SI Text

Model methodological details

Our simulation is an object oriented program consisting of two populations (P_0 and P_1). Each population contains a number of individuals, each of which has a genome made up of two chromosomes. These are represented as two binary arrays of length equal to the number of adaptive loci being simulated (N). The state of each position represents one of two alleles: one adapted to each population's habitat. Each locus additively contributes a set amount to the fitness (W) of an individual. Mosquitoes are selected for breeding based on their fitness relative to others in the population. The two populations are connected by migration (m_0 and m_1), described in more detail below.

Each time step of the simulation represents a generation and involves creating a new group of individuals from the current one. The size of the population after reproduction is defined by a linear function which increases from a minimum size (S_x) to a maximum size $100 \cdot S_x$ after five time steps. The population size is then decreased back to the minimum size after the tenth step.

The major steps for a generation are: 1) calculate the fitness of all individuals and 2) perform a "breeding" step the number of times necessary to create the offspring population.

Calculating fitness

Individuals are selected for breeding with a probability proportional to their fitness, so at the beginning of each time step the fitness of all individuals is calculated. For each locus, if the allele state is homozygous maladapted to the environment in which the individual is located, then that locus contributes 1 to the fitness of the organism. As the fitness of the whole organism is calculated additively, if a given individual is homozygous maladapted in all loci to the environment its fitness is N.

If the individual is homozygous adapted to the environment, then each locus contributes $1 + s$ to the individual's fitness, where s is the adaptive value of the locus (this value is the same for all loci but is varied between simulations). Thus the maximum fitness attainable is $N(1 + s)$.

If an individual is a heterozygote at a particular locus, then the fitness contribution of that locus to the summation is $1 + (s/2)$. The fitness calculation for a single individual may be written as follows

$$W = H(1 + s) + h(1 + \frac{s}{2}) + (N - [H + h]) \tag{1}$$

where W is the calculated fitness, H is the number of homozygous adapted loci, h is the number of heterozygous loci.

Breeding

The breeding function creates a new individual from two parents. The first parent is selected with probability proportional to its fitness from the population which is being stepped into the next generation. A second individual is chosen from the other population with probability m, where m is the migration rate. If a migrant is selected, the second parent is chosen randomly from the other population. With probability $1-m$ the second parent is from the same population, and is selected with probability proportional to its fitness. In either case the parents must be two distinct individuals. An individual may be selected for breeding multiple times or not at all during a particular breeding cycle.

One chromosome is chosen at random from each parent and combined in a new individual. At this point recombination may occur. Each gap between loci has the same probability (r) of being the site of a chromosomal break-point. A draw is made for each gap, in order from locus 0 to locus i_N. If a break occurs at locus i_j, then all alleles from i_j through i_N are swapped from one chromosome to another. Thus multiple break-points and compound recombination patterns are possible.

If a break point is chosen in an interval between loci that are within an inversion in that particular individual, then the recombination event is not carried out. This is the case if there is only one or if both chromosomes carry the inversion.

Chromosomal inversion

At a time t_i, an inversion is introduced to one chromosome of a single individual in the peripheral population. The number of consecutive loci marked as being in an inversion is a simulation parameter, n_i. All the loci within the inversion are adapted to the peripheral environment and there can be no crossing over in the gaps between them as noted above.

If the frequency of the inversion becomes zero after t_i, the simulation is stopped since inversions are not introduced at any other time and so the frequency cannot change from zero.

Experimental parameter sets and parallelization

Individual simulations were run on 11 worker cpus, tasked by a master program which sampled from the parameter space. The master program used Latin Hypercube Sampling (LHS) [1] to select parameter sets. To do this, each parameter range is divided into uniform subranges ("slices") numbering the overall number of simulations. To obtain a parameter set, we randomly chose one of the slices for each parameter without replacement. The actual value of each parameter sent to the simulation was the mid-point of the slice. This process is repeated until no slices remain and the requisite number of simulations are complete.

Model validation

We tested the model's response to several single parameters and compared simulation results against expectations from ref. 2 before varying them in tandem. We individually tested migration rate, recombination rate and number of loci captured by the inversion. In all cases, the model's behavior conformed to expectations, supporting the assumption that it captures the essential elements of the ecotypification hypothesis.

As an example, Figure 3 shows the results of some tests on migration rate. We used a peripheral population size of 500 individuals with two loci under selection. These individuals may breed with migrants, which are maladapted to the peripheral environment (fixed for the disadvantageous allele at all loci). An inversion is introduced encompassing both loci. We found that higher migration rates lead to a faster increase in inversion frequency as expected [2]: the selective advantage of not having recombination with maladapted (migrant) individuals is higher when there are more migrants. Inversion frequencies stabilize at varying levels depending on the degree of influx of wild-type individuals.

Factors affecting final inversion frequency

We performed a second analysis restricted to those simulations which ended in inversion polymorphism. With this test we created a conditional model that has a simple interpretation [3]. Those runs with $0 < p_f < 1$ yielded a data set of 385 simulations with inversion frequencies in the peripheral population that were approximately normally distributed ($\bar{p_f} = 0.57$, $\mathrm{SD}(p_f) = 0.17$; one sample Kolmogorov-Smirnov test, $D = 0.05$, $p = 0.25$). The inversion frequencies in the core population were significantly lower ($\bar{p_f} = 0.0006$, SD(p_f) = 0.001; Mann-Whitney U test, $W \sim 0$, $p < 0.001$) and not normally distributed (one sample Kolmogorov-Smirnov test, $D \sim 1$, $p < 0.001$) compared to those in the peripheral population.

Linear regression of the predictors on p_f for the peripheral population showed m_1, s and n_i to be significant determinants of final inversion frequency (Table 4).

References

[1] Blower, S & Dowlatabadi, H. (1994) Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example. *Int Stat Rev* 2:229 − 243.

[2] Kirkpatrick, M & Barton, N. (2006) Chromosome inversions, local adaptation and speciation. *Genetics* 173:419 − 434.

[3] Welsh, A, Cunningham, R, Donnelly, C, & Lindenmayer, D. (1996) Modelling the abundance of rare species: statistical models for counts with many zeros. *Ecol Model* 88:297 − 308.