BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 16 2008, pages 1821–1822
doi:10.1093/bioinformatics/btn317

Genetics and population analysis

ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth

Brian W. Lambert[1,*], Joseph D. Terwilliger[2,3] and Kenneth M. Weiss[1,*]

[1]Department of Anthropology, Penn State University, University Park, PA, [2]Department of Psychiatry, Genetics and Development, Columbia Genome Center, Columbia University and [3]Division of Medical Genetics, New York State Psychiatric Institute, New York, NY, USA

Received on April 18, 2008; revised and accepted on June 17, 2008

Advance Access publication June 19, 2008

Associate Editor: Martin Bishop

ABSTRACT

Many important problems in biology involve complex traits affected by multiple coding or regulatory parts of the genome. How well the underlying genetic architecture can be inferred by statistical methods such as mapping and association studies are active research areas. *ForSim* is a flexible forward evolutionary simulation tool for exploring the consequences of evolution by phenotype, whereby demographic, chance, behavioral and selective effects mold genetic architecture. Simulation is useful for exploring these issues as well as the choice of study design inferential methods.

Contact: bwl1@psu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

We have today, at best, a generic understanding of the distributional characteristics even of the most important parameters that generate the genetic architecture of complex traits, such as allelic effects on phenotypes, or their effects on evolutionary fitness. *ForSim* can simulate a wide range of user-constructed plausible models to test consistency with empirical data, and help optimize study designs to infer the underlying architecture from samples.

The most important simulation tool in the last 20 years for evolutionary processes as well as genetic inference has been backward, or coalescent, simulation (Hudson and Kaplan, 1988). Coalescent approaches are fast but limited in terms of the complexity of scenarios they can simulate, such as selection, complex genetic architecture, penetrance, environmental effects, recombination, population structure, and they typically assume evolutionary equilibrium.

A more realistic approach is forward simulation. A starting ancestral population is simulated forward in time from some starting time to the present (Hey, 2004; Hoggart *et al.*, 2007; Peng *et al.*, 2007). Forward simulation substantially raises the demand for memory and CPU time, but hardware is rapidly improving and can handle an even more flexible simulation of the evolution of complex traits, when many genes contribute. Comparisons can be made between neutrally evolving and selected traits as well as the effects of demographic complexity, and changing environments. It is the core aspects of genetic architecture that are deeply conserved in nature: the number of genes, nature of gene pathways, etc. While these are stable, what controls trait diversification are allelic effects that are laid upon this underlying structure.

*To whom correspondence should be addressed.

2 METHODS

ForSim simulates evolution naturally. Almost every level of this process and many aspects of the genetic architecture of traits and populations can be controlled by the user (Table 1). *ForSim* allows users to define the number, lengths and location of genes and chromosomes, the genetic contributions and interactions, environmental effects and other conditions. Multigenic traits and simple networks of genetic contributions can be specified. An arbitrary number of populations, their local environments, phenotype-based natural selection, gene flow and mate choice, as well as time-variable changes in these processes, can all be defined by the user. Phenotypes are determined by user-specified quantitative effects of genes and environments, and relative fitness is based on user-specified criteria.

A mutation is stochastically assigned a phenotypic effect governed by a user-specified probability and it can be neutral, lethal, negative or positive. The effect amount is determined from a user-parameterized gamma distribution. The effect of a gene is comprised of the sum of effects of the polymorphisms its haplotypes contain. Users define the number of phenotypes and the contributions of genes to these phenotypes through traditional algebraic syntax such that 'PhenA = (G1 + G2)*G3^2' meaning the trait is the product of the contribution of gene G3^2 and the sum of genes G1 and G2. Genes may contribute to any number of phenotypes, and when a gene is specified to affect more than one trait, the traits become correlated. For example, defining PhenB = (G4 + G2)*G3^2 will cause PhenA and PhenB to be correlated. The user can specify random universal or family-specific environmental contributions that can be specific to each phenotype and can vary over populations and time to simulate epidemiologically or ecologically important effects.

Natural selection can be modeled as a deterministic truncation or probabilistic process, based on a user-specified function relating fitness of the metric phenotype of an individual to its distance from the population mean or some user-specified optimal phenotype value (which may be changed during the simulation). Phenotype means and variances can be traced through generations in the post-run output data, from which multivariate phenotype and genotype analysis can easily be done.

Probabilities of mate choice within and between populations are user specifiable and can be phenotype based to test selective migration and assortative mating. Specified population size is reached and maintained by logistic growth scaled by adjusting expected family size (sibship size is stochastic).

