We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,300

Open access books available

172,000

International authors and editors

190M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

The Sensitiveness of Expected Heterozygosity and Allelic Richness Estimates for Analyzing Population Genetic Diversity

María Eugenia Barrandeguy and María Victoria García

Abstract

Genetic diversity comprises the total of genetic variability contained in a population and it represents the fundamental component of changes since it determines the microevolutionary potential of populations. There are several measures for quantifying the genetic diversity, most notably measures based on heterozygosity and measures based on allelic richness, i.e. the expected number of alleles in populations of same size. These measures differ in their theoretical background and, in consequence, they differ in their ecological and evolutionary interpretations. Therefore, in the present chapter these measures of genetic diversity were jointly analyzed, highlighting the changes expected as consequence of gene flow and genetic drift. To develop this analysis, computational simulations of extreme scenarios combining changes in the levels of gene flow and population size were performed.

Keywords: allelic richness, computational simulations, gene diversity, molecular markers, population genetics

1. Introduction

Genetic diversity comprises the total of genetic variability contained in a population and it represents the raw material for evolutionary changes since it determines the microevolutionary potential of populations.

The most popular measure of genetic variation is the average heterozygosity expected in Hardy–Weinberg equilibrium. Nei [1] called this measure as gene diversity index, and defined it as either the average proportion of heterozygotes per locus in a randomly mating population or the probability that two alleles randomly and independently selected from a gene pool will represent different alleles.

Expected heterozygosity at n loci within a population is calculated, as:

$$H_e = 1 - \sum_{i=1}^n p_i^2 \quad (1)$$

Being p_i the allele frequency. Since this index has been formulated entirely in terms of alleles and genotypic frequencies, its treatment is biologically the most direct [2]. Expected heterozygosity can be applied to any population of all

organisms (sexual or asexual, diploid or non-diploid) independently of the number of alleles at a given locus or the pattern of evolutionary forces [1].

The total number of alleles at a locus has also been used as a measure of genetic variation and is an important measure of the long-term evolutionary potential of populations [3]. The major drawback of the number of alleles is that, unlike heterozygosity, it is highly dependent on sample size. Therefore, samples sizes must be equal in order to obtain meaningful comparisons between samples because of the presence of many alleles at low frequencies in natural populations. In this way, the allelic richness estimator (r) can avoid this problem owing to this estimator represents a measure of allelic diversity that takes into account the sample size [4]. By means of rarefaction method, the r estimator calculates the expected number of alleles at a locus for a fixed sample size, considering generally the smallest sample size in a series of sampled populations [5].

1.1 Loss of genetic diversity in reduced sized populations

The starting question for analyzing the effect of reduced sized populations on genetic diversity levels is how population size (N) influence on the allele and genotype frequencies. In case that Hardy–Weinberg principle assumption of infinite population size being violated, genetic drift will occur in populations. Genetic drift is a stochastic sampling process that determines what alleles will constitute the gene pool in the next generation. Fragmentation and isolation due to habitat loss and landscape modification can reduce the population size of many species of plants and animals throughout the world hence understand genetic drift and its effects is extremely important for biodiversity conservation [3].

The implementation of molecular biology techniques for differentiation of individuals directly at DNA level allows inferring genetic diversity parameters in real populations even these parameters were defined prior to the development of DNA-based molecular markers. In addition, technological development of capillary electrophoresis has improved the resolution power for allele identification and advances in computer power has allowed the analysis of a huge number of highly polymorphic loci simultaneously in a simply and quickly manner.

1.2 Molecular markers as workhorses for genetic diversity studies

A molecular marker is known as any specific DNA fragment that may or may not correspond to coding regions of the genome [6] and is representative of differences at the genomic level [7]. In case that a molecular marker shows segregation according to the Mendelian laws of inheritance, it can also be defined as a genetic marker and it provides genetic information [6]. Molecular markers offer advantages over conventional alternatives based on phenotype, since contrary to morphological data, molecular data are stable and detectable in all tissues without being related to the development, differentiation, growth, or defense state of the cell and they are not influenced by environmental effects [7, 8].

