

## PROGRAM NOTE

**KERNELPOP, a spatially explicit population genetic simulation engine**

ALLAN E. STRAND\* and JAMES M. NIEHAUS†

\*Department of Biology, Grice Marine Lab, College of Charleston, Charleston, SC 29424, USA,

†Department of Computer Science, North Carolina State University, Raleigh, NC 27695-8206, USA

**Abstract**

Individual-based, spatially explicit models provide a mechanism to understand distributions of individuals on the landscape; however, few models have been coupled with population genetics. The primary benefits of such a combination is to assess performance of population-genetic estimators in realistic situations. KERNELPOP represents a flexible framework to implement almost any arbitrary population-genetic and demographic model in a spatially explicit context using a variety of dispersal kernels. Estimates of type I error associated with genome scans in metapopulations are provided as an illustration of this software's utility.

*Keywords:* dispersal, genome-scan, individual-based, neutral model

*Received 24 January 2007; revision accepted 13 April 2007*

The utility of population-genetic markers for inferring demographic parameters in natural populations is well established. Examples include estimation of population size (Roman & Palumbi 2003), mating systems (Ritland & Jain 1981), and dispersal characteristics (Hardesty *et al.* 2006). Each makes simplifying assumptions about nature to ensure mathematical tractability. In some cases, population-genetic estimators may be robust to violations of assumptions, but this is not a guarantee. For example, the shortcomings of Wright's (1951) relationship between  $F_{ST}$  and  $N_e m$  are well established (Whitlock & McCauley 1999). Theoreticians who implement new estimators typically use numerical simulations to test their performance. Unfortunately, no single simulation can examine the robustness of an estimator in idiosyncratic conditions associated with many empirical studies. A framework that simplified such testing would allow empiricists to assess the ability of a particular technique to estimate the quantity of interest in their system. Genome scans provide one example of why such a detailed understanding may be important. This approach uses  $F_{ST}$  as a means to assess selection upon loci associated with markers (Beaumont & Nichols 1996). Essentially, these studies perform outlier analyses upon  $F_{ST}$  estimates from large numbers of loci. Those loci that exhibit extreme values of among-population

divergence are considered to be evolving non-neutrally. Because of the resources required to further examine loci identified by this method, simulations may provide a way to fine-tune the false discovery rate in such studies.

Such a simulation framework could be easily applied to other problems in population biology. For example, there are an increasing number of efforts to use mechanistic models of dispersal to predict the distribution of individuals as well as genotypes on landscapes (Galindo *et al.* 2006). First, these efforts require mechanistic understanding of dispersal obtained from either physical models of the medium in which dispersal occurs or a well-resolved phenomenological description of dispersal. Second, some form of spatially explicit simulation approach is required to model stochasticity associated with environmentally coupled model predictions.

Because of their inherent flexibility, individual, or agent-based models represent a powerful approach to developing a general-purpose population genetic simulation engine. Indeed, EASYPop 3 (Balloux 2001) has been used extensively to model evolution of neutral markers. Previously we developed a software, RMETASIM, that included much of the functionality of EASYPop, but allows more flexibility in specifying within-population demography and among-population dispersal models (Strand 2002; Strand & Niehaus 2006). Here we describe an individual-based simulation environment that includes the functionality of RMETASIM, but adds spatial realism and flexible definitions of dispersal

Correspondence: Allan E. Strand, Fax: 843-953-9199; E-mail: strand@cofc.edu

models via probability density functions (PDF). A table in the documentation shipped with `KERNELPOP` summarizes the differences and similarities between `KERNELPOP`, `RMETASIM`, and `EASYPOP`.

The software introduced here is designed to be a flexible, spatially explicit, discrete-time simulation engine to allow biologists developing and using population-genetic estimators to test their performance and robustness under realistic conditions. Although it implements potentially complex simulations, `KERNELPOP` is intended to be relatively easy to use and portable to most major operating systems.

