American Journal of Animal and Veterinary Sciences 4 (4): 136-141, 2009
ISSN 1557-4555

Application of Statistical Procedures for Analysis of Genetic Diversity in Domestic Animal Populations

[1]M.R. Nassiry, [2]A. Javanmard and [2]Reza Tohidi
[1]Department of Animal Science, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran
[2]Department of Animal Science, Faculty of Agriculture, University Putra Malaysia,
43400 UPM Seldgang, Selangor, Malaysia

Abstract: Problem statement: A wide range of studies for the assessment of genetic diversity in livestock breed were conducted using genetic distance. For high-accuracy and unbiased estimation sampling methods, criteria of choosing type of DNA markers, distance measurement strategies, cluster analysis will be important for any genetic diversity projects. **Approach:** Main objective of this short review is focusing on application statistical procedures and methods in analysis of genetic diversity data in animals. **Results:** There is no simple strategy to address for best and effectively genetic diversity results by the way regarding to some important factors can make reliable results for next analysis. **Conclusion:** There is still a distinct need for developing comprehensive and user-friendly statistical packages that facilitate an integrated analysis of different data sets for generating reliable information about genetic relationships, genome diversity, and favorable allele variation. Equally important and perhaps more challenging, is the concerted and planned utilization of genome information in animal breeding programs on the basis of knowledge accrued from studies on genetic diversity.

Key words: Genetic diversity, statistical procedure, distance method, clustering methods

INTRODUCTION

Genetic diversity: Livestock breeding is important strategy for supporting our future requirement for best response against different environments[12]. Genetic conservation also is significant powerful tools for keep long term genetic relationships of animals. An essential first step in management farm and wild animal genetic resource is recognizing of genetic diversity parameters for make any decisions[33]. Diversity can be defined in a number of different way can we recognizes difference of two individuals. Genetic diversity is a platform for future genetic response[18]. Considering genetic diversity in agriculture populations not only the capacity to evolve with changing environment but also the capacity of copy with changing market requirement[5], Thus genetic diversity is seen as insurance against future changes[34]. The phenotypic difference are the result of genetic diversity and environmental difference more than a third of about 600 documented livestock breeds are under risk of extinction and up to two percent of the breed go extinct every year, thus one to two breed are lost per week[31]. The key question is which breeds should be chosen to assure the highest genetic diversity within species for the future[11]. The genetic composition of a population is usually described in terms of allele frequencies number of alleles and heterozygosity[24]. A wide range of studies for the assessment of genetic diversity in livestock breed were conducted using genetic distance[5]. For genetic distances the genetic difference between populations are assessed based on differences between allele frequencies at several loci[8]. Genetic distance is used to classify and elucidate the evolutionary relationship between populations such as species, which have been diverging for long period.

Genetic markers: The largest part of animal diversity is hidden, because it is genetic diversity. Hidden genetic variation is even more extensive than that observed through the phenotype so much therefore it is virtually impossible for two individuals in a population to have the same genotype at all loci, this genetic variation can detected through molecular technologies[12]. Modern technologies used in genetics enable us to measure this type of variability. Molecular genetic markers can be used to examine a group of individuals or populations to estimate various diversity measures and genetic distance. In principle genetic diversity can be measured on the

Corresponding Author: M.R. Nassiry, Department of Animal Science, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran

basis of polymorphic characters occurring at the different system (morphological, biochemical, protein), but DNA markers are very powerful tools for study of genetic diversity[12]. In practice, there is very litter information on the population, reproduction, adaptation and disease resistance potential of the most livestock breed, in this situation genetic information can provide valuable estimates of genetic diversity within and between populations. With regard to genetic diversity studies, molecular markers can be subdivided into two categories[15]. The first category comprises the allelic informative or codominate markers, such as microsatellite marker (SSRs) and Restriction Fragment Length Polymorphism (RFLP). The second category comprises the non-informative or dominates markers such as Amplified Fragment Length Polymorphism (AFLP) and Random Amplified Polymorphic DNA (RAPD). Genetic loci used in genetic distancing should be informative, manning they should display sufficient polymorphism, for a correct estimation of genetic distances[19]. It is important number of allele for loci for example about SSR markers; this should have at last 4 different alleles[9]. With regard to genetic diversity studies, microsatellite markers are very interesting tools because of their codominate mode of heritance, their high degree of reproducibility, their high level of polymorphism and therefore their high discriminative power.

