MLNe: Simulating and Estimating Effective Size and Migration Rate from Temporal Changes in Allele Frequencies

Jinliang Wang

Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom

Left running head: J Wang

Right running head: Infer effective size and migration rate

Key words: effective size, migration rate, inbreeding, genetic drift, markers, allele frequency, temporal samples

Corresponding author:

Jinliang Wang

Institute of Zoology

Regent's Park

London NW1 4RY

United Kingdom

Tel: 0044 20 74496620

Fax: 0044 20 75862870

Email: jinliang.wang@ioz.ac.uk

Abstract

In studies of molecular ecology, conservation biology and evolutionary biology, the current or recent effective size (N_e) of a population is frequently estimated from the marker genotype data of two or more temporally spaced samples of individuals taken from the population. Despite the developments of numerous Bayesian, likelihood and moment estimators, only a couple of them can use both temporally and spatially spaced samples of individuals to estimate jointly the effective size (N_e) of and the migration rate (m) into a population. In this note I describe new implementations of these joint estimators of N_e and m in software MLNe which runs on multiple platforms (Windows, Mac, Linux) with or without a graphical user interface (GUI), has an integrated simulation module to simulate genotype data for investigating the impacts of various factors (such as sample size and sampling interval) on estimation precision and accuracy, exploits both Message Passing Interface (MPI) and openMP for parallel computations using multiple cores and nodes to speed up analysis. The program does not require data pre-processing and accepts multiple formats of a file of original genotype data and a file of parameters as input. The GUI facilitates data and parameter inputs and produces publication-quality output graphs, while the non-GUI version of software is convenient for batch analysis of multiple datasets as in simulations. MLNe will help advance the analysis of temporal genetic marker data for estimating N_e of and m between populations, which are important parameters that will help biologists for the conservation management of natural and managed populations. MLNe can be downloaded free from the website http://www.zsl.org/science/research/software/.

Introduction

Effective population size, N_e, is a key concept in population genetics introduced by Wright (1931) and developed by many others, mainly Crow and Kimura (1970). It is defined as the size of an idealized Wright–Fisher population (Fisher 1930; Wright 1931), which would give the same rate of inbreeding or genetic drift in allele frequencies as in the population in question (Crow and Kimura 1970). It not only determines the genetic stochasticity of a population, but also affects the efficacy of all systematic evolutionary forces (such as migration and selection) acting on a population. Therefore, this pivotal parameter has found wide applications in studies in evolutionary biology, molecular ecology, conservation biology and selective breeding of domesticated species. In the conservation management of endangered species, for example, measuring N_e and its trend over time helps to monitor the genetic status and its changes (Schwartz et al. 2007) and thus to inform effective management of the species.

Many methods have been developed to estimate the current, historical and ancient effective population size from marker genotype data (Wang 2005). For conservation, the most relevant is usually the current or recent effective size, which largely explains the genetic variation within and between the current populations of a species and is informative about the conservation planning and management of the species. Among the various methods proposed for estimating the current N_e (Luikart et al. 2010; Wang et al. 2016), temporal methods are the first developed (Krimbas and Tsakas 1971), well tested and widely applied methods. They are so named because they measure and use the allele frequency changes between temporally spaced samples as information for the strength of genetic drift and thus the average N_e of the population during the temporal interval. The methods were subsequently further developed by many others (e.g. Nei and Tajima 1981; Pollak 1983; Waples 1989), and were extended to use more powerful statistical techniques such as likelihood or Bayesian methods (Williamson and Slatkin 1999; Anderson et al. 2000; Wang 2001; Berthier et al. 2002; Beaumont 2003; Laval et al. 2003; Hui and Burt 2015). These sophisticated methods are more flexible (e.g. in allowing for any number of temporal samples), can use marker information more efficiently (e.g. in handling rare alleles), and thus usually provide more accurate N_e estimates than the simple moment methods (Wang 2001).

