Genetics & Applications

Institute for Genetic Engineering and Biotechnology University of Sarajevo

Notes Open access

iMAF - Index of Major Allele Frequency

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.vol2iss2pp78-81

***Correspondence**

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

Received

November, 2018

Accepted

December, 2018

Published

December, 2018

Allele frequency is the relative frequency of an allele at a particular locus in a population (Gillespie et al., 2004). It is the basis of all population genetic analyses. Two parameters of allelic frequencies are special indicators: the major allele frequency when it comes to STR loci and the minor allele frequency when it comes to SNPs and other biallelic loci. The fluctuations of allele frequencies within and among populations have evolutionary significance, where selective pressure seeks to eliminate deleterious alleles, by favoring new genetic variants which bring an adaptive advantage. These effects of selection can directly modify major and minor allele frequencies by affecting the allele's distribution. In the cases of adaptively neutral loci, genetic drift has a significant role on the distribution of allelic frequencies.

The importance of major and minor allele frequencies in genetic diversity estimation were subjects of various studies (De la Cruz & Raska, 2014; Kaur et al., 2014; Engelsma et al., 2013).

Major and minor allele frequencies are one of indicators of genetic diversity and polymorphism at the observed locus within a wildlife populations and varieties (Singh et al., 2013). Serre and Pääbo (2004) observed the change of allelic frequencies at >350 microsatellites as evidence of gradients of human genetic diversity within and among continents. SNP alleles with extremely low frequencies are often associated with increased susceptibility to various diseases (Kido et al., 2018; Park et al., 2011; Cross et al., 2010).

Of course, when it comes to SNP markers and other biallelic loci, a simple ratio between minor and major allele frequencies may be worthwhile indicator of distribution of allele frequencies at given locus. Similarly, in the case of microsatellite markers and other multiallelic loci, the relation between the allele frequency, under the assumption of equal frequencies of all detected alleles at given locus and major allele frequency could be indicator of allele frequencies distributions and genetic diversity as well. Pojskic and Kalamujic (2015) 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.

In accordance with the above we propose index of major allele frequency (*iMAF*) as simple measure of distribution of allele frequencies at given locus and serve as one of the indicators of genetic diversity. It is expressed as $IfM = (1/A_N)/fM$, where A_N is the number of the detected alleles and fM is the frequency of

```
#############################################################################
#                 IMAF -  Index of Major Allele Frequency                   #
#                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                   #
#############################################################################

#File manipulation
message("R script and CSV input data file must be in the same directory")
message("Select CSV input data file")
cfi <- file.choose(new = FALSE)
cfd <-dirname(cfi)
setwd(cfd)
getwd()

#Import data
mydata <- read.csv2(cfi)
a <- mydata[[1]]
loc <- as.character(a)
b <- mydata[[2]]
bb <- factor(b)
fM <- as.numeric(sub(",", ".", paste(bb), fixed = TRUE))
print(fM)
c <- mydata[[3]]
cc <- factor(c)
An <- as.numeric(sub(",", ".", paste(cc), fixed = TRUE))
print(An)

k<-round(1/An,3)
print(k)
iM <- round(k/fM,3)
print(iM)

#Z value
iMo<-rep(c(1),times=ncol(t(iM)))
z<-(iM-iMo)/sqrt((iM*(1-iM)/2))

#P value
pval <-pnorm(-abs(z))
```

Figure 1. Presentation of iMAF R script code

major allele at the given locus. The range of this index is 0-1, where 0 represents absence of polymorphism at a locus, while 1 indicates the absence of major allele and that all the alleles have the same frequency. A Z-score of $P<0.01$ is considered as statistically significant. Indices with values close to 0 have a bad prediction, and those close to 1 have a better prediction in terms of distribution of allele frequencies of a given locus in the observed population. A higher value of the index indicates a better genetic potential of the given population in terms of diversity. The above mentioned calculations can be made using *iMAF* R script (Figure1) http://www.ingeb.unsa.ba/popgen/R/imaf/imaf.html within R version 3.5.1. (R Core Team, 2013). After introducing the input data (stored as csv file) and

