J. Dairy Sci. 105:9810–9821
https://doi.org/10.3168/jds.2022-22143

Identifying influential sires and distinct clusters of selection candidates based on genomic relationships to reduce inbreeding in the US Holstein

Y. Steyn,[1]* Y. Masuda,[1] S. Tsuruta,[1] D. A. L. Lourenco,[1] I. Misztal,[1] and T. Lawlor[2]
[1]Department of Animal and Dairy Science, University of Georgia, Athens 30602
[2]Holstein Association USA, Brattleboro, VT 05302

ABSTRACT

High relatedness in the US Holstein breed can be attributed to the increased rate of inbreeding that resulted from strong selection and the extensive use of a few bulls via reproductive biotechnology. The objectives of this study were to determine whether clustering could separate selected candidates into genetically different groups and whether such clustering could reduce the expected inbreeding of the next generation. A genomic relationship matrix composed of 1,145 sires with the most registered progeny in the breed born after 1985 was used for principal component analysis and k-means clustering. The 5 clusters reduced the variance by 25% and contained 171 (C1), 252 (C2), 200 (C3), 244 (C4), and 278 (C5) animals, respectively. The 2 most predominant families were C1 and C2, while C4 contained the most international animals. On average, C1 and C5 contained older animals; the average birth year per cluster was 1988 (C1), 1996 (C2 and C3), 1999 (C4), and 1990 (C5). Increasing to 10 clusters allowed the separation of the predominant sons. Statistically significant differences were observed for indices (net merit index, cheese merit index, and fluid merit index), daughter pregnancy rate, and production traits (milk, fat, and protein), with older clusters having lower merit for production but higher for reproduction. K-means clustering was also used for 20,099 animals considered as selection candidates. Based on the reduction in variance achieved by clustering, 5 to 7 clusters were appropriate. The number of animals in each cluster was 3,577 (C1), 3,073 (C2), 3,302 (C3), 5,931 (C4), and 4,216 (C5). The expected inbreeding from within or across cluster mating was calculated using the complete pedigree, assuming the mean inbreeding of animals born in the same year when parents are unknown. Generally, inbreeding was highest within cluster mating and lowest across cluster mating. Even when 10 clusters were used, one cluster always gave low inbreeding in all scenarios. This suggests that this cluster contains animals that differ from all other groups but still contains enough diversity within itself. Based on lower across cluster inbreeding, up to 7 clusters were appropriate. Statistically significant differences in genomic estimated breeding values were found between clusters. The rankings of clusters for different traits were mostly the same except for reproduction and fat. Results show that diversity within the population exists and clustering of selection candidates can reduce the expected inbreeding of the next generations.

Key words: k-means, genetic diversity, principal component analysis

INTRODUCTION

The purebred dairy breed populations have undergone strong selection for similar traits. The wide use of AI has led to a substantial genetic improvement in production of the US Holstein (Capper et al., 2009; Capper and Cady, 2020). However, the use of a relatively small number of bulls has increased the relatedness within the breed. By 2015, all AI bulls in North America could be traced back to only 2 bulls born in 1880 (Yue et al., 2015). In fact, (Makanjuola et al., 2020b) found an average inbreeding coefficient of 7.74% when estimated through traditional pedigree methods but between 15% and 31% when based on genomic information. The loss of genetic diversity caused by inbreeding has already led to inbreeding depression in both production and reproduction in dairy populations (Bjelland et al., 2013). This lack of variation can also hinder the ability of populations to adapt to change (Markert et al., 2010), which is a growing concern in the face of climate change and consumer preferences.

An anticipated advantage of genomic selection was to reduce the rate of inbreeding per generation by allowing the accurate identification of the best animals instead of the best families (Daetwyler et al., 2007). While it appears that this has been successful in the American Angus (Lozada-Soto et al., 2021), the rate

Received March 30, 2022.
Accepted July 19, 2022.
*Corresponding author: yvette.steyn@uga.edu

of inbreeding in the North American Holstein has increased considerably since the application of genomic selection. The annual increase in inbreeding was 0.11 percentage points from 2000 to 2008 (before genomic selection) but increased to 0.36 percentage points from 2013 to 2021 (CDCB, 2021). An increase in the rate of inbreeding after the implementation of genomic selection has also been observed outside the US, such as the French (Doublet et al., 2019) and Dutch-Flemish (Doekes et al., 2018) Holstein-Friesian.

Phenotypically similar animals can still be genetically diverse. Genetic evaluation has allowed the estimation of genetic merit, but even animals with identical breeding values will still have different clusters of genes. Identifying distinct groups within the Holstein breed can aid the mating of animals with underlying genomic differences and avoid the breedwise fixation of alleles. Breeding less related animals leads to a decrease in the average relatedness in the population (Wellmann and Bennewitz, 2019). The objectives of this study were to identify specific sires that have contributed to differences in the population, determine whether selected candidates can be clustered into genetically different groups, and establish whether this clustering can be used to reduce the expected inbreeding in the next generation.

MATERIALS AND METHODS

Data

Data were provided by the Council on Dairy Cattle Breeding (**CDCB**) and the Holstein Association USA; therefore, Animal Care and Use Committee approval was not required for this research. Genotypes were available for the US Holstein population up to 2014. The number of animals in the pedigree was 9,817,252, which contained 330,837 sires and 5,471,039 dams. The average progeny per sire was 29 with a maximum of 58,266 for sire Marshfield Elevation Tony (Mars). The phenotypic data contained type traits and totaled 10,067,745 records. After removal of unmapped and sex chromosomes, 58,990 SNP markers remained. Genotypes were available for 569,404 animals. Breeding values for net merit index (**NMI**), cheese merit index (**CMI**), fluid merit index (**FMI**), daughter pregnancy rate (**DPR**), milk yield, fat yield, and protein yield were available for a subset of animals.

Clustering of Sires

Family clustering within the breed was investigated using only sires of animals born after 1985. Of the 2,000 sires with the most progeny born in 1986 or later, 1,145 were genotyped. The number of progeny per sire ranged from 312 to 49,146 and the birth year of the sires ranged from 1962 to 2009. Six countries were represented based on the registration number, including the US (988 animals), Canada (139), Germany (7), Italy (6), the Netherlands (3), and Great Britain (1). Animals with foreign registration numbers may include animals that were the result of embryo transfers of US animals, which means that they would genetically be American. The genomic relationship matrix (**G**) was obtained using the formula

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\ \Sigma p_i \left(1 - p_i\right)},$$

where **Z** is a matrix of SNP content centered around the current allele frequencies, and p_i is the current allele frequency of SNP i (VanRaden, 2008). Principal component analysis was performed on **G** to visualize potential clustering and identify those more distant from the majority. The first 3 principal components are presented in Figure 1. The animals falling within the extreme ends of the first 12 principal components were considered as potential key sires that contributed to genetic variation within this group of sires.

