Lietuvos matematikos rinkinys
Proc. of the Lithuanian Mathematical Society, Ser. A
DOI: 10.15388/LMR.A.2016.10

ISSN 0132-2818
Vol. **57**, 2016
pages 53–58

A comparative analysis of mathematical methods for effective population size estimation

Alma Molyte, Alina Urnikyte, Vaidutis Kučinskas

Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University
Santariškių str. 2, LT-08661 Vilnius, Lithuania
E-mail: alma.molyte@mf.vu.lt, alina.urnikyte@mf.vu.lt, vaidutis.kucinskas@mf.vu.lt

Abstract. In this paper two mathematical methods, McEvoy and Mezzavilla–Ghirotto, were investigated, they are devoted to determine the effective population size. Comparative analysis of two methods, for estimating effective population size, was performed.

Keywords: effective population size, genotyping, Mezzavilla–Ghirotto method, McEvoy method.

1 Introduction

The effective population size (N_e) is one of the most important population genetic parameters revealing the historical demographic features of the population such as bottleneck and growth rates [1, 10]. By definition, N_e is a measure of the number of independent breeding individuals in an ideal population and is much lower than the actual census size N [1, 10]. Different genetic models based on genetic markers are used to estimate N_e [8]. The appropriate mathematical methods are developed for different genetic models to obtain information from genetic marker data to estimate N_e. Using the same data, different methods could yield considerably different estimates of N_e because of time-scales of each method. In this study, we compare N_e estimates obtained by two different genetic models (McEvoy and Mezzavilla–Ghirotto, MG) based on linkage disequilibrium (the non-random association between genetic loci, LD) between densely spaced single nucleotide polymorphisms (SNPs) data [2, 3]. Our interest was to determine the accuracy of each method that uses different mathematical approaches to estimate N_e. For the analysis, we used Illumina 770K HumanOmniExpress-12v1.0 array data of 295 unrelated individuals of the Lithuanian population.

2 Material and methods

2.1 Sample and genotyping

The data set consisted of 295 Lithuanian population samples. Genomic DNA was extracted from whole venous blood using either the phenol-chloroform extraction method or the automated DNA (extraction platform TECAN Freedom EVO (TECAN Group Ltd., Männedorf, Switzerland), based on paramagnetic particle method. DNA concentration and quality were measured by NanoDropR ND-1000 spectrophotometer (NanoDrop Technologies Inc., US).

SNP genotyping of 295 samples was performed with Illumina 770K HumanOmni Express-12v1.1 array (Illumina, San Diego, CA, USA) at the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University, Lithuania using the standard Illumina Infinium®HD Assay Ultra protocol recommended by the manufacturer (Catalog # WG-901-4005). Genotyping data quality control was performed according to the standard recommendations by the manufacturer. Individuals with call rate $< 98\%$ and Standard Deviation (SD) of Log R ratio > 0.3 were excluded from further analysis. GenomeStudio v2011.1 program (Illumina, USA) was used to distinguish the genotypes from the sample and to export the data in PED/MAP format.

For the Ne estimation, PLINK data file (binary format) was obtained using PLINK v1.07 program [6]. Individuals or SNPs with $> 10\%$ missing data, minor allele frequency (MAF) < 0.01 and Hardy–Weinberg equilibrium (HWE) test P-value of less than 10^{-4} were excluded. After quality control 1 of 296 individuals removed for low genotyping (MIND > 0.1). After frequency and genotyping pruning, there left 568040 SNPs.

This study is a part of the LITGEN project, which was approved by the Vilnius Regional Research Ethics Committee 235 No. 158200-05-329-79, date: 2011-05-03. The written informed consent was received from all participants of the study.

2.2 N_e estimation

The effective population size (N_e) was calculated with *NeON* package, which is developed for the free R environment. Data analysis consists of three steps:

1. Computation correlation coefficient of linkage disequilibrium between markers.

Linkage disequilibrium is the non-random association between alleles of two or more loci and can arise from marker proximity or from selection bias.

