosomes (LC) grids and theta values (θ)

network of interacting actors shaping biological variation. Hence, computationally efficient simulators that can explore complex demographic scenarios over large NRUP regions are likely to play an increasing role in modern evolutionary anthropology. With LC panels in the order of millions and multi-disciplinary prior knowledge becoming common place, hypothesis-driven and model-based approaches will require tools with increasing time-efficiency and flexibility.

The high number of recent simulators, their growing specialization, their generally low friendliness, the lack of a directory where one can find out what exactly a program can and cannot do, might make first approaches to the world of simulation an hard task. Nonetheless, users should trust that, once they have become accustomed to some basic concepts, the quality of the products will depend on their own skills and inventiveness rather than on the ability to manipulate computational tools.

Coalescent-based approaches enable the production of huge amounts of samples in little time but sophisticated algorithms have to be coupled to partially compensate their limited flexibility. Time-forward approaches more closely

fit the complexity underlying the evolution of real populations but demand intensive computations.

To date, backward and forward strategies should be considered synergic rather than conflicting or alternative. When possible, it is advisable to check the adequacy of observed parameters to model assumptions under both time-simulation strategies. Their combined use has been recently applied to test human evolutionary models (Padhukasahasram *et al.,* 2008; Cyran & Myszor, 2008) and sometimes they have produced contrasting results.

Hopefully, the increasing charge of variable markers from population-scaled genomic surveys will be fully compensated by either, the rapid development of multi-core High Performance Computing (HPC) systems (Bader, 2004; Mode & Gallop, 2008; Kim & Wiehe, 2009), capable of managing data grids in the order of billions by dividing the main algorithm into a number of block units running simultaneously on different CPUs, or the implementation of the algorithms under graphic processing units (GPU) (see Owens *et al.,* 2007, Suchard & Rambaut, 2009). As soon as complex simulation can be done in acceptable times, the accuracy and flexibility of FiT applications should be preferred.

For now, it is wise to make a personal evaluation each time of which is the best compromise between efficiency and flexibility, bearing in mind that all models are approximations. In any case, the higher the accuracy of competing models, the lower their overlapping and the closer the inferred history is to the real population history, whatever the simulation philosophy. This is why the most promising contribution to human evolutionary studies should rely on an original *in-silico* treatment of well-integrated sources of evidence (genetic and non-genetic) other than from the refinement of existing computational tools.

Acknowledgements

S.T. and G.P. were supported by a research grant (ex60%) from the University of Pisa.

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. *Bioinformatics*, 21: 1733–1734.

Antao T., Beja-Pereira A. & Luikart G. 2007. MODELER4SIMCOAL2: a user-friendly, extensible modeler of demography and linked loci for coalescent simulations. *Bioinformatics*, 23: 1848–1850.

Avise J.C. 2004. *Molecular Markers, Natural History, and Evolution, second edition.* Sinauer Associates Inc., Sunderland, MA.

Bader D.A. 2004. Computational Biology and High-Performance Computing. *Commun. A.C.M.*, 47: 35–41.

Balloux F. 2001. EASYPOP (version 1.7): a computer program for population genetics simulations. *J. Hered.*, 92: 301-302.

Beaumont M.A., Zhang W. & Balding D.J. 2002. Approximate Bayesian computation in population genetics. *Genetics*, 162: 2025–2035.

Beaumont M.A., Nielsen R., Robert C., Hey J., Gaggiotti O., Knowles L., Estoup A., Panchal M., Corer J., Hickerson M., Sisson S.A., Fagundes N., Chikhi L., Beerli P., Vitalis R., Cornuet J.M., Huelsenbeck J., Foll M., Yang Z.H., Rousset F., Balding D. & Excoffier L. 2010. In defence of model-based inference in phylogeography. *Mol. Ecol.,* 19:436–446.

Cann R.L., Stoneking M. & Wilson A.C. 1987. Mitochondrial DNA and human evolution. *Nature*, 325: 31–36.