The temporal approach was also extended to analyse both temporally and spatially separated samples from a metapopulation for estimates of both effective size of and migration rate between populations in both moment and likelihood frameworks (Wang and Whitlock 2003). The changes in allele frequencies of a population over time due to genetic drift (finite N_e) and migration are distinguishable because drift-caused changes are purely at random while migration-caused changes are systematic (directional). Although moment (e.g. Do et al. 2014) or likelihood (e.g. Hui and Burt 2015) estimators of N_e from temporal genotype data were implemented in software available for empirical data analysis, they all

assume a single isolated population. The software MLNe implements multiple moment estimators (Nei and Tajima 1981; Wang and Whitlock 2003) and likelihood methods (Wang 2001; Wang and Whitlock 2003) for estimating N_e for a single isolated population and for estimating N_e and migration rate (m) jointly for a metapopulation. The software was published online for free download, but no document describing the software was published.

In this computer note, I describe MLNe in a new version (2.0) much improved over the original version (1.0). I will focus on introducing the new features of the new version, leaving the methodological details in the original publications (Wang 2001; Wang and Whitlock 2003).

Graphical User Interface (GUI)

A graphical user interface in Visual Basic is added to MLNe version 2.0. It helps Windows users of MLNe to input data, to set parameter values, to conduct data analysis, and to visualize analysis results in tables and publication-quality graphs. The GUI for setting up a new project for analysing an empirical dataset is shown in Figure 1, and an example profile log-likelihood curve drawn by MLNe is shown in Figure 2. Similarly, it also draws profile log-likelihood curves for migration rate if the migration model is opted for. For both isolation and migration models, it also draws stacked bar charts for estimated allele frequencies at each locus of each temporal sample. These and other analysis results can also be viewed in tables.

Although GUI has many advantages, it also has some disadvantages. For example, GUI is inconvenient for making analysis of multiple datasets in a batch mode. In a simulation study, however, usually many replicate datasets need to be analysed and it is desirable to call MLNe directly from a program for conducting the analyses. In the new version of MLNe, the code for the computational kernel is in Fortran 2003 and has been compiled for platforms linux, Mac and Windows 10. It can be run in the x-terminal of linux and Mac, and in MS-DOS of Windows. When ran in non-GUI mode, MLNe reads a genotype data file and a corresponding parameter file, runs the analysis, and outputs the results in a few files.

A hurdle in applying the previous version (1.0) of MLNe is that it requires the counts of each allele at each locus in each temporal sample as the data. This means the raw genotype data must be pre-processed in two steps. First, the unique alleles observed for each locus across temporal samples must be identified from the genotype data. Second, the copies of each unique allele at each locus of each temporal sample must be counted. Both processes need some coding except for an extremely small dataset. The new version 2.0 completes the data pre-processing automatically. It accepts the raw genotype data in three optional formats (i.e. a genotype is encoded by two integer numbers, by integers 0, 1 or 2 indicating the number of reference alleles for diallelic markers, and by the GenePop format) and an input file for analysis parameters.

Simulations

MLNe has a built-in simulation module that can be used to simulate genotype data with user-defined parameters such as the isolation or migration models, true value of N_e, number of individuals sampled at each time point, the sampling interval for temporal samples, and number and polymorphisms of marker loci. On obtaining values of these parameters through the GUI (Figure 3), MLNe initiates an individual based forward simulation in the Wright-Fisher model to generate genotype data and outputs them to a file. It also generates a corresponding input file of parameters for analysing the genotype data. The two files are then used by MLNe to get estimates of N_e (and m for the migration model).

In addition to investigating factors affecting the power and accuracy of the temporal approaches, simulations are also valuable to optimizing the experimental design. It is useful, for example, to determine the suitable sample intensities (number of markers, number and the temporal interval of samples, and sample sizes) to yield accurate N_e estimates. Before initiating a project, one can use simulations to generate data in conditions similar to those of the conceived project, and to analyse the simulated data to get a feel of the estimation power and accuracy. For this same reason, simulations are also valuable for training and educational purposes.

The simulation module is capable of simulating genotype data for hundreds of thousands of individuals at hundreds of thousands of loci (see an example below). However, it is worth noting that the simulation assumes free recombination among loci, which is apparently violated when many genomic markers are simulated for any species. In the presence of linkage, temporal methods could yield estimated 95% confidence intervals that are too conservative (i.e. too narrow), although they are expected to yield good point estimates of N_e regardless of the linkage among markers.