K-means clustering (Hartigan and Wong, 1979) with 5 and 10 clusters was performed on G using the k-means package in R. This is an iterative procedure that aims to minimize within-cluster sum of squares. This method can be sensitive to initial values, and therefore 50 iterations were performed. K-means clustering was also performed on the pedigree matrix corresponding to these genotyped animals (A_{22}). The relationships within A_{22} were obtained using the full pedigree information, thus all registered animals regardless of age or sex were accounted for. The 3 oldest animals as well as the 3 with the most progeny in each cluster were considered key sires and are presented in Table 1. Figures 2 and 3, respectively, show the distribution of birth year and proportion of animals per country within each cluster using k-means clustering on G with 5 clusters. Breeding values were available for 1,125 animals. Analysis of variance and Tukey's honest significant difference test determined significant differences between clusters.

Clustering Selected Candidates

A subset of animals was chosen to represent those that may have been selected from the available selection candidates at the time (~2014). A total of 3,902 genotyped sires of animals born after 2010 with a minimum of 25 progeny were identified as male selected candidates in 2014. Among the females, 16,197 geno-

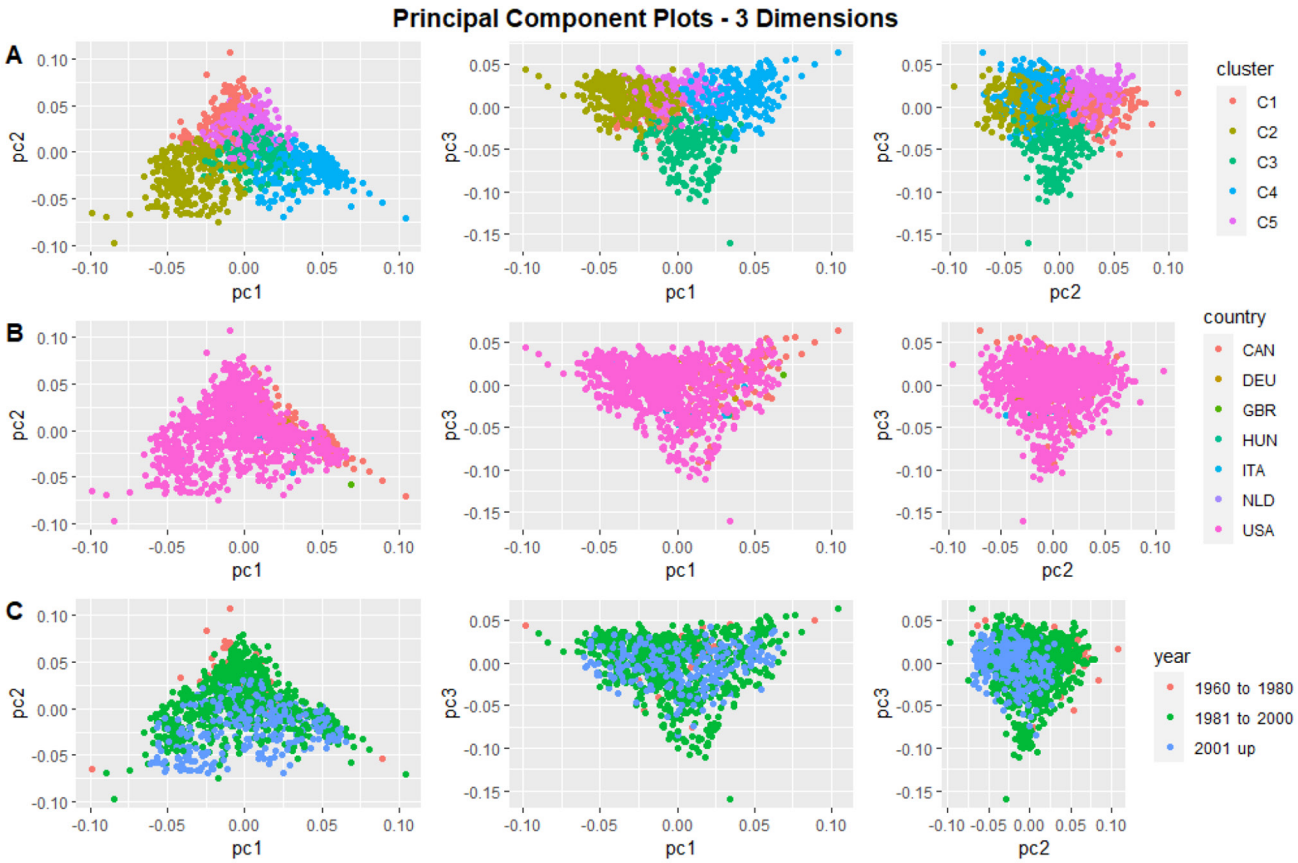

Figure 1. The first 3 principal components (pc) based on the genomic relationship matrix (**G**) of 1,145 sires. Animals are color-coded according to clusters (k-means with 5 clusters), country, or birth year groups (A, B, and C, respectively).

Table 1. The most important sires identified using different clustering methods on the 1,145 most popular sires of animals born after 1985[1]

Cluster	PCA	K-means 5 (**G**)	K-means (A_{22})	K-means 10 (**G**)
1	Mark, Altagrand	Mark, Mandingo, Tesk, Valiant	Mark, Mandingo, Tesk, Chief, Glendell, Conductor	Melvin, Mandingo, Tesk, Ivanhoe Chief
2	Bell, Elton, Bitzie, Celsius, Marshall, Elegant	Bell, Durham, Belltone	Bell, Durham, Belltone, Jesse	Bell, Mathie, Belltone, Cinnamon
3	Blackstar	Blackstar, Emory, Integrity, Chairman	Blackstar, Emory, Integrity, Chairman, Ivanhoe Chief	Blackstar, Emory, Integrity, Wayne
4	Starbuck, Storm, Rudolph, Aerostar	Starbuck, Outside, Morty, Aerostar	Starbuck. Outside, Elevation, Mars, Elav Mars	Starbuck, Encore, Morty, Astronaut
5	Valiant, Leadman, Formation	Ned Boy, Leadman, Cleitus, Tradition	Ned Boy, Levi, Enhancer, Bootmaker, Astronaut, Jet Stream	Ned Boy, Leadman, Cleitus, Bootmaker
6	Rotate, Melwood, Altamelwood			Oman, Melwood, Wister, Glendell
7	Elevation, Tradition, Cleitus			Outside, Million, Shottle, Prelude
8				BW Marshall, Toystory, Altabellwood
9				Mark, Highlight, Roebuck, Chief
10				Durham, Emerson, Mr. Sam, Elton

[1]The clustering using principal components analysis (PCA) looked at the animals on the outer edges of the first 12 principal components, resulting in 7 groups. K-means clustering was used with 5 or 10 clusters based on genomic relationship matrix (**G**), and 5 clusters based on pedigree relationships (A_{22}). The sires listed are the 3 sires with the most progeny in the cluster and the oldest (if different from the 3 with the most progeny).

