ASAS-CSAS Annual Meeting & Trade Show
Animal Science and Technology: Ensuring Food Security
July 8-12, 2017 • Baltimore, Maryland

# 161: Population Structure in a Thai Multibreed Dairy Cattle Population

Thawee Laodim*, Mauricio A. Elzo†, Skorn Koonawootrittriron* and Thanathip Suwanasopee*
*Department of Animal Science, Faculty of Agriculture, Kasetsart University, Bangkok 10900, Thailand
† Department of Animal Sciences, University of Florida, Gainesville, FL 32611, USA

## ABSTRACT

Accounting for population structure is important to help identify SNP associated with production traits in domestic animals particularly in multibreed populations. Models used to identify relevant SNP in multibreed populations utilize genetic groups usually constructed based on expected breed fractions. However, these groups may not appropriately account for structural differences due to SNP allelic frequencies. Thus, the objectives of this study were to construct genetic groups using SNP marker information, obtain genetic distances between genetic groups, and determine the correspondence between SNP-based and breed-fraction based genetic groups. The study included 2,661 cattle (89 bulls and 2,572 cows) from 304 farms located in Central, Northern, Northeastern, and Southern regions of Thailand, with complete pedigree that had been genotyped with GeneSeek Genomic Profiler 9K. Only SNP with minor allele frequency higher than 0.01, call rate larger than 90%, P-value of Hardy-Weinberg equilibrium lower than 0.0001, and $r^2$ value of linkage disequilibrium lower than 0.2 were included in this study (n = 5,005). A principal component analysis was used to obtain eigenvectors that were subsequently utilized to assign animals to genetic groups using a k-means clustering algorithm. Computations were performed using the discriminant analysis of principal component (DAPC) program of R-package adegenet. The optimum number of genetic groups in this population based on the lowest Bayesian Information Criterion (BIC) value was 28. Genetic distances among these SNP-based genetic groups were estimated using Nei's genetic distance. The DAPC scatterplot of the first and second principal components showed four genetic groups clearly separated, and 24 genetic groups were very close to each other forming a "super cluster". Conversely, Nei's genetic distances among the 28 groups revealed 3 clusters, one containing group 23, a second one including groups 1, 2, 12, 20, 21,25, and 27, and a third cluster with the remaining groups. There was almost no correspondence (r = 0.00025) between breed composition of animals and their allocation to SNP-based genetic groups. In fact, SNP-based genetic groups contained animals of a wide range of Holstein fractions, and animals with Holstein fractions above 90% were represented in all SNP-based genetic groups. Thus, the DAPC algorithm was effective at identifying structural differences among animals based on gene frequencies regardless of their breed origin. However, genetic distances between these groups showed a different clustering pattern compared to the one obtained with the DAPC scatterplot of the first and second principal components.

## INTRODUCTION

Single nucleotide polymorphisms (SNP) play an important role in livestock genetic evaluation programs because it can increase the accuracy of estimated breeding values (EBV) for economically important traits. Additionally, SNP can help identify genes related to economically important traits across the genome in genome-wide association studies (GWAS). However, identification of SNP genotypes associated with milk production in GWAS depend on population structure and breeds of animals (Purfield et al., 2015). Thus, accounting for population structure is important to properly identify SNP associated with production traits in Thai multibreed populations. Previous models used to identify relevant SNP in the Thai multibreed dairy population accounted for population structure using the expected breed composition of animals. Breeds were defined as Holstein (H) and other breeds (O), thus animal breed composition was explained in terms of H and O fractions (Koonawootrittriron et al., 2009). However, animal expected breed composition may not appropriately account for structural differences due to SNP allelic frequencies of animals. ***Thus, the objectives of this study were to construct genetic groups using SNP marker information, obtain genetic distances between genetic groups, and determine the correspondence between SNP-based and breed-fraction based genetic groups.***

## MATERIALS AND METHODS

**Animals.** Animals used in this study (n = 2,661; 89 sires and 2,572 cows) were from 304 farms located in Central, Northeastern, Northern, and Southern Thailand. Most animals (98%) were crossbred and over 92% of them had Holstein fractions above 75% due to an ongoing upgrading program to Holstein. Other breeds represented in the Thai multibreed dairy population to a lesser extent than Holstein were Jersey, Brown Swiss, Red Danish, Sahiwal, Red Sindhi, Brahman, and Thai Native.

