Identification of SNP markers associated with milk and fat yields in multibreed dairy cattle using two genetic group structures

Thawee Laodim[a], Mauricio A. Elzo[b], Skorn Koonawootrittriron[a*], Thanathip Suwanasopee[a], and Danai Jattawa[a]

[a] Department of Animal Science, Kasetsart University, Bangkok 10900, Thailand
[b] Department of Animal sciences, University of Florida, Gainesville, FL 32611-0910, USA

[*] Department of Animal Science, Faculty of Agriculture, Kasetsart University, Bangkok 10900, Thailand; Tel: +66 2 5791120; Fax: +66 2 5791120; Email: agrskk@ku.ac.th (Skorn Koonawootrittriron)

Abstract

The objective of this research was to determine the correspondence between significant SNP for first-lactation 305-d milk and 305-d fat yields and associated genes from mixed models accounting for population structure using all additive relationships among animals and genetic groups constructed with either SNP genotypic information or with expected breed composition in the Thai multibreed dairy cattle population. The dataset contained 2,410 MY and 912 FY from 2,410 first-lactation cows with complete pedigree information genotyped with GeneSeek Genomic Profiler 9K. SNP genotypes located in autosomes and the X chromosome, with call rates larger than 90%, minor allele frequencies (MAF) larger than 0.01, and P-values for Hardy-Weinberg Equilibrium tests larger than 0.00001 were used in the research. Significant SNP for MY and FY were identified using two mixed models that differed only in their definition of genetic groups. Model 1 (M1) defined genetic groups in terms of breed composition and model 2 (M2) in terms of SNP genotypic information. Fixed effects in M1 and M2 were contemporary group (herd-year-season), genetic group, heterosis, and calving age. Random effects were animal additive genetic and residual. Significant SNP markers were used to identify genes using R package Map2NCBI. Molecular function and biological processes of genes identified by significant SNP markers located inside or within 2,500 bp of these genes were obtained via program PANTHER. Both models yielded identically high correlations between number of significant SNP and number of genes per chromosome for MY (r = 0.97) and FY (r = 0.99). Over 60% of genes associated with MY and FY were involved in binding and catalytic activities. Similarly, over 50% of genes associated with MY and FY participated in cellular and metabolic processes. Larger numbers of significant SNP and genes were identified with M2 for MY and with M1 for FY. However, considering both traits, M1 identified more significant SNP and genes than M2 for MY and FY in this Thai multibreed dairy population. Genes associated with MY and FY were primarily involved in binding and catalytic activities as well as in cellular and metabolic processes. Genes identified to be important for MY and FY in the Thai multibreed population differed substantially from those identified in *Bos taurus* breeds in temperate environments indicating the need to continue to conduct studies with high-density genotyping chips that identify sets of genes relevant to MY and FY in populations of different breed composition under a variety of environmental conditions.

Key words: genetic group structures, significant SNP, genome, dairy cattle, tropical regions

1. Introduction

Single nucleotide polymorphisms (SNP) play an important role in livestock genetic evaluation programs because they can help increase the accuracy of animal genomic predictions and genomic selection for economically important traits (Zhang et al., 2014). Additionally, SNP markers can help identify genes affecting economically important traits across the genome in genome-wide association studies (GWAS). However, identification of SNP genotypes associated with milk production in GWAS depends on population stratification (Price et al., 2006; Ma et al., 2012) and breeds of animals (Bush and Moore, 2012; Purfield et al., 2015). Shin and Lee (2015) suggested that mixed model methodology could be used to account for the effect of population stratification in GWAS. In addition, genomic differences between cattle breeds may affect the significance of specific SNP genotypes (Purfield et al., 2015). Within-breed GWAS have identified sets of significant SNP for dairy production traits in various dairy breeds including Holstein (Reven et al., 2014; Nayeri et al., 2016), Jersey (Reven et al., 2014), Nordic Red (Iso-Touru et al., 2016), and Brown Swiss (Guo et al., 2012) under temperate environmental conditions. However, these within-breed sets of significant SNP identified under temperate conditions will likely differ from sets of significant SNP associated with dairy production traits in the Thai multibreed dairy population under tropical environmental conditions. The Thai multibreed dairy population is the product of an upgrading mating strategy of multiple *Bos taurus* and *Bos indicus* breeds to Holstein aimed at producing animals with high milk yield (primarily due to their Holstein fraction) and high adaptability to tropical environment conditions (due to their native and other *Bos indicus* breed fractions). Most animals (91%) in the Thai multibreed population are 75% Holstein or greater and some animals have as many as eight different cattle breeds represented in them (Koonawootrittriron et al., 2009; Ritsawai et al., 2014).

The only genome-wide association study for milk production traits in the Thai multibreed dairy cattle population accounted for population structure using the expected breed composition of animals (Yodklaew et al., 2014). Breeds were defined as Holstein (H) and other breeds (O), thus animal breed composition was explained in terms of H and O fractions. However, animal expected breed composition may not entirely account for structural differences due to SNP allelic frequencies of animals. Alternatively, principal components analysis of genotypic data successfully corrected for population stratification in GWAS (Price et al., 2006; Ma et al., 2012). In addition, discriminant analysis of principal components of genotypic data (DAPC; Jombart et al., 2010) was shown to be an effective method to identify genetic groups of related individuals in a population. Thus, the objective of this research was to determine the correspondence between significant SNP for first-lactation 305-d milk and 305-d fat yields and associated genes from mixed models accounting for population structure using all additive relationships among animals and genetic groups constructed with either DAPC of genotypic information or with expected breed composition in the Thai multibreed dairy cattle population.

2. Materials and methods

2.1. Animals and management

Animals used in this research (n = 2,661; 89 sires and 2,572 cows) were from 310 farms located in four regions of Thailand (Central, Northeastern, Northern, and Southern). Climate seasons were categorized as winter (November to February), summer (March to June), and rainy (July to October). Cows were housed in open barns and provided with fresh grass (30 to 40 kg/day for fresh grass via cut and carry), concentrate (5 to 10 kg/day or 2 kg of concentrate per 1 kg/milk produced), and mineral supplement. Concentrate (crude protein: 14 to 22%; nitrogen-free extract: 63 to 83%) was provided after milking in the morning (4:30 to 7:00 a.m.) and in the afternoon (14:30 to 16:30 p.m.). Water was available at all times. Instead of cut and carry, some farms placed cows directly on grass pastures. Available grasses included Napier grass (*Pennisetum purpureum*), Guinea grass (*Panicum maximum*), Ruzi grass (*Brachiaria ruziziensis*), or Para grass (*Brachiaria mutica*). As the quantity of fresh grass decreased in summer and winter, farmers fed cows agricultural byproducts (rice straw, pineapple waste, and sweet corn cob or husk), hay, and silage. Cows artificially inseminated with either pre-chosen Holstein or Holstein crossbred bulls or with semen from other bulls available at the time of the insemination (Koonawootrittriron et al., 2009).

2.2. Traits

First-lactation monthly test-day milk and fat yields as well as milk samples were collected monthly from 2,410 cows between 1997 and 2014. Cows were the progeny of 442 sires and 2,235 dams. Monthly test-day milk and fat records were used to compute 305-d milk yields (MY; kg) and 305-d fat yields (FY; kg) using a test-interval procedure (Sargent et al., 1968; Koonawootrittriron et al., 2002). The dataset contained 2,410 MY and 912 FY first-lactation records.

2.3. Tissue sampling and SNP genotyping

Blood and semen samples were collected from 2,661 animals (89 sires and 2,572 dams) that had complete pedigree and phenotypic information. Genomic DNA was extracted from whole blood using a MasterPure™ DNA Purification kit for blood version II (EPICENTRE® Biotechnologies, USA) and from frozen semen using a GenElute™ Mammalian Genomic DNA Miniprep Kit (Sigma®, USA). The quantity and quality of extracted DNA were assessed with a Thermo Scientific NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc., Wilmington, DE, USA). The minimum concentration of DNA in samples was 15 ng/µl with an absorbance ratio of approximately 1.8 at 260/280 nm. Dried DNA subsamples of 50 µl were sent to GeneSeek for genotyping with GeneSeek Genomic Profiler 9K BeadChip (GeneSeek Inc., Lincoln, NE, USA).