Sampling: Optimal sampling strategies will support to next reliable results. Sampling is part of statistical practice concerned with the selection of individual observations individual observations intended to yield some knowledge about populations[17]. Basically, sampling strategies of animal population would be very difficult because various factors including total population size, migration, inbreeding, selection will affect reliability of next following results. Since sampling methods are begin point of any genetic diversity investigations thus understanding of statistical methods which can allow more reliable results will perfectable. Structure of our population, mating system, number of allele per locus, frequency of alleles per locus are some of factors for consideration for The sampling frame must be representative of whole population. With any form of sampling, there is a risk that the sample may not adequately represent the population but random sampling enable an appropriate sample size to be chosen. There are two types of random variables: categorical and numerical, categorical random variable yield response such as present and absent. Numerical response is such as your height in centimeters[30]. Usually, however, the true genotype frequencies are not known, estimation of minimum sample size for detecting alleles in population is very important when our population show complete hemozygosity it means that we need minimum sample size because most of new allele can be found per homozygous individual. When allele frequencies in the population are know and also population is under HW equilibrium, in the other hand, if the alleles are randomly associated in the genotype, minimum sample size is equal to one half the minimum sample sizes than complete hemozygosity. Generally, N = 25 sampled animal are taken to be a minimum requirement, with that 2N = 50 drawing of alleles per locus are performed, which should give a reasonably reliable estimate of allele frequencies[30].

Genetic diversity of sample can be described by quantifying allelic richness and allelic evenness of the sample[18].

Important parameters in genetic diversity analysis: Quantification of genetic diversity depends on some parameters: Hardy-Weiberg equilibrium Test (HWT), Polymorphism, Average number of alleles per locus Effective number of alleles, Average expected Heterozygosity (He), Shannon index.

Hardy-Weiberg equilibrium test (HWT): Hardy-Weiberg equilibrium explains that the both gene and genotype frequencies will be constant from generation to subsequent next generations[13]. Hardy-Weiberg assumption is under following consideration: Diploid, sexual reproduction, Random mating, no selection, no mutation and no immigration[14]. Deviation from HWT indicates that one or some of mentioned factors make disequilibrium from this test. Chi-square test is useful for determining whether the allelic frequencies are in HW equilibrium. The statistical test follows this formula[7]:

$$HWT = \sum (O_i - E_i)^2 / E_i$$

When:

HWT = Statistical test
O_i = Observed frequencies
E_i = Expected frequencies
df = Degree of freedom

If $X^2_{cal} \leq X^2_{tab}$ then H_0 hypothesis is accepted, it mean that allele frequencies for loci in a given population are HWT equilibrium, if $X^2_{cal \geq} X^2_{tab}$ then H_0 hypothesis is rejected[10].

Polymorphism: A polymorphic gene is usually defined as one for which the most common alleles has a

frequency of less than 0.95. Genetic loci use in genetic distance should be informative, meaning they should display sufficient polymorphism, for a correct estimation of genetic distance it is important number of allele for loci, for example about SSR markers, this should have at least 4 different alleles.

Average number of alleles per locus: These measures provide complementary information to that polymorphism:

$$N = \left(\frac{1}{k}\right)\sum_{i=1}^{k} n_i$$

When:
k = Number of loci
n_i = Number of alleles detected by locus

This parameter has best application in codominate markers.

Effective number of alleles: The measure explain about the number of alleles that would be expected in a locus in each population:

$$A_e = \left(\frac{1}{\sum_{a=1}^{k} p_a^{\ 2}}\right)$$

where, $p_a^{\ 2}$ is the frequency of the a^{th} of k alleles. By taking allele frequencies into account, this descriptor of allelic richness is less sensitive to rare alleles. This parameter also play fundamental role for verification of our sampling strategy. If the figure obtained the second time is less than the first estimated number. This could mean that our sampling strategies need revising.

Average expected Heterozygosity (H_e): Average expected heterozygosity is the probability that at a single locus a diploid organism any two alleles, chosen at random, are different from each other. It is an indicator of genetic diversity in a population:

$$H_e = 1 - \frac{1}{m}\sum_{i=1}^{m}\sum_{a=1}^{k} p^2_a$$

where, m number of analyzed loci.

Range of this parameter from 0-1 and it is maximized when there are many alleles at equal frequency.

Shannon index[32]**:** The measure explain about gene diversity, when Shannon index is near 1 then we can conclude that heterozygosity in our population is high and also we can compare Shannon index when it calculated for two loci, if one primer was higher amount of Shannon Index than other primers, it means that primers is suitable for genetic diversity studied in that breeds or populations.