Carvajal-Rodríguez A. 2008a. Simulation of genomes: a review. *Curr. Genomics*, 9: 155–159.

Carvajal-Rodríguez A. 2008b. GENOMEPOP: a program to simulate genomes in populations. *BMC Bioinformatics*, 9: 223.

Carvajal-Rodríguez A. 2010. Simulation of Genes and Genomes Forward in Time. *Current Genomics*, 11: 58-61

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*, 24:2713-2719.

Cyran K.A. & Myszor D. 2008. Coalescent vs. time-forward simulations in the problem of the detection of past population expansion. *Int. J. Appl. Math. Inf.,* 1:10-17.

Destro-Bisol G., Jobling M.A., Rocha J., Novembre J., Richards M.B., Mulligan C., Batini C. & Manni F. 2010. Molecular anthropology in the genomic era. *J. Anthropol. Sci.,* 88:93-112.

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

Excoffier L. & Foll M. 2011. FASTSIMCOAL: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, 27:1332-1334.

Fagundes N.J., 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.*, 104:17614–17619.

Fu Y.X. 2006. Exact coalescent for the Wright-Fisher model. *Theor. Popul. Biol.*, 69: 385–394.

Gabor TM., Czabarkaa E., MurvaiaJ. & Sherrya S.T. 2004. The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations. *Genetics*, 166:351–372.

Hoggart C.J., Chadeau-Hyam M., Clark T.G., Lampariello R., Whittaker J.C., De Iorio M. & Balding D.J. 2007. Sequence-level population simulations over large genomic regions. *Genetics*, 177: 1725–1731.

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, 18:337-338.

Ingman M., Kaessmann H., Pääbo S. & Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408:708-713.

Jobin M.J. & Mountain J.L. 2008. REJECTOR: software for population history inference from genetic data via a rejection algorithm. *Bioinformatics*, 24: 2936–2907.

Jobling M.A. & Tyler-Smith C. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.*, 4: 598-612.

Kaplan N.L., Darden T. & Hudson R.R. 1988. The coalescent process in models with selection. *Genetics*, 120: 819–829.

Kim Y. & Wiehe T. 2009. Simulation of DNA sequence evolution under models of recent directional selection. *Brief. Bioinform.*, 10: 84–96.

Kimura M. & Crow J.F. 1964. The number of alleles that can be maintained in a finite population. *Genetics*, 49: 725–738.

Kingman J.F.C. 1982. The coalescent. *Stochastic Process. Appl.*, 13:235–248.

Kitano H. 2002. Systems biology: a brief overview. *Science*, 295:1662–1664.

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. *Bioinformatics*, 20:2485–2487.

Lohmueller K.E., Bustamante C.D. & Clark A.G. 2010. The Effect of Recent Admixture on Inference of Ancient Human Population History. *Genetics*, 185: 611–622.

Lopes J.S., Balding D. & Beaumont M.A. 2009. PopABC: a program to infer historical demographic parameters. *Bioinformatics*, 25:2747–2749.

Lowe A., Harris S. & Ashton P. 2004. *Ecological genetics: design, analysis, and application.* Blackwell Publishing, Oxford.

Marjoram P., Molitor J., Plagno V. & Tavaré S. 2003. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. U.S.A.,* 100: 15324–15328.

Merlitti D. & Tofanelli S. 2009. ASHEs An application to simulate Haplotype Evolution. http://ashes.codeplex.com/

Mode C.J. & Gallop R.J. 2008. A review on Monte Carlo simulation methods as they apply to mutation and selection as formulated in Wright-Fisher models of evolutionary genetics. *Math. Biosci.*, 211: 205–225.

Moran P.A. 1976. Wandering distribution and the electrophoretic profile. II. *Theor. Popul. Biol.*, 10: 145–149.

Moritz C., Dowling T.E. & Brown W.M. 1987. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Ann. Rev. Ecol. Syst.*, 18: 269–292.

