



Nonparallel genome changes within subpopulations over time contributed to genetic diversity within the US Holstein population

Y. Steyn, T. Lawlor, Y. Masuda, S. Tsuruta, A. Legarra, D. Lourenco, I. Misztal

► To cite this version:

Y. Steyn, T. Lawlor, Y. Masuda, S. Tsuruta, A. Legarra, et al.. Nonparallel genome changes within subpopulations over time contributed to genetic diversity within the US Holstein population. *Journal of Dairy Science*, 2023, 106 (4), pp.2551-2572. 10.3168/jds.2022-21914 . hal-04089687

HAL Id: hal-04089687

<https://hal.inrae.fr/hal-04089687>

Submitted on 5 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



Nonparallel genome changes within subpopulations over time contributed to genetic diversity within the US Holstein population

Y. Steyn,^{1*} T. Lawlor,² Y. Masuda,¹ S. Tsuruta,¹ A. Legarra,³ D. Lourenco,¹ and I. Misztal¹

¹Department of Animal and Dairy Science, University of Georgia, 425 River Road, Athens 30602

²Holstein Association USA Inc., Brattleboro, VT 05302

³GenPhySE, INRA, INPT, ENVT, Université de Toulouse, Castanet-Tolosan 31520, France

ABSTRACT

Maintaining genetic variation in a population is important for long-term genetic gain. The existence of subpopulations within a breed helps maintain genetic variation and diversity. The 20,990 genotyped animals, representing the breeding animals in the year 2014, were identified as the sires of animals born after 2010 with at least 25 progenies, and females measured for type traits within the last 2 yr of data. K-means clustering with 5 clusters (C1, C2, C3, C4, and C5) was applied to the genomic relationship matrix based on 58,990 SNP markers to stratify the selected candidates into subpopulations. The general higher inbreeding resulting from within-cluster mating than across-cluster mating suggests the successful stratification into genetically different groups. The largest cluster (C4) contained animals that were less related to each animal within and across clusters. The average fixation index was 0.03, indicating that the populations were differentiated, and allele differences across the subpopulations were not due to drift alone. Starting with the selected candidates within each cluster, a family unit was identified by tracing back through the pedigree, identifying the genotyped ancestors, and assigning them to a pseudogeneration. Each of the 5 families (F1, F2, F3, F4, and F5) was traced back for 10 generations, allowing for changes in frequency of individual SNPs over time to be observed, which we call allele frequencies change. Alternative procedures were used to identify SNPs changing in a parallel or nonparallel way across families. For example, markers that have changed the most in the whole population, markers that have changed differently across families, and genes previously identified as those that have changed in allele frequency. The genomic trajectory taken by each family involves selective sweeps, polygenic changes, hitchhiking, and epistasis. The replicate frequency spectrum was used

to measure the similarity of change across families and showed that populations have changed differently. The proportion of markers that reversed direction in allele frequency change varied from 0.00 to 0.02 if the rate of change was greater than 0.02 per generation, or from 0.14 to 0.24 if the rate of change was greater than 0.005 per generation within each family. Cluster-specific SNP effects for stature were estimated using only females and applied to obtain indirect genomic predictions for males. Reranking occurs depending on SNP effects used. Additive genetic correlations between clusters show possible differences in populations. Further research is required to determine how this knowledge can be applied to maintain diversity and optimize selection decisions in the future.

Key words: K-means, clustering, polygenic adaptation, selection sweeps, epistasis

INTRODUCTION

Understanding the population structure of a breed is critical in revealing its genetic diversity and the changes occurring within its genome over time. Stratification of a single population into more distinct subpopulations allows for the identification of SNP that change in frequency in a uniform way across all subpopulations, and those that change uniquely within one or more subpopulations. Without stratification, the pooling of all animals together masks these family-specific changes. Recently, an abundance of genomic information on different species undergoing adaptive responses to environmental change or selection for different agricultural goals has become available. This has led to new ideas on evaluating adaptation and understanding the genetic architecture of traits (Csilléry et al., 2018; Barghi et al., 2020; Buffalo and Goop, 2020; Meuwissen et al., 2020; McGaugh et al., 2021; Rowan et al., 2021).

