https://doi.org/10.48048/tis.2022.2045

Hierarchical Clustering on Principal Components of Microsatellite Frequency Allele Data from Indonesian Buffaloes

Ferdy Saputra[1,*], Anneke Anggraeni[1] and Cece Sumantri[2]

[1]*Indonesian Research Institute for Animal Production, Bogor, Indonesia*
[2]*Department of Animal Production and Technology, Faculty of Animal Science, IPB University, Bogor, Indonesia*

([*]Corresponding author's e-mail: ferdy44saputra@gmail.com)

Received: 3 September 2020, Revised: 21 May 2021, Accepted: 3 June 2021

Abstract

Buffalo is a livestock that is used for meat, milk, draught animals, and religious ceremonies in Indonesia. Buffalo genetic information has not been obtained optimally for the development of buffalo breeding programs. Allele frequency is essential information to determine the genetic diversity of a population. This study investigates the use of one of the multivariate analyzes, hierarchical clustering on principal component (HCPC). The data used were microsatellite allele frequency data of 199 swamp buffaloes and 12 river buffaloes in 8 population (7 populations swamp buffalo and 1 population of river buffalo). Furthermore, the data is processed using factoextra and FactoMineR package in R 4.0.0. The results found that the ILSTS61 and ILSTS17 loci could be used as genetic markers to determine the genetic relationship of Indonesian buffalo. From the study, it is concluded that the HCPC method with allele frequency data can be used to analyze genetic relationships in Indonesian buffalo. The PC (Principle Component) value can describe which loci determines the genetic relationship.

Keywords: Multivariate, Microsatellite, Allele frequency, Indonesian buffaloes

Introduction

Genetic diversity in indigenous breeds is a major concern considering the necessity of preserving what may be a precious and irreplaceable richness, regarding new productive demands [1]. Conservation should be based on a deep knowledge of the genetic resources of the specific breed [2,3]. Therefore, it is important to try to characterize genetically indigenous breeds. Genes affecting polygenic traits and characterizing milk or meat productions are difficult to identify. The maintenance of genetic diversity in livestock species requires the adequate implementation of conservation priorities and sustainable management programs, which should be based on comprehensive information regarding the structure of the populations, including sources of genetic variability among and within breeds [6]. Genetic diversity is an essential component for population survival, evolution, genetic improvement and adaptation to changing environmental conditions [7,8]. Molecular methods based on molecular markers, such as RAPD, RFLP and microsatellites, are useful tools to study the genetic variations [9]. Short tandem repeats known as microsatellites are widely used as molecular markers of choice for genetic studies. Advantages of microsatellites are high degree of polymorphism due to existence of several alleles at each locus, their large number, distribution throughout the genome, high level of polymorphism, neutrality with respect to selection, codominant inheritance and easy automation of analytical procedures [9,10].

Buffaloes are livestock that is used to produce meat, milk, and draught animals. In Southeast Asia, buffalo are used in rice cultivation activities, especially in Indonesia. The genetic diversity of buffalo observed phenotypically is very diverse [11]. Genetic diversity analysis usually uses genetic markers such as mtDNA, Y chromosome, microsatellites, or functional genes. In Indonesia, there are 13 breeds of swamp buffalo (Gayo, Jawa, Kalang Kalsel, Kalang Kaltim, Kuntu, Moa, Pampangan, Simeulue, Sumatra-Barat, Sumatra-Utara, Sumbawa and Toraja) and 1 breed of river buffalo (Murrah) [12].

The diversity of buffalo in Indonesia has been carried out based on the cytochrome oxidase subunit I [13], Cytochrome B [14,15] and microsatellite [5,16]. DNA sequences and body size of livestock can be used to determine the genetic relationships of different populations. Indonesian buffalo based on morphometrics has been carried out by Anggraeni *et al*. [17]. Based on the microsatellite approach and

morphometrics, it produces a large number of clusters in Indonesian buffalo. Multivariate analyes such as principle component analysis (PCA) and principle coordinate analysis (PCoA) were the statistical methods implemented for genetic markers [18]. The multivariate method that can be used for genetic markers is Hierarchical Clustering on Principal Components (HCPC). HCPC is one of multivariate analysis that involved the application of objective clustering techniques to the principal components analysis results [19]. HCPC had similar result with PCA and Neighbor-Joining [20]. HCPC is also used to identify different patterns of the SARS-CoV-2 epidemic across Italian regions [21].

Genetic relationships of different populations and closely related species can be inferred using allele frequency data [22]. Genetic distance is usually calculated based on allele frequency data. Furthermore, the phylogenic tree is described based on genetic distance data. In addition to methods for reconstructing phylogenetic trees, the multivariate analysis itself has been widely used in analyzing genetic diversity in livestock. This study aims to analyse the use of allele frequency data using Hierarchical Clustering on Principal Components (HCPC).