Although there are several type of molecular markers the ideal genetic marker must be reliably measurable, exhibit highly variable loci, be codominant, and be densely distributed throughout the genome. The microsatellite markers also called Simple Sequence Repeat (SSRs) meet all these requirements [9]. SSRs are monotonous repeats of short nucleotide motifs of 1 to 6 base pairs (e.g., cgtcgtcgtcgtcgt, which can be represented by $(cgt)_n$ where $n = 5$). These repetitive elements can be found interspersed in the three eukaryotic genomes: nucleus (SSRs), mitochondria (mtSSRs) and chloroplasts (cpSSRs) [10]. The different SSRs alleles are mainly generated through simple repeat addition and subtraction mechanisms that occur

with equal probability [11], and they are rarely found in coding regions [9]. SSRs are informative and practical markers because they provide information about the amount and distribution of genetic diversity and the processes that determine the genetic structure and variation within and between natural populations [12]. Regarding methodological concerns, they present high stability with high intra- and inter-laboratory repeatability and they can be implemented in low complexity laboratories using external sequencing services. A limitation for SSRs implementation is that the sequence of repetitive flanking region is required to the development of specific primers although the cross transference of primers between closely related species is usually successful. SSRs have become the most widely used DNA marker in population genetics for genome mapping, molecular ecology, and conservation studies [3]. Despite the fact that massive sequencing methods to identify single nucleotide polymorphisms (SNPs) have gained prominence, microsatellites continue to be widely used tool because the analysis of generated data is simple and easily comparable with previous studies.

1.3 Simulations as a tool for predicting what is expected under certain conditions

Simulations help to recreate the stochastic process that accompanies the transmission of genes from parents to offspring because they recreate the movement of alleles under a model with same conditions several times. In addition, using different model conditions can help to disentangle sampling effects and scale dependencies, as well as historical influences of gene flow.

Any model (analytical, simulation, and otherwise) makes simplifying assumptions, excepting that it be “an entire reconstruction of the actual system—whereupon it ceases to be a model” [13].

The focus of this chapter is define the simplest model that show the effects of population size and gene flow on contemporary levels of genetic diversity, attending to the influence that multiplicity and abundance play on the classic genetic diversity estimators.

2. Materials and methods

2.1 Simulations

In order to test the effect of population size and gene flow on the magnitude of genetic diversity parameters simulated genetic data were obtained using IBDsim program [14]. This program simulates genetic data under isolation by distance model using a backward simulation strategy at population level. Stepping Stone Model was considered which assumes discrete populations, discrete number of generations, genetic drift within each population, and migration between adjacent or spatially proximal population [15–17] being m the total dispersal rate in one dimension [18]. Four different scenarios were simulated considering a population composed by a square grid of 6 x 6 subpopulations. Those scenarios combine two subpopulation sizes (n): 100 or 20 diploid individuals and two migration rates (m): 0.5 or 0.005, respectively (**Table 1**). The four combinations of n and m allowed to obtain scenarios that show expected genetic diversity with low or high levels of gene flow in population of small or large populations. Scenarios A-C and A-D allowed to evaluate the consequences of high or low levels of gene flow on the diversity parameters in populations of high size, respectively while scenarios B-C and B-D allowed to evaluate the consequences of high or low levels of gene flow on diversity

Population size (<i>n</i>)	Migration rate (<i>m</i>)	
	0.5	0.005
100	A - C	A - D
20	B - C	B - D

Table 1.
*Four simulated scenarios combining population size (*n*) and migration rate (*m*).*

parameters in populations of small size, respectively. Each data set was composed by 180 diploid individuals sampled from nine subpopulations. To avoid edge effects, a two-dimensional lattice was represented in a torus [18]. At grid edges, we used ‘absorbing’ boundaries in IBDSim whereby ‘the probability mass of going outside the lattice is equally shared on all movements inside the lattice’ [19]. The total simulated population was kept constant, but samples were taken from a smaller area of 3 x 3 subpopulations with 20 individuals per node. This sampling strategy was implemented in order to restrict the sampling design to a relatively small geographical area in order to work at a local geographical scale [19]. Each individual was characterized by a multilocus genotype defined by ten nuclear microsatellite loci of a two base pair repeated motif with a mutation rate (μ) of 10^{-3} with two to 20 alleles per locus. From each scenario, 10 data sets were simulated.

2.2 Analysis of simulated data

Expected heterozygosity (*He*) was estimated using Nei’s gene diversity index (1) [1] and allelic richness (*r*) was estimated using a rarefaction method. Both estimators were calculated for each subpopulation (nine in each data set) under each scenario (four) and for each repetition (10 in each scenario) obtaining as result 360 estimations of each genetic diversity measures. These estimations were developed using FSTAT software [20]. Means of *He* and *r* were estimated for each scenario. In order to determine if differences between means were statistically significant a standard *t*-test of means was implemented. Differences between means was considered statistically significant if the chance occurrence of such statistic was 5 percent or less ($p < 0.05$). This test was implemented using Microsoft Excel software.