F-statistics[31]**:** fixation indices F_{is}, F_{st}, F_{it} were used to analyze of partitioning of genetic variation, fixation indices are measures of standardized variances in allele frequencies that detect departure from HWT caused by biased inbreeding, biased outbreeding or population subdivision and drift. The subscript I, S and t refer to individual, subpopulation and total population. The F statistics is a measure of the difference between the mean heterozygosity among the subdivision is a population and potential frequency of heterozygote if all members of population mixed freely and none assertively.

F_{is}: F_{is} detects inbreeding individuals relative to subpopulation (within individual within populations). This parameter can range from -1 to 1 indicating maximal inbreeding and outbreeding respectively. A positive F_{is} value indicates inbreeding as the observed heterozygosity is lower than the expected heterozygosity:

$$F_{st} = \left[\frac{\text{obeserved}(\text{mean}(\text{Hpop1} + \text{Hpop2}))}{(2)(\text{meanA}(\text{pop1} + \text{pop2})(\text{meana}(\text{pop1} + \text{pop2})}\right]$$

F_{st} is considered to be the most informative statistic for examining the overall level of genetic divergence among subpopulations.

F_{st} detects inbreeding in subpopulations relative to the total population, when F_{st} equal zero it means that subpopulations are identical allele frequency and fixed for different allele. Range of this parameter always is positive and F_{st} is better for population estimates in cases where a high level of gene flow is present. F_{st} was 0-0.05 small, 0.05-0.15 moderate and 0.15-0.25,>0.025 very large:

$$F_{st} = 1\left[\frac{\text{exp ected (mean}(\text{Hpop1} + \text{Hpop2})}{\text{Total (Hmix population)}}\right]$$

F_{it}: This parameter can range from -1 to 1 indicating maximal inbreeding and outbreeding respectively:

$$F_{it} = 1 - \left[\frac{\text{oberseved(mean(Hpop1 + Hpop2)}}{\text{Expected(meanHpop1 + Hpop2)}} \right]$$

Gst coefficient: This parameter is measure of differentiation in term of alleles per locus in two or more populations. It rang from 0-1. A negative value may be obtained if a error was mad for sampling or an inappropriate system was used. If g_{st} is significant, it means that high percentage of genetic diversity is distributed among populations:

$$gst = 1 - \left(\frac{h_s}{h_t} \right)$$

Where:
h_s = Population diversity
h_t = Total diversity

Measurement of genetic distance: Many type of estimation of genetic distance are available, if two populations are homozygous for different genes at a particular locus, the distance is the maximum. For qualitative characters, distance between two individual is score as 0 and 1, for quantitative characters, the distance between two individual is calculated as the different in the trait values.

Euclidean distance is the most commonly measurement for estimating of genetic distance by morphological data:

$$d_E = \sqrt{\sum_{i=1}^{m} \sum_{a=1}^{k} \left(p_{ila} - P_{jla} \right)^2}$$

d_E between studies is difficult and rang of this value between zero and $\sqrt{2}$ m. Where p_{ila} the frequency of allele à at locus L for individual P_{jla}. The frequency of allele à at locus L for individual j, m the number of loci and K. the number of alleles of alleles at locus L.

Various genetic distance measures have been proposed for analysis of molecular marker data, for example, we can use these distances for analysis of SSR, RAPD, AFLP, PBR DNA markers studies. For dominate markers, the total number of bands is conventionally set as the number of analyzed loci. For codominate markers, genetic similarity between two individuals number of alleles per locus determined for total collection, is in general higher than two, Opposite to the 1- and 0- allele for dominant markers. Generally, genetic distance in codominate markers are based on allele frequencies.

If we assume that a = 3, b = 1, c = 3 and d = 2 then:

- Nei and Li or Dice : $\frac{a}{\left[a + (b + c)/2 \right]} = 0.42$
- Jaccard $\frac{a}{(a + b + c)} = 0.42$
- Sokal and Sneath $\frac{a}{\left[a + 2(b + c) \right]} = 0.27$
- Roger and Tanimoto $\frac{(a + d)}{\left[a + d + 2(b + c) \right]} = 0.38$

The Jaccard coefficient only count bands present for either individual, double absences are treated as missing data. If false-positive or false negative data occur, the index estimate tends to be biased. Nei and Li coefficient counts the percentage of shard bands among two individuals and gives more weight to those bands they are present in both.