Because individual-based models are slow, we implemented the actual simulation engine in C++ for speed and the user interface in R (R Development Core Team 2005) for accessibility. Furthermore, because R is a freely available statistical language that runs on multiple platforms, this implementation results in a simulation environment for all major operating systems. `KERNELPOP` has been added to the Comprehensive R Archive Network (<http://cran.r-project.org>), under the GPL. Once R is downloaded on the user's computer, `KERNELPOP` and its online documentation can be downloaded and installed automatically using the R function `install.packages("KERNELPOP")`.

A landscape in `KERNELPOP` is rectangular in extent with arbitrary length and width. This landscape is composed of (at least potentially) suitable habitats and unsuitable intervening areas. Habitats are rectangular and can have any length and width up to the full extent of the landscape. Habitats cannot overlap; continuous spatial distributions can be simulated either by constructing the entire landscape out of a single habitat or by placing multiple habitats adjacent to one another. Each habitat has its own local characteristics including demography, carrying capacity, and extinction rate. These characteristics can also change during the course of a simulation either using the mechanisms originally built into the C++ library (see Strand (2002)) or by using functions implemented in R.

Within population dynamics in `KERNELPOP` are functionally similar to `RMETASIM`'s. Briefly, population dynamics are determined by three vital stage-transition matrices. The first matrix,  $S$ , determines the rate of survival and growth from one life stage to another. The second matrix,  $R$ , determines the mean reproductive output of individuals within each life stage drawn from a Poisson distribution. The final matrix,  $M$ , encodes the probability that male gametes (e.g. those in pollen) come from a particular life stage. Together these matrices should accommodate any arbitrary stage-based demographic model.

Each individual has its own coordinates in the overall landscape. In addition, depending upon the habitat in which they are located, they may have different vital rates defined by demographic stage matrices that describe their respective habitats.

Population size is regulated by implementing a carrying capacity within each habitat,  $K_h$ . If the population size exceeds  $K_h$ ,  $N_h - K_h$  individuals are chosen randomly to be killed, where  $N_h$  is the current census size in habitat  $h$ .

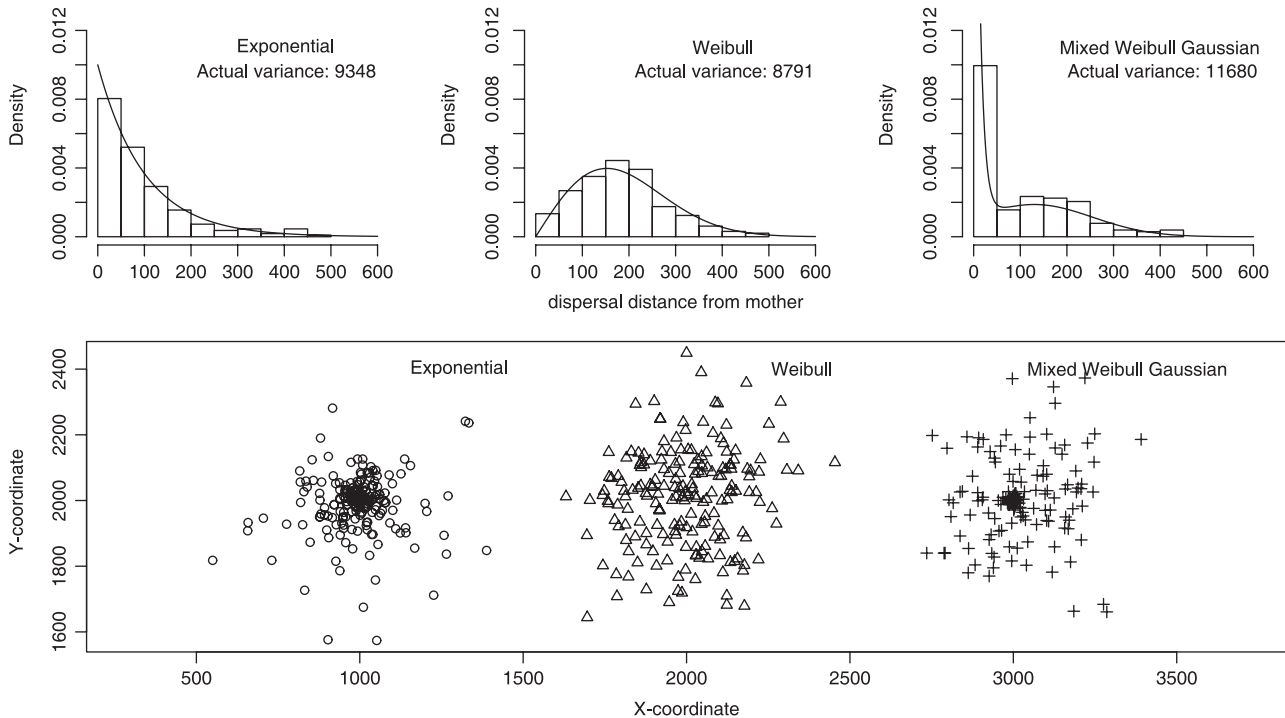
Pollen (e.g. male gamete) dispersal is determined by four types of parameters: i) selfing rate, ii) proportion of total pollen production allocated to each demographic class, iii) the dispersal kernel associated with each demographic class, and iv) the parameters of that dispersal kernel. For each mother  $m(i)$  in demographic stage  $i$  with coordinates  $x_m, y_m$ , the probability that she receives pollen from a father,  $f(j)$ , in demographic stage  $j$  with coordinates  $x_f, y_f$  can be summarized as