Two estimators of LD are computed [9]:

- D raw difference in frequency between the number of **AB** pairs and the expected number: $D = \rho_{AB} - \rho_A\rho_B$,
- r_{LD} correlation coefficient between the markers $r_{LD} = \frac{-D}{\sqrt{\rho_A \cdot \rho_a \cdot \rho_B \cdot \rho_b}}$, where

  ρ_A – is defined as the observed probability of allele **A** for marker 1,

  $\rho_a = 1 - \rho_A$ – is defined as the observed probability of allele **a** for marker 1,

  ρ_B – is defined as the observed probability of allele **B** for marker 2,

  $\rho_b = 1 - \rho_B$ – is defined as the observed probability of allele **b** for marker 2,

  ρ_{AB} – is defined as the probability of the marker allele pair **AB**.

  The default parameters are the genotyping rate higher than 98%, a rate of individual missing data lower than 10%, a window of 500 kilobases and 9999 SNPs [4].

2. Each pair of markers is binned into recombination distance categories. The second step is estimated for each category.

The estimate of effective population size (N_e) is based on the recombination or genetic distance between SNPs. For each pair of SNPs separated by < 0.25 centiMorgan (cM), for each population separately, we described LD levels by the correlation (r_{LD}) and squared correlation (r_{LD}^2) in genotype frequencies.

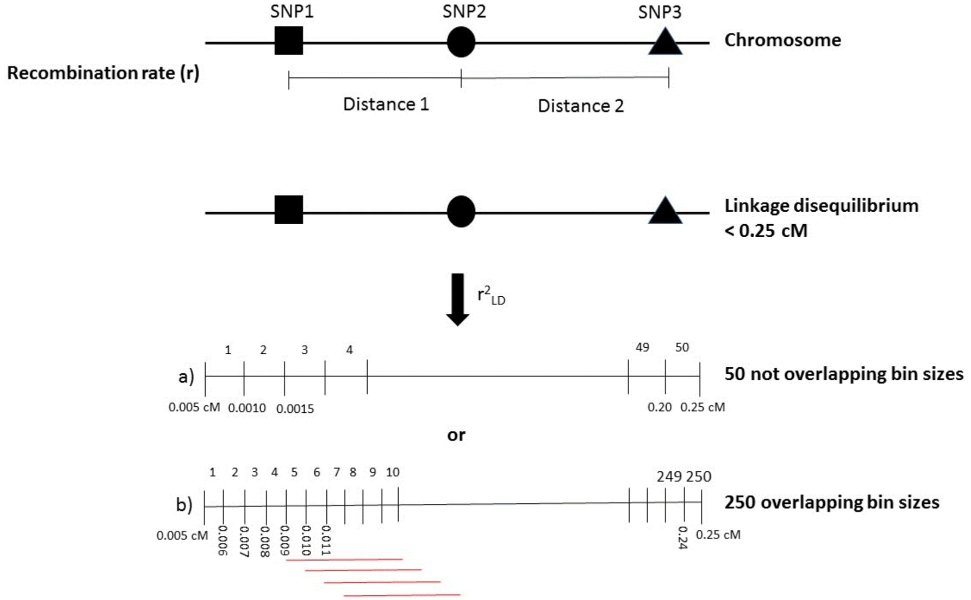

Fig. 1. Genetic models representation to estimate the effective population size based on linkage disequilibrium on different conditions: (a) McEvoy method; (b) Mezzavilla–Ghirotto method.

The effective population size was estimated by formula:

$$N_e \approx 1/(4 \cdot c) \cdot \left[\left(1/r_{LD}^2\right) - 2\right],$$

where c is distance between genetic markers in morgans. r_{LD} can be positive or negative. All individual r_{LD}^2 were adjusted: $r_{LD}^2 - (1/n)$, where n is the sample size, prior to the calculation of N_e [7].