Loc	fM	An
D3S1358	0,275	9
TH01	0,28	7
D21S11	0,227	15
D18S51	0,19	15
PENTA E	0,1615	18
D5S818	0,3805	8
D13S317	0,355	8
D7S820	0,278	8
D16S539	0,309	8
CSF1PO	0,32	8
PENTA D	0,2255	11
vWA	0,2595	10
D8S1179	0,3145	10
TPOX	0,561	8
FGA	0,2115	17

Figure 2. Format of input data (Loc–locus; fM-major allele frequency; An – number of alleles)

```
> source("Z:/iMAF/iMAF.R")
R script and CSV input data file must be in the same directory
Select CSV input data file
 [1] 0.2750 0.2800 0.2270 0.1900 0.1615 0.3805 0.3550 0.2780 0.3090 0.3200 0.2255 0.2595 0.3145
[14] 0.5610 0.2115
 [1]  9  7 15 15 18  8  8  8  8  8 11 10 10  8 17
 [1] 0.111 0.143 0.067 0.067 0.056 0.125 0.125 0.125 0.125 0.125 0.091 0.100 0.100 0.125 0.059
 [1] 0.404 0.511 0.295 0.353 0.347 0.329 0.352 0.450 0.405 0.391 0.404 0.385 0.318 0.223 0.279
 [1] 0.04293 0.08327 0.01440 0.02777 0.02619 0.02171 0.02750 0.05897 0.04325 0.03879 0.04293 0.03694
[13] 0.01918 0.00415 0.01150
      Loc        fM       An   k        iM       P
 [1,] "D3S1358" "0.275"  "9"  "0.111" "0.404" "0.04293"
 [2,] "TH01"    "0.28"   "7"  "0.143" "0.511" "0.08327"
 [3,] "D21S11"  "0.227"  "15" "0.067" "0.295" "0.0144"
 [4,] "D18S51"  "0.19"   "15" "0.067" "0.353" "0.02777"
 [5,] "PENTA E" "0.1615" "18" "0.056" "0.347" "0.02619"
 [6,] "D5S818"  "0.3805" "8"  "0.125" "0.329" "0.02171"
 [7,] "D13S317" "0.355"  "8"  "0.125" "0.352" "0.0275"
 [8,] "D7S820"  "0.278"  "8"  "0.125" "0.45"  "0.05897"
 [9,] "D16S539" "0.309"  "8"  "0.125" "0.405" "0.04325"
[10,] "CSF1PO"  "0.32"   "8"  "0.125" "0.391" "0.03879"
[11,] "PENTA D" "0.2255" "11" "0.091" "0.404" "0.04293"
[12,] "vWA"     "0.2595" "10" "0.1"   "0.385" "0.03694"
[13,] "D8S1179" "0.3145" "10" "0.1"   "0.318" "0.01918"
[14,] "TPOX"    "0.561"  "8"  "0.125" "0.223" "0.00415"
[15,] "FGA"     "0.2115" "17" "0.059" "0.279" "0.0115"
Statistical significance level at P<0.01
[1] "File (output.txt) and graphs (plot1.jpg; plot2.jpg) are saved in Z:/iMAF"
>
```

Figure 3. View of RStudio console with results

output.txt - Notepad

File Edit Format View Help

Loc	fM	An	k	iM	P
D3S1358	0.275	9	0.111	0.404	0.04293
TH01	0.28	7	0.143	0.511	0.08327
D21S11	0.227	15	0.067	0.295	0.0144
D18S51	0.19	15	0.067	0.353	0.02777
PENTA E	0.1615	18	0.056	0.347	0.02619
D5S818	0.3805	8	0.125	0.329	0.02171
D13S317	0.355	8	0.125	0.352	0.0275
D7S820	0.278	8	0.125	0.45	0.05897
D16S539	0.309	8	0.125	0.405	0.04325
CSF1PO	0.32	8	0.125	0.391	0.03879
PENTA D	0.2255	11	0.091	0.404	0.04293
vWA	0.2595	10	0.1	0.385	0.03694
D8S1179	0.3145	10	0.1	0.318	0.01918
TPOX	0.561	8	0.125	0.223	0.00415
FGA	0.2115	17	0.059	0.279	0.0115

Figure 4. Format of output results (Loc- locus; fM-major allele frequency; An-number of alleles at given locus; k-frequency according to assumption that all detected alleles at given locus have same value; iM-index of major allele frequency; P-P value for index of major allele frequency)

executing R code, script creates two files. The first one is textual file with the result of calculation (Figure 4) and the other is graphic file which contains barplot graph (Figure 5).