**Tissue sampling and genotyping.** Blood and semen samples were collected from 2,661 animals that had complete pedigree and phenotypic information. DNA was extracted from whole blood samples using a MasterPure$^{TM}$ DNA Purification kit for blood version II (EPICENTRE® Biotechnologies, USA), and from frozen semen using a GenElute$^{TM}$ Mammalian Genomic DNA Miniprep Kit (Sigma®, USA). DNA quantity and quality were assessed with a Thermo Scientific NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc., Wilmington, DE, USA). Genotyping was done by GeneSeek Inc. (Lincoln, NE, USA) using the Genomic Profiler 9K BeadChip. SNP genotypes located in the 29 autosomes and the X chromosome with minor allele frequencies higher than 0.01, call rates larger than 90%, P-values for Hardy-Weinberg equilibrium lower than 0.0001, and $r^2$ values of linkage disequilibrium lower than 0.2 were included in this study (n = 5,005).

**Genetic group based on discriminant analysis of principal components (DAPC) analysis.** Assignment of animals to genetic groups using genotypic information was performed using discriminant analysis of principal components (DAPC) with R-package adegenet (Jombart and Collins, 2015). The DAPC identified 2,000 principal components that explained approximately 98% of the variation among the 5,005 SNP in the 2,661 animals (**Figure 1A**). These 2,000 principal components were utilized to assign animals to genetic groups using a k-means clustering algorithm. Then, an optimum number of genetic clusters was determined using the lowest Bayesian Information Criterion (BIC) value from a set of clustering models with 1 to 100 genetic clusters (**Figure 1B**). Lastly, the 2,000 principal components obtained in the initial PCA analysis were used in a DAPC analysis to reexamine the assignment of animals in the Thai population to genetic clusters.

**Genetic distance analysis.** Genetic distances among the SNP-based genetic groups were estimated using Nei's genetic distance. Then, genetic distance values were used to construct a phylogenetic tree using a neighbor-joining method in software MEGA (Tamura et al., 2013).

ASAS-CSAS Annual Meeting & Trade Show
Animal Science and Technology: Ensuring Food Security
July 8 - 12, 2017 • Baltimore, Maryland

# 161: Population Structure in a Thai Multibreed Dairy Cattle Population

**Thawee Laodim*, Mauricio A. Elzo†, Skorn Koonawootrittriron* and Thanathip Suwanasopee***

*Department of Animal Science, Faculty of Agriculture, Kasetsart University, Bangkok 10900, Thailand

† Department of Animal Sciences, University of Florida, Gainesville, FL 32611, USA

## RESULTS AND DISCUSSION

**Figure 1A** shows the cumulative variance explained by retained principal components as their number increased from 1 to 2500. A total of 2000 principal components accounting for 98% of the variance were utilized to assign animals to genetic groups using a k-means clustering algorithm. **Figure 1B** shows the value of BIC depending on number of genetic clusters ranging from 1 to 100. The optimum number of genetic clusters in the Thai dairy cattle population was 28 (smallest BIC value; highlighted with a red oval). The scatterplot of the first and second principal components of the DAPC analysis showed four genetic groups clearly separated (3, 13, 20, and 21), and 24 genetic groups were very close to each other forming a "super cluster" (**Figure 2**). This finding was in agreement with the five distinct genetic groups identified in the Nguni cattle population of South Africa using SNP genotypic information (Wang et al., 2015). It may be possible that genetic groups 20 and 21 represent clusters that recently diverged from the super cluster (Jonker et al., 2013). Thus, because of the proximity of genetic groups 20 and 21 to the "super cluster", they were combined into a single genetic group.