In other hand, Nei coefficient puts more weight to shared bands than the coefficient of Jaccard. When our population is line, Nei and Jaccard coefficient lead to identical ranking, but in hybrid population, it seem that result will be different.

Clustering methods: Cluster analysis is the grouping of objects into different categories or class based on similarities between items in order to minimize variation within and maximize variation between categories[1]. For cluster method, we must consider that what kind of reproduction system we have in population and also we must know about levels of heterozygosity and which genetic characters we want to analysis[18]. Three main clustering methods are about:

- Nearest neighbor (simple matching)
- Furthest neighbor
- Unweighted Pair Group Method using Arithmetic Average (UPGMA)

SM consider absence corresponds to homozygous loci, it can be used with dominate marker (RAPD, AFLP) because absence could corresponds to homozygous recessives. UPGMA is most commonly method for cluster analysis, UPGMA can only be used when the evolutionary rate is nearly same for all groups included in the study, when studying the genetic diversity of germplasm collection, SM method should be preferred above the UPGMA clustering method, because genetic difference among accessions in germplasm are dominantly determined by selection and breeding rather than by evolutionary forces.

Validation of a single cluster: Resampling is a term used in statistics for bootstrapping and permutation[6] these procedures can be used in genetic diversity studies to assign confidence to the presence of clusters in a dendrogram.

Bootstrapping is a statistical method for estimating the sampling distribution of a estimator by sampling with replacement from the original sample[4], major purpose of bootstrapping is deriving robust estimates of standard errors and confidence intervals of population parameters.

A permutation test is type of statistical significant test in which a reference distribution is obtained by calculating all possible values of the test statistic under rearrangements the tables on the observed data points.

Molecular data analysis software: Many software programs for molecular population genetics studies have been developed for personal computer[16]. Four important software for analysis of population genetics are TFPGA, Arlequin[3], GENEPOP and POPGENE by using these software we can calculate observed and expected heterozygosity, percent polymorphic loci, Hardy Weinberg test, Nei distance and UPGMA clustering methods. TFPGA, Arlequin and population program are available in windows environment GEEPOP can used DOS operation system.

CONCLUSION

There is still a distinct need for developing comprehensive and user-friendly statistical packages that facilitate an integrated analysis of different data sets for generating reliable information about genetic relationships, genome diversity, and favorable allele variation. Equally important, and perhaps more challenging, is the concerted and planned utilization of genome information in animal breeding programs on the basis of knowledge accrued from studies on genetic diversity.

REFERENCES

1. Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press, New York, ISBN: 0120576503, pp: 359.
2. Excoffier, L., G. Laval and S. Schneider, 2005. Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evolut. Bioinform. Online, 1: 47-50. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2658868/
3. Brown, J.K.M., 1994. Boostrap hypothesis test for evolutionary tree and other dendrogram. Proc. Natl. Acad. Sci. USA., 91: 12293-12297. http://www.pnas.org/content/91/25/12293.full.pdf
4. Cowley, P.H., 1992. Resampling methods for computation-intensive data analysis in ecology and evaluation. Annul. Rev. Ecol. Syst., 23: 405-447. http://www.jstor.org/stable/2097295
5. Emigh, T.H., 1980. A comparison of tests for Hardy-Weigberg equilibrium. Biometrics, 36: 627-642. http://www.citeulike.org/user/yangjustinc/article/4860967
6. Gillet, E.M., 1999. Minimum Sample Sizes for Sampling Genetic Marker Distributions. In: Which Marker for Which purpose? Molecular Tools for Biodiversity. Gillet, E.M. (Ed.). http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm
7. Guo, S.W. and E.A. Thompson, 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics, 48: 361-372. http://www.ncbi.nlm.nih.gov/pubmed/1637966
8. Hall, J.G., 2004. Livestock Biodiversity Genetic Resource for Farming of Future. Blackwell Science, ISBN: 0632054999, pp: 280.
9. Hanotte, O. and Jianlian, 2005. Genetic characterization of livestock populations and its use in conservation decision making. http://www.fao.org/biotech/docs/hanotte.pdf
10. Hartl, D.L. and A.G. Clark, 1989. Principles of Population Genetics. 2nd Edn., Sinauer Associates, Sunderland,MA., pp: 652.
11. Hernandez, J.L. and B.S. Weir, 1989. A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics, 45: 53-70. http://www.ncbi.nlm.nih.gov/pubmed/2720060
12. Karp, A., D.S. Ingram and P.G. Issac, 1997. Molecular Tools for Screening Biodiversity: Plant and Animals. 1st Edn., Springer, London, ISBN: 10: 0412638304, pp: 528.
13. Labate, J.A., 2000. Software for population genetic analyses of molecular marker data. Crop Sci., 40: 1521-1528. http://crop.scijournals.org/cgi/content/abstract/40/6/1521
14. Mariette, V. Lecorre and A. Kremer, 1999. Sampling within the Genome for Measuring within-Population Diversity. In: Which DNA Marker for Which Purpose? Gillet, E.M. (Ed.). http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/
15. Mohammadi, S.A. and B.N. Prasanna, 2003. Analysis of genetic diversity in crop plants-Salient statistical tools and considerations. Crop Sci., 43: 1235-1248.