$$\Pr(m(i)_{x_m, y_m}, f(j)_{x_f, y_f}, \lambda, \sigma, \gamma) = \begin{cases} \text{if } m(i) = f(j) & \begin{cases} \text{if } \text{Un}[0,1] \leq s & 1 \\ & > s & \gamma_j g_j(m(i)_{x_m, y_m}, f(j)_{x_f, y_f}, \lambda_j, \sigma_j) \end{cases} \\ m(i) \neq f(j) & \gamma_j g_j(m(i)_{x_m, y_m}, f(j)_{x_f, y_f}, \lambda_j, \sigma_j) \end{cases} \quad (1)$$

In this equation,  $\text{Un}[0,1]$  is a random draw from a uniform distribution on the interval  $(0,1)$ . The male fecundity for individuals in each stage is specified by  $\gamma$ , a vector of length equal to the number of demographic stages in the landscape ( $k$ ). The PDF describing the pollen dispersal kernel associated with males in demographic stage  $j$  is given by  $g_j()$ , and the scale and shape parameters associated with each stage's dispersal kernel are given by  $\lambda$  and  $\sigma$ , respectively. For each mother, the probability of mating with all possible males is determined. Fertilization occurs when one male is picked at random from this set with a probability determined by equation 1.

Propagule (e.g. seed) dispersal is similar to pollen dispersal in that the fecundity of a demographic stage, a dispersal kernel associated with that stage, and parameters specific to that stage determine the distribution of propagule rain. Each seed-dispersal event is composed of a direction, chosen from a uniform distribution on the interval  $(0, 2\pi)$ , and a distance determined by random draw from the PDF and its parameters associated with the mother's demographic stage. Currently, three probability distributions are available to describe both pollen and seed dispersal: exponential, Weibull, and a mixture between Weibull and left-truncated Gaussian. Figure 1 illustrates for seeds how any combination of these distributions can be used within a single population during the course of a simulation.

`KERNELPOP` implements biparental inheritance of unlinked diploid loci and maternal inheritance of haploid loci. Currently, loci are selectively neutral. Three mutation models are implemented: infinite allele model, strict step-wise model, and a simple 1-parameter model of DNA nucleotide substitution.



**Fig. 1** Example of different seed-dispersal kernels in different demographic stages. Three mothers and their offspring locations are modelled. Each mother represents a demographic stage with a unique dispersal kernel. Circles correspond to an exponential kernel with mean 100. Triangles correspond to a Weibull distribution with shape 2 and scale 215.9. Crosses correspond to a mixed distribution between a Weibull with scale 10 and shape 1 and a left-truncated Gaussian with mean 130 and standard deviation 125. The mixing parameter results in 50% of dispersal events drawn from the Weibull distribution. These parameters were chosen to result in similar variance in dispersal distance among mothers. The actual distribution of distances along with the intended probability density functions are indicated in the panels above each mother's location on the map.