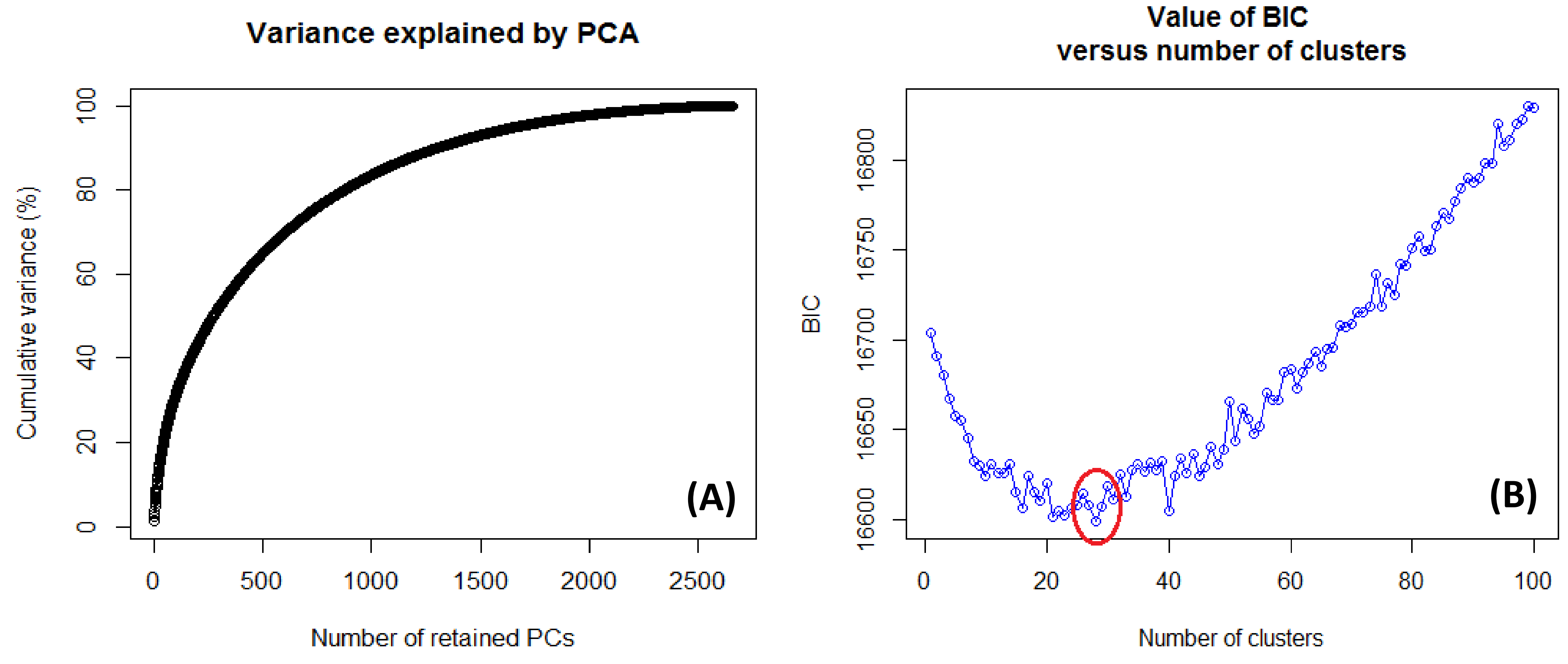

**Figure 1** Variance explained by retained principal components **(A)** and Bayesian Information Criterion (BIC) values for models with 1 to 100 genetic clusters obtained with a k-means clustering algorithm with 2,000 principal components **(B)**