The additive genetic model does an excellent job of allowing breeders to change the phenotypic average of a population toward a desired goal. However, it does not expose the genetic complexity and diversity that help maintain the genetic variation that allows for current

Received February 1, 2022.

Accepted October 3, 2022.

*Corresponding author: yvette.steyn@uga.edu

and future genetic change. The breeding value is the sum of all markers affecting the trait and only considers additive effects. The different combinations may be near-infinite; thus, populations may show nonparallel changes in gene frequencies. Additionally, nonadditive effects add more complexity. These near-infinite possibilities to achieve the same genetic merit or phenotype lead to genetic redundancy. Genetic redundancy is a phenomenon where an excess of beneficial variants exist, which allows multiple genetic pathways to achieve the same phenotype (Goldstein and Holsinger, 1992; Nowak et al., 1997). Therefore, populations that have been separated and selected for the same trait, may have undergone different changes in allele frequencies (AF). This is due to different sets of loci responding differently to the same selection pressure. Heterogeneous change in AF among subpopulations is an indication of genetic redundancy (Barghi et al., 2019).

Genetic redundancy is caused by multiple factors. One or more genes may serve the same function and therefore, the absence of expression in one may not affect the phenotype (Pickett and Meeks-Wagner, 1995). Redundancy can also occur because highly polygenic traits are influenced by many genes that each have a relatively small contribution to the phenotype, hence the infinitesimal model (Fisher, 1918). Each allele would then be expected to slowly change by subtle shifts instead of selective sweeps of a few genes (Höllinger et al., 2019). Additionally, many genes not directly involved in obvious biological pathways of trait expression may collectively explain more variation in traits than those with more direct involvement (core genes), reflecting the omnigenic nature of traits (Boyle et al., 2017). It has been shown that up to 70% of trait variance can be attributed to trans-chromosomal effects through peripheral genes that affect the expression of core genes (Liu et al., 2019). These trans-chromosomal effects are partly due to pleiotropy (where genes are involved in the expression of more than one trait) and epistasis (where the expression of one gene influences the expression of another). In the US Holstein population, the percentage of interchromosomal epistatic effects varied from 1.9 to 84.2%, depending on the trait (Prakapenka et al., 2021). Because many traits of economic importance in livestock are highly polygenic and omnigenic, a combination of selective sweeps and subtle shifts can be expected.

The global dairy industry is dominated by a few breeds, particularly the Holstein. Concern has been expressed that artificial insemination has resulted in the widespread use of semen from a handful of bulls (Yue et al., 2015), which can lead to higher inbreeding and genetic similarities worldwide. Although this genetic connectedness can be advantageous for genetic evalu-

ations and the similarity of animals provides a more uniform and predictable product, it may be problematic for long-term genetic improvement and adaptability. Although inbreeding can increase the frequency of favorable genes for traits under selection, it leads to the decrease in performance of other traits, in particular fertility and overall health (Pryce et al., 2014), as well as the loss of rare alleles that could be of importance in the future. Maintaining genetic diversity is crucial for a population to adapt to changing environments, such as climate change and consumer preferences.

The objectives of this study were to investigate the amount of stratification occurring within the US Holstein population and observe nonparallel changes that have contributed to the differences in these subpopulations.

MATERIALS AND METHODS

Data

The study was based on already available data; therefore, ethical approval was not required. 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 most progenies for a sire was 58,266. This sire was Marshfield Elevation Tony (Mars). The average number of progenies per sire was 29. The data file contained only type traits and totaled 10,067,745 records. Genotypes were available for 569,404 animals. Imputed genotypes from various SNP panels were obtained from the Council on Dairy Cattle Breeding. Markers were not removed from these imputed genotypes based on allele frequency or Hardy-Weinberg equilibrium deviations. After removal of unmapped and sex chromosomes, 58,990 SNP markers remained.

