multiNe: the many facets of effective population size.

Rebecca B Harris[1], Christina Ewers-Saucedo[2], Lucy Li[3], Julia A Palacios[4], and George Shirreff[5]

[1]Department of Biology, University of Washington, Seattle, WA 98122
[2]University of California at Davis, Davis, CA
[3,5]Department of Infectious Disease, Imperial College London, London, UK
[4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138

1 Abstract

2 Introduction

Effective population size (N_e) is a key parameter in understanding the evolutionary trajectory of a population. Wright (1931) formally described N_e as the number of breeding individuals in an ideal population that show the same rate of genetic drift as the population being studied (Wright, 1931). Since then, it has become the primary parameter of concern for evolution, ecology, and conservation biology studies. Estimates of N_e encompass the effects of both demographic and genetic processes in finite populations, thus effectively quantifying the rate and timing of molecular evolution (Caballero, 1994).

Despite its wide popularity, the calculation of N_e in natural populations remains challenging and multiple methods have been developed to estimate N_e indirectly from genetic data. Here, we distinguish between two fundamentally different ways to estimate N_e...

On the one hand, coalescent theory provides a means to estimate Ne over evolutionary times using either summary statistics, such as the number of segregating sites **?**, number of alleles (Ewens 1972), heterozygosity (Kimmel *et al*, 1998), variance of the number of microsatellite repeats (Kimmel *et al*, 1998), or using the shape of the genealogy itself (Kingman, 1982). The parameter calculated is theta, the mutation-rate scaled effective population size. Knowing the mutation rate allows us then to convert theta into N_e. Several of these theta calculators are available in the R package pegas (Paradis, 2010).

Functionality

Phy2Sky

Coalescent theory states that the rate at which the lineages in a phylogeny coalesce is inversely proportional to the effective population size. This relationship allows the estimation of changes in N_e over time to be based on changes in the branching pattern of the genealogy, which can be visualized in skyline plots. Here, we extend the coalescent method found in the ape package (Paradis *et al*, 2004) by allowing for phylogenies with heterochronous dated tips. This will be of particular value in epidemiological studies where there is precise information about the sample collection time.

The `Phy2Sky` function takes as its first argument a `multiPhylo` object, the typical output of widely used Bayesian phylogenetic programs. We provide a function, `burninfrac`, to discard a user-defined proportion of the raw posterior distribution as burn-in. Given a set of trees, `Phy2Sky` will output the end times and N_e estimates of each coalescent interval for each tree. Alternatively, results from multiple trees can be merged and `Phy2Sky` will generate a table with all possible events across all trees. Users may then plot the median N_e skyline with the 95% confidence intervals.

Branch-lengths can be converted from substitutions per site to time units by defining the clock rate in the scaling option of `Phy2Sky`. Certain users may want to fix the skyline to reflect specific sampling dates. We provide the tools to extract sampling dates from tip names and use these in the skyline output plot.

Skyline data are usually presented by a step function, consisting of only vertical and horizontal lines, implying that for a given period the best estimate for the effective population size is a certain value. We provide numerous functions to manipulate the plotting of these stepwise graphs. For some analyses, it may be preferred that points on the graph are joined by straight sloping lines, implying that the effective population size during this period changes in a regular fashion. We also allow users to combine neighbouring time intervals until a minimum time interval cutoff is met. [Fig #? Have figure to show these types of graphs?] On the other hand, Ne can be estimated by observing and quantifying deviations from exceptions of infinitely large populations.

Inferring Ne from multiple genealogies

Here, we implement a method to infer Ne from multiple genealogies (assuming independence) using integrated nested Laplace approximation (INLA).

New coalescent simulators

(Palacios & Minin, 2013)

More functions

(Hein *et al*, 2005)

Implementation of existing methods

The following estimators calculate N_e of a population over the past few generations, whereas coalescent estimators generally integrate N_e over the past N_e to $4 * Ne$ generations, depending of the mode of inheritance of the genetic marker employed.

Temporal effective population size

Temporal N_e, also called variance N_e, is based on the premise that finite populations experience genetic drift. This drift results in changes in allele frequencies from one generation to the next which are inversely proportional to N_e (cite Nei and Tajima 1981). To implement this method, genotypic data from a minimum of one locus sampled at two or more generations (assumed to be non-overlapping) is needed. While other programs have implemented this calculation, this is the first time to our knowledge that temporal N_e has been implemented in an R package. Given an a set of loci sampled from known generations, the `varNE` function will calculate the point estimate for Ne for each possible comparison. Confidence intervals are obtained using jackknifing. However, for many systems, obtaining samples across multiple generations may be unrealistic.

Linkage Disequilibrium

To accommodate studies where only one time sample is available, we implement a method to calculate N_e based on the magnitude of linkage disequilibrium (LD). Generally, smaller populations are expected to give rise to higher LD than larger populations. Waples (2006) demonstrated that low frequency alleles will bias estimates of N_e. Therefore, we specify the lowest allele frequency to be retained in the dataset. While Waples and Do (2008) suggest critical values between 0.05 and 0.01, this value is dataset dependent and we leave it to the user to determine the proper cut-off (Waples & Do, 2010, 2008). The `LDNe` function requires users to characterize their system as mating randomly, or monogamously. In most cases, random mating may be more appropriate, as it refers to the lifetime mating pattern.

2.1 Existing Ne estimators in R

NB package is a multiple sample maximum likelihood estimator(**?**).

References

Caballero A (1994) Developments in the prediction of effective population size. *Heredity*, **73 (Pt 6)**, 657–679.

Hein J, Schierup MH, Carsten W (2005) *Gene genealogies, variation, and evolution: a primer in coalescent theory*. Oxford University Press, Oxford, 1st edn..

Kimmel M, Chakraborty R, King JP, Bamshad MJ, Watkins WS, Jorde LB (1998) Signatures of population expansion in microsattelite repeat data. *Genetics*, **148**, 1921–1930.

Kingman JFC (1982) On the genealogy of large populations. *Journal of Applied Probability*, **19**, 27–43.

Palacios JA, Minin VN (2013) Gaussian Process-Based Bayesian Nonparametric Inference of Population Size Trajectories from Gene Genealogies. *Biometrics*, **69**, 8–18.

Paradis E (2010) Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics*, **26**, 419–420.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Waples RS, Do C (2008) LDNE: A program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, **8**, 753–756.

Waples RS, Do C (2010) Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: A largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, **3**, 244–262.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.