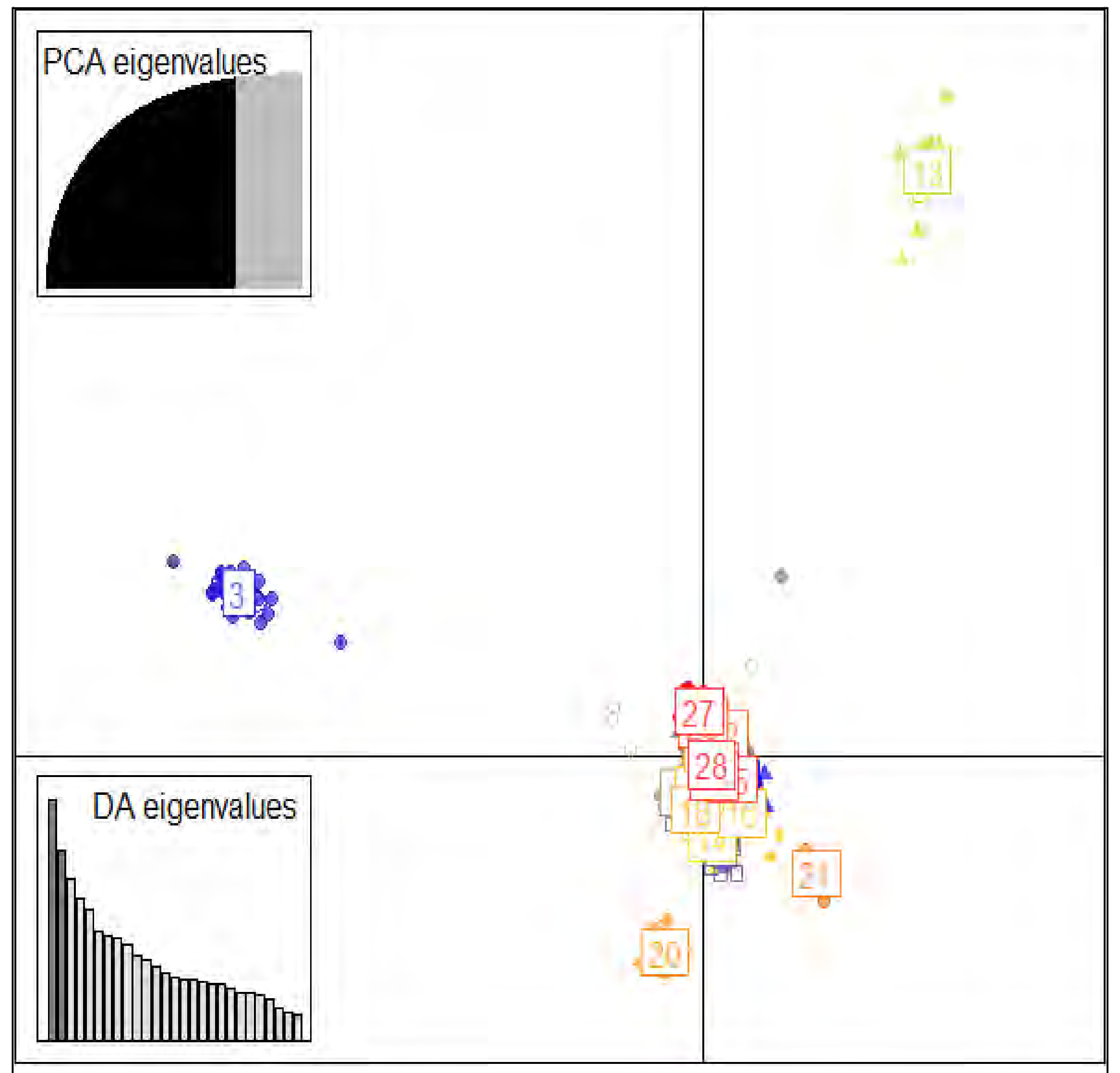

**Figure 2** Scatterplot of the first and second principal components of DAPC

Nei's genetic distances among the 28 groups revealed 3 clusters, one containing group 23, a second one including groups 1, 2, 12, 20, 21,25, and 27, and a third cluster with the remaining groups (**Figure 3**). The Holstein fraction of animals in each of the 28 genetic clusters (**Figure 4**). The correlation between genetic cluster and Holstein fraction was close to zero (r = 0.00025) indicating that there was no correspondence between breed composition of animals and their allocation to SNP-based genetic groups. The SNP-based genetic clusters contained animals of a wide range of Holstein fractions, and animals with Holstein fractions above 90% were represented in all SNP-based genetic clusters.

