Genetics & Applications

Institute for Genetic Engineering and Biotechnology University of Sarajevo

Notes Open access

ALRATIO - R script for the analysis of relation between the effective and the detected number of alleles

Naris Pojskić[1*]

[1] *Laboratory for Bioinformatics and Biostatistics, University of Sarajevo - Institute for Genetic Engineering and Biotechnology, 71000 Sarajevo, Bosnia and Herzegovina*

DOI: 10.31383/ga.vol3iss1pp77-80

***Correspondence**

E-mail: naris.pojskic@ingeb.unsa.ba

Received

June, 2019

Accepted

June,2019

Published

June, 2019

It is widely accepted that understanding the heterogeneity of a population is important in assessment of the vulnerability of a conservation unit (Frankham et al., 2002). Standard measures such as estimation of heterozygosity, deviations from Hardy–Weinberg equilibrium, effective population size, inbreeding coefficients are widely used. Minor, but very important elements of these measures are allelic diversity, effective number of alleles and allelic richness which characterize the extent of genetic diversity. Allelic diversity (A_n) represents an average number of alleles per locus determined by direct count. When more than one locus is considered, it is calculated as a number of alleles averaged over loci expressed as *k/l* where *k* is the total number of alleles determined at all the observed loci and *l* is the number of loci (Frankham et al., 2002). The effective number of alleles (A_e) is a measure that shows the number of alleles required to ensure the same level of heterozygosity under the assumption of balanced allele frequency and low influence of rare alleles. It is expressed as $1/\Sigma p_i^2$ where p_i is the frequency of the i^{th} allele (Allendorf et al., 2013). Assessing allelic richness (A_E) implies using the rarefaction method to estimate the number of alleles expected in samples of specified size (Foulley et Ollivier, 2006). It is based on the assumption that higher number of alleles is expected in large samples and rarefaction solves this problem (Kalinowski, 2004). This measure of genetic diversity is indicative of a population's long-term potential for adaptation (Greenbaum et al., 2014).

These three measures are used less often than heterozygosity as a genetic diversity measure, but their usefulness for identifying populations in need of for conservation is asserted. When we consider the loss of genetic variation in small populations we usually refer to lower levels of heterozygosity and allelic diversity. Even bottleneck of low intensity may have significant impact on loss of allelic diversity. Generally, bottlenecks have a stronger effect on allelic diversity than on heterozygosity (Allendorf et al., 2013).

In light of the above, in the estimates of the genetic diversity of a conservation unit, the predictions based on alleles count calculations are essential. The estimates based solely on the number of alleles may lead to inaccurate conclusions. Allelic diversity shows the number of alleles, while effective number

of alleles provides information about contribution of alleles to heterozigosity. The difference between the detected and effective number of alleles indicates the presence of rare alleles which with high probability, will be lost in next few generations. Pojskic et Kalamujic (2015) in their study discuss the practical applications of such approach in the analysis of feral populations of brown trout in the Neretva river and its tributaries as a model. Significance of allele number reduction (effective number of alleles) can contribute to the assessment of the genetic status of a conservation unit. In accordance with the above, we propose allele ratio as adequate measure of relation between effective (A_e) and detected number of alleles (A_n). It is expressed as $R=A_e/A_n$ where A_e is effective number of alleles and A_n is the number of the detected alleles. The range of this ratio is 0-1, where lower values represent larger difference, while higher values indicates smaller difference between effective and detected number of alleles. A Z-score of $P<0.01$ is considered as statistically significant. The abovementioned calculations can be made using ALRATIO R script http://www.ingeb.unsa.ba/popgen/R/alratio/alratio.html (Figure 1.). When R code is executed with previously introduced input data as .csv extension file (Figure 2.), three files are created (Figure 3.).