Flexible models and multiple methods

Methods under two population genetics models, isolation and migration, are implemented in MLNe. The isolation model is the one assumed in nearly all temporal methods since the seminal work of Krimbas and Tsakas (1971). In this model, a population is assumed to be isolated without immigration from other populations during the period between the first and last sample taken from it. Therefore, the changes in allele frequency at a neutral marker locus in the relatively short sampling period (thus mutations are negligible) would come solely from genetic drift and reflect the average N_e of the population during the period. Using both a moment estimator (Nei and Tajima 1981) and a likelihood estimator (Wang 2001), MLNe analyses the data and yields N_e estimates with 95% confidence interval estimates as demonstrated in Figure 2.

The migration model removes the restrictive assumption of a single isolated population and considers a metapopulation consisting of a small (focal) population and an

infinitely large source population providing immigrants into the focal population (Wang and Whitlock 2003). The methods are robust to violations of the assumption and can be applied approximately to a finite source population composing of one or more small subpopulations (Wang and Whitlock 2003). MLNe implements a moment estimator and a likelihood estimator developed in Wang and Whitlock (2003) to estimate the N_e of and the immigration rate (m) into the focal population jointly.

For both models, MLNe allows for and uses any number of temporal samples in the estimation. For more than 2 samples, the average N_e and m over the entire sampling period are estimated directly by the likelihood method, while N_e and m for each sampling period are estimated by the moment estimator and their harmonic and arithmetic means over multiple sampling periods are reported respectively.

Parallel computation

When genomic data of many markers are used to estimate N_e of a large (say, N_e in tens of thousands) population, the likelihood estimator becomes computationally demanding and may take a long time to complete an analysis. This is especially so for a metapopulation in the migration model, where both N_e and m are inferred jointly. To speed up the analysis, MLNe uses both Message Passing Interface (MPI) and openMP to make parallel runs of the data with multiple processes and multiple threads per process. Both numbers of MPI processes and openMP threads per process are determined by a user according to the data size and computer capacity. While MPI processes use multiple nodes of a computer with distributed memory or multiple cores of a computer with shared memory, openMP threads use cores of a single node with shared memory. Roughly, the computational efficiency depends on the total number of parallel threads, which is the product of the number of MPI processes and the number of openMP threads per MPI process.

To demonstrate the computational speedup of applying MPI and openMP and the capacity of MLNe to handle large genomic data sampled from a large population, I simulated data from an isolated large population of N_e=60000 using the simulation module. Two small samples separated by 10 generations, each containing only 50 individuals, were taken from the population and each sampled individual was genotyped at 100000 SNP loci. The data were analysed by MLNe with a maximal N_e set at 100000, using 2,3,4,5,6 nodes of a linux cluster. Each node of the cluster has two 20-core Intel Xeon Gold 6248 2.5GHz processors with 192 gigabytes of 2933MHz DDR4 RAM, and each physical core has two logical cores by hyperthreading. Therefore, each node has 2×20×2=80 logical cores, which are used as openMP threads in MLNe. The total number of parallel threads used in analysing the data is thus 80n, where n (=2,3,4,5,6) is the number of nodes or the number of MPI processes. The time taken for analysing the data using different number of nodes (threads) is compared in Figure 4.

The example in Figure 4 shows that (1) MLNe has the capacity to handle genomic data in estimating N_e of an extremely large population and (2) running time decreases with an increase in the total number of parallel threads used in an analysis. The speedup by parallelization does not increase linearly with an increasing number of threads, perhaps due to the communication cost among threads (openMP) and processes (MPI).

Conclusion

MLNe is a powerful software implementing multiple population genetics models (migration and isolation) and multiple statistical methods of each model for estimating N_e (and m for th emigration model) from any number of temporally spaced samples of individuals, with each individual genotyped at either a few microsatellite loci or many thousands of SNPs. It can be run on multiple computer platforms with or without a GUI, and has a built-in simulation module for generating simulated temporal genotype data. It uses both MPI and openMP for parallel computation to use multiple computer nodes of distributed memory and multiple cores within a node of shared memory. As a result, it can handle genomic data for estimating the N_e and migration rate of very large populations. It could hopefully become a valuable tool for conservation genetics research and teaching.