Materials and methods

Allele frequency data used are derived from Saputra *et al*. [16], consisting of 199 swamp buffaloes and 12 river buffaloes collection of Laboratory of Animal Molecular Genetics, Faculty of Animal Science, IPB University. Allele frequency data were also analyzed using Hierarchical Clustering on Principal Components (HCPC). HCPC was carried out using the FactoMineR [23]. The results were visualized using the factoextra [24] packages in the R 4.0.0 [25].

Results and discussion

Principal Component 1 (PC1) with an eigenvalue of 3,934 or in other words, 39.34 % of the variation can be explained by PC1 (**Table 1**). Moreover, 26.23 % of the variation can be explained by Principal Component 2 (PC2). Therefore, about 65.57 % of the variation can be explained by PC1 and PC2 simultaneously.

Table 1 Eigenvalue of the principal component from frequency allele data.

Principal component	Eigenvalue
PC1	3.934298e+00
PC2	2.622625e+00
PC3	1.486106e+00
PC4	5.246031e-01
PC5	4.211559e-01
PC6	1.121251e-02
PC7	1.341476e-31

The variables that correlate the most with the PC1 are ILSTS61-Allele B (20.308), ILSTS17-Allele A (15.439), ILSTS17-Allele B (15.439), and ILSTS61-Allele C (13.179) (**Table 2**). The first principal component is positively correlated with all 4 of these variables. On the other hand, The variables that correlate the most with the PC2 are ILSTS61-Allele A (27.374), ILSTS61-Allele E (14.619), ILSTS17-Allele A (12.372), and ILSTS17-Allele B (12.372). A larger PC value can better explain variations in the data. Therefore, ILSTS61 and ILSTS17 loci can be used to describe the genetic relationships of buffalo in Indonesia.

Table 2 Principal components values from frequency allele data.

Variable (Loci)	PC1	PC2
CSSM66-Allele A	11.605	9.183
CSSM66-Allele B	11.605	9.183
ILSTS61-Allele A	0.143	27.374
ILSTS61-Allele B	20.308	0.053
ILSTS61-Allele C	13.179	6.173
ILSTS61-Allele D	0.419	8.671
ILSTS61-Allele E	11.863	14.619
ILSTS17-Allele A	15.439	12.372
ILSTS17-Allele B	15.439	12.372

The genetic relationship of Indonesian buffalo using the HCPC method was found in 4 clusters: Cluster 1 (River buffalo from North Sumatra), cluster 2 (Aceh, North Sumatra, and Riau), cluster 3 (Central Java, West Nusa Tenggara, and South Sulawesi), cluster 4 (Banten) (**Figure 1**). That results is different from the results of using the Neighbor-Joining method conducted by Saputra *et al*. [16], which found 3 clusters where buffalo from Banten grouped with Central Java, West Nusa Tenggara, and Sulawesi. The difference is due to different approaches. However, other multivariate methods, such as PCA and PCoA, have been widely used to analyze genomic data [26,27].

Buffaloes from Aceh, North Sumatra and Riau become 1 cluster because these areas are on the same island, namely Sumatra island. South Sulawesi, West Nusa Tenggara and Central Java became 1 cluster possibly due to the domestication routes to Sulawesi Island, Nusa Island, and Java Island. Banten buffalo are in a different cluster, possibly because Banten is an area on the island of Java which is adjacent to the island of Sumatra. This finding also supported Colli *et al*. [28] suggestion about 2 route of post-domestication migration to Indonesia.