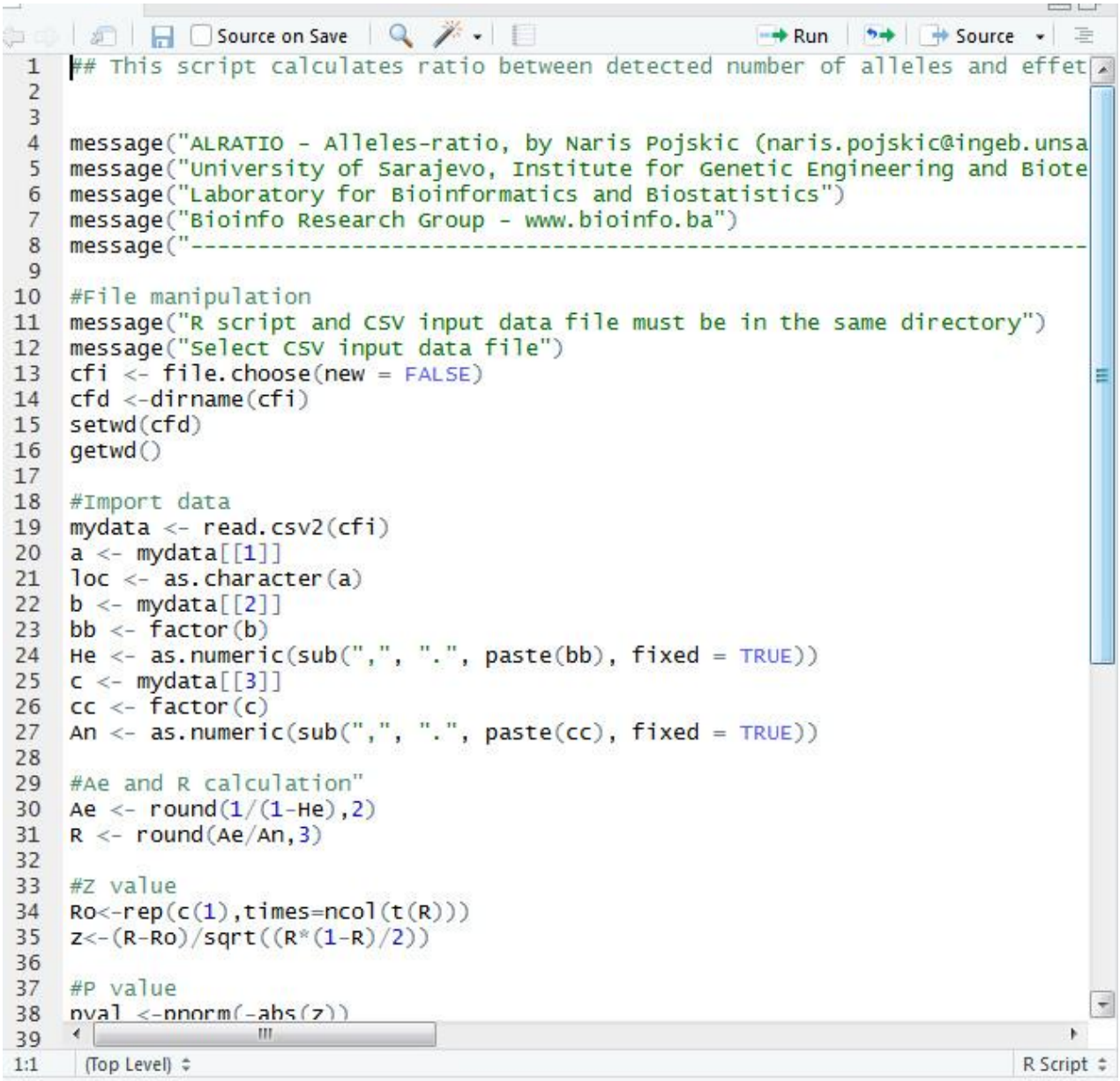

Figure 1. Presentation of ALRATIO R script code

```
> source("Z:/programiranje/Rcode/alratio/alratio.r")
ALRATIO - Alleles-ratio, by Naris Pojskic (naris.pojskic@ingeb.unsa.ba)
University of Sarajevo, Institute for Genetic Engineering and Biotechnology
Laboratory for Bioinformatics and Biostatistics
Bioinfo Research Group - www.bioinfo.ba
------------------------------------------------------------------------------
R script and CSV input data file must be in the same directory
Select CSV input data file
     Loc       He        An    Ae      R         Z         P
[1,] "CypG30"  "0.3784"  "11"  "1.61"  "0.146"  "-3.42"   "0.00031"
[2,] "CypG24"  "0.828"   "15"  "5.81"  "0.387"  "-1.78"   "0.03755"
[3,] "Ca1"     "0"       "1"   "1"     "1"      "NaN"     "NaN"
[4,] "Rru2"    "0.4624"  "4"   "1.86"  "0.465"  "-1.517"  "0.06464"
[5,] "Z21908"  "0.1464"  "2"   "1.17"  "0.585"  "-1.191"  "0.1168"
[6,] "Mean"    "0.363"   "6.6" "1.57"  "0.238"  "-2.53"   "0.0057"
[7,] "Mean*"   "0.4538"  "8"   "1.83"  "0.229"  "-2.595"  "0.00473"
Statistical significance level at P<0.01
Mean* = values without monomorphic loci
[1] "File (output.txt) and graphs (plot1.jpg; plot2.jpg) are saved in Z:/programiranje/Rcode/alratio"
>
```

Figure 3. View of RStudio console with results

Loc	He	An
CypG30	0,3784	11