References

Anderson EC, Williamson EG, Thompson EA. 2000. Monte Carlo evaluation of the likelihood for Ne from temporally spaced samples. *Genetics*. 156:2109–2118.

Beaumont MA. 2003. Estimation of population growth or decline in genetically monitored populations. *Genetics*. 164:1139–1160.

Berthier P, Beaumont MA, Cornuet JM, Luikart G. 2002. Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetics*. 160:741–751.

Crow JF, Kimura M. 1970. *An introduction to population genetics theory*. New York: Harper and Row.

Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. 2014. NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Mol Ecol Resour*. *14*:209-214.

Fisher RA. 1930. *The genetical theory of natural selection*. Oxford: Oxford University Press.

Hui TYJ, Burt A. 2015. Estimating effective population size from temporally spaced samples with a novel, efficient maximum-likelihood algorithm. *Genetics*. 200:285-293.

Krimbas CB, Tsakas S. 1971. The genetics of *Dacus oleae* V. Changes of esterase polymorphism in a natural population following insecticide control: selection or drift? *Evolution*. 25:454–460.

Laval G, SanCristobal M, Chevalet C. 2003. Maximum-likehhood and Markov chain Monte Carlo approaches to estimate inbreeding and effective size from allele frequency changes. *Genetics*. 164:1189–1204.

Luikart G, Ryman N, Tallmon DA, Schwartz MK, Allendorf FW. 2010. Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conser Genet.* 11:355-373.

Nei M, Tajima F. 1981. Genetic drift and estimation of effective population-size. *Genetics*. 98: 625–640.

Pollak E. 1983. A new method for estimating the effective population size from allele frequency changes. *Genetics*. 104:531–548.

Schwartz MK, Luikart G, Waples RS. 2007. Genetic monitoring as a promising tool for conservation and management. *Trends Ecol Evol***.** 22:25-33.

Waples RS. 1989. A generalised approach for estimating effective population size from temporal changes in allele frequency. *Genetics*. 121:379–391.

Wang J. 2001. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genet Res*. 78:243–257.

Wang J. 2005. Estimation of effective population sizes from data on genetic markers. *Philos Trans R Soc Lond B Biol Sci.* 360:1395-1409.

Wang J, Whitlock MC. 2003. Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*. 163:429–446.

Wang J, Santiago E, Caballero A. 2016. Prediction and estimation of effective population size. *Heredity*. 117:193-206.

Williamson EG, Slatkin M. 1999. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*. 152:755–761.

Wright S. 1931. Evolution in Mendelian populations. *Genetics.* 16:97–159.

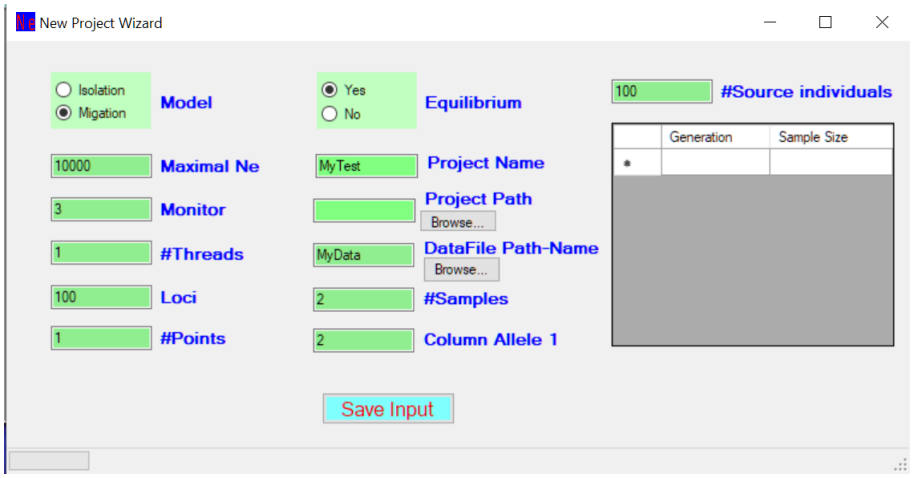

Figure 1. New project wizard.