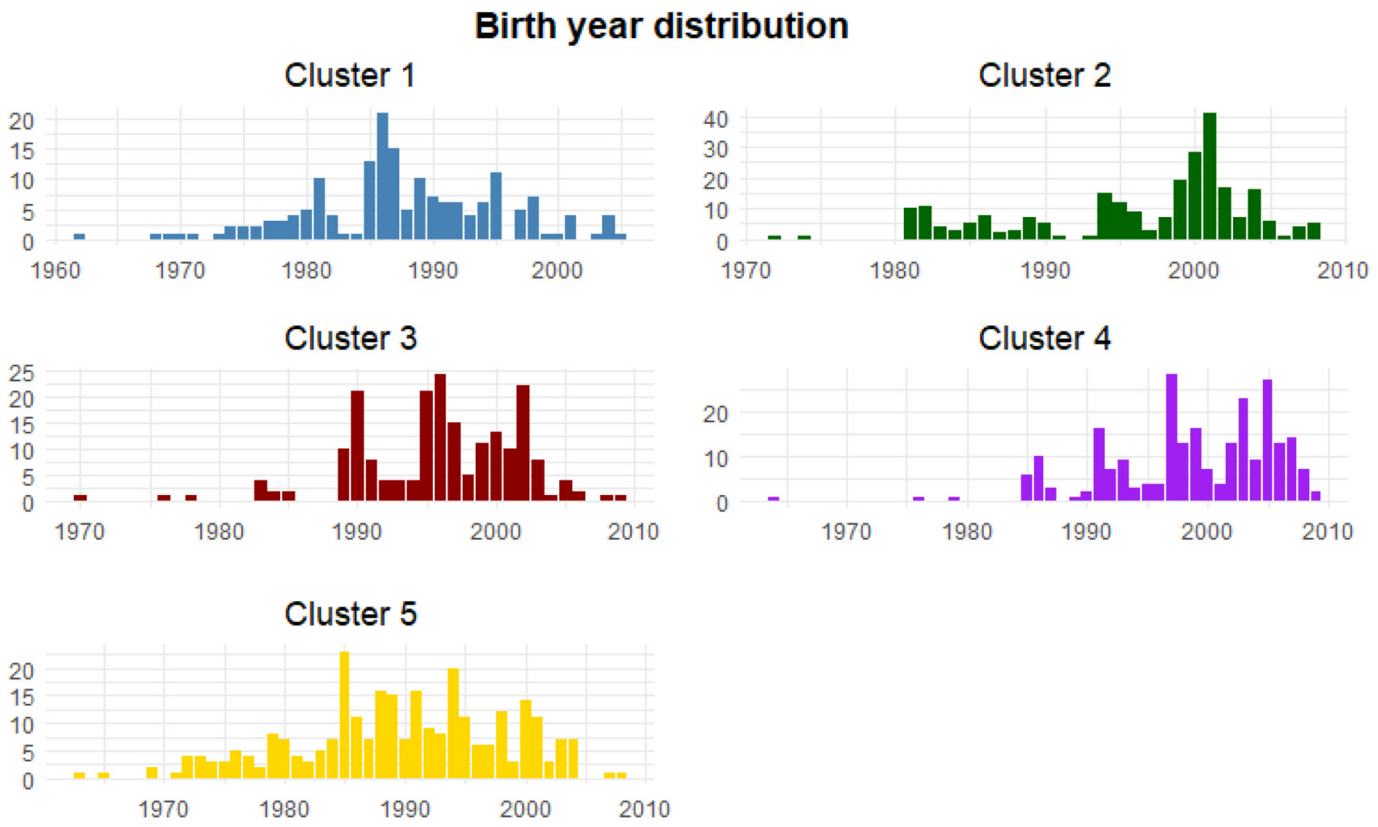

Figure 2. The distribution of birth year within each cluster of the 1,145 most popular sires of animals born after 1985.

typed animals measured for type traits after 2012 were chosen as candidates. K-means clustering was applied to G using up to 10 clusters with the built-in k-means R package. This clustering method reduces the sum of squares between data points by calculating the sum based on the distance to the nearest cluster center. Adding more clusters reduces overall variance. The point where the reduction appears to reach a plateau is an estimate of the appropriate number of clusters. On our data, 5 to 7 clusters were appropriate based on Figure 4. We chose to use 5 clusters (C1, C2, C3, C4, and C5). The first 3 principal components using **G** are presented in Figure 5.

Hypothetical mating was performed within and across clusters of these selection candidates with the INBUPGf90 software package within the BLUPF90 software suite (Misztal et al., 2014). The expected pedigree-based inbreeding of offspring was calculated for every possible mating between the sires of a cluster and the dams of each cluster. The complete available pedigree information of the Holstein population was used in a recursive algorithm, assuming nonzero inbreeding for unknown parents (Aguilar and Misztal, 2008). The recursive algorithm makes use of the tabular method to obtain expected relationships based on parents. Animals must be sorted based on year of birth so that parents precede their progeny. The method to calculate inbreeding coefficient (F_x) for each animal x is $F_x = 0.5R_{sd}$, where R_{sd} is the numerator relationship between the sire (s) and dam (d). The calculation is recursive and involves tracing the ancestors back and computing the relationship between parents. The mean inbreeding of animals born in the same year is used when parents are unknown. It is possible to calculate the expected inbreeding between a specific bull and cow in our study when the relationship between the parents is known.

The genetic merit of the bulls was compared across cluster using ANOVA. Breeding values for selection indices, DPR, and yield traits were available for most bulls.

RESULTS AND DISCUSSION

Clustering of Influential Sires

Clear clustering within the 1,145 sires was not observed in the principal component analysis (**PCA**) plot (Figure 1). However, sires from Canada could be

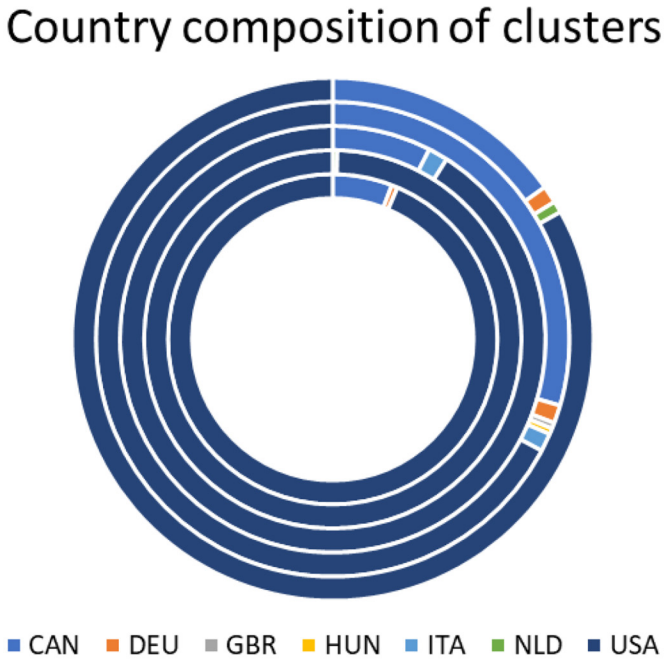

Figure 3. The proportion of sires from different countries in each cluster when the 1,145 most popular sires of animals born after 1985 are used. The innermost circle is cluster 1, and the outermost is cluster 5. CAN = Canada; DEU = Germany; GBR = Great Britain; HUN = Hungary; ITA = Italy; NLD = the Netherlands.

