Genomic estimates of effective population size with overlapping generations

Reference Number: 6348

May 21, 2014

Purpose of the visit

The effective population is a crucial parameter that allows us to estimate the chances of extinction of a population. Several methods have been developed to estimate its value, from temporal methods based on the increase in inbreeding from samples taken at different times to the use of molecular markers in one sample to calculate it from the excess of heterozygosity or linkage disequilibrium. This latter method is based on the work by Sved [4] and can also be used to infer the ancestral effective population [3]. However, Sved assumed a population with discrete, non-overlapping generations. When this method is blindly used in populations with overlapping generations, it has been shown that the estimates of Ne from LD are biased [2]. The aim of this visit was to develop an extension of Sved's method for populations with overlapping generations to obtain a formula as simple as possible that gives us a less biased Ne.

Description of the work carried out during the visit

Analytical formulae were developed to obtain population size in terms of linkage disequilibrium in cases with overlapping generations. We extended the classical and well known formula obtained by Sved [4]

$$E\left(r_{eq}^2\right) \approx \frac{1}{1+4Nc},$$

to account for overlapping generations. In this formula, N is the population size, c is the recombination rate and $E\left(r_{eq}^2\right)$ is the expected linkage disequilibrium at equilibrium. Sved's analysis is based on $Q_t = E[r^2]$, where Q_t is the probability that two gametes chosen at random are identical by descent for two genes A and B at map distance c.

In the first scenario studied, each cohort at time$t$ is produced with contributions from up to two cohorts back (i.e., from t-1 and t-2). The cohort at t-1 and the cohort at t-2 have different contributions to the cohort at time t, being those contributions w_1 and w_2. We define also $Q_{i,j}$ as the probability of two

gametes chosen at random of cohorts i and j to be identical by descent in the same way. Following these definitions we can write

$$\begin{aligned} Q_t &= w_1^2 (1-c)^2 \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) Q_{t-1} \right] \\ &+ w_2^2 (1-c)^2 \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) Q_{t-2} \right] \\ &+ 2 w_1 w_2 (1-c)^2 Q_{t-1,t-2}, \end{aligned}$$

where Q_i denotes ibd probabilities within individuals of cohort i and $Q_{i,j}$ ibd probabilities between cohorts i and j. After some algebra, we can obtain the convergence at equilibrium as

$$Q_\infty = E\left(r_{eq}^2\right) = \frac{\frac{1}{2N}(1-c)^2 \left[w_1^2 + w_2^2 + \frac{2 w_1^2 w_2 (1-c)^2}{1 - w_2 (1-c)^2} \right]}{1 - \left(1 - \frac{1}{2N}\right)(1-c)^2 \left[w_1^2 + w_2^2 + \frac{2 w_1^2 w_2 (1-c)^2}{1 - w_2 (1-c)^2} \right]}$$

In order to test this formula, we obtained empirical values of Q_∞ at equilibrium by using a Monte Carlo algorithm. We used the multinomial distributions that govern the changes on haplotypic frequenies. Then, every generation we perform the following steps:

1. Sample the number of recombinant gametes from a binomial distribution with parameters N and c. Alternatively, we used a Poisson distribution with parameter $\lambda = Nc$ for small c values.

2. For recombinant gametes, we sampled the realized amount for each haplotype from a multinomial distribution with parameters N and the haplotypic frequencies at gametic equilibrium

3. For non-recombinant gametes, we sampled from a multinomial distribution with the current haplotypic frequencies.

4. Sum the amounts of steps 2 and 3, update haplotypic frequencies and calculate the new $Q = d^2/(p_A p_B (1-p_A)(1-p_B))$.

Repeating steps 1 to 4, we reached an equilibrium. We performed the algorithm over several replicates, and those replicates that went to fixation for any allelic frequency where replaced with another unfixed replicate at random.

Description of the main results obtained

We tested the above formula against empirical results from simulations. The effective population size was calculated by using the method developed by Felsenstein [1] for overlapping scenarios. From Felsenstein's effective population size, we calculated the $E\left(r_{eq}^2\right)$ by using Sved's approach. This procedure does not

require linkage disequibrium to estimate the effective population size, but we used it to test our estimates which are indeed based on linkage disequibrium.

Table 1 show the results of our analytical approach, the empirical results obtained from the multinomial based algorithm and the Felsenstein's based method. We used a big cohort size $N = 100000$ and a small recombination rate $c = 10^{-5}$ and different ratios $w_1 : w_2$.

$w_1 : w_2$	1:0	0.7:0.3	0.5:0.5	0.3:0.7
Analytical	0.2000	0.1612	0.1428	0.1282
Empirical	0.1960	0.155	0.133	0.12
Felsenstein's	0.2000	0.1589	0.1379	0.1206

Table 1: r^2 obtained with several patterns of overlapping generations. Cohort size $N = 10^5$ and $c = 10^{-5}$. Monte carlo results at generation 1 million.

The empirical algorithm seems to agree pretty well with Felsenstein's approach, while our analytical seems to be slightly biased. An upwards bias in the estimation of r^2_{eq} corresponds to a downwards bias in the estimation of N_e.

Future collaboration with host institution

We expect to continue collaboration to extend the analytical formula for an arbitrary number of cohorts, to perform tests with other scenarios varying N and c and to prepare a paper containing all these results.

Projected publications / articles resulting or to result from the grant

Our objective is to submit our research for publication within the coming months. This collaboration has been possible thanks to the funding obtained through Congenomics and this will be acknowledged in the manuscript.

References

[1] Felsenstein J (1971) Genetics 68:581-597

[2] Robinson and Moyer (2013) Evol Apps 6:290-302

[3] Tenesa A et al. (2007) Genome Res. 17: 520-526

[4] Sved JA (1971) Theor. Pop. Biol. 2:125-141