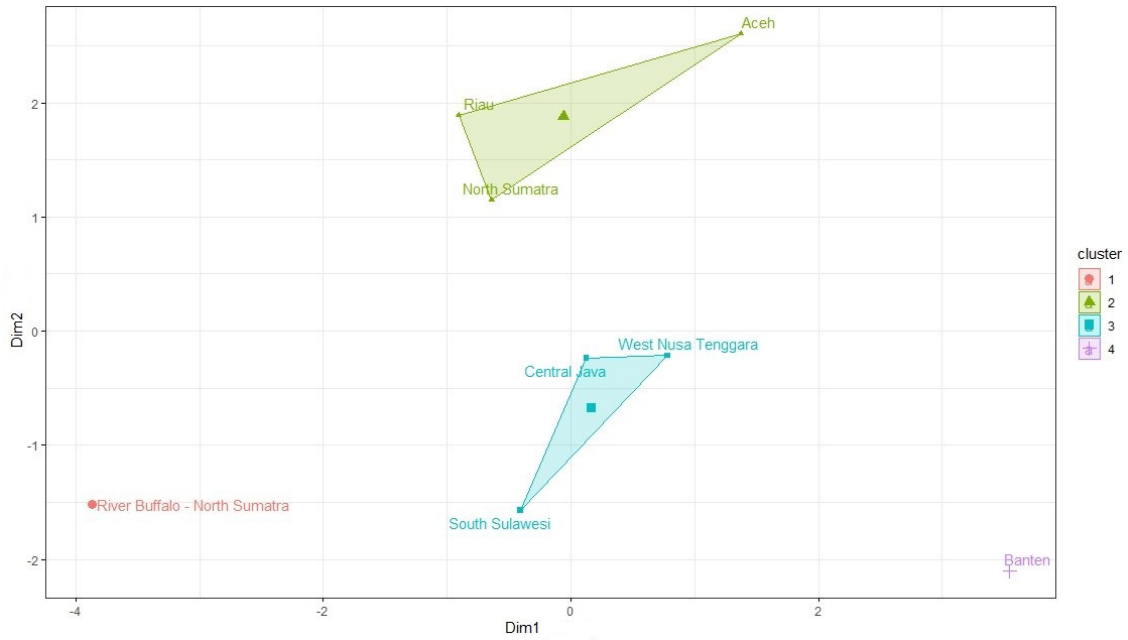

Figure 1 Hierarchical clustering on principal components of Indonesian buffaloes.

Conclusions

This study suggests that allele frequency data can be used to determine genetic relationships among population using HCPC method. The ILSTS61 and ILSTS17 loci can be used as genetic markers to determine the genetic relationship of Indonesian buffalo based on the PC1 and PC2 values. We also suggested to observe PC value, where the PC value can determine which the loci is influential in determining a genetic relationship.

References

[1] S Shamsalddini, MR Mohammadabadi and AK Esmailizadeh. Polymorphism of the prolactin gene and its effect on fiber traits in goat. *Russ*. *J. Genet*. 2016; **52**, 405-8.

[2] P Zamani, M Akhondi, MR Mohammadabadi, AA Saki, A Ershadi, MH Banabazi and AR Abdolmohammadi. Genetic variation of Mehraban sheep using two intersimple sequence repeat (ISSR) markers. *Afr*. *J*. *Biotechnol*. 2011; **10**, 1812-7.

[3] HK Koopaei, MRM Abadi, SA Mahyari, AR Tarang, P Potki and A Esmailizadeh. Effect of DGAT1 variants on milk composition traits in Iranian Holstein cattle population. *Anim. Sci. Pap. Rep.* 2012; **30**, 231-40.

[4] MN Ruzina, TA Shtyfurko, MR Mohammadabadi, OB Gendzhieva, T Tsedev and GE Sulimova. Polymorphism of the BOLA-DRB3 gene in the Mongolian, Kalmyk, and Yakut cattle breeds. *Russ. J. Genet.* 2010; **46**, 456-63.

[5] FG Gooki, M Mohammadabadi, MA Fozi and M Soflaei. Association of biometric traits with growth hormone gene diversity in Raini Cashmere goats. *Walailak J*. *Sci*. *Tech*. 2019; **16**, 499-508.

[6] M Pasandideh, MR Mohammadabadi, AK Esmailizadeh and A Tarang. Association of bovine PPARGC1A and OPN genes with milk production and composition in Holstein cattle. *Czech J*. *Anim*. *Sci*. 2015; **60**, 97-104.

[7] MR Nassiry, FE Shahroodi, J Mosafer, A Mohammadi, E Manshad, S Ghazanfari, MR Mohammadabadi and GE Sulimova. Analysis and frequency of bovine lymphocyte antigen (BoLA-DRB3) alleles in Iranian Holstein cattle. *Russ*. *J*. *Genet*. 2005; **41**, 664-8.

[8] FG Gooki, MR Mohammadabadi and MA Fozi. Polymorphism of the growth hormone gene and its effect on production and reproduction traits in goat. *Iran*. *J*. *Appl*. *Anim*. *Sci*. 2018; **8**, 653-59.

[9] MTV Ebrahimi, MR Mohammadabadi and AK Esmailizadeh. Using microsatellite markers to analyze genetic diversity in 14 sheep types in Iran. *Arch*. *Anim*. *Breed*. 2017; **60**, 183-9.

[10] MR Mohammadabadi, A Torabi, M Tahmourespoor, A Baghizadeh, A Esmailizadeh and A Mohammadi. Analysis of bovine growth hormone gene polymorphism of local and Holstein cattle breeds in Kerman province of Iran using polymerase chain reaction restriction fragment length polymorphism (PCR-RFLP). *Afr*. *J*. *Biotechnol*. 2010; **9**, 6848-52.