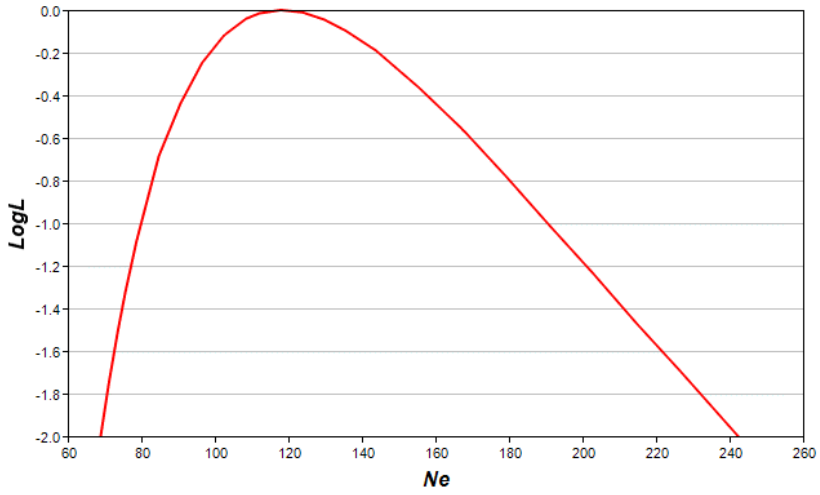

Figure 2. Profile log-likelihood curve generated by MLNe for a simulated data set. It shows the maximum likelihood estimate of N_e is 121, and the 95% confidence interval is {65, 241}.

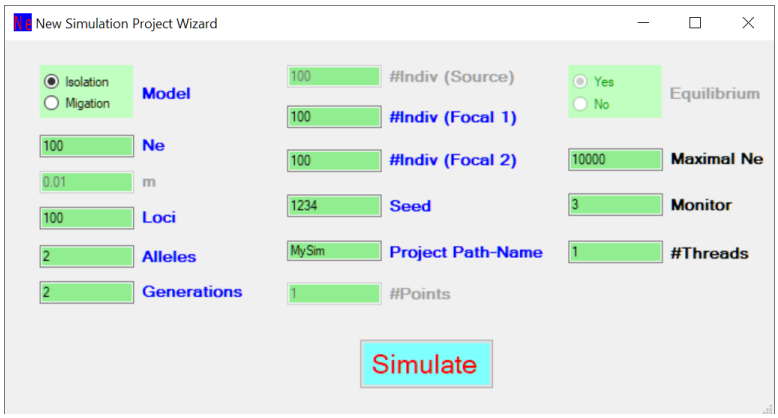

Figure 3. New simulation project wizard.

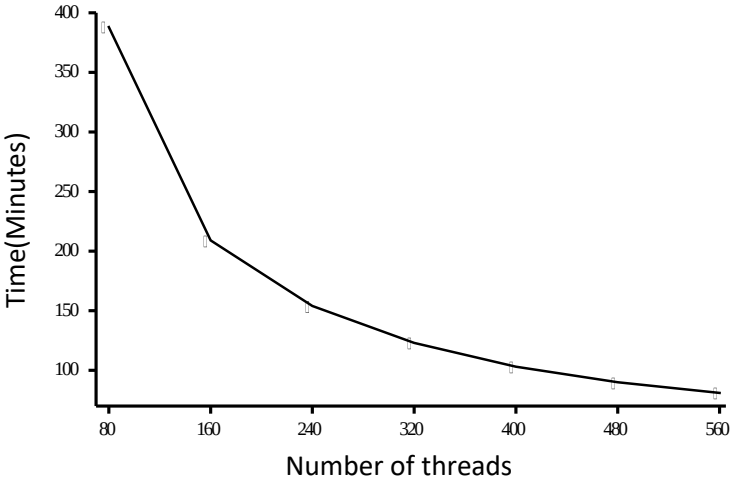

Figure 4. Running time (minutes) as a function of the number of parallel threads (*x* axis) for an example dataset.