SNP genotypes located in the 29 autosomes and the X sex chromosome were included in this study. Quality control and pruning of SNP genotypes was performed via PLINK version 1.9. (Purcell et al., 2007; Purcell and Chang, 2017). SNP genotypes with call rates lower than 90%, minor allele frequencies (MAF) lower than 0.01, and P-values for Hardy-Weinberg Equilibrium tests lower than 0.00001 were excluded. After applying these quality control criteria, 7,720 SNP were left for this research.

2.4. Genome-wide association analyses

2.4.1. Genetic group definitions for genome-wide association models

Genetic groups in the Thai multibreed dairy population were explained using two different approaches, one based on animal expected breed composition, and another one based on information from a panel of 5,005 SNP markers.

2.4.1.1. Genetic groups based on expected breed composition

There were eight breeds represented in animals from the Thai population to various extents (Holstein, Jersey, Brown Swiss, Red Danish, Sahiwal, Red Sindhi, Brahman, and Thai Native). However, due to the existing upgrading program to Holstein, 98% of the animals in the dataset were crossbred and over 92% of them had Holstein fractions above 75%. Thus, for the purposes of the statistical analysis, breeds were defined as Holstein and Other breeds (Koonawootrittriron et al., 2009), where Other breeds comprised all breeds except for Holstein. Consequently, genetic groups based on animal breed composition for GWAS were defined as linear functions of expected Holstein and Other breeds fractions of animals in the dataset.

2.4.1.2. Genetic groups based on SNP genotypic information

Assignment of animals to genetic groups using genotypic information was performed using discriminant analysis of principal components (DAPC; Jombart et al., 2010) with R-package adegenet (Jombart and Collins, 2015). SNP genotypes from 2,661 animals with a linkage disequilibrium r^2 value lower than 0.2 were included in this analysis (n = 5,005). This editing was done to increase the chance of identifying animals belonging to different subpopulations within the Thai multibreed dairy population. The DAPC identified 2,000 principal components that explained approximately 98% of the variation among the 5,005 SNP in the 2,661 animals. These 2,000 principal components were utilized to assign animals to genetic groups using a k-means clustering algorithm. Then, an optimum number of 28 genetic clusters in this population was determined using the lowest Bayesian Information Criterion (BIC) value from a set of clustering models with 1 to 100 genetic clusters (Figure 1A). Lastly, the 2,000 principal components obtained in the initial PCA analysis were used in a DAPC to reexamine the assignment of animals in the Thai population to these 28 genetic clusters. Figure 2 shows a DAPC scatterplot containing three distinct groups of genetic clusters, where group 1 includes cluster 3, group 2 has cluster 13, and group 3 captured the remaining 26 clusters. These three groups were used as the set of subclass SNP-based genetic groups for the genome-wide association analysis.

2.4.2. Genome-wide association models

Software Qxpak.5 (Pérez-Enciso and Misztal, 2011) was utilized to identify significant SNP for MY and FY using two mixed models that differed in their approach to account for genetic groups in the Thai multibreed population. Model 1 (M1) explained genetic groups as a linear function of the expected breed composition of animals using a regression approach. Model 2 (M2) explained genetic groups using a discriminant analysis of principal components approach to assign animals to distinct genetic groups based on SNP genotypic information. Models 1 and 2 can be represented as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e},$$

where **y** was a vector of cow MY and FY, **b** was a vector of fixed effects that included contemporary groups (herd-year-season) subclass effects, genetic groups defined as linear functions of expected H and O fractions (M1) or as three subclass SNP-based genetic groups (M2), heterosis regression effects as functions of heterozygosities (computed as expected fraction of H in the sire times expected fraction of O in the dam plus expected fraction of O in the sire times expected fraction of H in the dam), calving age regression effects, and SNP genotype (11, 12, and 22) subclass effects, **a** was a vector of random animal additive genetic effects, and **e** was a vector of random residuals. Incidence matrices **X** and **Z** related MY and FY records to elements of vector **b** and **a**, respectively.

Significant SNP for MY and FY were those that were significant at P < 0.001 (F-test) and above the threshold provided by the false discovery rate (FDR) when SNP markers were ordered from lowest to highest P-value. Thus, number of a true positive SNP was equal to 1 - FDR times the number of significant SNP at P < 0.001 for each trait.

The FDR was calculated as follows (Bolormaa et al., 2013):

$$\mathbf{FDR} = \frac{[\mathbf{P}(1 - \frac{\mathbf{n}}{\mathbf{N}})]}{[\frac{\mathbf{n}}{\mathbf{N}}(1 - \mathbf{P})]},$$

where **P** was equal to 0.001, the probability value used to detect the significance of each SNP marker, **n** was the number of SNP that were significant at P < 0.001, and **N** was the total number of SNP tested (n = 7,720).

2.4.3. Description of genes in terms of molecular function and biological processes

The position of significant SNP markers for MY and FY in base pairs was used to locate genes or nearby genes in the UMD *Bos taurus* 3.1 assembly of the bovine genome at the National Center for Biotechnology Information (NCBI) using R package Map2NCBI (Hanna and Riley, 2014). This research focused on molecular functions and biological processes for genes identified by SNP by M1 and M2 located inside or within 2,500 bp of these genes. The molecular function and biological processes of all NCBI genes associated with MY and FY were searched in the *Bos taurus* Gene Ontology database (Mi et al., 2013) using program PANTHER (http://www.pantherdb.org/), where molecular function refers to the biochemical activity of a gene product, and biological processes are determined by functional activities of multiple gene products (The Gene Ontology Consortium, 2000).

3. Results and discussion

3.1. Genetic groups based on SNP genotypic information

Figure 1A presents the value of the Bayesian Information Criterion (BIC) for models with 1 to 100 genetic clusters in the Thai dairy cattle population. The optimum number of genetic clusters in this population, indicated by the smallest BIC value, was 28. Figure 1B shows the Holstein fraction of animals in each of the 28 genetic clusters. 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 (Figure 1B). 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. This indicated that the SNP-based genetic clusters were likely produced by differences in gene frequencies among the up to eight *Bos taurus* (Holstein, Jersey, Brown Swiss, Red Danish) and *Bos indicus* (Sahiwal, Red Sindhi, Brahman, and Thai Native) breeds represented in animals from the Thai multibreed population. Thus, the SNP-based genetic clusters appeared to have accounted for differences in the genetic background of animals beyond those explained by additive genetic relationships among animals in the pedigree (Martin-Burriel et al., 2011; Wang et al., 2015).

The scatterplot of the first and second principal component of the DAPC analysis showed four distinct genetic groups with one cluster each (3, 13, 20, and 21), and a fifth genetic group formed a super cluster formed by 24 clusters very close to each other (Figure 2). It is possible that genetic groups 20 and 21 represent clusters that may have recently diverged (Jonker et al., 2013) from the super cluster. However, because of the proximity of clusters 20 and 21 to the other 22 clusters in the super cluster, all of them were combined into a single genetic group.

3.2. Number of significant SNP and genes associated with MY and FY

Table 1 shows the number of significant SNP for MY and FY at P-value ≤ 0.001 after correcting for FDR based on distance between SNP genotypes and genes in the NCBI database in M1 and M2. False discovery rates for MY (M1: 1.87%; M2: 1.77%) were less than half the values for FY (M1: 3.63%; M2: 5.94%). This occurred because the number of significant SNP for MY was substantially larger than the corresponding number for FY (nearly twice as large for M1 and over three times as large for M2; Table 2). Lower FDR and larger number of significant SNP for MY than for FY were also obtained in straightbred Holstein and Jersey (Bolormaa et al., 2010; Pryce et al., 2010) and buffalo populations (Venturini et al., 2014).