The genomic relationship matrix (\mathbf{G}) was obtained using the formula $\mathbf{G} = \frac{\mathbf{MM}'}{2 \sum p_i (1 - p_i)}$, where \mathbf{M} is a matrix of SNP content centered by twice the current AF, and p_i is the current allele frequency for SNP i (VanRaden, 2008).

Clustering the Selected Candidates

Potential selected candidates in 2014 (i.e., the breeding animals in 2014 that were selected by genomic testing) were identified as sires of animals born after 2010 with at least 25 progenies (3,902 animals), and cows that were recorded for type traits in 2013 or 2014 (16,197 animals). The registration number of animals represented 14 countries, including Australia, Austria, Canada, the Czech Republic, Germany, Denmark,

Table 1. The average expected inbreeding of offspring resulting from hypothetical mating within cluster and across cluster¹

Cluster ID	Predominant sire in cluster	Sire birth date	Inbreeding within cluster	Inbreeding across cluster
1	Planet	2003	0.22	0.11
2	Goldwyn	2000	0.20	0.11
3	Shottle	1999	0.18	0.12
4	Multiple sires	—	0.10	0.10
5	O Man	1998	0.17	0.11

¹The expected inbreeding if animals were mated at random is 0.12.

Finland, France, Great Britain, Hungary, Italy, the Netherlands, Sweden, and the United States. K-means clustering (Hartigan and Wong, 1979) of the genomic relationship matrix identified 5 clusters of animals (C1, C2, C3, C4, and C5) to characterize the genetic diversity of the Holstein. K-means clustering aims to increase across-cluster variation while decreasing within-cluster variation. Up to 10 clusters were explored. Based on the reduction in sum of squares as more clusters are used, 5 to 7 clusters could be reasonable for identifying groups that are genetically more different and of sufficient size to allow further evaluations. Many clusters would reduce the number of selected candidates in each cluster, which would in turn reduce the number of animals used to estimate SNP effects and calculate AF. The number of animals in each cluster was 3,577 (C1), 3,073 (C2), 3,302 (C3), 5,931 (C4), and 4,216 (C5). A principal component analysis was performed on the genomic relationship matrix to visualize the separation of clusters. The plot is presented in Figure 1. The animals are labeled according to the cluster that was assigned to them using K-means clustering.

Hypothetical matings were performed within and across clusters with the INBUPGF90 software package within the BLUPF90 software suite (Misztal et al., 2014). Expected inbreeding of offspring was calculated for every possible mating between a specific group of sires and specific group of dams. Solutions were based on the complete pedigree information of the Holstein population assuming nonzero inbreeding for unknown parents (Aguilar and Misztal, 2008). The average expected inbreeding of animals when mating within-cluster, and of all animals in across-cluster scenarios, are presented in Table 1.

The additive genetic correlations between clusters were estimated following an adjustment to the procedure proposed by Duenk et al. (2020). This compares the breeding values of populations when expressed within different genetic backgrounds. Stature was used as a trait to achieve this goal. Five different indirect genomic predictions (**IGP**) of male animals from each cluster were obtained, one where SNP effects based on females of the same cluster were used, and the rest where SNP effects based on females of the other clus-