Nei M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York, NY.

Nielsen R. & Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, 158:885-896.

Ohta T. & Kimura M. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res*., 22: 201–204.

Owens J.D., Luebke D., Govindaraju N., Harris M., Kruger J., Lefohn A.E. & Purcell T.J. 2007. A survey of general-purpose computation on graphics hardware. *Comput. Graph. Forum*, 26:80-113.

Padhukasahasram B., Marjoram P., Wall J.D., Bustamante C.D. & Nordborg M. 2008. Exploring population genetic models with recombination using efficient forward-time simulations. *Genetics*, 178:2417-2427.

Sala A., Marino M., Arguelles C., Fenocchio A. & Corach D. 2004. Uniparentally inherited genetic markers as tools for ethnic and geographical origin detection of forensic samples. *Int. Congr. Ser.*, 1261: 625–627.

Sisson S.A., Fan Y. & Tanaka M.M. 2007. Sequential Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. U.S.A.*, 104:1760–1765.

Spencer C.C. & Coop G. 2004. SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, 20:3673-5.

Suchard M.A. & Rambaut A. 2009. Many-core algorithms for statistical phylogenetics. *Bioinformatics*, 25:1370–1376.

Takahata N., Satta Y. & Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.*, 48:198–221.

Takezaki N. & Nei M. 2008. Empirical Tests of the Reliability of Phylogenetic Trees Constructed With Microsatellite DNA. *Genetics*, 178:385-392.

Templeton A.R. 2010a. Correcting approximate Bayesian computation. *Trends Ecol. Evol.* 25:488-9.

Templeton A.R. 2010b. Coherent and incoherent inference in phylogeography and human evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 107:6376–6381.

Thomson R., Pritchard J.K., Shen P., Oefner P.J. & Feldman M.W. 2000. Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proc. Natl. Acad. Sci. U.S.A.*, 97:7360-7365.

1000 Genomes Project Consortium, Durbin R.M., Abecasis G.R., Altshuler D.L., Auton A., Brooks L.D., Durbin R.M., Gibbs .RA., Hurles M.E. & McVean G.A. 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073.

Underhill P.A., Shen P., Lin A.A., Jin L., Passarino G., Yang W.H., Kauffman E., Bonné-Tamir B., Bertranpetit J., Francalacci P., Ibrahim M., Jenkins T., Kidd J.R., Mehdi S.Q., Seielstad M.T., Wells R.S., Piazza A., Davis R.W., Feldman M.W., Cavalli-Sforza L.L. & Oefner P.J. 2000. Y chromosome sequence variation and the history of human populations. *Nat. Genet.*, 26:358-361.

Wakeley J. & Takahashi T. 2003. Gene genealogies when the sample size exceeds the effective size of the population. *Mol. Biol. Evol.*, 20:208–213.

Watterson G.A. 1974. The sampling theory of selectively neutral alleles. *Adv. Appl. Probab.*, 6: 463–488.

Wright S. 1951. The genetical structure of populations. *Ann. Eugen.*, 15: 323–354.

Applications of PopABC

Czellar V. & Ronchetti E. 2010. Accurate and Robust Tests for Indirect Inference. *Biometrika*, 97:621–630.

Palero F., Lopes J., Abelló P., Macpherson E., Pascual M. & Beaumont M.A. 2009. Rapid radiation in spiny lobsters (*Palinurus spp*) as revealed by classic and ABC methods using mtDNA and microsatellite data. *BMC Evol. Biol.*, 9: 263.

Applications of BayeSSC

Balakrishnan C.N. & Edwards S.V. 2009. Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the zebra finch (*Taeniopygia guttata*). *Genetics*, 181: 645–660.

Fabre V., Condemi S. & Degioanni A. 2009. Genetic evidence of geographical groups among Neanderthals. *PLoS One*, 4:e5151.