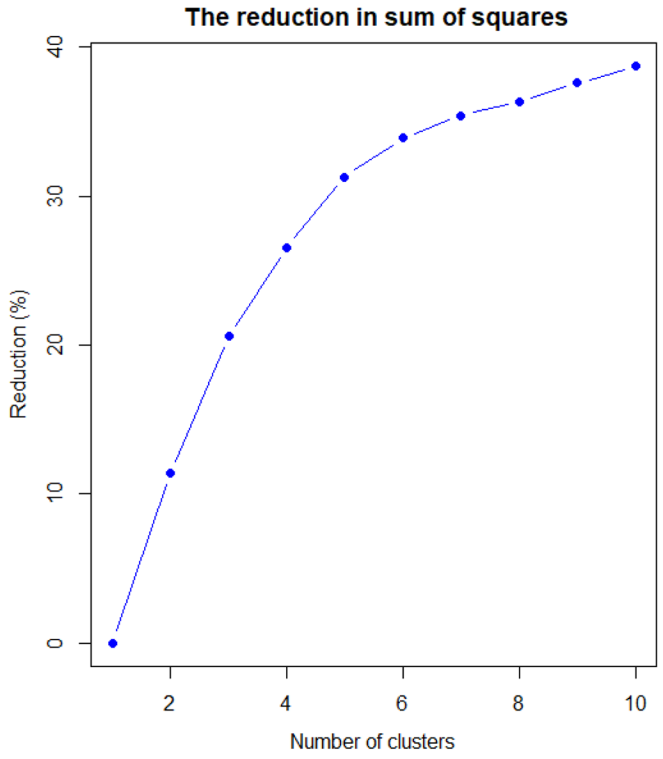

Figure 4. The reduction in the sum of squares achieved with k-means clustering as the number of clusters specified increases. The k-means clustering was based on the 20,099 selected candidates.

observed in 1 plane. Animals further away from the center were older bulls, while younger bulls appeared in the center.

It has been found that the 2 most influential bulls to Holstein US sires were Round Oak Rag Apple Elevation (Elevation) and Pawnee Farm Arlinda Chief (Chief). Up to 99% of AI bulls born after 2010 traced back to these animals (Yue et al., 2015; Dechow et al., 2020). Three of the bulls that appear more distant in the PCA, namely Tomar Blackstar (Blackstar), SWD Valiant (Valiant), and Walkway Chief Mark (Mark), are key male descendants of Chief. Madawaska Aerostar (Aerostar) is an important son of Hanoverhill Starbuck (Starbuck), who is the son of Elevation. The third most important bull identified by previous studies is Pennstate Ivanhoe Star (Ivanhoe Star), the sire of Carlin M. Ivanhoe Bell (Bell). Aerostar, along with others in its group, is a Canadian-born bull. As the families became more related to each other, distinct animals could not be observed clearly with more principal components. Table 1 contains a summary of bulls identified as genetically more different compared with the majority based on the observation of the first 12 principal components.

Cross-validation can be used to determine the usefulness of genomic prediction. Studies have shown that the accuracy obtained from cross-validation was lowest when using k-means as clustering (Saatchi et al., 2012; Boddhireddy et al., 2014; Baller et al., 2019). The accuracy of genomic predictions relies on the relationships between the training and target populations (Habier et al., 2010; Clark et al., 2012; Pszczola et al., 2012), which suggests that k-means clustering is successful at separating groups that are more related to each other but less related to other clusters. The principal component plot in Figure 1 shows that animals in the same cluster are closer together. The number of animals per cluster based on the G-matrix was 171 (C1), 252 (C2), 200 (C3), 244 (C4), and 278 (C5). The clustering method reduced variance by 25%. The number of animals based on the pedigree matrix was 125 (C1), 181 (C2), 263 (C3), 270 (C4), and 306 (C5). While the clustering of many animals changed, the same influential sires were identified using genomic information or only pedigree information. These differences in cluster size based on G or A_{22} could be due to incomplete pedigrees, the difference between identity by state (**IBS**) and identity by descent (**IBD**), and the lack of clear separation of animals that are closely related based on either method. The pedigree relationships are based on expected relationships between animals using only recorded relatives

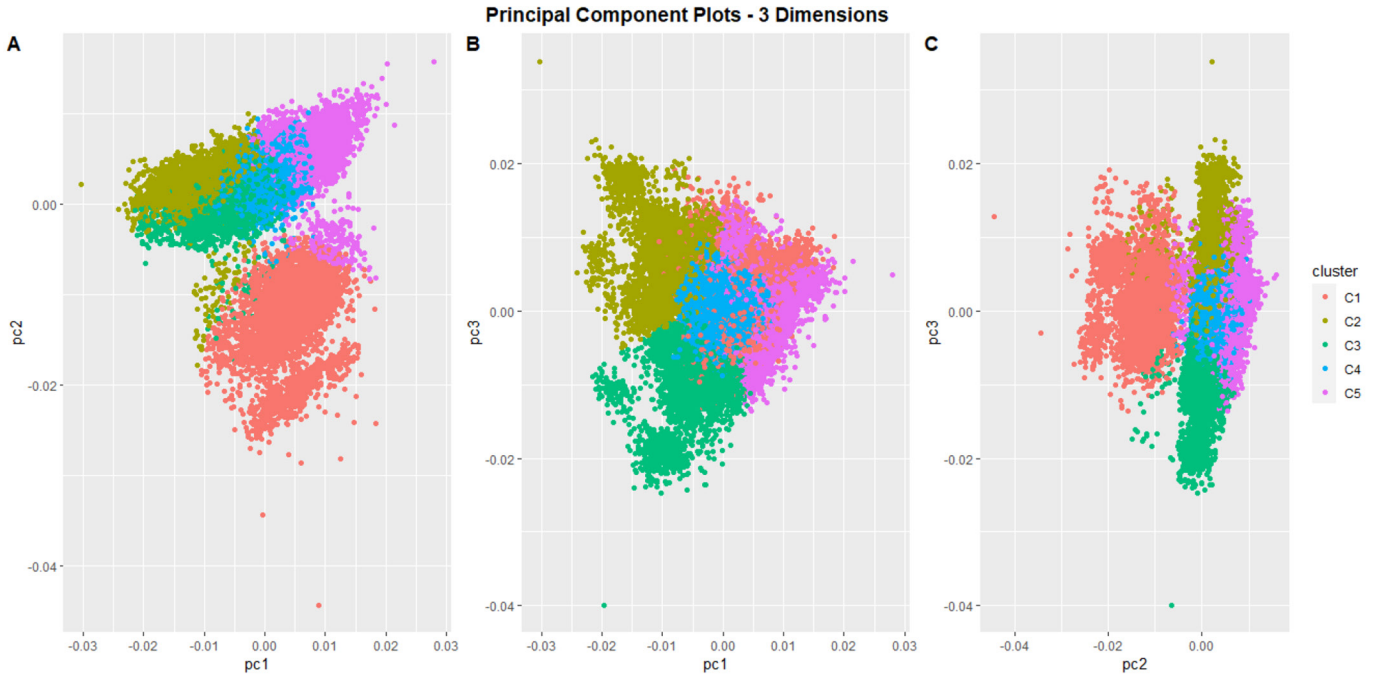

Figure 5. Principal component (pc) analyses plots for 3 dimensions showing the clustering results of the 20,099 selected candidates.