CypG24	0,828	15
Ca1	0	1
Rru2	0,4624	4
Z21908	0,1464	2
Mean	0,363	6,6
Mean*	0,4538	8

Figure 2. Format of input data (Loc- locus; He - expected heterozygosity; An – number of alleles at given locus)

The first one is text file with the result of calculation (Figure 4.). It has tabular view with name of locus, values of expected heterozygosity, detected (A_n) and effective number of alleles (A_e), ratio (R) with Z and its P estimation. The others two are graphic files, which contain barplot graph (Figure 5), as well as scatter diagram (Figure 6). The ALRATIO script was tested using data from microsatellite loci analyzed in Dalmatian barbelgudgeon (Aulopyge huegelii Heckel, 1841; Kalamujic Stroil et al., 2019). The results are shown in Figure 3. and Figure 4. The *CypG30* locus has statistically significant deviation of effective number of alleles in comparison with detected number of alleles. The ratio for *Ca1* locus could not be estimated since it is monomorphic. Other loci did not show statistically significant deviation.

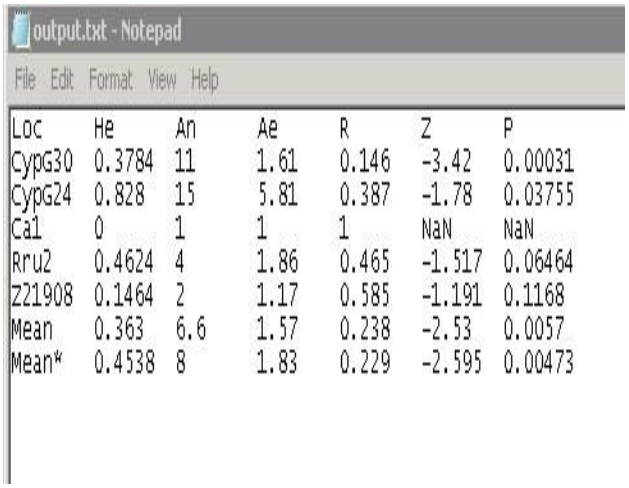

Figure 4. Format of output results (Loc- locus; *He* - expected heterozygosity; *An* - detected number of alleles at given locus; *Ae* - effective number of alleles; *R* - ratio between effective and detected number of alleles; *Z* - Z statistics; *P* - P value for ratio)