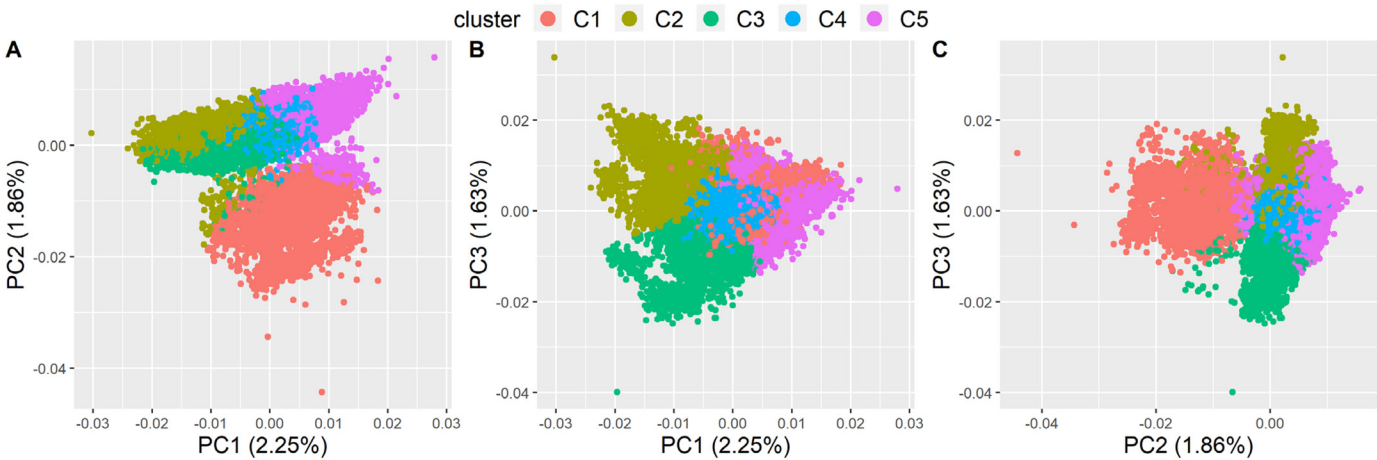


Figure 1. Principal component (PC) analysis plots for 3 dimensions showing the clustering results of selected candidates (generation 10). The analyses were based on the genomic relationship matrix. The colors are assigned to the clusters that resulted from a K-means clustering of the genomic relationship matrix.

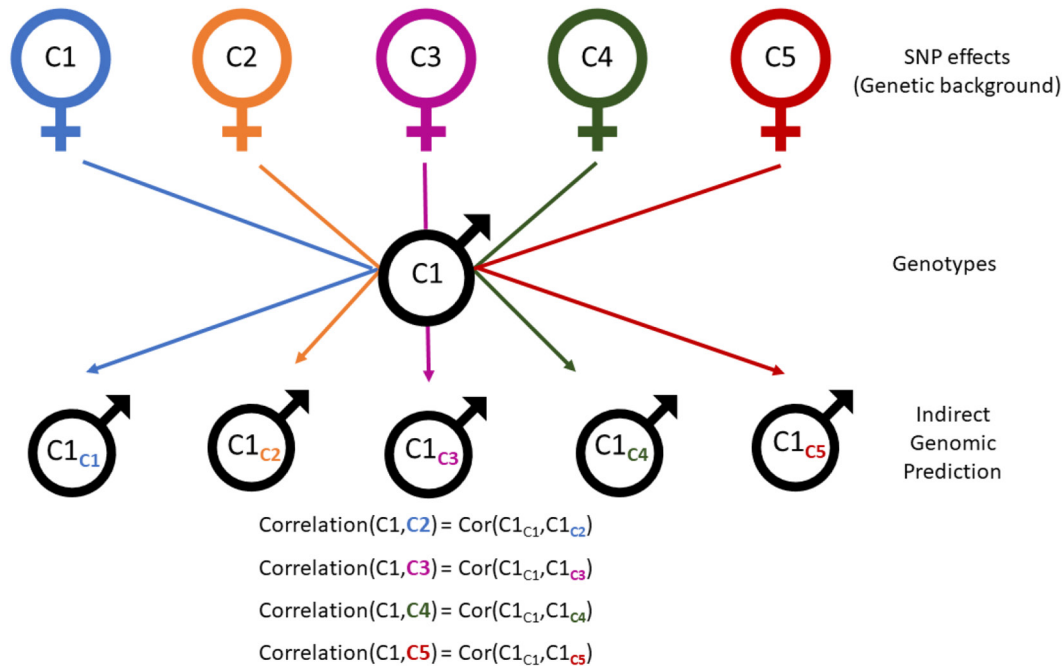


Figure 2. The genetic correlations between clusters are based on the expression of the genotypes of one population within the genetic background of another. Here, 5 indirect genomic predictions are estimated for males of correlation 1 (C1) based on SNP effects of females of each cluster.