http://cat.inist.fr/?aModele=afficheN&cpsidt=14906916
16. Moritz, 2002. Molecular systematics. Natl. Acad. Sci. USA., 78: 5913-5916.
17. Nei, M. and R.K. Chesser, 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet., 47: 253-259. http://www.ncbi.nlm.nih.gov/pubmed/6614868
18. Nei, M., 1972. Genetic distance between populations. Am. Nat., 106: 283-292.
19. Nei, M., 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics, 89: 583-590. http://www.genetics.org/cgi/reprint/89/3/583
20. Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York, ISBN: 10: 0231063210, pp: 512.
21. Oldenbork, J.K., 1999. Genebanks and the conservation of farm animals genetic resources. DLO Institute for Animal Science and Health. Lelystand. The Netherlands, pp: 119
22. Raymond, M.L. and F. Rousset, 1995. An exact test for population differentiation. Evolution, 49: 1280-1283. http://www.evolutionhumaine.fr/michel/publis/pdf/raymond_1995_evolution.pdf
23. Rogers, J.S., 1972. Measures of genetic similarity and genetic distance. Stud. Genet., 7: 145-153.
24. Scherf, B.D., 2000. World Watch List for Domestic Animal Diversity. FAO. Rome.
25. Shannon, C.E. and W. Weaver, 1949, The Mathematical Theory of Communication. University of IIlinois Press, Urbana.
26 Simianer, H., 2005. Decision making in livestock conservation. Ecol. Econ., 53: 559-572. DOI: 10.1016/j.ecolecon.2004.11.016
27. Smith, C., 1984. Genetic aspect of conservation in farm livestock. Livestock Prod. Sci., 2: 37-48. DOI: 10.1016/0301-6226(84)90005-8
28. 36. Sokal, R. and F.J. Rohlf, 1995. Biometry. 3rd Edn., W.H. Freeman and Co., New York.
29. Vos, P., R. Hogers, M. Bleeker, M. Reijans and M. Zabeau *et al*., 1995. AFLP: A new technique for DNA fingerprinting. Nucleic Acids Res., 23: 4407-4414. http://www.ncbi.nlm.nih.gov/pubmed/7501463
30. Ward, J.H. Jr., 1963. Hierarchical grouping to optimize an objective function. J. Am. Satit. Assoc., 58: 236-244. http://www.jstor.org/stable/2282967
31. Weight, S., 1950. The genetical structure of populations. Nature, 166: 247-249. http://www.ncbi.nlm.nih.gov/pubmed/15439261
32. Weir, B.S., 1990. Genetic Data Analysis: Methods for Discrete Population Genetic Data. Sinauer Associated. Inc., Sunderland, Massachusetts, ISBN: 0878938710, pp: 337.
33. Weir, B.S., 1996. Genetic Data Analysis II: Methods for Discrete Population Genetic Data. 2nd Edn., Sinauer Associates, Inc., Sunderland, Massachusetts, ISBN: 10: 0878939024, pp: 445.
34. Weir, B.S. and C.C. Cockerham. 1984. Estimating f-statistics for the analysis of population structure. Evolution, 38: 1358-1370. http://banyan.usc.edu/research/ref1/Weir1984
36. Wright, S., 1978. Evolution and the Genetics of Populations, Variability Within and Among Natural Populations. University of Chicago Press, Chicago.
37. Yeh, F., R.C. Yang, T.J.B. Boyle, Z.H. Ye and J.X. Mao, 2000. POPGENE. The User Friendly Shareware for Population Genetic Analysis. Version 1.32.Molecular Biotechnology Center. University of Alberta. Canada.