It is created by a number of recombination distances with incremental upper boundaries of 0.005 cM up to 0.25 cM and calculates the r_{LD}^2 for each pair of markers in each recombination distance category. To do this, we applied two different methods. One of them is McEvoy, which considers 50 not overlapping bins (Fig. 1(a)). The other method is the Mezzavilla–Ghirotto (MG), which is with 250 overlapping bins with a step of 0.001 cM from 0.005 to 0.25 cM (Fig. 1(b)).

This is the main difference of these two methods, McEvoy and Mezzavilla–Ghirotto (Fig. 1(a), (b)).

The effective population size value calculated in each bin corresponds to the effective population size at a specific moment in the past, i.e. $1/(2c)$ generation ago [2], with c calculated as the mean value in each recombination distance category [5]. We obtain a data frame with the values of the effective population size and the corresponding time in the past, for each bin and chromosome.

If a population is constant in size or grows linearly, then N_e is approximately true $E(r_{LD}^2) \approx 1/(2 + 4N_e(t)c)$ for the N_e, t generations ago, where $t = 1/(2c)$ [2, 3].

3. Obtain the demographic function of your population and the long-term N_e *with its confidence interval.*

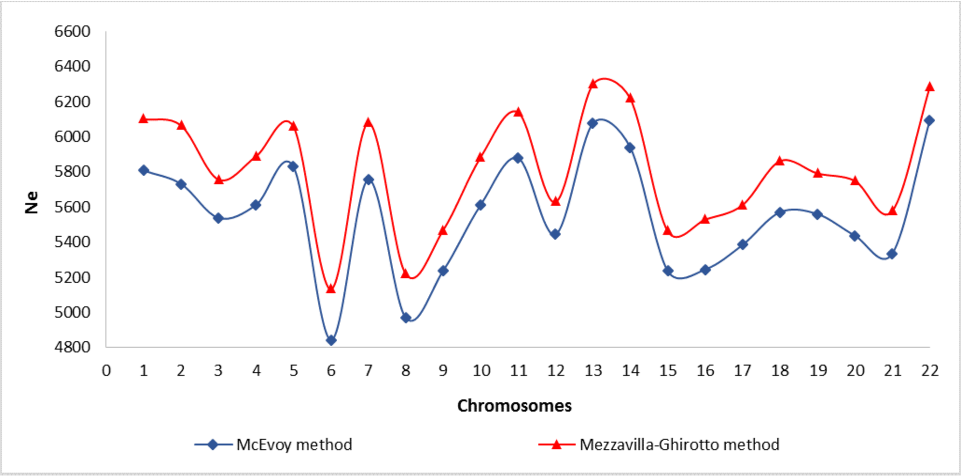

Fig. 2. Effective population size calculated for each chromosome.

The long term N_e is calculated as the harmonic mean of effective population size along the generations in the past, i.e. in each recombination distance category [4].

To determinate the statistical significance of differences in our N_e, we estimated its ninety-five percent confidence intervals. The confidence interval of the long-term N_e is calculated using each chromosome as a replica.

The median of effective population size is calculated from all autosomes estimated for each temporal point, depending on the Mezzavilla–Ghirotto method with 250 temporal points and the McEvoy method with 50 temporal points.

3 Results

The aim of this study was to discover differences between the McEvoy and Mezzavilla–Ghirotto methods employed for N_e calculation. We employed both methods to estimate the effective population size of the Lithuanian population. The effective population size for each chromosome performing 250 or 50 permutations, depending on the method used, with each chromosome was calculated as well (Fig. 2). The data was not normally distributed and therefore the nonparametric Mann–Whitney U test was used to establish statistically significant differences between the groups of N_e. P value of < 0.05 was considered statistically significant.

The results of the data analysis show that the effective population size obtained by the McEvoy method is 5481 and the 95% confidence interval (CI) for the population median is [5384; 5563], while the result obtained by the Mezzavilla–Ghirotto method is 5722, CI [5679; 5755]. The differences between the N_e median values among the different methods were statistically significant ($p < 0.001$). However, the effective population size obtained by these methods, aren't statistically significant with all chromosomes (Table 1, Fig. 2).