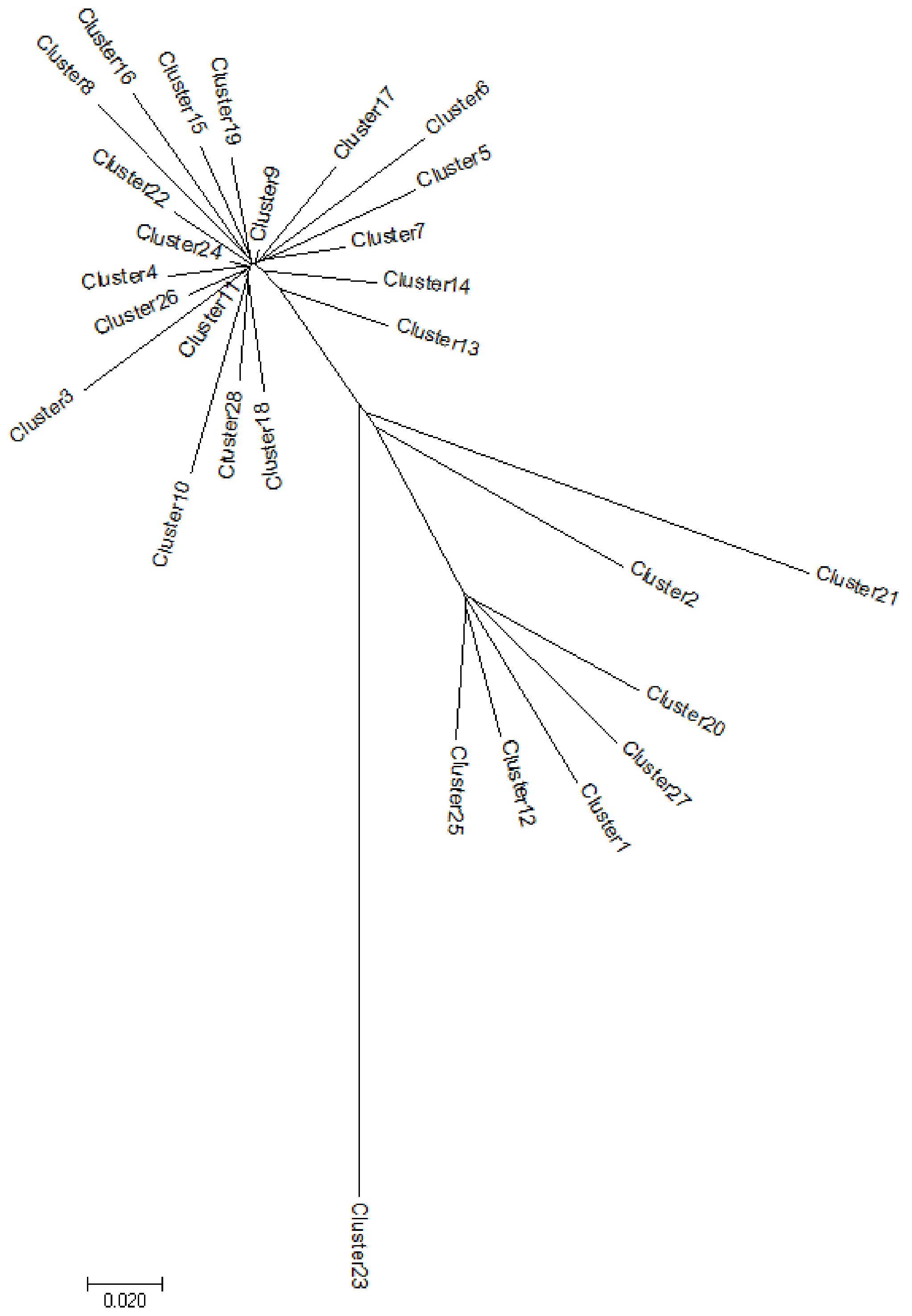

**Figure 3** Nei's genetic distances among genetic clusters 1 to 28

## FINAL REMARKS

- The DAPC algorithm was effective at identifying structural differences among animals based on gene frequencies regardless of their breed of origin
- Nei's genetic distances between the 28 genetic clusters showed a different clustering pattern from the one obtained with the DAPC scatterplot of the first and second principal components