Model 1 identified lower number of significant SNP for MY (385) than M2 (406), but the reverse occurred for FY, where numbers of significant SNP were 199 for M1 and 120 for M2. Perhaps the number of significant SNP identified by Model 2 could have been higher if animals in this study would have been genotyped with a higher-density chip than 9k. Utilization of a higher-density chip may have uncovered a larger number of SNP markers. However, the two models yielded nearly identical percentages of significant SNP markers for MY and FY located inside genes, within 2,500 bp, between 2,500 and 5,000 bp, between 5,000 and 25,000 bp, and beyond 25,000 bp (Table 1). Thus, although M1 and M2 differed in the number of significant SNP identified, the proportion of SNP in each distance category was similar in both models. The largest fraction of the significant SNP were either inside genes or within 2,500 bp of genes in the NCBI database for MY (42% for M1; 42% for M2) and FY (40% for M1; 39% for M2).

Numbers of significant SNP ($P \leq 0.001$ corrected for FDR) for MY and FY per chromosome from M1 and M2 are shown in columns 2 to 5 in Table 2. Significant SNP for MY and FY were found in all autosomes and the X chromosome. Numbers of significant SNP per chromosome for MY ranged from 2 (chromosomes 24, 25, and 27) to 28 SNP (chromosome 20) for M1, and from 2 (chromosomes 24 and 27) to 29 SNP (chromosome 20) for M2. Narrower ranges of numbers of SNP per chromosome existed for FY due to their lower number of significant SNP in models 1 and 2. Numbers of significant SNP per chromosome for FY ranged from 1 (chromosome 16, 22, and 25) to 15 SNP (chromosome 6) for M1, ranged from 0 (chromosome 29) to 12 SNP (chromosome 6) for M2.

Number of genes associated with MY and FY per chromosome are shown in columns 6 to 9 in Table 2. Genes associated with significant SNP were also present in all chromosomes. A wide distribution of genes associated with MY and FY across chromosomes was also found previously in Holstein (Edwards et al., 2015). Numbers of genes associated with MY per chromosome ranged from 2 (chromosomes 24, 25, and 27) to 27 genes (chromosome 9) for M1 and ranged from 2 (chromosomes 24 and 27) to 24 genes (chromosomes 2 and 9) for M2. Corresponding numbers of genes associated with FY per chromosome ranged from 1 (chromosomes 16, 22, and 25) to 15 genes (chromosome 6) for M1 and ranged from 0 (chromosome 29) to 12 genes (chromosome 6) for M2. The chromosomes with the largest number of genes associated with MY were chromosomes 2 (24 for M2) and 9 (27 for M1 and 24 for M2) and the corresponding chromosome for FY was chromosome 6 (15 for M1 and 12 for M2).

Models 1 and 2 yielded identically high correlation values between number of significant SNP and number of genes per chromosome for MY ($r = 0.97$; P-value < 0.00001) and FY ($r = 0.99$; P-value < 0.00001). These high correlations indicated that the vast majority of significant SNP for MY or FY in M1 and M2 pointed at a single gene within each chromosome. This was likely the result of the low number of SNP used in this study (7,720) relative to the approximately 20,000 genes (Michelizzi et al., 2011) present in the bovine genome.

3.3. Molecular function and biological processes associated with genes identified by SNP inside or within 2,500 bp of these genes

Table 3 shows the number of genes associated with MY ($P < 0.001$; FDR = 1.87 for M1 and 1.77 for M2) and FY ($P < 0.001$; FDR = 3.63 for M1 and 5.94 for M2) identified by SNP genotypes inside genes and within 2,500 bp of genes in the NCBI database with M1 and M2. The number of genes associated with MY was similar for M1 (151) and M2 (158). Conversely, a larger number of genes for FY was identified with M1 (78) than with M2 (46). Information on all the significant SNP markers located either inside or within 2,500 bp associated genes are in csv attachment file 1 for MY with M1, file 2 for MY with M2, file 3 for FY with M1, and file 4 for FY with M2. Information contained in these tables include SNP name, chromosome, position in bp, P-value, gene name, gene description, distance from gene in bp, molecular function, and biological processes.

Figure 3A shows the proportion of genes associated with MY, and Figure 4A shows the proportion of genes associated with FY based on molecular function. Approximately 65% of genes associated with MY and FY were involved in either binding activities (MY: 34% for M1 and 33% for M2; FY: 35% for M1 and 31% for M2) and (or) catalytic activities (MY: 32% for M1 and 33% for M2; FY: 28% for M1 and 31% for M2). The remaining 35% of genes associated with MY and (or) FY were involved in structural molecule, transporter, receptor, signal transducer, channel regulator, and (or) signal transducer activities (Figures 3A and 4A).

Large fractions of genes associated with MY involved in cell binding activities were also reported in Holstein (Yang et al., 2009; Yang et al., 2015a), and Sahiwal cattle (Janjanam et al., 2014). Binding activities include DNA binding, RNA polymerase binding, and transcription-factor binding (Tripathi et al., 2013) that are essential for molecular interactions between cells such as cell-cell signaling, cell adhesion, and signal transduction related to growth development and remodeling of the mammary gland (Janjanam et al., 2014; Yang et al., 2009). The fractions of genes associated with MY involved in catalytic activities in the Thai multibreed dairy population (Figure 3A) were similar to previous findings in Holstein (Yang et al., 2009; Yang et al., 2015a), and Sahiwal cattle (Janjanam et al., 2014). Catalytic activities and other molecular activities such as cell structure and transporter, receptor, signal transducer, channel regulator, and signal transducer activities are essential for biological processes related to milk and fat production in cattle, particularly in cells of the mammary gland (Faria et al., 2012; Ghorbani et al., 2015; Janjanam et al., 2014).

The proportions of genes identified by M1 and M2 involved in various biological processes are shown in Figure 3B for MY and Figure 4B for FY. Approximately 55% of the genes associated with MY and FY identified by M1 and M2 were involved in either cellular processes (MY: 28% for M1 and 29% for M2; FY: 32% for M1 and M2) and (or) metabolic processes (MY: 19% for M1 and 21% for M2; FY: 24% for M1 and 20% for M2). The other 45% of genes related to MY or FY were involved in developmental, multicellular organismal, cellular component organization or biogenesis, localization, biological regulation, response to stimuli, immune system, biological adhesion, reproduction, and locomotion processes (Figures 3B and 4B). Cellular and metabolic processes, and to a lesser extent with cell communication, transport, and biogenesis processes were also found to be important for milk yield and milk components in Sahiwal and Holstein cattle (Dai et al., 2017; Janjanam et al., 2014). Genes involved in cellular and metabolic processes such as energy storage, glycolysis, and glycogen metabolism were found to be essential for cell proliferation in the mammary gland during pregnancy and lactation in Holstein (Weikard et al., 2012; Yang et al., 2015b).

Table 4 shows the top 20% of SNP markers and associated genes in common for MY located inside or within 2,500 bp of these genes (n = 23) and FY (n = 7) across models ordered by P-value from lowest to highest after applying FDR within each trait and model. Genes associated with MY were located in chromosomes 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 20, 22, 26, and X (Table 4) and genes associated with FY were distributed in seven different chromosomes (4, 10, 15, 17, 22, 28, and X; Table 4). Except for one gene associated with FY (DNAH11; Cochran et al., 2013), none of the genes identified by the top 20% of SNP in common across models to be associated with MY or FY in this population were found to be associated with these same traits in various *Bos taurus* breeds under temperate environmental conditions (Holstein: Jiang et al., 2010; Raven et al., 2014; Nayeri et al., 2016; Nordic Red: Iso-Touru et al., 2016; Jersey: Raven et al., 2014; and Braunvieh: Maxa et al., 2012).