This approach is applicable to haplotype frequencies as well, where *fM* is major haplotype frequency and A_N is number of detected haplotypes.

The *iMAF* was tested using genotype profiles from Bosnian and Herzegovinian reference STR database (Pilav et al., 2017; 15 autosomal STR loci, 1000 individuals) as input data (Figure 2). The index

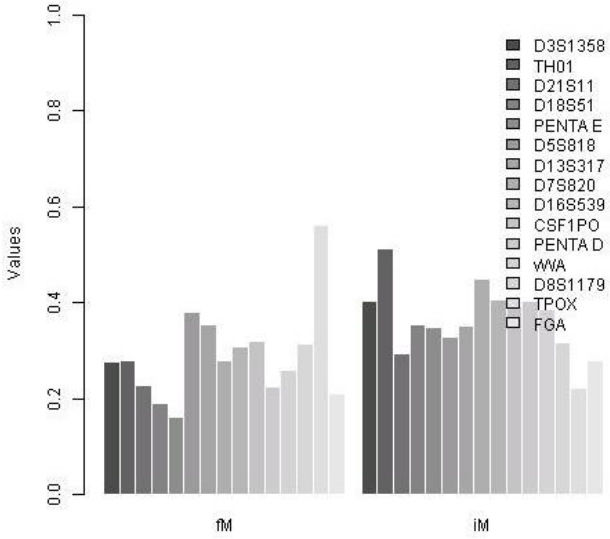

Figure 5. View of index of major allele frequency plot

shows that the major allele frequency of TPOX locus has statistically significant higher proportion than expected assuming equal frequencies of all the detected alleles at a given locus (Figure 3, Figure 4, Figure 5). This indicates that the most frequent allele has a tendency to increase its presence in the population, which may results in the loss of alleles with smaller frequencies.

We can conclude that the index of major allele frequency certainly can enabled additional overview of the distribution of allelic frequencies of a given locus indicating possible ways of changing the frequency. It can help to predict current and further

genetic status of wildlife population or agricultural variety.

References

Cross DS, Ivacic LC, Stefanski EL, McCarty CA (2010) Population based allele frequencies of disease associated polymorphisms in the Personalized Medicine Research Project. BMC Genet, 11:51.

De la Cruz O, Raska P (2014) Population structure at different minor allele frequency levels. BMC Procc, 8(1):S55.

Engelsma KA, Veerkamp RF, Calus MPL, Windig JJ (2013) Consequences for diversity when animals are prioritized for conservation of the whole genome or of one specific allele. J Anim Breed Genet, 131(1):1-10.

Gillespie JH (2004) Population genetics: a concise quide (2nd ed). The Johns Hopkins University Press, Baltimore, Md.

Kaur S, Cogan NOI, Forster JW, Paull JG (2014) Assessment of Genetic Diversity in Faba Bean Based on Single Nucleotide Polymorphism. Diversity, 6(1):88-101.

Kido T, Sikora-Wohlfeld W, Kawashima M, Kikuchi S, Kamatani N, Patwardhan A, Chen R, Sirota M, Kodama K, Hadley D, Butte AJ (2018) Are minor alleles more likely to be risk alleles? BMC Med Genomics, 11(1):3.

Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF Jr, Chatterjee N (2011) Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci USA, 108(44):18026-18031.

Pilav A, Pojskic N, Ahatovic A, Dzehverovic M, Marjanovic D (2017) Allele frequencies of 15 STR loci in Bosnian and Herzegovinian population. Croat Med J, 58(3):250-256.

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.

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/ (accessed on 29th November 2018).

Serre D, Pääbo S (2004) Evidence for Gradients of Human Genetic Diversity Within and Among Continents. Genome Res, 14(9):1679-1685.

Singh N, Choudhury DR, Singh AK, Kumar S, Srinivasan K, Tyagi RK, Singh NK, Singh R (2013) Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. PLoS ONE, 8(12):e84136.