Ghirotto S, Mona S, Benazzo A, Paparazzo F, Caramelli D & Barbujani G. 2010. Inferring genealogical processes from patterns of Bronze-Age and modern DNA variation in Sardinia. *Mol. Biol. Evol.*, 27:875–886.

Haak W., Balanovsky O., Sanchez J.J., Koshel S., Zaporozhchenko V., Adler C.J., Der Sarkissian C.S., Brandt G., Schwarz C., Nicklisch N., Dresely V., Fritsch B., Balanovska E., Villems R., Meller H., Alt K.W., Cooper A. & Members of the Genographic Consortium. 2010. Ancient DNA from European early Neolithic farmers reveals their near eastern affinities. *PLoS Biol.*, 8:e1000536.

Valdiosera C.E., García-Garitagoitia J.L., Garcia N., Doadrio I., Thomas M.G., Hänni C., Arsuaga J.L., Barnes I., Hofreiter M., Orlando L. & Götherström A. 2008. Surprising migration and population size dynamics in ancient Iberian brown bears (*Ursus arctos*). *Proc. Natl. Acad. Sci. U.S.A.*, 105: 5123–5128.

Zárate S., Pond S.L., Shapshak P. & Frost S.D. 2007. Comparative study of methods for detecting sequence compartmentalization in human immunodeficiency virus type 1. *J. Virol.*, 81: 6643–6651.

Applications of SimCoal and SerialSimCoal

Belle E.M., Ramakrishnan U., Mountain J.L. & Barbujani G. 2006. Serial coalescent simulations suggest a weak genealogical relationship between Etruscans and modern Tuscans. *Proc. Natl. Acad. Sci. U.S.A.*, 103: 8012–8017.

Chan Y.L., Anderson C.N. & Hadly E.A. 2006. Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA. *PLoS Genet.*, 2:e59.

Eckert A.J. & Carstens B.C. 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. *Mol Phylogenet. Evol.*, 49:832–842.

Kayser M, Lao O. & Stoneking M. 2008. Reply to Hedrick. *Am. J. Hum. Genet.*, 83: 140–142.

Moya O., Contreras-Díaz H.G., Oromí P. & Juan C. 2007. Phylogeography of a ground beetle species in La Gomera (Canary Islands): the effects of landscape topology and population history. *Heredity*, 99: 322–330.

Rodrigo A.G., Tsai P. & Shearman H. 2009. On the Use of Bootstrapped Topologies in Coalescent-Based Bayesian MCMC Inference: A Comparison of Estimation and Computational Efficiencies. *Evol. Bioinform. Online*, 5: 97–105.

Applications of EasyPop

Balloux F., Amos W. & Coulson T. 2004. Does heterozygosity estimate inbreeding in real populations. *Mol. Ecol.*, 13: 3021–3031.

Balloux F., Lehmann L. & de Meeus T. 2003. The population genetics of clonal and partially clonal diploids. *Genetics*, 164: 1635–1644.

Corander J. & Marttinen P. 2006. Bayesian identification of admixture events using multilocus molecular markers. *Mol. Ecol.*, 15:2833–2843.

Hardy O.J., Charbonnel N., Fréville H. & Heuertz M. 2003. Microsatellite allele sizes: A simple test to assess their significance on genetic differentiation. *Genetics*, 163: 1467–1482.

Koskinen M.T., Haugen T.O. & Primmer C.R. 2003. Contemporary fisherian life-history evolution in small salmonid populations. *Nature*, 419: 826-830.

Fisher M.C., Rannala B., Chaturvedi V. & Taylor J.W. 2002. Disease surveillance in recombining pathogens: multilocus genotypes identify sources of human Coccidioides infections. *Proc. Natl. Acad. Sci. U.S.A.*, 99: 9067–9071.

Goudet J., Perrin N. & Waser P. 2002. Tests for sex-biased dispersal using bi-parentally inherited genetic markers. *Mol. Ecol.*, 11: 1103–1114.