The genes in the top 20% for MY in the Thai *Bos taurus-Bos indicus* multibreed population (Table 4) were UFM1 specific ligase 1 (UFL1), pecanex homolog (PCNX), cadherin 18 (CDH18), laminin subunit alpha 4 (LAMA4), integrin subunit alpha 9 (ITGA9), unc13 homolog 3-like (UNC13C), mitogen-activated protein kinase 15 (MAP3K15), peptidoglycan recognition protein 4 (PGLYRP4), pleckstrin homology domain containing M3 (PLEKHM3), G protein-coupled receptor 160 (GPR160), spermatogenesis associated 16 (SPATA16), calcium voltage-gated channel auxiliary subunit gamma 2 (CACNG2), GLIS family zinc finger 3 (GLIS3), protein tyrosine phosphatase, receptor type E (PTPRE), OCA2 melanosomal transmembrane protein (OCA2), iodothyronine deiodinase 1 (DIO1), dynein axonemal heavy chain 11 (DNAH11), alkylglycerone phosphate synthase (AGPS), mitogen-activated protein kinase 5 (MAP3K5), far upstream element binding protein 3 (FUBP3), polypeptide N-acetylgalactosaminyltransferase-like 6 (GALNTL6), kynurenines (KYNU), and EPH receptor A5 (EPHA5). The corresponding genes for FY (Table 4) were ankyrin-3 (LOC100337251), LPS responsive beige-like anchor protein (LRBA), sortilin related receptor 1 (SORL1), C1GALT1 specific chaperone 1 (C1GALT1C1), gephyrin (GPHN), contactin associated protein-like 2 (CNTNAP2), and unc-51 like kinase 4 (ULK4).

Products of these genes were important for binding (4 genes), catalytic (5 genes), signal transducer (3 genes), transporter (3 genes), receptor (3 genes), and channel regulator functions (1 gene; Table 4). In addition, the products of these genes were involved in developmental (4 genes), cellular (9 genes), biological regulation (3 genes), metabolic (11 genes), response to stimulus (4 genes), biogenesis (1 gene), and localization (3 genes) processes (Table 4). Seventeen genes had unknown molecular function and fourteen genes were had no biological process associated with them. Considering the numbers of genes per molecular function and biological process, products of genes in the top 20% for MY and FY in the Thai dairy population with binding, catalytic, transducer, transporter, and receptor functions were involved primarily in metabolic and cellular biological processes and secondarily in developmental, biological regulation, and response to stimulus processes. These biological processes were important for both MY and FY.

As indicated above, DNAH11 was the only gene previously reported to be associated with one of the two dairy traits here (FY) in another cattle population (US Holstein; Cochran et al., 2013) under temperate conditions. In addition, gene DNAH11 was also reported to be associated with daughter pregnancy rate, cow conception rate, and heifer conception rate in US Holstein (Ortega et al., 2016). Six other genes in the top 20% for MY and FY in the Thai dairy population were found to be associated with other traits in various dairy and beef cattle breeds. Gene MAP3K5 was associated with lactation persistency in Canadian Holstein (Do et al., 2017) and calf birth weight in US Holstein (Cole et al., 2013). Gene KYNU was associated with somatic cell count in Canadian Holstein (Chen et al., 2015). Gene EPHA5 was associated with feed conversion ratio in Brazilian Nellore cattle (Santana et al., 2016). Gene ITGA9 was associated with respiration rate during climatic stress in US Angus, Simmental, and Piedmontese (Howard et al., 2014). Gene GALNTL6 was associated with cull-cow carcass weight in Ireland Holstein-Friesian (Doran et al., 2014) and with myristic saturated fatty acid content in Brazilian Nellore (Lemos et al., 2016). Gene LAMA4 was associated with meat quality traits in Mongolian Simmental (Xia et al., 2016) and marbling score in Korean Hanwoo cattle (Sudrajad et al., 2016). Lastly, gene CNTNAP2 was associated with linolenic acid in Brazilian Nellore (Lemos et al., 2016).

Milk yield and FY in the Thai multibreed population were influenced by a set of genes that had not (except for one for FY) been previously reported to be associated with these traits in *Bos taurus* populations under temperate environmental conditions. This may have occurred because of the lower density of the chip used here (9k) vs. other studies (50k, 80k, and 770k), differences in gene frequencies in the Thai *Bos taurus* and *Bos indicus* multibreed dairy population and gene frequencies in Holstein and other *Bos taurus* breeds. The vastly different environmental conditions in Thailand (tropical climate, open-housing, nutrition based on local roughage, concentrate and byproducts) may have affected the expression of genes relevant to MY and FY. Allowing for differences in identification of genes important for MY and FY due to SNP marker density, results here indicated that the sets of genes important for MY and FY under tropical conditions in Thailand may be substantially different from those in other dairy populations under temperate environments. Thus, it is likely that the combined effect of genetic and environmental factors determined different sets of genes to be more relevant in each dairy population-environment combination. This points out the need to continue to conduct studies that identify sets of genes relevant to MY and FY in populations of different breed composition under a variety of environmental conditions.

4. Conclusions

Considering both MY and FY, model 1 (genetic groups based on expected breed composition) identified more significant SNP and genes than model 2 (genetic groups based on SNP genotypic information). However, both models exhibited high correlations between number of significant SNP and number of genes per chromosome for MY and FY. These genes were primarily involved in binding and catalytic activities as well as in cellular and metabolic processes. Nearly all genes associated with MY and FY in the Thai multibreed population were not previously reported in temperate *Bos taurus* populations, perhaps due to differences in gene frequencies and the lower density chip used here vs. the higher-density chips used elsewhere. Thus, this study will need to be repeated with a higher-density chip and a larger population to confirm the identity of the set of genes influencing MY and FY in Thailand.

Acknowledgments

The authors thank the Royal Golden Jubilee Ph.D. program (Grant No. PHD/0040/2558) of the Thailand Research Fund for giving a scholarship to the first author, the projects of Development of a Dairy Genetic-Genomic Evaluation System in Thailand (P-11-00116) and the Increasing Genetic Potential of Thai Cattle using Genomic Selection for supporting a genomic dataset, and the University of Florida for supporting the training of the first author. We also thank Thai dairy farmers, dairy cooperatives, and private organizations for their participation and support.

References

Bolormaa, S., Pryce, J.E., Hayes, B.J., Goddard, M.E., 2010. Multivariate analysis of a genome-wide association study in dairy cattle. J. Dairy Sci. 93, 3818-3833.

Bolormaa, S., Pryce, J.E., Kempe, K., Savin, K., Hayes, B.J., Berendse, W., Zhang, Y., Reich, C.M., Mason, B.A., Bunch, R.J., Harrison, B.E., Reverter, A., Herd, R.M., Tier, B., Grase, H-U., Goddard, M.E., 2013. Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in *Bos taurus*, *Bos indicus*, and composite beef cattle. J. Anim. Sci. 91, 3088-3104.

Bush, W.S., Moore, J.H., 2012. Genome-wide association studies. PLoS Comput. Biol. 8, e1002822.

Chen, X., Cheng, Z., Zhang, S., Werling, D., Wathes, D.C., 2015. Combining genome wide association studies and differential gene expression data analyses identifies candidate genes affecting mastitis caused by two different pathogens in the dairy cow. Open J. Anim. Sci. 5, 358-393.

Cochran, S.D., Cole, J.B., Null, D.J., Hansen, P.J., 2013. Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle. BMC Genet. 14, 49.

Cole, J.B., Waurich, B., Wensch-Dorendorf, M., Bickhart, B.M., Swalve, H.H., 2013. A genome-wide association study of calf birth weight in Holstein cattle using single nucleotide polymorphisms and phenotypes predicted from auxiliary traits. J. Dairy. Sci. 37, 3156-3172.

Dai, W., Chen, Q., Wang, Q., White, R.R., Liu, J., Liu, H., 2017. Complementary transcriptomic and proteomic analyses reveal regulatory mechanisms of milk protein production in dairy cows consuming different forages. Sci. Rep. 7, 44234.