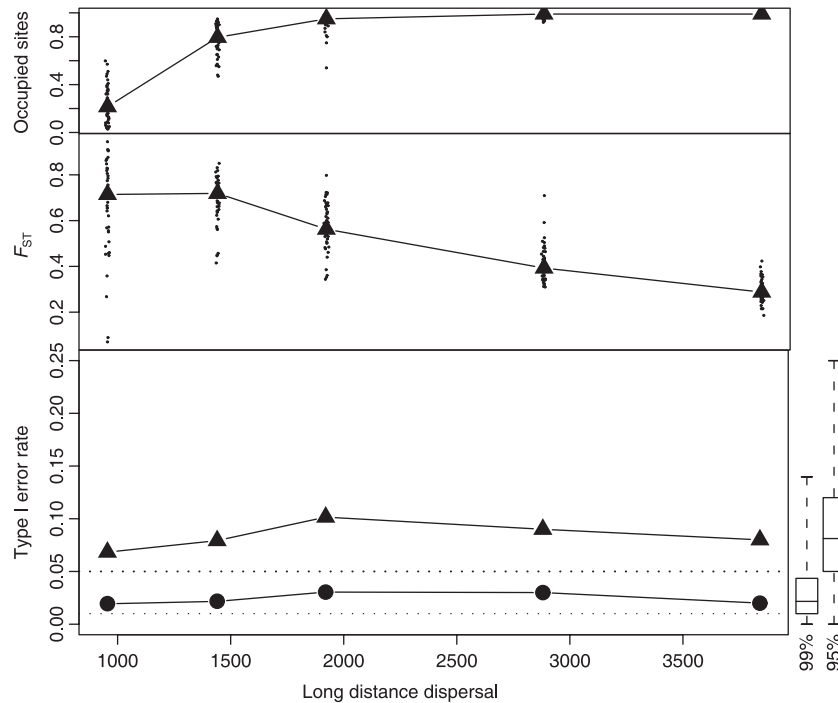
Because *KERNELPOP* is intended primarily as a simulation engine, it performs limited analyses of genotypic data. *KERNELPOP*, however, leverages the power of other R packages, in particular *ade4* and *ape* (also available from <http://cran.r-project.org>), to perform additional analyses such as estimation of  $4N_e\mu$  and  $\phi_{ST}$  from simulated data sets. To facilitate more detailed analysis of simulation results, *KERNELPOP* outputs a number of formats used by analytical programs such as *GENEPOP* and *ARLEQUIN*.

Beaumont & Nichols (1996) developed a method to detect outliers in the distribution of  $F_{ST}$  values using coalescent simulation of a large number of neutral loci and subpopulations. The results of these simulations are used to develop a heterozygosity-corrected confidence envelope for  $F_{ST}$ ; loci that fall outside of this envelope might be under selection. Beaumont & Nichols (1996) analysed the performance of their technique under several simulated population scenarios, including a simple invasion model (Slatkin 1993), and showed that their approach had good performance.

We implemented a simple test of the type I error rate of Beaumont & Nichols's (1996) approach under a more realistic invasion model with extinction–recolonization dynamics.

In each replicate, we simulated 100 randomly located and sized demes upon a square landscape. The demography within each deme was that of an annual plant with a persistent local seedbank. Carrying capacity was proportional to the area of each deme; on average, population sizes were ~1000 per deme. We ran sets of replicated simulations where we varied the amount of long-distance dispersal and per-deme extinction. Long-distance dispersal (LDD) is defined as the 95% ile of all dispersal distances. The genetic simulation included 60 neutral loci evolving under a strict stepwise mutation model. To mimic a range expansion, each simulation was initialized with only the centremost deme containing 1000 individuals. Each locus was initialized with five alleles at equal frequencies. Three extinction rates and five long-distance dispersal rates were simulated. For each of the 15 combinations, 20 replicate simulations were run for 150 years.