Landguth E.L., Cushman S.A., Murphy M.A. & LuiKart G. 2010. Relationships between migration rates and landscape resistance assessed using individual-based simulations. *Mol. Ecol. Resour.,* 10: 854–862.

Applications of ASHEs

Batini C., Ferri G., Destro-Bisol G., Brisighelli F., Luiselli D., Sánchez-Diz P., Rocha J., Simonson T., Brehm A., Montano V., Elwali N.E., Spedini G., D'Amato M., Myres N., Ebbesen P., Comas D. & Capelli C. 2011. Signatures of the pre-agricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. *Mol. Biol. Evol.*, 28(9):2603-13.

Capelli C., Onofri V., Brisighelli F., Boschi I., Scarnicci F., Masullo M., Ferri G., Tofanelli S., Tagliabracci A., Gusmao L., Amorim A., Gatto F., Kirin M., Merlitti D., Brion M., Verea A.B., Romano V., Cali F. & Pascali V. 2009. Moors and Saracens in Europe: estimating the medieval North African male legacy in southern Europe. *Eur. J. Hum. Genet.*, 17:848–852.

Tofanelli S., Bertoncini S., Castrì L., Luiselli D., Calafell F., Donati G. & Paoli G. 2009a. On the origins and admixture of Malagasy: new evidence from high resolution analyses of paternal and maternal lineages. *Mol. Biol. Evol.*, 26: 2109–2124.

Tofanelli S., Ferri G., Bulayeva K., Caciagli L., Onofri V., Taglioli L., Bulayev O., Boschi I., Alu` M., Barni F., Rapone C., Beduschi G., Luiselli D., Cadenas A.M., Awadelkarim K.D., Mariani-Costantini R., Elwali N.E., Verginelli F., Pilli E., Herrera R.J., Gusmao L., Paoli G. & Capelli C. 2009b. J1-M267 Y lineage marks climate-driven pre-historical human displacements. *Eur. J. Hum. Genet.*, 17:1520-1524.

Applications of MS (a selection)

Cox M.P., Morales D.A., Woerner A.E., Sozanski J., Wall J.D. & Hammer M.F. 2009. Autosomal resequence data reveal late stone age signals of population expansion in Sub-Saharan African foraging and farming populations. *PLoS One*, 4: e6366.

De Giorgio M., Jakobsson M. & Rosenberg N.A. 2009. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc. Natl. Acad. Sci. U.S.A.*, 106:16057–16062.

Hammer M.F., Mendez F.L., Cox M.P., Woerner A.E. & Wall J.D. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet.,* 4: e1000202.

Hodgkinson A. & Eyre-Walker A. 2010. The genomic distribution and local context of coincident SNPs in Human and Chimpanzee. *Genome Biol. Evol.*, 2:547–557.

Labuda D., Lefebvre J-F., Nadeau P. & Roy-Gagnon M-H. 2010. Female-to-male breeding ratio in modern humans: an analysis based on historical recombinations. *Am. J. Hum. Genet.*, 86: 353–363.

Lambert C.A., Connelly C.F., Madeoy J., Qiu R., Olson M.V. & Akey J. 2010. Highly punctuated patterns of population structure on the X chromosome and implications for African evolutionary history. *Am. J. Hum. Genet.*, 86: 34–44.

Marth G., Schuler G., Yeh R., *et al.,* 2003. Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc. Natl. Acad. Sci. U.S.A.,* 100: 376–381.

Tarazona-Santos E., Fabbri C., Yeager M., Magalhaes W.C., Burdett L., Crenshaw A., Pettener D. & Chanock S.J. 2010. Diversity in the glucose transporter-4 gene (SLC2A4) in humans reflects the action of natural selection along the old-world primates evolution. *PLoS One*, 5: e9827

Voight B.F., Adams A.M., Frisse L.A., Qian Y., Hudson R.R. & Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl. Acad. Sci. U.S.A.,* 102: 18508–18513.

Wall J.D., Lohmueller K.E. & Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.*, 26: 1823–1827.