Do, D.N., Bissonntte, N., Lacasse, P., Miglior, F., Sargolzaei, M., Zhao, X., Ibeagha-Awemu, E.M., 2017. Genome-wide association analysis and pathways enrichment for lactation persistency in Canadian Holstein cattle. J. Dairy Sci. 100, 1955-1970.

Doran, A.G., Berry, D.P., Creevey, C.J., 2014. Whole genome association study identifies regions of the bovine genome and biological pathways involved in carcass trait performance in Holstein-Friesian cattle. BMC Genomics 15, 837.

Edwards, S.M., Thomsen, B., Madsen, P., Sorensen, P., 2015. Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle. Genet. Sel. Evol. 47, 60.

Faria, D., Schlicker, A., Pesquita, C., Bastos, H., Ferreira, A.E.N., Albrecht, M., Falcao, A.O., 2012. Mining GO annotations for improving annotation consistency. PLoS ONE 7, e40519.

Ghorbani, S., Thamoorespur, M., Nejad, A.M., Nasisi, M.R., Asgari, Y., 2014. Analysis of the enzyme network involved in cattle milk production using graph theory. Mol. Biol. Res. Comm. 4, 93-103.

Guo, J., Jorjani, H., Carlborg, O., 2012. A genome-wide association study using international breeding-evaluation data identifies major loci affecting production traits and stature in the Brown Swiss cattle breed. BMC Genet. 13, 82.

Hanna, L.L.H, Riley, D.G., 2014. Mapping genomic markers to closest feature using the R package Map2NCBI. Livest. Sci. 162, 59-65.

Howard, J.T., Kachman, S.D., Snelling, W.M., Pollak, E.J., Ciobanu, D.C., Kuehn, L.A., Spangler, M.L., 2014. Beef cattle body temperature during climatic stress: a genome-wide association study. Int. J. Biometeorol. 58, 1665-1672.

Iso-Touru, T., Sahana, G., Guldbrandtsen, L.M.S., Vilkki, J., 2016. Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genet. 17, 55.

Janjanam, J., Singh, S., Jena, M.K., Varshney, N., Kola, S., Kumar, S., Kaushik, J.K., Grover, S., Dang, A.K., Mukesh, M., Prakash, B.S., Mohanty, A.K., 2014. Comparative 2D-DIGE proteomic analysis of bovine mammary epithelial cells during lactation reveals protein signatures for lactation persistency and milk yield. PLoS ONE 9, e102515.

Jiang, L., Liu, J., Sun, D., Ma, P., Ding, X., Yu, Y., Zhang, Q., 2010. Genome wide association studies for milk production traits in Chinese Holstein population. PLoS ONE 5, e13661.

Jombart, T., Collins, C., 2015. A tutorial for Discriminant Analysis of Principal Components (DAPC) using adegenet 2.0.0. Imperial College London MRC Centre for Outbreak Analysis and Modelling, London, UK. (http://adegenet.r-forge.r-project.org/files/tutorial-dapc.pdf).

Jombart, T., Devillard, S., Balloux, F., 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11, 94.

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., Prins, H.H.T., 2013. Genetic consequences of breaking migratory traditions in barnacle geese *Branta leucopsis.* Mol. Eco. 22, 5835-5847.

Koonawootrittriron, S., Elzo, M.A., Thongprapi, T., 2009. Genetic trends in a Holstein x other breeds multibreed dairy population in Central Thailand. Livest. Sci. 122, 186-192.

Koonawootrittriron, S., Elzo, M.A., Tumwasorn, S., 2002. Multibreed genetic parameters and predicted genetic values for first lactation 305-d milk yield, fat yield, and fat percentage in a *Bos taurus* × *Bos indicus* multibreed dairy population in Thailand. Thai. J. Agric. Sci. 35, 339-360.

Lemos, M.V.A., Chiaia H.L.J., Berton, M.B., Feitosa, L.B., Aboujiaound, C., Camargo, G.M.F., Pereira, A.S.C., Albuquerque, L.G., Ferrinho, A.M., Mueller, L.F., Mazalli, M.R., Furlan, J.J.M., Carvalhiro, R., Gordo, D.M., Tonussi, R., Espigolan, R., de Oliveira Silva, R.F., de Oliveira, H.N., Duckett, S., Aguilar, I., Baldi, F., 2016. Genome-wide association between single nucleotide polymorphisms with beef fatty acid profile in Nellore cattle using the single step procedure. BMC Genomics 17, 213.

Ma, L., Wiggans, G.R., Wang, S., Sonstegard, T.S., Yang, J., Crooker, B.A., Cole, J.B., Van Tassell, C.P., Lawlor, T.J., Da, Y., 2012. Effect of sample stratification on dairy GWAS results. BMC Genomics 13, 536.

Maxa, J., Neuditschko, Russ, I., Forster, M., Medugorac, I., 2012. Genome-wide association mapping of milk production traits in Braunvieh cattle. J. Dairy Sci. 95, 5357-5364.

Martin-Burriel, I., Rodellar, C., Canon, J., Cortes, O., Dunner, S., Landi, V., Martinez-Martinez, A., Gama, L.T., Ginja, C., Penedo, M.C.T., Sanz, A., Zaragoza, P., Delgado, J.V., 2011. Genetic diversity, structure, and breed relationships in Iberian cattle. J. Anim. Sci. 89, 893-906.

Mi, H., Muruganuian, A., Casagrande, J.T., Thomas, P.D., 2013. Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. 8, 1551-1566.

Michelizzi, V.N., Wu, X., Dodson, M.V., Michal, J.J., Zambrana-Varon, J., McLean, D.J., Jiang, Z., 2011. A global view of 54,001 single nucleotide polymorphisms (SNPs) on the Illumina BovineSNP50 BeadChip and their transferability to water buffalo. Int. J. Biol. Sci. 7, 18-27.

Nayeri, S., Sagolzaei, M., Abo-Ismial, M.K., May, N., Miller, S.P., Schenkel, F., Moore, S.S., Stothard, P., 2016. Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. BMC Genet. 17, 75.

Ortega, M.S., Denicol, A.C., Cole, J.B., Null, D.J., Hansen, P.J., 2016. Use of single nucleotide polymorphisms in candidate genes associated with daughter pregnancy rate for prediction of genetic merit for reproduction in Holstein cows. Anim. Genet. 47, 288-297.

Pérez-Enciso, M., Misztal, I., 2011. Qxpak.5: Old mixed model solutions for new genomics problems. BMC Bioinformatics 12, 202.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904-909.

Pryce, J.E., Bolormaa, S., Chamberlain, A.J., Bowman, P.J., Savin, K., Goddard, M.E., Hayes, B.J., 2010. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. J. Dairy Sci. 93, 3331-3345.

Purcell, S., Chang, C., 2017. PLINK 1.9. https://www.cog- genomics.org/plink2.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559-575.

Purfield, D.C., Bradley, D.G., Evans, R.D., Kaerney, F.J., Berry, D.P., 2015. Genome-wide association study for calving performance using high-density genotypes in dairy and beef cattle. Genet. Sel. Eval. 47, 47.

Raven, L-A., Cocks, B.G., Hayes, B.J., 2014. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. BMC Genomics 15, 62.

Ritsawai, P., Koonawootrittriron, S., Jattawa, D., Suwanasopee, T., Elzo, M.A. 2014. Fraction of cattle breed and their influence on milk production of Thai dairy cattle. In: Proceeding of 52rd Kasetsart conference, Kasetsart University, Bangkok, Thailand.

Santana, M.H.A., Junior, G.A.O., Cesar, A.S.M., Freua, M.C., Gomes, R.C., Silva, S.L., Leme, P.R., Fukumasu, H., Carvalho, M.E., Ventura, R.V., Coutinho, L.L., Kadarmideen, H.N., Ferraz, J.B.S., 2016. Copy number variations and genome-wide associations reveal putative genes and metabolic pathways involved with the feed conversion ratio in beef cattle. J. Appl. Genet. 57, 495-504.