In addition, the spread and skew of both estimated parameters in all simulations by each scenario was shown using box and whisker plots that display a five-number summary: minimum, maximum, median, upper and lower quartiles. The central rectangle spans the first quartile to the third quartile, or the interquartile range (IQR). A segment inside the rectangle shows the median while whisker to the left and to the right show the locations of the minimum and maximum. These estimations were calculated using Microsoft Excel software.

3. Results

Combination of *n* and *m* allowed analyze the effect of population size and genetic isolation among population on genetic diversity estimators based on all differences between scenarios parameters estimations were statistically significant (Table 2). Scenarios A-C and A-D which consider large population size the allelic richness and the expected heterozygosity were higher than scenarios B-C and B-D which consider small population size (Figure 1). However, allelic richness showed lower values than heterozygosity in smaller populations comparing with large

	A-C	A-D	B-C	B-D
A-C	—	9.05511E-11	2.27959E-75	2.20212E-68
A-D	6.23453E-15	—	3.10501E-68	3.01124E-66
B-C	4.77563E-87	1.35851E-69	—	8.60895E-19
B-D	9.19086E-97	1.10061E-81	4.24449E-15	—

Table 2.
Pairwise t-test results between scenarios. Below diagonal p values of t-test applied for allelic richness (r) means and above diagonal p values of t-test applied for expected heterozygosity (He) means.

Figure 1.
Allelic richness (r) and expected heterozygosity (He) means by scenario.

populations with the same migration rate (A-C vs. B-C and A-D vs. B-D, respectively) (**Figure 1**). **Figure 2** shows box and whisker plots of r and He parameters for all simulated populations in the fourth scenarios. Despite the overlapping in simulated data from same population size and differences in the migration rates (A-C vs. A-D and B-C vs. B-D, respectively) differences in median values among all scenarios were detected. In addition, these plots show higher spread of r than He (**Figure 2**). In the comparison of means and median values between scenarios considering high levels of gene flow ($m = 0.5$) with differences in population size (A-C vs. B-C) and low levels of gene flow with differences in population size (A-D vs. B-D) r showed higher reduction than He (**Table 3**). Furthermore, the reduction was higher for r than the reduction for He between scenarios considering large population size with differences in migration rates (A-C vs. A-D). However, the reduction was higher for He than the reduction for r between scenarios considering small population size with differences in migration rates (B-C vs. B-D) (**Table 4** and **Figure 3**).

4. Discussion

Genetic diversity is a pre requisite for population adaptation to environmental changes [12]. Large populations of naturally outbreeding species usually have extensive genetic diversity, but genetic diversity is usually reduced in populations and species of conservation concern [12]. Theoretical analyses based on simulations give information for understanding empirical results.

Figure 2. Box and whisker plots for allelic richness (r) and expected heterozygosity (He) by scenario.

Parameter	Statistic	A-C vs B-C	A-D vs B-D
r	Mean	2.769 (42.24%)	2.575 (43.69%)
	Median	2.900 (43.94%)	2.600 (54.93%)
He	Mean	0.201 (25.77%)	0.246 (32.98%)
	Median	0.202 (26.77%)	0.248 (33.20%)

Table 3. Reduction of allelic richness (r) and expected heterozygosity (He) as consequence of changes in population size with high levels of gene flow ($m = 0.5$) (A-C vs. B-C) and in populations with low levels of gene flow ($m = 0.005$) (A-D vs. B-D). Reduction percentage are showed between brackets.

The total allele number by locus is a complementary measure of genetic diversity because it is more sensitive to loss of genetic variation as consequence of small population size than heterozygosity. In this way, r becomes in an important measure

Parameter	Statistics	A-C vs A-D	B-C vs B-D
r	Mean	0.662 (10.10%)	0.468 (12.36%)
	Median	0.700 (11.31%)	0.400 (10.81)
He	Mean	0.034 (4.35%)	0.079 (13.64%)
	Median	0.037 (4.72%)	0.083 (14.26%)

Table 4.
Reduction of allelic richness (r) and expected heterozygosity (He) as consequence of changes in gene flow levels in large populations ($n = 100$) (A-C vs. A-D) and in small populations ($n = 20$) (B-C vs. B-D). Reduction percentage are showed between brackets.