ters were used. This is therefore the genetic expression of an animal when expected to perform within the genetic background of different groups. Figure 2 illustrates how IGP were estimated based on the genetic background of different clusters and correlated. The model was, as described by Tsuruta et al. (2021), for only one trait. Cluster-specific SNP effects were calculated from GEBV estimated with GBLUP using BLUP-F90IOD2 with only females from each cluster. The SNP effects were estimated for each cluster separately based on these GEBV using POSTGSF90 with the formula $\hat{\mathbf{a}}_c = \lambda \mathbf{DZ}_c' \mathbf{G}_c^{-1} (\text{GEBV}_c)$ (VanRaden, 2008; Wang et al., 2012), where $\hat{\mathbf{a}}_c$ is a vector of estimated SNP effects for cluster c , λ is the ratio of SNP to additive genetic variance, \mathbf{D} is a diagonal matrix of weights for SNP (in this case an identity matrix), \mathbf{Z}_c included the SNP effects of cluster c , GEBV_c is the GEBV of females in cluster c , and \mathbf{G}_c^{-1} is the inverse genomic relationship matrix containing female animals in cluster c . The IGP were obtained with the formula $\text{IGP}_j^c = \mathbf{z}_j' \hat{\mathbf{a}}_c$, where IGP_j^c is the IGP of animal j within the genetic background of cluster c , and \mathbf{z}_j is a vector of SNP content of animal j centered by twice the current AF. The additive genetic correlations between clusters were calculated as the Pearson correlation between the IGP of male animals in one cluster when using SNP effects of its own cluster,

and that of another cluster. The method provides 2 different correlations because one refers to the IGP of population 1 based on the SNP effects of population 2, and the other to the IGP of population 2 based on the SNP effects of population 1. These are different due to the different AF in the respective clusters. Correlations between breeding values without taking the reliability into account will underestimate genetic correlations. Therefore, the method of Calo et al. (1973) was used to adjust correlations. The adjustment factor was calculated as

$$\sqrt{\frac{\left(\sum_i^n Rel_{iA}\right) \left(\sum_i^n Rel_{iB}\right)}{\sum (Rel_{iA} Rel_{iB})}}, \text{ where } n \text{ is the number}$$

of male animals in generation 10 of family A, Rel_{iA} is the reliability of the indirect genomic prediction of animal i when using SNP effects from family A, and Rel_{iB} is the reliability of animal i when using the SNP effects of family B. Individual reliability for IGP was obtained using the PREDf90 software package (Misztal et al., 2014). This software computes the reliability of indirect predictions based on the prediction error covariance of SNP effects which in turn is backsolved from prediction error covariance of genotyped animals (Gualdrón Duarte et al., 2014; Lourenco et al., 2019; Legarra et al., 2022). Animals from each cluster were ranked according to their IGP based on different SNP effects.