(IBD). Incomplete pedigrees and Mendelian sampling impair the estimation of true relationships based on IBD. The genomic relationship matrix detects any similarities between animals (IBS) and is independent of the available pedigree information. Clustering based on k-means forces animals to be assigned to only 1 specific cluster, which is a challenge when animals from different clusters share strong similarities.

The age distribution within each cluster is presented in Figure 2. It shows that C1 and C5 consisted of older bulls on average, compared with the rest, with the bull popularity in C1 peaking in 1987 (mean and median birth year) and C5 showing a more extended popularity period from 1985 to 2000 (with mean and median of 1990). The smallest cluster was C1, which contained the key sires Mark and Valiant, who have Chief as the primary progenitor. The largest cluster was C5 and contained Sweet Haven Tradition (Tradition), Rothrock Tradition Leadman (Leadman), and Bismay Tradition Cleitus (Cleitus) as key sires, representing multiple US breeding families. In C2, the periods of popularity peaked around 1985 and 2001 (mean of 1996 and median of 1999). That cluster had Chairman and Blackstar as key sires, with Ivanhoe and No-Na-Me Fond Matt (Fond Matt) as primary progenitors. Cluster 3 showed peak periods around 1990, 1995, and 2004, while C4 showed peaks around 1999 and 2005. The proportion of animals born after 1999 in each cluster was 6% (C1), 50% (C2), 31% (C3), 49% (C4), and 16% (C5). The key sire of C3, Bell, was a carrier of complex vertebral malformation and bovine leukocyte adhesion deficiency. Although he was also popular in other countries, he was almost exclusively used in the United States. The key sires of C4 are Canadian-born Hanoverhill Starbuck (Starbuck), Comestar Outside (Outside), Aerostar, and Stouder Morty (Morty). As reflected in Figure 3, this was the most international group.

When the number of clusters was increased from 5 to 10, the important sons of key sires were separated from C1, C3, and C4. A disadvantage of k-means clustering is the subjective choice of the number of clusters used. Because the total number of sires is 1,145, more clusters may become unreasonable. Animals will inevitably be separated regardless of whether they should be considered different groups or not.

Trait Differences

Statistically significant differences ($P < 0.05$) were found between groups for all traits, but many pairwise comparisons showed no differences (Table 2). Apart from protein, C1 and C5 were not different and generally had the lowest values for indices and production traits but the highest for reproductive rate. These clusters contained a larger number of older bulls. Historic bulls are expected to compare unfavorably to modern animals based on net merit and production traits (De

Table 2. The average breeding values per cluster for the 1,145 most popular sires of animals born after 1985

	Trait[1]						
Cluster	NMI ($)	CMI	FMI	DPR (%)	Milk (kg)	Fat (kg)	Protein (kg)
C1	−484[a]	−496[a]	−458[a]	0.85[a]	−503[a]	−20[a]	−18
C2	−289[b]	−295[b]	−278[bc]	−0.36[bc]	−222[bc]	−10[b]	−8[a]
C3	−358[c]	−372[c]	−329[bd]	−0.09[bd]	−264[bd]	−15[cd]	−11[b]
C4	−321[bc]	−327[bc]	−308[cd]	−0.55[cd]	−293[cd]	−13[bc]	−10[ab]
C5	−429[a]	−438[a]	−410[a]	0.64[a]	−420[a]	−18[ad]	−15

[a–d]Groups with corresponding letters did not show statistically significant differences at $P < 0.05$.
[1]NMI = net merit index; CMI = cheese merit index; FMI = fluid milk index; DPR = daughter pregnancy rate.

Vries, 2017) due to the remarkable progress that has been made over generations of selection. Due to strong selection of production traits, unfavorable correlations with fitness and reproductive traits (Berry et al., 2014) have created challenges for the dairy industry. A recent study aimed to reincorporate lost genetics from older bulls to improve diversity and fitness. The new progeny of these old bulls showed average, or better performance for DPR but below average for production traits (Dechow et al., 2020).

Clustering of Selected Candidates

Five clusters reduced the variance by 31%, while 7 clusters reduced variance by 35%. Most comparisons for this study were based on 5 clusters. The number of animals in each cluster was 3,577 (C1), 3,073 (C2), 3,302 (C3), 5,931 (C4), and 4,216 (C5). Each cluster contained females of 227 (C1), 296 (C2), 336 (C3), 894 (C4), and 328 (C5) sires. Of these sires, 47%, 41%, 37%, 37%, and 34% also appeared in the same cluster for C1, C2, C3, C4, and C5, respectively. The proportion of the females of each cluster that were sired by males in their own cluster was 85% (C1), 70% (C2), 64% (C3), 62% (C4), and 80% (C5). Some of these sires have daughters in more than one cluster. The proportion of daughters from sires in one cluster that are in other clusters are 3% (C1), 19% (C2), 35% (C3), 28% (C4), and 28% (C5). Table 3 shows the bulls with the most daughters in its own cluster, along with the number of daughters in each other cluster. Ensenado Taboo Planet (Planet) and 2 of his sons have the most daughters in C1, which shows a strong association between Planet and C1. Braedale Goldwyn (Goldwyn) and his sons are associated with C2, while Shottle and his sons are associated with C3. Cluster 4 is not dominated by any particular sire, although Comestar Outside (Outside) is a maternal grandsire of the bull with the most daughters in C4, Ronelee Toystory Domain (Domain), and a sire of the bull with the second most daughters, England-Ammon Million (Million). The 3 sires with the most progeny in C5 are all sons of O-bee Manfred Justice (Oman).

Table 3. The sires within a cluster with the most daughters in its own cluster, along with the number of daughters in each other cluster[1]

				Number of daughters in each cluster				
Bull's cluster	Bull name	Sire of bull	MGS of bull	C1	C2	C3	C4	C5
C1	Shamrock	Planet[2]	Shottle	521	11	4	0	0
	Observer	Planet[2]	Oman	488	0	0	0	7
	Planet[2]	Taboo	Amel	369	0	0	0	0
C2	Goldwyn[2]	James	Storm	0	310	0	0	0
	GW Atwood	Goldwyn[2]	Durham	0	274	0	0	2
	Gold Chip	Goldwyn[2]	Shottle	1	160	6	0	3
C3	Shottle[2]	Mtoto	Aerostar	0	0	434	0	0
	Beacan	Shottle[2]	BW Marshall	7	6	134	1	14
	Hill	Shottle[2]	Boliver	3	2	103	2	2
C4	Domain	Toystory	Outside	12	38	101	249	39
	Million	Outside	BW Marshall	1	19	48	123	4
	Colt P-Red	Lawn-Boy P-Red	Bolton	4	20	28	104	6
C5	Altaiota	Oman[2]	Juror	6	8	6	0	389
	Man-O-Man	Oman[2]	Altaaaron	0	0	1	0	268
	Freddie	Oman[2]	Die-Hard	11	0	0	0	260

[1]The sire and maternal grandsire (MGS) of each bull is included.
[2]Animals that appear to characterize each specific cluster.