[11] Y Zhang, L Colli and JSF Barker. Asian water buffalo: Domestication, history and genetics. *Anim. Genet.* 2020; **51**, 177-91.

[12] FAO. Domestic Animal Diversity Information System (DAD-IS). Available at: http://www.fao.org/dad-is/en/, accessed June 2020.

[13] F Saputra, Jakaria and C Sumantri. Genetic variation of mtDNA cytochrome oxidase subunit I (COI) in local swamp buffaloes in Indonesia. *Media Peternakan* 2013; **36**, 165-70.

[14] M Rusdin, DD Solihin, A Gunawan, C Talib and C Sumantri. Genetic variation of eight Indonesian swamp-buffalo populations based on cytochrome b gene marker. *Trop. Anim. Sci. J.* 2020; **3**, 1-10.

[15] A Sukri, M Amin, A Winaya and A Gofur. Substitution and haplotype diversity analysis on the partial sequence of the mitochondrial DNA Cyt b of Indonesian swamp buffalo (*Bubalus bubalis*). *Biol. Med. Nat. Prod. Chem*. 2014; **3**, 59-63.

[16] F Saputra, Jakaria, A Anggraeni and C Sumantri. Genetic diversity of indonesian swamp buffalo based on microsatellite markers. *Trop. Anim. Sci. J.* 2020; **43**, 191-6.

[17] A Anggraeni, C Sumantri, L Praharani and E Andreas. Genetic distance estimation of local swamp buffaloes through morphology analysis approach. *Jurnal Ilmu Ternak dan Veteriner* 2011; **16**, 199-210.

[18] T Jombart, D Pontier and AB Dufour. Genetic markers in the playground of multivariate analysis. *Heredity* 2019; **102**, 330-41.

[19] M Argüelles, C Benavides and I Fernández. A new approach to the identification of regional clusters: hierarchical clustering on principal components. *Appl. Econ*. 2014; **46**, 2511-9.

[20] F Saputra, T Sartika, A Anggraeni, ABL Ishak, Komarudin and N Pratiwi. Multivariate analysis of five chicken breed in Indonesia based on microsatellite allele frequency. *Livest. Anim. Res.* 2021; **19**, 48-53.

[21] A Maugeri, M Barchitta, G Basile and A Agodi. Applying a hierarchical clustering on principal components approach to identify different patterns of the SARS-CoV-2 epidemic across Italian regions. *Sci Rep*. 2021; **11**, 7082.

[22] N Takezaki, M Nei and K Tamura. POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with windows *Interface. Mol. Biol. Evol.* 2010; **27**, 747-52.

[23] F Husson, J Josse, S Le and J Mazet. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. R package version 1.31.4. 2015; **24**, 1-18.

[24] A Kassambara and F Mundt. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. 2020.

[25] R Core T. R: A Language and Environment for Statistical Computing, Available at: https://cran.r-project.org/bin/windows/base/old/, accessed May 2020.

[26] W Zhang, X Gao, Y Zhang, Y Zhao, J Zhang, Y Jia, B Zhu, L Xu, L Zhang, H Gao, J Li and Y Chen. Genome-wide assessment of genetic diversity and population structure insights into admixture and introgression in Chinese indigenous cattle. *BMC. Genet.* 2018; **19**, 114.

[27] PP Agung, F Saputra, MSA Zein, AS Wulandari, WPB Putra, S Said and J Jakaria. Genetic diversity of Indonesian cattle breeds based on microsatellite markers. *Asian Australas. J. Anim. Sci.* 2019; **32**, 467-76.

[28] L Colli, M Milanesi, E Vajana, D Iamartino, L Bomba, F Puglisi, M Del Corvo, EL Nicolazzi, SSE Ahmed, Jesus R V Herrera, L Cruz, S Zhang, A Liang, G Hua, L Yang, X Hao, F Zuo, S-J Lai, S Wang, R Liu, Y Gong, M Mokhber, Y Mao, F Guan, A Vlaic, B Vlaic, L Ramunno, G Cosenza, A Ahmad, I Soysal, EÖ Ünal, M Ketudat-Cairns, JF Garcia, YT Utsunomiya, PS Baruselli, MEJ Amaral, R Parnpai, MG Drummond, P Galbusera, J Burton, E Hoal, Y Yusnizar, C Sumantri, B Moioli, A Valentini, A Stella, JL Williams and P Ajmone-Marsan. New insights on water buffalo genomic diversity and post-domestication migration routes from medium density SNP chip data. *Front. Genet.* 2018; **9**, 53.