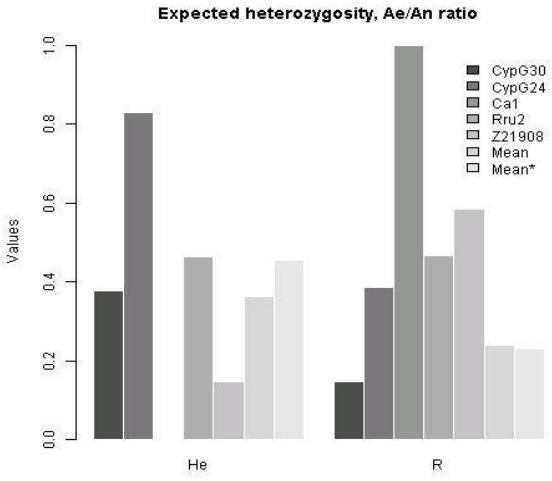

Figure 5. View of ratio plot

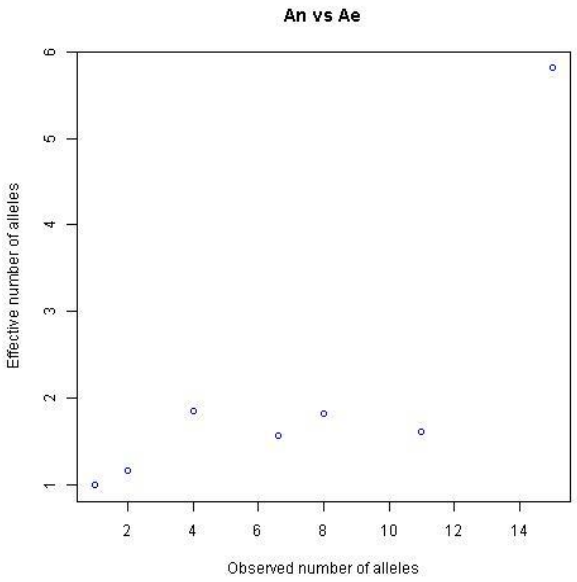

Figure 6. Scatter diagram between effective and detected number of alleles

We can conclude that the ratio between the effective and detected number of alleles can reveal the significance of the differences between them. The statistically significant value of such a measure indicates the proportion of alleles that do not really participate in genetic diversity of a population, which can be used for the prediction of genetic status of feral populations or agricultural varieties.

References

Allendorf FW, Luikart G, Aitken SA (2013) Conservation and the genetics of populations. John Wiley & Sons, UK.

Foulley JL, Ollivier L (2006) Estimating allelic richness and its diversity. Livestock Science, 101 (1–3): 150-158.

Frankham R, Ballou JD, Briscoe DA (2002) Introduction to Conservation Genetics. Cambridge University Press, New York, United States of America.

Greenbaum G, Templeton AR, Zarmi Y, Bar-David S (2014) Allelic Richness following Population Founding Events – A Stochastic Modeling Framework Incorporating Gene Flow and Genetic Drift. PLoS ONE, 9(12): e115203. doi:10.1371/journal.pone.0115203.

Kalamujić Stroil B, Mušović A, Škrijelj R, Dorić S, Đug S, Pojskić N (2019) Molecular-genetic diversity of the endangered Dalmatian barbelgudgeon, *Aulopyge huegelii* from the Buško Blato reservoir. Genetica, https://doi.org/10.1007/s10709-019-00069-z.

Kalinowski ST (2004) Counting alleles with rarefaction: private alleles and hierarchical sampling designs. Conservation Genetics, 5:539-543.

Pojskic N, Kalamujic B (2015) Simulations based on molecular-genetic data in detection of expansion *Salmo trutta* allochtonous population in the Neretva River's tributaries. 27th International Congress for Conservation Biology / 4th European Congress for Conservation Biology, Montpellier, France, pp. 539-540.