Table 4. The average breeding values for the different groups using only male animals among the selected candidates[1]

	Trait						
Group	NMI ($)	CMI	FMI	DPR (%)	Milk (kg)	Fat (kg)	Protein (kg)
C1	98	99	97	−0.51[ab]	144	2.33	4.37
C2	−135	−134	−138	−0.83[a]	−165	−3.79[a]	−4.56
C3	−40	−42	−35	−0.16[cd]	−23[a]	−0.59	−1.49[a]
C4	−87	−91	−80	−0.40[bc]	−48[a]	−3.53[a]	−2.23[a]
C5	38	47	20	0.07[d]	36	3.89	3.36

[a–d]Groups with corresponding letters did not show statistically significant differences at $P < 0.05$.
[1]NMI = net merit index; CMI = cheese merit index; FMI = fluid milk index; DPR = daughter pregnancy rate.

Trait Differences

Table 4 shows the average breeding values for the males in each cluster. Pairwise differences were statistically significant for all indices, though many were not for DPR. Overall, C1 was better for all indices and traits except DPR and fat. For DPR, C5 was best along with C3. Cluster 3 tended to be intermediate for traits and indices. The poorest genetic merit overall tended to be C2, although DPR was not significantly different from C1 and fat was not significantly different from C4. Although C4 differed significantly for all indices compared with other clusters, most traits for C4 were not significantly different from at least 1 other cluster. Based on these results, the main line for reproduction and fat was C5, which ranked second-best for other traits and indices. Cluster 2 was the worst for all indices and traits.

Expected Inbreeding

Our clustering methods have the practical potential to manage diversity by mating animals across clusters. The expected inbreeding from random mating between all selected candidates is 0.121. Among the expected inbreeding when 5 clusters were used, 35% of the 20 combinations across clusters resulted in inbreeding greater than 0.121 (maximum of 0.129), while only 1 within cluster mating (C4 sires with C4 dams) was less (0.101). The other within cluster mating ranged from 0.168 to 0.215. Figure 6 shows the distribution of expected inbreeding when mating sires of C1 with dams of each other cluster. The distribution for within cluster mating (C1 sires with C1 dams) was shifted toward the right and had a wider distribution. A similar pattern was observed for all other mating combinations except for C4. Figure 7 shows that the expected inbreeding for all combinations were lower with narrower distributions.

This lower within cluster inbreeding for C4 was unexpected because clustering is expected to group animals that more related to each other together. In our study, all average mating combinations with C4 gave low inbreeding. Across cluster mating scenarios with C4 (whether sires or dams from C4) resulted in inbreeding ranging from 0.099 (C1 sires with C4 dams) to 0.108 (C3 sires to C4 dams). The unexpected results from C4, the largest cluster, may suggest that the cluster was genetically further removed from the other clusters, but still contained enough genetic diversity to allow low inbreeding levels within the cluster. This also suggests that more than 5 clusters can be used for the sake of mating strategies. However, even when using up to 10 clusters, one cluster always resulted in low inbreeding levels within and across cluster compared with all others. This group may contain all remaining animals that do not fit into distinct groups. As shown in Table 4, C4 is not dominated by a specific sire since those with the most daughters in C4 also have more daughters in other clusters.

Increasing the number of clusters for mating purposes resulted in higher within cluster inbreeding levels (up to 0.266 with 9 clusters) and higher maximum across cluster inbreeding (up to 0.212 with 9 clusters). Using more than 7 clusters also resulted in more across cluster combinations with high inbreeding levels (over 0.157). Based on these results, up to 7 clusters could be reasonable in this population.

Breeding programs need to consider inbreeding when making mating decisions. In particular, recent inbreeding is more detrimental than ancient inbreeding (Makanjuola et al., 2020b; Lozada-Soto et al., 2021). Our clustering method can successfully reduce the expected inbreeding of future generations compared with random mating. It is unknown how much difference this strategy can make on a genomic level since estimates reported here are based on pedigree information. Studies have found that inbreeding measured with genomic information can be more than double the estimate based on pedigree (Makanjuola et al., 2020a, Lozada-Soto et al., 2021). Focusing merely on inbreeding will not be the best approach to achieve genetic improve-

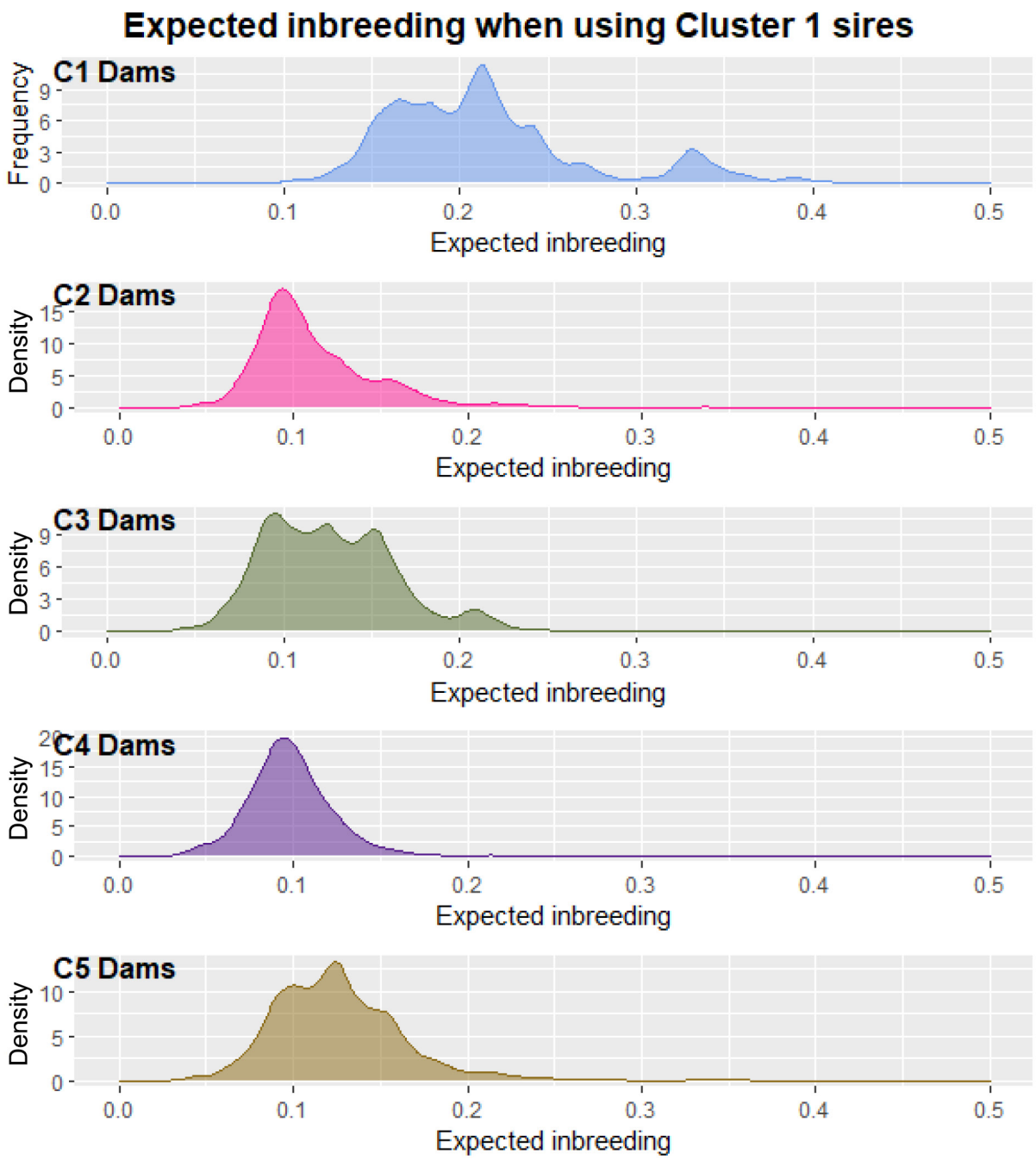

Figure 6. The expected inbreeding (calculated based on pedigree and all possible pairwise mating) when C1 sires are mated to dams of each cluster (C1, C2, C3, C4, C5). All animals are of generation 10 (G10).

ment. Based on expected inbreeding, C4 was the most suited to mate with animals from any other cluster to achieve lower expected inbreeding. However, average breeding values showed that this does not result in the best performing males. With the exception of DPR, it ranks fourth based on all 3 indices and yield traits.