Sargent, F.D., Lytton, V.H., Wall, J.R.O.G., 1968. Test interval method of calculating dairy herd improvement association records. J. Dairy Sci. 51, 170-179.

Shin, J., Lee, C., 2015. A mixed model reduces spurious genetic associations produced by population stratification in genome-wide association studies. Genomics 105, 191-196.

Sudrajad, P., Sharma, A., Dang C.G., Kim, J.J., Kim, K.S., Lee, J.H., Kim, S., Lee, S.H., 2016. Validation of single nucleotide polymorphisms associated with carcass traits in a commercial Hanwoo population. Asian-Australas. J. Anim. Sci. 29, 1541-1546.

The Gene Ontology Consortium, 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25-29.

Tripathi, S., Christie, K.R., Balakrishnan, R., Huntley, R., Hill, D.P., Thommesen, L., Blake, J.A., Kuiper, M., Laegreid, A., 2013. Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. Database, bat062.

Venturini, G.C., Cardoso, D.F., Baldi, F., Freitas, A.C., Asplicueta-Borquis, R.R., Santos, D.J.A., Camargo, G.M.F., Stafuzza, N.B., Albuquerque, L.G., Tonhati, H., 2014. Association between single-nucleotide polymorphisms and milk production traits in buffalo. Genet. Mol. Res. 13, 10256-10268.

Weikard, R., Golgammer, T. Brunner, R.M., Kuehn, C., 2012. Tissue-specific mRNA expression patterns reveal a coordinated metabolic response associated with genetic selection for milk production in cows. Physiol. Genomics 44, 728-739.

Xia, J., Qi, X., Wu, Y., Zhu, B., Xu, L., Zhang, L., Gao, X., Chen, Y., Li, J., Gao, H., 2016. Genome-wide association study identifies loci and candidate genes for meat quality traits in Simmental beef cattle. Mamm. Genome 27, 246-255.

Yang, Y., Zheng, N., Zhao, X., Zhang, Y., Han, R., Ma, L., Zhao, S., Li, S., Guo, T., Wang, J., 2015a. Proteomic characterization and comparison of mammalian milk fat globule proteomes by iTRAQ analysis. J. Proteomics 26, 34-43.

Yang, J., Jiang, J., Liu, X., Wang, H., Guo, G., Zhang, Q., Jiang, L., 2015b. Differential expression of genes in milk of dairy cattle during lactation. Anim. Genet. 47, 174-180.

Yang, Y-X., Cao, S-Z., Zhang, Y., Zhao, X-X., 2009. Proteomic approach analysis of mammary membrane proteins expression profiles in Holstein cows. Asian-Australas J. Anim. Sci. 22, 885-892.

Yodklaew, P., Koonawootrittriron, S., Elzo, M.A., Suwanasopee, T., 2014. Genome-wide association study for milk yield, fat yield, and age at first calving of dairy cattle in Thailand. Thai J. Anim. Sci. 1, 301-304.

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., Li, J., Simianer, H., 2014. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS ONE 9, e93017.

Table 1 Number of the significant SNP for milk yield and fat yield at P-value ≤ 0.001 after correcting for false discovery rate based on distance between SNP genotypes and genes in the NCBI database

Distance between SNP and gene	Milk yield[1]		Fat yield	
	M1	M2	M1	M2
Inside gene	138	144	71	41
≤ 2,500 bp	25	28	9	6
2,500 bp < distance ≤ 5,000 bp	9	14	8	5
5,000 bp < distance ≤ 25,000 bp	59	60	31	16
distance > 25,000 bp	154	160	80	52
Total	385	406	199	120
False discovery rate (%)	1.87	1.77	3.63	5.94

[1]M1 = model with genetic groups based on expected breed composition; M2 = model with genetic groups based on SNP genotypic information

Table 2 Number of significant SNP and number of genes associated with milk yield and fat yield by chromosome at P-value ≤ 0.001 after correcting for false discovery rate

Chromosome	Number of significant SNP (n)				Number of genes (n)			
	Milk yield[1]		Fat yield		Milk yield		Fat yield	
	M1	M2	M1	M2	M1	M2	M1	M2
1	19	25	10	6	17	22	8	4
2	23	26	7	2	22	24	7	2
3	10	11	2	3	10	11	2	3
4	14	15	12	9	14	15	11	8
5	13	17	4	3	11	15	4	3
6	21	19	15	12	21	19	15	12
7	19	20	6	5	19	20	6	5
8	16	13	12	5	15	13	11	4
9	27	24	3	1	27	24	3	1
10	8	10	14	9	8	10	13	8
11	21	24	9	6	19	23	8	5
12	16	14	6	4	14	12	6	4
13	5	5	4	1	5	5	4	1
14	17	16	10	3	16	15	9	3
15	8	8	10	4	8	8	10	4
16	12	13	1	1	12	13	1	1
17	10	5	8	5	10	5	8	5
18	14	17	12	6	14	17	11	6
19	5	7	4	2	5	7	4	2
20	28	29	8	9	20	20	8	9
21	20	22	7	4	12	14	7	4
22	4	6	1	2	3	5	1	2
23	4	6	4	1	4	6	4	1
24	2	2	7	7	2	2	6	7
25	2	4	1	2	2	4	1	2
26	14	12	4	1	14	12	3	1
27	2	2	3	2	2	2	3	2
28	7	8	7	2	7	8	6	2
29	3	3	2	0	3	3	2	0
X	21	23	6	3	21	23	6	3
Total	385	406	199	120	357	377	188	114

[1]M1 = model with genetic groups based on expected breed composition; M2 = model with genetic groups based on SNP genotypic information

Table 3 Number of genes associated with milk yield and fat yield based on position of SNP genotypes inside and within 2,500 bp to genes in the NCBI database

Model[1]	Number of genes	
	Milk yield	Fat yield
M1	151	78
M2	158	46
Genes in common in M1 and M2	125	42

[1]M1 = model with genetic groups based on expected breed composition; M2 = model with genetic groups based on SNP genotypic information

Table 4 Top 20% of SNP[1] in common across models for milk yield and fat yield

Trait[2]	SNP	Chr[3]	Position	Gene name	Gene description	Molecular function	Biological process
Milk yield	Hapmap59494-rs29020429	9	54062560	*UFL1*	UFM1 specific ligase 1		
	BTA-78317-no-rs	10	83050810	*PCNX*	pecanex homolog (Drosophila)		developmental
	ARS-BFGL-BAC-34293	20	53571599	*CDH18*	cadherin 18	binding	cellular, developmental, multicellular organismal
	Hapmap29482-BTA-146449	9	38804655	*LAMA4*	laminin subunit alpha 4		
	ARS-USMARC-Parent-DQ990832-rs29015065	22	11038205	*ITGA9*	integrin subunit alpha 9		
	ARS-BFGL-NGS-119158	10	56015779	*UNC13C*	unc13 homolog 3-like		
	ARS-BFGL-NGS-20636	X	130773887	*MAP3K15*	mitogen-activated protein kinase kinase kinase 15	catalytic, signal transducer	biological regulation, cellular, metabolic, response to stimulus
	BTA-67160-no-rs	3	17195924	*PGLYRP4*	peptidoglycan recognition protein 4		
	Hapmap51953-BTA-48787	2	96570370	*PLEKHM3*	pleckstrin homology domain containing M3		
	Hapmap39920-BTA-43352	1	98078072	*GPR160*	G protein-coupled receptor 160		
	BTB-01086542	1	95279975	*SPATA16*	spermatogenesis associated 16		metabolic, response to stimulus
	ARS-BFGL-NGS-44080	5	75361846	*CACNG2*	calcium voltage-gated channel auxiliary subunit gamma 2	binding, channel regulator, receptor, signal transducer, transporter	biological regulation, cellular, multicellular organismal, response to stimulus
	ARS-BFGL-NGS-102255	8	40747782	*GLIS3*	GLIS family zinc finger 3		developmental
	ARS-BFGL-NGS-6343	26	47837750	*PTPRE*	protein tyrosine phosphatase, receptor type E	catalytic, receptor	cellular, metabolic
	ARS-BFGL-NGS-112825	2	487391	*OCA2*	OCA2 melanosomal transmembrane protein	transporter	
	BTB-00144037	3	92896187	*DIO1*	iodothyronine deiodinase 1		