Families

To observe change over time, the transmission of SNP markers from one animal to another should be traced back from the current animals to the oldest ancestors. We created 5 pseudofamilies based on each cluster. First, we assigned selected candidates to clusters as described before (C1, C2, C3, C4, and C5). Within each cluster, we traced back 10 generations to form 5 families (F1, F2, F3, F4, and F5) of up to 11 nondiscrete generations. Generation 10 (G10) was the selected candidates (the clusters) and generation 0 (G0) was the oldest animals detected by tracing back the pedigree. Note that because generations overlap, one sire might be a parent of an animal in G7 and also a parent of an animal in G8. Thus, this animal will be present in both G7 and G8 of the same family; assignment of animals to generations is ambiguous as pedigrees overlap. This concept is visually explained in Figure 3. In this figure, we show only sires and maternal grandsires for simplicity, but both sires and dams were included in our study. The main sire represented in a particular family is in larger font, bold, and italicized: Planet is the main sire in family 1, Goldwyn in family 2, Shottle in family 3, and O Man in family 5. Family 4 does not have a clear major sire. O Man appears in G8 of both families 1 and 5. Shottle is present in G10 of family 3, and G8 of families 2, 3, and 5. Goldwyn appears in G8 and G10 of family 2, and G7 of family 3. Mtoto occurs in 3 of the 5 families in G7 and G6. Sires that are repeated across families or generations are in bold and corresponding colors. As generations are traced further back, more animals are in common to more than one family. Generations 9 and 10 have 5 animals and one animal per family. Generation 8 contains 2 animals per family, a total of 10 entries, but only 7 animals (O Man, Taboo, Goldwyn, Shottle, Alta-Baxter, Bacculum-Red, and Bolton). Generation 7 has 3 animals per family, 15 total entries, but 11 animals (Manfred, Majic, Amel, Durham, James, Mtoto, Goldwyn, Blitz, Rubens, Alta-Hershel, and Alta-Aaron). Generation 6 has 5 animals per family, a total of 25 entries, but 16 animals (Elton, Blackstar, Cubby, Mark, Choice, Prelude, Alta-Grand, Storm, Aerostar, James, Emory, Mtoto, Alta-Astre, Rudolph, Alta-Luke, and Convincer). Families 1, 2, 4, and 5 each include 12 animals (from G10 to G6), but family 3 includes 11 animals because Mtoto occurs in both G7 and G6 of the same family. There are 7 animals that are unique to only family 1 (Observer, Taboo, Majic, Amel, Blackstar, Mark, and Choice), whereas there are 6 animals unique to only family 2 (Airlift, GW Atwood, Durham, James, Alta-Grand,

and Storm). Similarly, 6 are unique to only family 3, 9 to family 4, and 3 in family 5.

Grouping of animals into different families is highly dependent upon the most recent ancestors. Ancestors varying in their occurrence in different families alter the gene flow, resulting in different AF across families. The number of genotyped animals in each generation of each family are presented in Table 2. The total number of unique genotyped animals for each family is 7,411 (family 1), 6,373 (family 2), 6,700 (family 3), 11,041 (family 4), and 8,333 (family 5). The oldest animal in all families was born in 1952 (Osborndale Ivanhoe).

Changes in Allele Frequencies

Changes in AF for the whole breed were calculated from the differences in AF between each generation for all families combined. Specific within-family allele frequency changes were determined by starting with the G10 animals within each cluster and tracing backward. Five different procedures were used to identify SNPs changing in a parallel or nonparallel way across families to observe visually. These procedures included allele frequency changes for specific genes known to have changed substantially in the US population, the largest regression coefficient when regressing AF on generation, the variance and range in the absolute difference in AF between G0 and G10 across the 5 families, and those SNPs identified by the Lewontin and Krakauer's test (Lewontin and Krakauer, 1973), which is defined below.

Selected Genes

The *DGAT1* gene on chromosome 14 was observed over time due to its known significant genetic effect on milk production (Thaller et al., 2003; Barbosa da Silva et al., 2010). Additionally, *AVEN* (chromosome 10), *SPATA6* (chromosome 3), *ERBB4* (Chromosome 2), *SKIV2L* (chromosome 23), and *USP13* (chromosome 1) were chosen based on results from Ma et al. (2019), which showed that these genes are among those that have changed the most in the US Holstein population. The chosen genes were neither on the same chromosome nor the sex chromosomes.