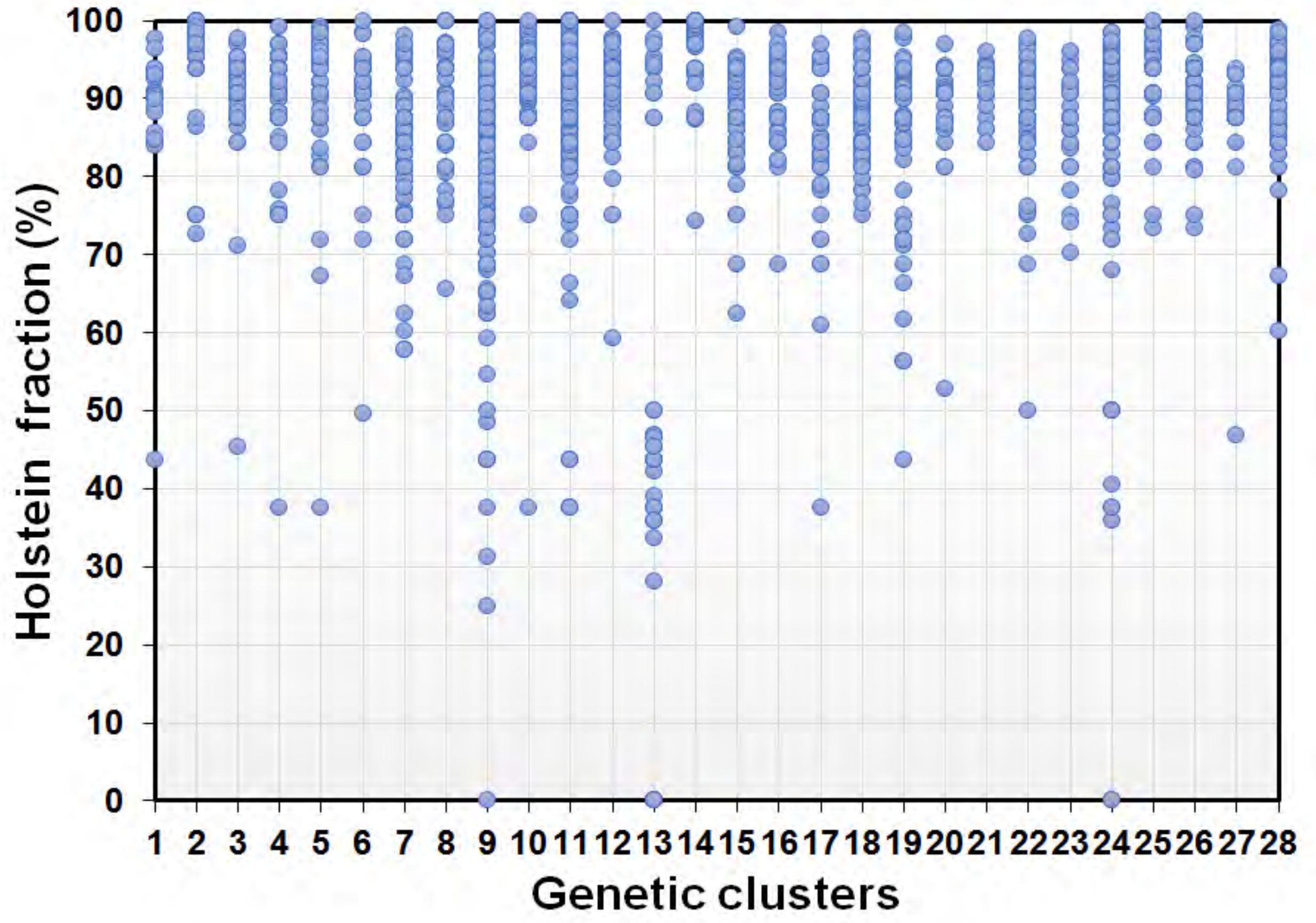

**Figure 4** Percent Holstein of animals in genetic clusters 1 to 28

## REFFERRENCES

Jombart, T. and Collins, C. 2015. Imperial College London MRC Centre for Outbreak Analysis and Modelling, London, UK.

Jonker, B.M., Kraus, R.H.S., Zhang, Q., van Hooft, P., Larsson, K., van der Jeugd, H.P., Kurvers, R.H.J.M., van Wieren, S.E., Loonen, M.J.J.E., Crooijmans, R.P.M.A., Ydenberg, R.C., Groenen, M.A.M. and Prins, H.H.T. 2013. Mol. Eco. 22, 5835-5847.

Koonawootrittriron, S., Elzo, M.A., Thongprapi, T. 2009. Livest. Sci. 122, 186-192.

Purfield, D.C., Bradley, D.G., Evans, R.D., Kaerney, F.J., Berry, D.P. 2015. Genet. Sel. Eval. 47, 47.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S. 2013. Mol. Biol. Evol. 30, 2725-2729.

Wang, M.D., Dzama, K., Hefer, C.A. and Muchadeyi, F.C. 2015. BMC Genomics 16, 894.