[1]Top 20% of SNP ordered by P-value from lowest to highest after applying FDR values for MY and FY in M1 and M2
[2]M1 = model with genetic groups based on expected breed composition; M2 = model with genetic groups based on SNP genotypic information
[3]Chr = chromosome

Table 4 (Continued)

Trait[2]	SNP	Chr[3]	Position	Gene name	Gene description	Molecular function	Biological process
Milk yield	ARS-BFGL-NGS-70466	4	30770711	*DNAH11*	dynein axonemal heavy chain 11	catalytic, structural molecule	cellular component organization or biogenesis, cellular, localization, reproduction
	BTB-01112800	2	19338223	*AGPS*	alkylglycerone phosphate synthase	catalytic	metabolic
	Hapmap33532-BTA-84282	9	75597064	*MAP3K5*	mitogen-activated protein kinase kinase kinase 5	catalytic, signal transducer	biological regulation, cellular, metabolic, response to stimulus
	ARS-BFGL-NGS-103520	11	100924099	*FUBP3*	far upstream element binding protein 3	binding, catalytic	cellular, developmental, localization, metabolic, multicellular organismal
	Hapmap54974-rs29015318	8	4270697	*GALNTL6*	Polypeptide N-acetylgalactosaminyl transferase-like 6	catalytic	metabolic
	BTB-01390865	2	53931759	*KYNU*	kynurenines	catalytic	metabolic
	Hapmap27307-BTC-043200	6	82605943	*EPHA5*	EPH receptor A5		
Fat yield	ARS-USMARC-Parent-EF034087-no-rs	28	16097749	*LOC100337 251*	ankyrin-3		
	BTB-01311082	17	7527510	*LRBA*	LPS responsive beige-like anchor protein		
	ARS-BFGL-NGS-57210	15	32637662	*SORL1*	sortilin related receptor 1	binding, receptor, transporter	localization, metabolic
	ARS-BFGL-NGS-39335	X	4699555	*C1GALT1C1*	C1GALT1 specific chaperone 1	catalytic	cellular, metabolic
	BTA-76281-no-rs	10	78905418	*GPHN*	gephyrin		cellular, metabolic
	BTB-01367046	4	111863574	*CNTNAP2*	contactin associated protein-like 2		
	ARS-BFGL-NGS-74971	22	14218256	*ULK4*	unc-51 like kinase 4		

[1]Top 20% of SNP ordered by P-value from lowest to highest after applying FDR values for MY and FY in M1 and M2

[2]M1 = model with genetic groups based on expected breed composition; M2 = model with genetic groups based on SNP genotypic information

[3]Chr. = chromosome

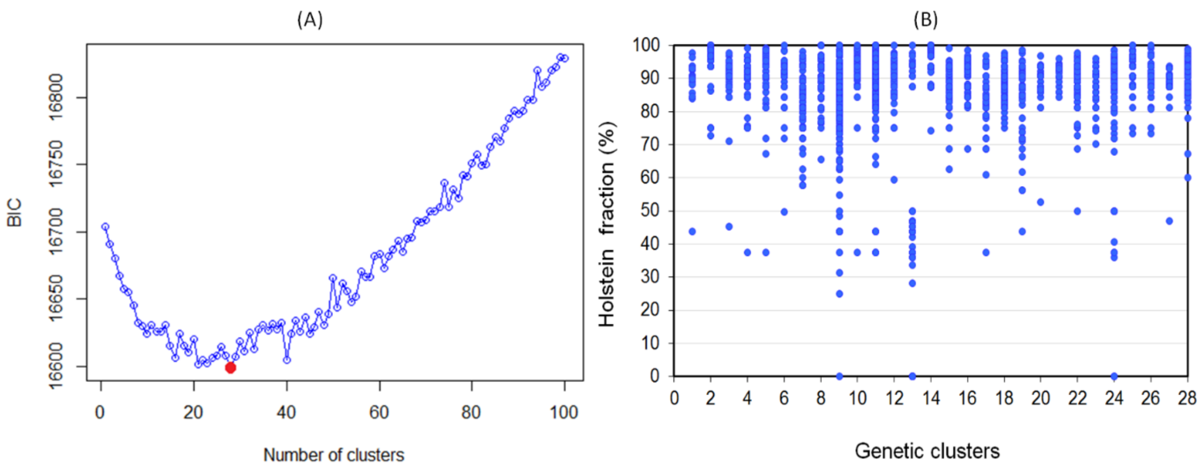


Figure 1 Bayesian Information Criterion (BIC) values for models with 1 to 100 genetic clusters by a K-means algorithm with 2,000 principal components from 2,661 animals. Figure 1A indicates that the optimal number of genetic clusters in the population to be 28 (lowest BIC value). Figure 1B shows the percent Holstein of animals in genetic clusters 1 to 28.

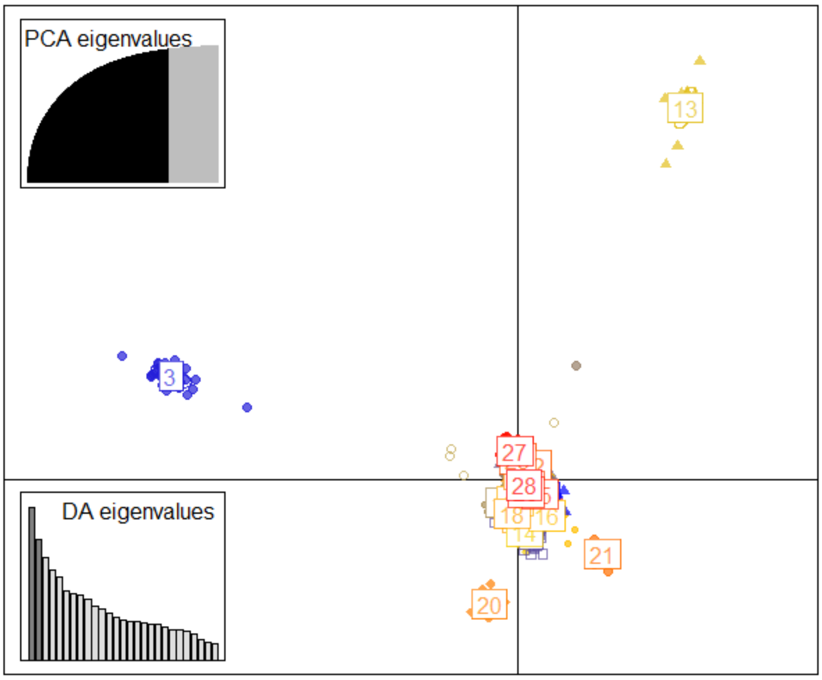


Figure 2 Scatterplot of the first and second principal components of the DAPC in a Thai multibreed dairy cattle population. The top left inset shows the PCA eigenvalues corresponding to the 2,000 PCA eigenvectors used in the DAPC analysis. The bottom left inset shows the DA eigenvalues from the DAPC analysis with 28 genetic clusters.