Change in Allele Frequencies Over Generations

All families were combined to determine the allele frequency change over time for the overall population. The allele frequency was calculated for each generation and regressed over generations to obtain a slope for each SNP marker. The formula was $y_{kl} = \beta_{0k} +$

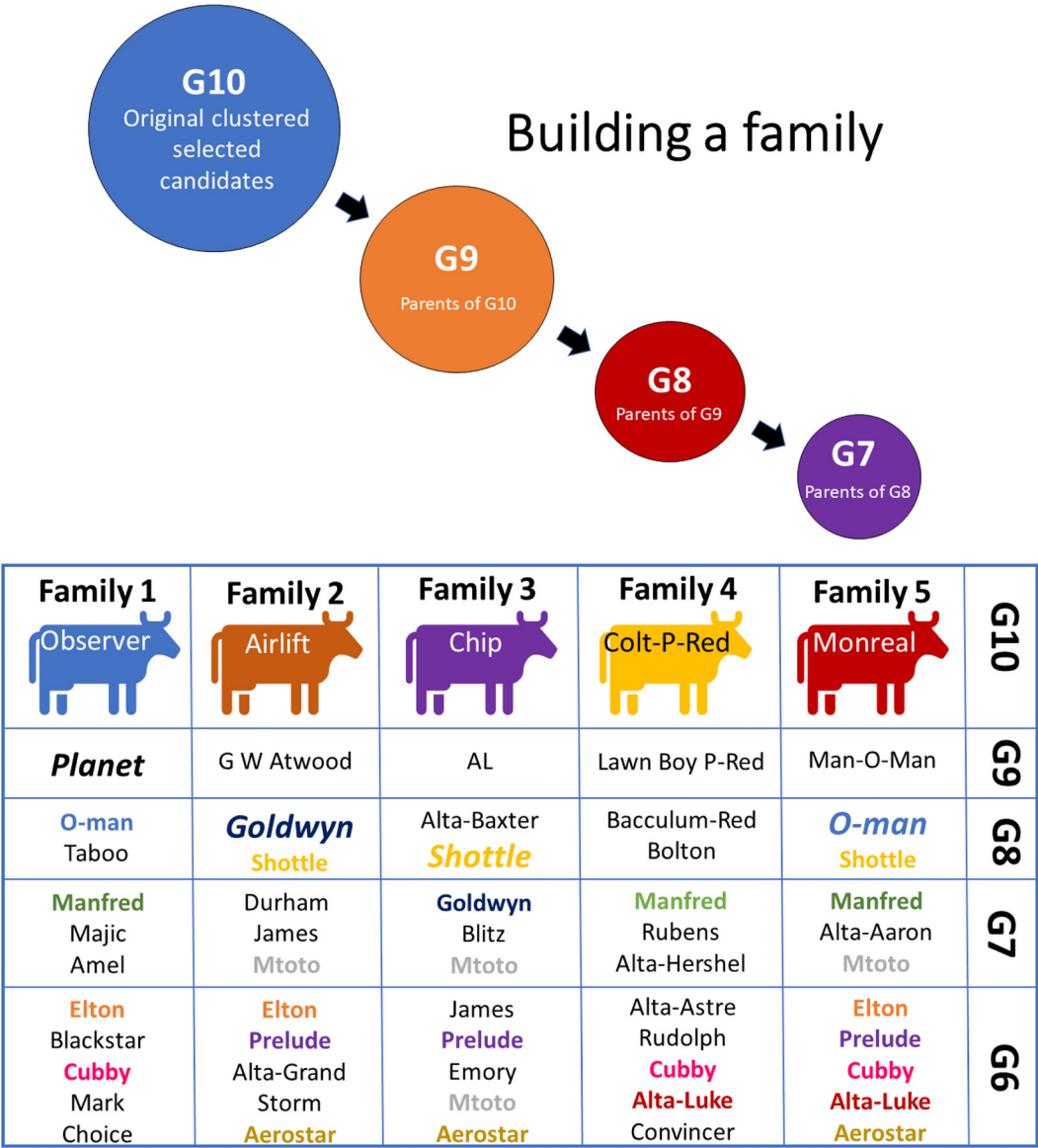


Figure 3. Families were built by tracing pedigrees back 10 times. The original clustered selected candidates are in generation 10 (G10), whereas generation 9 (G9) contained parents of G10, generation 8 (G8) contained parents of G9, and so on, until generation 0 (G0). A sire line is illustrated to explain overlapping generations and common ancestors. Colors indicate sires that are repeated across families or generations. Sires in large, bold italics are the main sires represented in that family. They are also present in G10.