Editor, Giovanni Destro-Bisol

Appendix - Features of seven applications which simulate NRUP haplotypes evolving under backward-in-time (BiT) and forward-in-time (FiT) models.

ABC	Approximate Bayesian Computation	MD	Mismatch distribution
ASD	Average square distance	MFS	Mutation Frequency Spectrum
DHS	Haplotype sharing distance	Mono	Cross-platform framework
FST	Wright's Fixation index	MPD	Mean pairwise differences
FWC	Weir & Cockerman's FST	Ne	Population effective size
G2	Pearson's G2	Nm	Number of migrants
GUI	Graphic User Interface	p	Allele/haplotype frequency
H	Nei's haplotype diversity	RST	Shriver's genetic distance
HS	Mean haplotype diversity	S	number of segregating sites
HT	Total haplotype diversity	Sh	Shannon's index
IAM	Infinite alleles model	SMM	Stepwise mutation model
IBD	Isolation by distance	SSM	Stepping-stones model
IM	Isolation with Migration model	Tl	Tree length
IsM	Island model	TMRCA	Time since the most recent common ancestor
JC	Jukes & Cantor model	VarL	Variance of alleles length
k	Number of diverse alleles/haplotypes	W-F	Wright-Fisher model
K2p	Kimura 2 parameter model	WPD	Walsh's probability distribution
KurL	Kurtosis of alleles length	π	Nucleotide diversity
MCMC	Markov chain Monte Carlo	*imported from .txt files under a binary format	

	ASHES	**EASYPOP**	**POPABC**
Version	1.1	2.0.1	1.0
Language	C#	C	C
Interface	GUI	Command-line	ASCII-based menu Command-line
Operating system	Windows, Unix (Mono)	Windows, MacOSX	Windows, Unix, MacOSX
License	GNU GPL	None	GNU
Time model	FiT	FiT	BiT
Reproductive model	W-F	W-F	Exact-coalescent
Random N generator	Knuth's Random	L'Ecuyer	–
MCMC	Yes	Yes	No
Inference framework	–	–	Bayesian
Source Reference	Merlitti & Tofanelli 2009	Balloux 2000	Lopes *et al.* 2009
URL	ashes.codeplex.com	www.unil.ch/dee	www.reading.ac.uk/~sar05sal/software.htm
INPUTS			
Max N populations	2	10,000	5
Loci x chromosomes	unlimited	unlimited	unlimited
DNA sequences	Yes*	No	Yes
STRs	Yes	Yes	Yes
Binary data	Yes	Yes	No
Linkage option	Full	Settable	Settable
Haplotype diversity	user-defined	2 states	random
Growth rates	Yes	No	No
Migration rates	Yes	No	Yes
Migr rate distribution	point	point	point
Migration model	Splitting	SSM,IsM,IBD	IM

Appendix (continued).

		ASHES	EASYPOP	POPABC
Mutation rates		Yes	Yes	Yes
Mut rate distribution		point	point	normal,lognormal
Mutation models		SMM/IAM	SMM/KAM	SMM/IAM
Historical events		No	No	No
Non random sampling		Settable	No	No
Max N Iterations		infinite	999	infinite
Input Datafile		Text (.txt) ARLEQUIN (.arp) EXCEL (.csv)	user-dependent	Text (.len) Convertible from GENEPOP or NEXUS files
Data generator		Yes	Yes	No
Data download		Yes	No	Yes
OUTPUTS				
Haplotype format		Yes	No	No
Data output		.xml	.dat .gen	–
Variables/parameters			.equ	.txt .dat .mut
Within groups	**DNA**	p,H,k,Ne	–	π,k,MFS,Sh,S
	Others	p,H,k,Ne	–	H,varL,kurL,Nm,k,Sh,S
Between groups	**DNA**	FST,FWC,DHS	–	–
	Others	FST,FWC,DHS	FST,HS,HT	–
Genealogy		No	.pdg	No
Real time output		Yes	No	No
Post-processing		R package, EXCEL	FSTAT, ARLEQUIN,GENEPOP	R package, EXCEL
Summary statistics		No	No	Yes
Model checking		No	No	No