The mean rank of the Effective population size calculated using the McEvoy method is 2783.27 and applying the Mezzavilla–Ghirotto method is 3352.69. The

Table 1. Effective population size for each method and their statistical significance.

Chromosome	McEvoy	Mezzavilla–Ghirotto	p-value
1	5807.5	6101.0	0.1075
2	5731.5	6063.0	0.1569
3	5538.0	5756.0	0.0414*
4	5611.5	5888.0	0.0745
5	5829.0	6057.0	0.0372*
6	4843.5	5132.5	2.116e−08**
7	5756.0	6079.0	0.0397*
8	4970.5	5219.5	0.0005**
9	5237.0	5467.0	0.0107*
10	5611.0	5882.5	0.0883
11	5879.0	6139.5	0.1484
12	5444.0	5628.0	0.0253*
13	6076.0	6297.0	0.1228
14	5939.0	6216.5	0.0263*
15	5236.5	5462.5	0.1656
16	5245.0	5529.0	0.0191*
17	5385.0	5611.5	0.0275*
18	5571.0	5861.5	0.0016*
19	5560.5	5791.5	0.0120*
20	5438.0	5749.0	0.0532
21	5335.5	5577.5	0.0679
22	6092.0	6281.5	0.0233*

Significant p-value ($p < 0.05$) with * and p-value ($p < 0.001$) with **.

higher N_e values are with the higher mean rank, i.e. when the Mezzavilla–Ghirotto method was used.

4 Conclusions

In this paper the McEvoy and Mezzavilla–Ghirotto methods for effective population size estimation have been investigated. The effective population size estimation can be performed by both methods, but it is necessary to take into consideration the recombination distance categories that create different methods. The method with more recombination distance categories gives the bigger N_e value and contrariwise with the less recombination distance categories gives smaller N_e value. We conclude that with the purpose to have a fine-scale recombination distance categories for N_e estimation the Mezzavilla–Ghirotto method should be used, if otherwise – the McEvoy method.

References

[1] B. Charlesworth. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.*, **10**:195–200, 2009.

[2] B.J. Hayes, P.M. Visscher, H.C. McPartlan, and M.E. Goddard. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.*, **13**(4):635–643, 2003.

[3] B.P. McEvoy, J.E. Powell, M.E. Goddard and P.M. Visscher. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res.*, **21**821–829, 2011.

[4] M. Mezzavilla and S. Ghirotto. Package for the effective population size estimation. *Package*, versija 1.0. 2013, 10 pp.

[5] M. Mezzavilla and S. Ghirotto. Neon: an *R* package to estimate human effective population size and divergence time from patterns of linkage disequilibrium between SNPS. *J. Comput. Sci. Syst. Biol.*, **8**(1):37–44, 2015.

[6] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira and et al. PLINK: a toolset for whole-genome association and population-based. *Am. J. Human Genet.*, **81**, 2007.

[7] A. Tenesa, P. Navarro, BJ. Hayes, D.L. Duffy, G.M. Clarke, M.E. Goddard and P.M. Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome Res.*, **17**:520–526, 2007.

[8] J. Wang. Estimation of effective population sizes from data on genetic markers. *Phil. Trans. Royal Soc., Ser. B*, **360**:1395–1409, 2005.

[9] G. Warnes, with contributions from G. Gorjanc, F. Leisch, M. Man. *Population Genetics*, 2013.

[10] S. Wright. Evolution in Mendelian populations. *Genetics*, **16**:97–159, 1931.

REZIUMĖ

Matematinių metodų lyginamoji analizė efektyvaus populiacijos dydžio nustatymui

A. Molytė, A. Urnikytė, V. Kučinskas

Darbe nagrinėjami du matematiniai metodai, t. y. McEvoy ir Mezzavilla–Ghirotto, kurie skirti nustatyti efektyvųjį populiacijos dydį. Atlikta gautų rezultatų lyginamoji analizė.

Raktiniai žodžiai: efektyvus populiacijos dydis, genotipavimas, Mezzavilla–Ghirotto metodas, McEvoy metodas.