Figure 3.
Plot of allelic richness (r) and expected heterozygosity (He) of nine populations at one simulation for each scenario.

for long-term evolutionary population potential [3]. We will represent this statement using a hypothetical situation: population A ($n = 100$) and population B ($n = 10$) (**Figure 4**). There, population B is a random sample from population A. Population B shows three out of eight alleles from population A because of the reduction in population size, which cause that only alleles present in a high frequency remain in the small population. It means that by chance the more frequent alleles have a highest probability to being contained in the gene pool of small population while the rare alleles shows low frequency and as consequence they have high probability to be lost. In this way, the genetic drift is operating and as consequence of this microevolutionary process, not all alleles of a population will be present in the next generation producing a sampling error. As results of this sampling error, the change in the allelic frequencies is at random and the action of genetic drift does not have pre-established direction. However, in the analyzed example (**Figure 4**) the estimated value of He changes from 0.719 to 0.620 as consequence of 10 times reduction of population size. This change could indicate that He is less sensitive to rare allele lost as consequence of population size reduction. We can explain it by means of other hypothetical situation: We consider four pairs of small populations that contain between eight and 10 alleles (**Figure 5**). At left side of **Figure 5**, four populations show one allele at high frequency and rare alleles increase successively their number step by step (a, b, c and d) while at right side in the same Figure, four populations show alleles at equal frequency that increase successively their number step by step (a, b, c and d). For each population r and He were estimated. In the step (a) both populations show two alleles ($r = 2$) but He was lower in the population at

Figure 4.

Changes in number of alleles (NA) and expected heterozygosity (He) as consequence of population size reduction.

left side than population at right side (0.18 vs. 0.50, respectively), being the alleles frequencies the unique difference between both populations. Successively, in the following steps (b, c and d) while the number of different alleles increases, He also increases in populations at both sides. However, in populations at the right side, since the alleles are equally frequent in all steps, He reaches the maximum values, while in the populations at left side, the new alleles show low frequencies (rare alleles) and He increases little by little. Finally, in the step (e) He reaches the maximum value although all alleles are rare because of they show the same frequency. Hence, the estimation of He is highly dependent on allele frequencies and its value will be determined in a greater extent by the presence of alleles at high frequency which usually show high probability to be proportionally maintained when population reduce its size.

The effects of changes in population size on genetic diversity estimators considering different gene flow levels were studied in the present chapter by means of simulations (A-C vs. B-C and A-D vs. B-D, respectively). As expected, reductions in r and He values were obtained between large and small populations. In case that r and He are used for detecting genetic diversity reduction, r is more sensitive than He to detect genetic diversity reduction independently gene flow levels (**Table 3**).

The effects of gene flow levels on genetic diversity estimators considering different population sizes were studied in the present chapter by means of simulations

Figure 5.
Changes in allelic richness (r) and expected heterozygosity (He) in small populations with increasing in number of different alleles: two, three, four, five and ten (a, b, c, d and e, respectively).

(A-C vs. A-D and B-C vs. B-D, respectively). In large populations, r is more sensitive than He to detect genetic diversity reduction as consequence of low gene flow level. On the other hand, in small populations He is more sensitive than r to detect genetic diversity reduction as consequence of low gene flow level (**Table 4**).

Gene flow is a microevolutionary process that maintain the genetic exchange among local populations increasing population genetic diversity [21]. Gene flow can be quantified by the parameter m , which describes the movement of each gamete or individual independently of population size [22]. As microevolutionary process, gene flow counteracts the genetic drift effect and the balance between gene flow

and genetic drift determine genetic diversity levels for neutral alleles. Genetic diversity is the basis for local adaptation and genetic drift could be understood as a threat for biodiversity because of it causes genetic diversity loss in natural populations. Current climate change and fragmentation of natural populations as consequence of anthropic impacts are calling to urgent collective and interdisciplinary actions from researchers. The study of genetic diversity levels is especially important for the management of endangered and valuable species. The focus in conservation biology is the maintenance of genetic diversity because of inbreeding and reduction in reproductive fitness is often associated with loss of genetic diversity [12]. Although the International Union for Conservation of Nature (IUCN) recognizes the need to conserve genetic diversity as one of three global conservation priorities [23] the genetic factors are not currently considered to assign the conservation status of species [24].

5. Conclusion

The comprehensive quantification of genetic diversity levels demand the estimation of r and He because of the sensitiveness of both estimators depends on allele multiplicity and frequencies. In this way, the estimation of r and He is recommended for genetics studies in populations that inhabit disturbed environments.

Acknowledgements

The authors wish to thank National Council of Scientific and Technical Research (CONICET, Argentina).

Conflict of interest

The authors declare no conflict of interest.