At the end of each replicate simulation, 24 demes were chosen at random. These samples were then subjected to the analyses implemented by the program *FDIST2* (slightly modified for batch operation). The parameters for the *FDIST2* analyses were 200 demes and 20 000 simple sequence



**Fig. 2** Results of simulations described in text. Differences in extinction rates among replicates are not indicated. Top panel: relationship between 95%ile for seed dispersal distance (LDD) and the proportion of sites occupied observed after 150 years. Solid line connects median values in all panels. Middle panel: LDD vs.  $F_{ST}$ . Points represent mean  $F_{ST}$  across 60 loci observed in each replicate simulation. Declining slope was confirmed with linear regression ( $P < 0.0001$ ,  $R^2 = 0.61$ ). Bottom panel: relationship between LDD and type I error,  $\alpha$ , in a  $F_{ST}$ -based genome scan. Triangles and circles correspond to 95% and 99% criteria for the genome scan, respectively. Initial rise in 95% criterion type I error with increasing LDD had weak support using linear regression (for LDD  $< 2000$ ,  $P < 0.01$ ,  $R^2 = 0.07$ ). Bottom and top dotted lines denote type I error rate equal to 0.01 and 0.05, respectively. Individual replicates have been removed for visual clarity. Boxplots on the lower right margin summarize the distribution of type I error in all replicates for each criterion. Note that the majority of replicates using the 99% criterion fall below a type I error rate of 0.05 and a majority of those based on a 95% criterion exceed this same level.

repeat loci as recommended by the authors. Loci identified as outliers using criteria based both on 95% and 99% confidence intervals were tallied and per-replicate estimates of type I error were recorded.

We used LDD and extinction rate as independent variables in our analysis of type I error. Only LDD had a detectable effect on type I error for the 95% criterion, and this effect diminished at higher rates of dispersal. At the same time, LDD had a clear effect on metapopulation characteristics like the proportion of sites occupied and  $F_{ST}$ . This result further supports the robustness to variation in demographic rates of the Beaumont & Nichols (1996) approach. Overall type I error rates were elevated, however. Mean type I error using 95% criteria across all simulated replicates was  $0.088 \pm 0.003$  and 75% of simulated replicates yielded type I error rates greater than 5% whereas nearly 25% of the type I errors using the 99% criterion were greater than 5% (Fig. 2). Proximally, this result suggests that a more strict criterion to avoid false discovery is appropriate when applying the *FDIST2* approach in expanding dynamic metapopulations.

More generally, we argue that a tool like *KERNELPOP* provides a flexible framework that would enable an assessment of this rate (or other population-genetic statistic) in an arbitrary biological system. This analysis took two days on a 2GHz dual core Pentium processor. Given the substantial investment in resources required to further evaluate loci identified in genome scans, the relatively short time required to investigate the statistical behaviour of an estimator seems well justified.

### Acknowledgements

Erik Sotka commented on an earlier version of this manuscript. This project was supported in part by NIH-BRIN#RR16461-01.

### References

- Balloux F (2001) *EASYPop* (version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, **92**, 301–302.
- Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **263**, 1619–1626.

- Galindo HM, Olson D, Palumbi S (2006) Seascape genetics: a coupled oceanographic-genetic model predicts population structure of Caribbean corals. *Current Biology*, **16**, 1622–1626.
- Hardesty B, Hubbell SP, Bermingham E (2006) Genetic evidence of frequent long-distance recruitment in a vertebrate-dispersed tree. *Ecology Letters*, **9**, 516–525.
- R Development Core Team (2005) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Ritland K, Jain SK (1981) A model for the estimation of outcrossing rate and gene frequencies using  $n$  independent loci. *Heredity*, **47**, 35–52.
- Roman J, Palumbi SR (2003) Whales before whaling in the North Atlantic. *Science*, **301**, 508–510.
- Slatkin M (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, **47**, 264–279.
- Strand AE (2002) METASIM 1.0: an individual-based environment for simulating population genetics of complex population dynamics. *Molecular Ecology Notes*, **2**, 373–376.
- Strand A, Niehaus J (2006) *RMETASIM: An Individual-Based Population Genetic Simulation Environment (version 1.1)*. Available at URL: [cran.r-project.org](http://cran.r-project.org)
- Whitlock M, McCauley D (1999) Indirect measures of gene flow and migration:  $F_{ST} \neq 1/(Nm + 1)$ . *Heredity*, **82**, 117–125.
- Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.