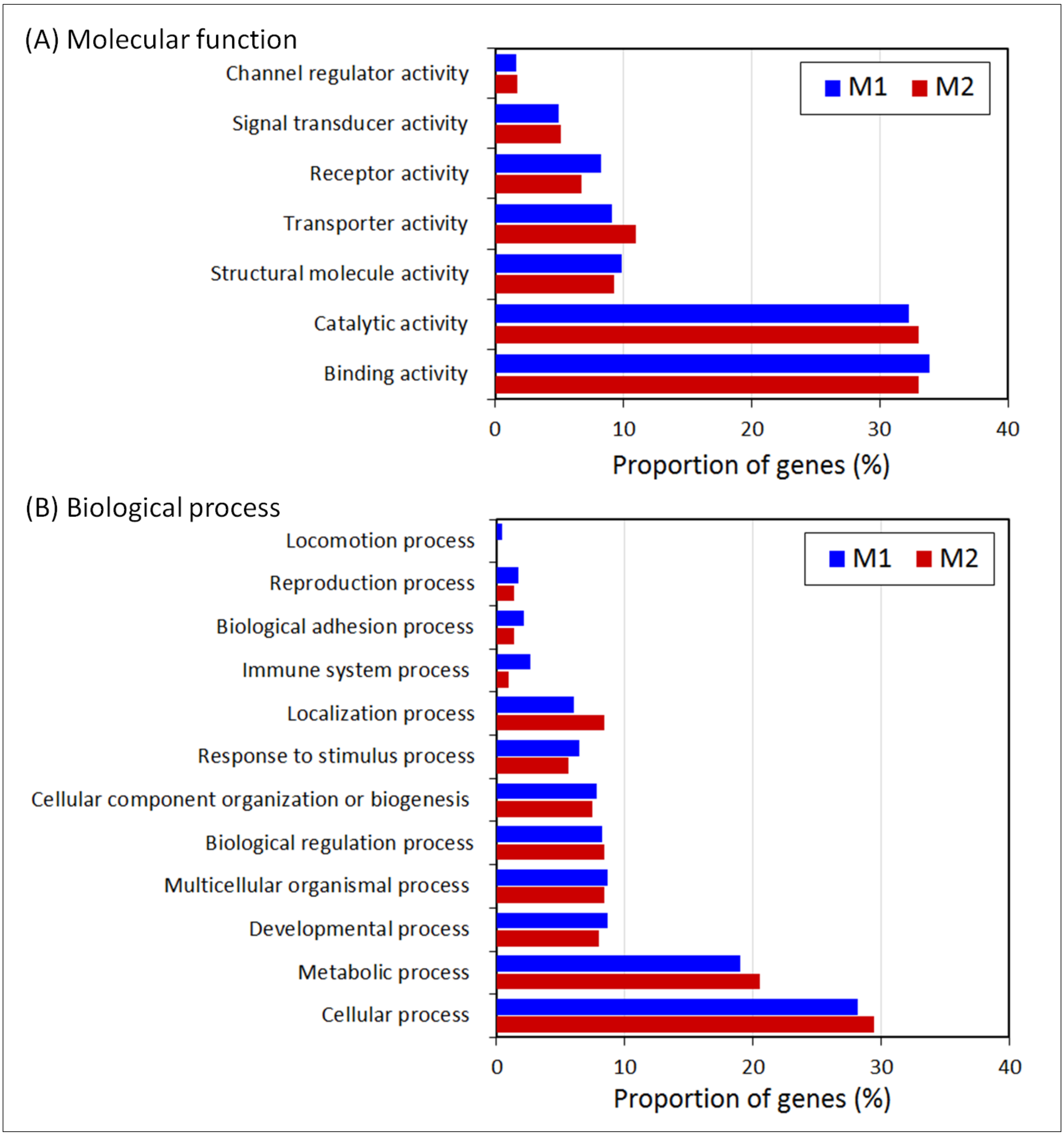
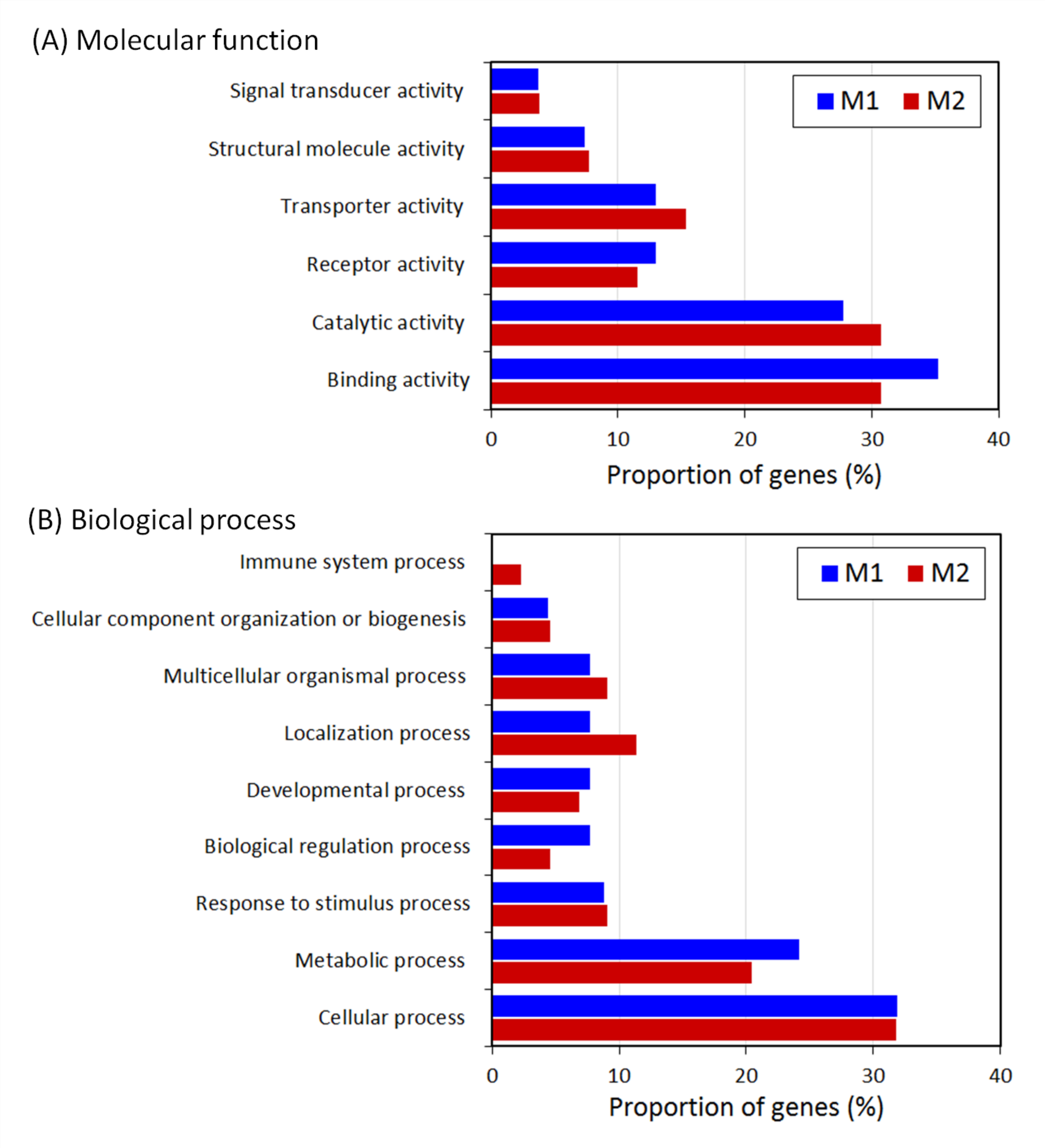


Figure 3 Proportion of genes associated with milk yield classified based on molecular functions and biological processes of gene products for a model with genetic groups based on expected breed composition (M1) and a model with genetic groups based on SNP genotypic information (M2). Figure 3A shows the proportion of genes associated with fat yield based on activities of gene products that occur at a molecular level. Figure 3B shows the proportion of genes associated with milk yield based on processes determined by activities of multiple gene products.

634 

635 **Figure 4** Proportion of genes associated with fat yield classified based on molecular
636 functions and biological processes of gene products for a model with genetic groups
637 based on expected breed composition (M1) and a model with genetic groups based
638 on SNP genotypic information (M2). Figure 4A shows the proportion of genes
639 associated with fat yield based on activities of gene products that occur at a
640 molecular level. Figure 4B shows the proportion of gene associated with fat yield
641 based on processes determined by activities of multiple gene products.

642