IntechOpen

Author details

María Eugenia Barrandeguy^{1,2*} and María Victoria García^{1,2}

1 Population and Quantitative Genetics, Department of Genetics, Faculty of Exact, Chemicals and Natural Sciences, National University of Misiones (FCEQyN – UNaM), Argentina

2 Institute of Subtropical Biology – Posadas Node (UNaM – CONICET), National Council of Scientific and Technical Research, Argentina

*Address all correspondence to: ebarran@fceqyn.unam.edu.ar

IntechOpen

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

- [1] Nei M. Mint: Analysis of Gene Diversity in Subdivided Populations. *Proceedings of the National Academy of Sciences*.1973; 70 (12) 3321-3323. DOI: 10.1073/pnas.70.12.3321
- [2] Nagylaki T. Mint: The expected number of heterozygous sites in a subdivided population. *Genetics*.1998; 149: 1599-1604.
- [3] Allendorf FW, Luikart GH. Conservation and the Genetics of Populations. Blackwell Publishing; 2007. 642 p.
- [4] El Mousadik A, Petit RJ. Mint: High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. *Theoretical and Applied Genetics*. 1996; 92: 832-839. DOI: 10.1007/BF00221895
- [5] Petit R, El Mousadik A, Pons O. Mint: Identifying Populations for Conservation on the Basis of Genetic Markers. *Conservation Biology*. 1998; 12(4): 844-855.
- [6] Ferreira M, Grattapaglia D. Introducao ao uso de marcadores moleculares em análise genética. EMBRAPA-CENARGEN; 1996. 220 p.
- [7] Agarwal M, Shrivastava N, Padh H. Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Reports* 2008; 27:617-631.
- [8] Marcucci Poltri S. Marcadores Moleculares aplicados a Programas de Mejoramiento Genético de Eucalyptus. In: Secretaría de Agricultura. Ganadería, Pesca y Alimentos editors. Mejores árboles para más forestadores, 2005. 241 p.
- [9] Karhu A. Evolution and applications of pine microsatellites. [thesis]. Faculty of Science. University of Oulu. Oulu. 52 p.
- [10] Tautz D, Renz M. Mint: Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acid Research*. 1984; 12(10):4127-4138.
- [11] Schlötterer C, Tautz D. Mint: Slippage synthesis of simple sequence DNA. *Nucleic Acids Research*. 1992; 20: 211-215.
- [12] Frankham R, Ballou JD, Briscoe D.A. *Introduction to Conservation Genetics*. Cambridge University Press, 2002. 617 p.
- [13] Epperson BK, Mcrae BH, Scribner K, Cushman SA, Rosenberg MS, Fortin MJ, James PM, Murphy M, Manel S, Legendre P, Dale MR. Mint: Utility of computer simulations in landscape genetics. *Molecular Ecology*. 2010; 19: 3549-3564.
- [14] Leblois R, Estoup A, Rousset F IBDsim: a computer program to simulate genotypic data under isolation by distance. 2008; *Molecular Ecology Resources* 9(1): 107-109. DOI: 10.1111/j.1755-0998.2008.02417.x
- [15] Kimura M. Mint: “Stepping stone” model of population. *Annu Rep Natio Inst Genet*. 1953; 3: 62-63.
- [16] Kimura M, Weiss GH. Mint: The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*. 1964; 49: 561-576.
- [17] Weiss G H, Kimura M. A mathematical analysis of the stepping stone model of genetic correlation. *Appl Probab*. 1965; 2: 129-149. DOI: 10.2307/3211879
- [18] Leblois R, Beeravolu C R, Rousset F. IBDsim version 2.0 User manual.

[19] Leblois R, Estoup A, Rousset F. Mint: Influence of mutational and sampling factors on the estimation of demographic parameters in a “continuous” population under isolation by distance. *Mol Biol Evol.* 2003; 20(4): 491-502. DOI: 10.1093/molbev/msg034

[20] Goudet J. Mint: FSTAT (vers.2.9.3.2): a computer program to calculate F statistics. *Heredity.* 1995; 86:485-486.

[21] Hartl, DL, Clark AG. Principles of population genetics. Sinauer Associates, Inc Publishers; 2007. 652 p.

[22] Slatkin M, Barton NH. A comparison of three indirect methods for estimating average levels of gene flow. *Evolution.* 1989; 43(7):1349-1368

[23] McNeely JA, Miller KR, Reid WV, Mittermeier RA, Werner TB. Conserving the world's biological diversity. IUCN, World Resources Institute, Conservation International, WWF-US, and the World Bank, 1990.

[24] Garner BA, Hoban S, Luikart G. Mint: IUCN Red List and the value of integrating genetics. *Conservation Genetics.* 2020; 21: 795-801. DOI: 10.1007/s10592-020-01301-6