Table 1. Current features of *ForSim*

* Specifiable duration of simulation (in synchronous generations)
* Single or multiple replicate simulations
* Multiple univariate or multivariate phenotypes
* Phenotype-based mate choice and migration between populations
* Multiple genes and chromosomes, of arbitrary number and length
* Diploid populations
* Mating with or without replacement
* Stochastic family size distribution and logistic population growth
* Phenotype-based natural selection of user-specified type
* Point mutation and recombination that can be sex-specific
* Stochastically determined mutation-specific allelic effects
* Environmental contribution to each individual's phenotypes

User-scriptable specifications or in-run changes:

* Mating, phenotype determination, migration, and selection
* Multiple populations with hierarchical (cladistic) splitting, and split-times, phenotype-based or random mate-choice and gene flow, environmental effects, population size, and natural selection regimes
* Pleiotropic genetic effects
* Multilocus phenotypes including networks and gene interaction
* Flexible natural selection criteria with stochastic fitness I/O
* Data saved at user-specified generations with real-time plotting of conditions and user-specification of the data to be saved
* Output suitable for standard human genetic analysis and mapping software (e.g. LINKAGE format)
* Output in rapid and easily parsed XML as well as plain text
* Graphical output in browser-readable SVG, and other formats

All these parameters are defined in a user-authored input file, with a block-structured syntax with entities (populations, chromosomes, etc.) named and their attributes defined. *ForSim* is distributed with several example input files as well as a syntax-highlighting EMACS mode to assist in editing and scripting. At the end of the simulation, complete genotype and phenotype data are saved for each individual in each population, LINKAGE-format pedigrees for each individual, along with quantitatively ascertained cases and controls are saved for SNP association tests, and parent-offspring trios for LD and haplotype analysis. The entire history of every allele can be saved so that analytic approaches can be developed on the basis of full information.

ForSim is not a model-fitting or empirical hypothesis-testing tool, but can test models to see if they generate plausible results compared to empirical data. For example, using well-known human genetic parameters for mutation, recombination, effective population size and trait prevalence, *ForSim* produces data consistent with the observed and/or theoretically expected values of nucleotide diversity, segregating SNPs in samples, sibling relative risk, linkage disequilibrium structure and so on as these are observed in human data (e.g. www.hapmap.org). Reasonable selection scenarios generate correspondingly reduced nucleotide and haplotype heterozygosity, increased haplotype length, etc. Statistical tests of results that include traits that were, and were not, subject to selection can determine whether the specified selection, migration and so on, can be detected in the results. Examples are in the *ForSim* user manual and Supplementary Material.

Runtime depends on the complexity of simulated conditions, especially the population size(s) and number of genes and generations. On a 2.8 Ghz Intel Pentium D processor with 2 GB RAM, a *ForSim* simulation of a population of 10 000 for 10 000 generations (roughly the age and effective population size of the human species), for a chromosome of 10 Mb containing 10 neutrally evolving genes (five of 20 Kb and five of 50 Kb), a mutation of $2.5 * 10^{-8}$/nt and one $cM = 10^6$ base pairs, with normally distributed environmental phenotypic noise, takes 28 min. Time scales up, sometimes non-linearly, with increase in these values, because SNP sojourn times are non-linear with respect to population size until equilibrium is established, and all current and previously fixed SNPs must be tracked. Small test runs can guide users in specifying the simplest, fastest running conditions that can adequately address their objectives.

3 DISCUSSION

ForSim can simulate simple controlled situations effectively and efficiently, but it can also simulate complex situations, that require more memory and runtime. To our knowledge, no other current programs are as general and flexible. The program is written in portable C++, with optional wrapper scripts for the preparation and presentation of data. These wrapper scripts, written in Ruby, automate graphical output via R (www.r-project.org), as well as haplotype and SNP-tagging analysis using Haploview and haplotype alignments via ClustalW. *ForSim* source code and user manual are available free upon request (www.anthro.psu.edu/weiss_lab/research.shtml#ForSim), with agreement to an open source non-commercialization license. It builds and runs under Linux/Unix/MacOS and should run under Windows with CygWin or MinGW. Updates and errata will be posted and sent to registered users.

ACKNOWLEDGEMENTS

Funding: This work was supported by NIH grant number MH63749.

Conflict of Interest: none declared.

REFERENCES

Hey,J. (2004) FPG–a computer program for forward population genetic simulation. Web document: available at http://lifesci.rutgers.edu/heylab/HeylabSoftware.htm#FPG (last accessed date July 14, 2008).

Hoggart,C.J. *et al.* (2007) Sequence-level population simulations over large genomic regions. *Genetics*, **177**, 1725–1731.

Hudson,R.R. and Kaplan,N.L. (1988) The coalescent process in models with selection and recombination. *Genetics*, **120**, 831–840.

Peng,B. *et al.* (2007) Forward-time simulations of human populations with complex diseases. *PLoS Genet.*, **3**, e47.