Additionally, strictly applying across cluster mating with more clusters may decrease the selection intensity, which will slow the response to selection. Optimal contribution selection is an alternative, common, and helpful strategy to make genetic progress while limiting inbreeding (Meuwissen, 1997; Clark et al., 2013; Olsen

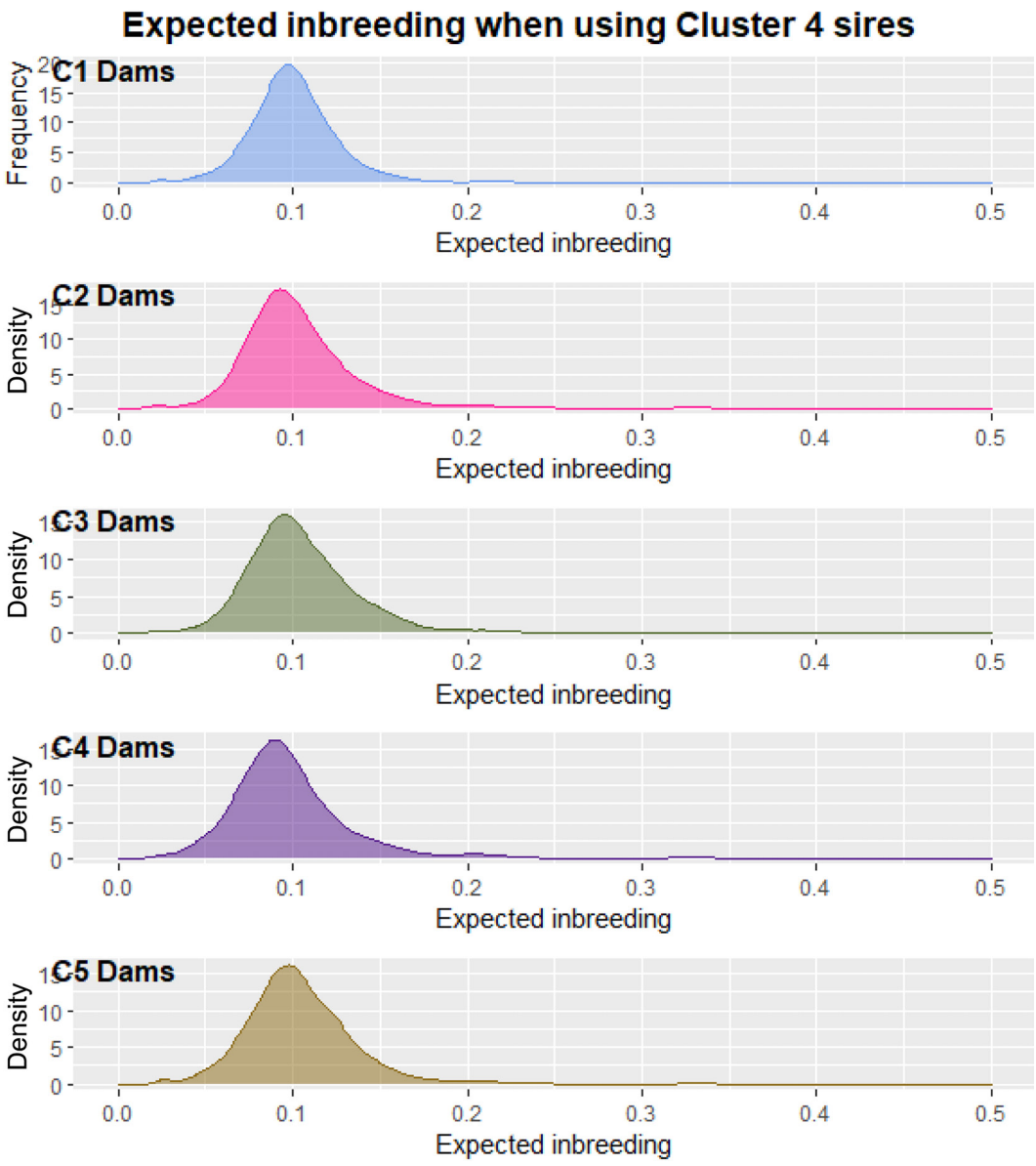

Figure 7. The expected inbreeding (calculated based on pedigree and all possible pairwise mating) when C4 sires are mated to dams of each cluster (C1, C2, C3, C4, C5). All animals are of generation 10 (G10).

et al., 2013; Meuwissen et al., 2020). In large populations, such as dairy, inbreeding restrictions based on pedigree did not lead to less genetic progress than inbreeding based on genomic information (Clark et al., 2013). Breeders can use commercial software to balance their breeding goals with a level of inbreeding they find acceptable.

Due to alternative strategies and good software available to US Holstein breeders, our method might not provide additional benefit. Pedigrees have been recorded for more than 100 years, a wide range of phenotypes and animal information has allowed traditional BLUP evaluations for decades, and genomic selection was incorporated in 2009. Our method could potentially be more useful to genotyped populations with considerably less information. This includes smaller livestock breeds, recently developed breeds, or countries that started recording recently. It can also be applied in conservation genetics where little is known about the population and strong selection is not applied to any particular trait.

CONCLUSIONS

Key sires within the US Holstein population were identified among sires with the most registered progeny born after 1985. Differences between clusters could be

observed in terms of international influence and time periods of use. Clusters with older bulls showed the worst performance for production but best for reproduction, which is expected due to strong selection over many generations. Clustering selection candidates into groups allows across cluster mating that can reduce the inbreeding of future generations. Based on the reduction of variance achieved by clustering and lower inbreeding across cluster, up to 7 groups may be present among the selection candidates. While our method can decrease or maintain the expected future inbreeding, it must be used in conjunction with selection based on genetic merit. Our method may be more beneficial for breeds or species where little is known about the population.

ACKNOWLEDGMENTS

This study received funding from the Holstein Association USA (Brattleboro, VT). The authors have not stated any conflicts of interest.

REFERENCES

Aguilar, I., and I. Misztal. 2008. Technical note: Recursive algorithm for inbreeding coefficients assuming nonzero inbreeding of unknown parents. J. Dairy Sci. 91:1669–1672. https://doi.org/10.3168/jds.2007-0575.