	BAYESSC	FASTSIMCOAL	MS	SIMCOAL
Version	1.0	1.1.2	Oct 2007	1.0
Language	C++	C++	C++	C++
Interface	ASCII menu	ASCII menu	Command-line	ASCII menu
Operating system	Windows,Unix,MacOSX	Windows,Linux,MacOSX	Unix,Linux,MacOSX	Windows, Linux
License	None	None	None	None
Time model	BiT	BiT	BiT	BiT
Reproductive model	Exact coalescent	n-coalescent under a Markov sequential algorithm	n-coalescent	n-coalescent
Random N generator	Mersenne Twister	–	Specified in rand1.c	–
MCMC	Yes	Yes	No	Yes
Inferential framework	Bayesian	Bayesian	–	–
Source Reference	Anderson et al. 2005	Excoffier & Foll 2011	Hudson 2002	Laval & Excoffier 2004
URL	www.stanford.edu/group/hadlylab/ssc.html	http://cmpg.unibe.ch/software/fastsimcoal/	http://home.uchicago.edu/rhudson1/source/mksamples.html	http://cmpg.unibe.ch/software/simcoal/

Appendix (continued).

	BAYESSC	FASTSIMCOAL	MS	SIMCOAL
INPUTS				
Max N populations	unlimited	unlimited	unlimited	unlimited
Loci x chromosomes	unlimited	unlimited	unlimited	unlimited
DNA sequences	Yes	Yes	Yes	Yes
STRs	Yes	Yes	No	Yes
Binary data	Yes	Yes	Yes	Yes
Linkage option	Full	Settable	Settable	Full
Haplotype diversity	random	random	random	random
Growth rates	Yes	Yes	Yes	Yes
Migration rates	Yes	Yes	Yes	Yes
Migr rate distribution	point	point	point	point
Migration model	SSM,IsM	SSM,IsM	SSM,IsM	SSM,IsM
Mutation rates	Yes	Yes	Yes	Yes
Mut rate distribution	uniform, gamma	uniform, gamma	point	uniform, gamma
Mutation models	SMM/IAM/JC/K2p	SMM/KAM/JC/K2p	IAM	SMM/KAM/JC/K2p
Historical events	Yes	Yes	Yes	Yes
Non random sampling	No	No	No	No
Max N Iterations	infinite	infinite	infinite	infinite
Input Datafile	Text (.par)	Text (.par)	Command line	Text (.par)
Data generator	Yes	Yes	Yes	Yes
Data download	No	No	No	No
OUTPUTS				
Haplotype format	Yes	Yes	Yes	Yes
Data output	–	.arp	Screen display or .txt	.arp .
Variables/parameters	.csv .gen	.gen .paup	–	.gen .paup
Within groups	π,Tajima's D,MD	TMRCA, MPD, TI, Tajima's D	TMRCA, MPD, TI, Tajima's D	TMRCA, MPD, TI
	H,k	TMRCA, MPD, TI	TMRCA, MPD, TI	TMRCA, MPD, TI
Between groups	F_{ST},H_S,H_T,PD,TMRCA	TMRCA, MPD, TI	TMRCA, MPD, TI	TMRCA, MPD, TI
	R_{ST},G^2,TMRCA	TMRCA, MPD, TI	TMRCA, MPD, TI	TMRCA, MPD, TI
Genealogy/Tree	.trees	.trees	Screen display or .txt	.trees
Real time output	No	No	No	No
Postprocessing	R package, ARLEQUIN,PAUP, EXCEL,TREEVIEW	ARLEQUIN, ARLSUMSTAT	TEXTPAD, PHYLIP	ARLEQUIN,PAUP, EXCEL,TREEVIEW
Summary statistics	Yes	Yes	No	No
Model checking	No	No	No	No