Baller, J. L., J. T. Howard, S. D. Kachman, and M. L. Spangler. 2019. The impact of clustering methods for cross-validation, choice of phenotypes, and genotyping strategies on the accuracy of genomic predictions. J. Anim. Sci. 97:1534–1549. https://doi.org/10.1093/jas/skz055.

Berry, D. P., E. Wall, and J. E. Pryce. 2014. Genetics and genomics of reproductive performance in dairy and beef cattle. Animal 8(Suppl 1):105–121. https://doi.org/10.1017/S1751731114000743.

Bjelland, D. W., K. A. Weigel, N. Vukasinovic, and J. D. Nkrumah. 2013. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. J. Dairy Sci. 96:4697–4706. https://doi.org/10.3168/jds.2012-6435.

Boddhireddy, P., M. J. Kelly, S. Northcutt, K. C. Prayaga, J. Rumph, and S. DeNise. 2014. Genomic predictions in Angus cattle: Comparisons of sample size, response variables, and clustering methods for cross-validation. J. Anim. Sci. 92:485–497. https://doi.org/10.2527/jas.2013-6757.

Capper, J. L., and R. A. Cady. 2020. The effects of improved performance in the US dairy cattle industry on environmental impacts between 2007 and 2017. J. Anim. Sci. 98:skz291. https://doi.org/10.1093/jas/skz291.

Capper, J. L., R. A. Cady, and D. E. Bauman. 2009. The environmental impact of dairy production: 1944 compared with 2007. J. Anim. Sci. 87:2160–2167. https://doi.org/10.2527/jas.2009-1781.

CDCB (Council on Dairy Cattle Breeding). 2021. Trend in inbreeding coefficients of cows for Holstein or Red & White, Calculated April 2021. Accessed Jun. 9, 2021. https://queries.uscdcb.com/eval/summary/inbrd.cfm.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44:4. https://doi.org/10.1186/1297-9686-44-4.

Clark, S. A., B. P. Kinghorn, J. M. Hickey, and J. H. van der Werf. 2013. The effect of genomic information on optimal contribution selection in livestock breeding programs. Genet. Sel. Evol. 45:44. https://doi.org/10.1186/1297-9686-45-44.

Daetwyler, H. D., B. Villanueva, P. Bijma, and J. A. Woolliams. 2007. Inbreeding in genome-wide selection. J. Anim. Breed. Genet. 124:369–376. https://doi.org/10.1111/j.1439-0388.2007.00693.x.

De Vries, A. 2017. Economic trade-offs between genetic improvement and longevity in dairy cattle. J. Dairy Sci. 100:4184–4192. https://doi.org/10.3168/jds.2016-11847.

Dechow, C. D., W. S. Liu, L. W. Specht, and H. Blackburn. 2020. Reconstitution and modernization of lost Holstein male lineages using samples from a gene bank. J. Dairy Sci. 103:4510–4516. https://doi.org/10.3168/jds.2019-17753.

Doekes, H. P., R. F. Veerkamp, P. Bijma, S. J. Hiemstra, and J. J. Windig. 2018. Trends in genome-wide and region-specific genetic diversity in the Dutch-Flemish Holstein-Friesian breeding program from 1986 to 2015. Genet. Sel. Evol. 50:15. https://doi.org/10.1186/s12711-018-0385-y.

Doublet, A.-C., P. Croiseau, S. Fritz, A. Michenet, C. Hozé, C. Danchin-Burge, D. Laloë, and G. Restoux. 2019. The impact of genomic selection on genetic diversity and genetic gain in three French dairy cattle breeds. Genet. Sel. Evol. 51:52. https://doi.org/10.1186/s12711-019-0495-1.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5. https://doi.org/10.1186/1297-9686-42-5.

Hartigan, J. A., and M. A. Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. Appl. Stat. 28:100–108. https://doi.org/10.2307/2346830.

Lozada-Soto, E. A., C. Maltecca, D. Lu, S. Miller, J. B. Cole, and F. Tiezzi. 2021. Trends in genetic diversity and the effect of inbreeding in American Angus cattle under genomic selection. Genet. Sel. Evol. 53:50. https://doi.org/10.1186/s12711-021-00644-z.

Makanjuola, B. O., C. Maltecca, F. Miglior, F. S. Schenkel, and C. F. Baes. 2020a. Effect of recent and ancient inbreeding on production and fertility traits in Canadian Holsteins. BMC Genomics 21:605. https://doi.org/10.1186/s12864-020-07031-w.

Makanjuola, B. O., F. Miglior, E. A. Abdalla, C. Maltecca, F. S. Schenkel, and C. F. Baes. 2020b. Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations. J. Dairy Sci. 103:5183–5199. https://doi.org/10.3168/jds.2019-18013.

Markert, J. A., D. M. Champlin, R. Gutjahr-Gobell, J. S. Grear, A. Kuhn, T. J. McGreevy Jr., A. Roth, M. J. Bagley, and D. E. Nacci. 2010. Population genetic diversity and fitness in multiple environments. BMC Evol. Biol. 10:205. https://doi.org/10.1186/1471-2148-10-205.

Meuwissen, T. H. E. 1997. Maximizing the response of selection with a predefined rate of inbreeding. J. Anim. Sci. 75:934–940. https://doi.org/10.2527/1997.754934x.

Meuwissen, T. H. E., A. K. Sonesson, G. Gebregiwergis, and J. A. Woolliams. 2020. Management of genetic diversity in the era of genomics. Front. Genet. 11:880. https://doi.org/10.3389/fgene.2020.00880.

Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica. 2014. Manual for BLUPF90 family of programs. in Athens: University of Georgia. Accessed Mar. 28, 2022. http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf.

Olsen, H. F., T. Meuwissen, and G. Klemetsdal. 2013. Optimal contribution selection applied to the Norwegian and the North-Swedish cold-blooded trotter: A feasibility study. J. Anim. Breed. Genet. 130:170–177. https://doi.org/10.1111/j.1439-0388.2012.01005.x.

Pszczola, M., T. Strabel, H. A. Mulder, and M. P. L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389–400. https://doi.org/10.3168/jds.2011-4338.

Saatchi, M., R. D. Schnabel, M. M. Rolf, J. F. Taylor, and D. J. Garrick. 2012. Accuracy of direct genomic breeding values for nation-

ally evaluated traits in US Limousin and Simmental beef cattle. Genet. Sel. Evol. 44:38. https://doi.org/10.1186/1297-9686-44-38.
VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.
Wellmann, R., and J. Bennewitz. 2019. Key genetic parameters for population management. Front. Genet. 10:667. https://doi.org/10.3389/fgene.2019.00667.
Yue, X.-P., C. Dechow, and W.-S. Liu. 2015. A limited number of Y chromosome lineages is present in North American Holsteins. J. Dairy Sci. 98:2738–2745. https://doi.org/10.3168/jds.2014-8601.

ORCIDS

Y. Steyn https://orcid.org/0000-0001-5467-9555
Y. Masuda https://orcid.org/0000-0002-3428-6284
S. Tsuruta https://orcid.org/0000-0002-6897-6363
D. A. L. Lourenco https://orcid.org/0000-0003-3140-1002
I. Misztal https://orcid.org/0000-0002-0382-1897
T. Lawlor https://orcid.org/0000-0002-4458-1025