$\beta_{1k}\mathbf{x}_1 + e_{kl}$, where \mathbf{y}_{kl} is the vector of AF for SNP k in generation l , β_{0k} is the mean allele frequency for SNP k , β_{1k} is the regression coefficient over generations for SNP k , \mathbf{x}_1 is a vector of generation l (0 to 10), and e_{kl} is the error associated with SNP k in generation l . The 20 SNP with the greatest change over time were identified. Of these SNP, 5 that were further apart (based on genomic location) were identified. The 5 SNP selected were those with the highest, 5th highest, 6th highest, 16th highest, and 19th highest regression coefficients.

Greatest Variance of Change Over Generations Within Families

The absolute difference between G10 and G0 was calculated within each family, providing an estimate for change in allele frequency over time. To identify SNP markers that have changed differently across families, the variance of these differences in the 5 families was calculated. The 20 SNP with the highest variance were identified. The 5 selected SNP had the highest, 2nd

Table 2. The number of genotyped animals per generation within each family¹

Generation	Family 1	Family 2	Family 3	Family 4	Family 5
G10	3,577	3,073	3,302	5,931	4,216
G9	1,513	1,471	1,478	2,830	1,859
G8	1,337	1,242	1,302	2,364	1,591
G7	1,043	975	1,004	1,666	1,161
G6	838	792	797	1,139	914
G5	645	582	608	839	669
G4	467	432	454	603	489
G3	336	304	310	426	346
G2	243	229	221	299	245
G1	189	171	171	223	183
G0	148	135	137	171	146

¹The most recent generation (G10) was used for clustering; their pedigrees were traced back 10 generations (G9 to G0).

highest, 3rd highest, 5th highest, and 19th highest variance.

Greatest Range of Change Over Generations Within Families

The range between the allele frequency change of the family with the least change, and that of the family with the greatest change was calculated. This is an additional measure to identify SNP markers that have changed differently across families. The top 20 SNP were identified, and 5 were chosen to avoid markers close to each other and eliminate those that were also among the top 20 based on variance. These 5 SNP had the 2nd highest, 5th highest, 7th highest, 9th highest, and 13th highest range.

Lewontin and Krakauer Test

Genetic drift and migration also contribute to changes in gene frequencies over time (Falconer and Mackay, 1996). The Lewontin and Krakauer test (\mathbf{T}_{LK}) aims to identify markers that changed due to selection, not drift or migration (Lewontin and Krakauer, 1973). The test uses a measure of genetic differences among subpopulations, namely the fixation index (\mathbf{F}_{st}) test as defined by Wright (1943). The F_{st} test is effectively the fraction of total genetic variance due to differences among subpopulations. Only SNP markers with a minor allele frequency >0.05 based on all families combined were used for this test. The AF of these SNP within G10 of each family were calculated. Let $\mathbf{p} = (\mathbf{p}_1, \dots, \mathbf{p}_5)$ be a vector of the AF of an individual SNP in each of the 5 families. The F_{st} for each SNP was calculated as

$$F_{st} = \frac{s_p^2}{\bar{p}(1 - \bar{p})},$$

where \bar{p} is the mean of vector \mathbf{p} and s_p^2 is the sampling variance of each SNP across families.

This is used in the T_{LK} formula (Lewontin and Krakauer, 1973):

$$T_{LK} = \frac{n-1}{\bar{F}_{st}} F_{st},$$

where n is the number of families and \bar{F}_{st} is the mean F_{st} of all markers. The T_{LK} follows a χ^2 distribution with $n - 1$ df. The P -values were obtained from this distribution based on the T_{LK} of each marker. The Bonferroni adjustment and the false discovery rate were used as measures of significance.

Nonparallel Changes

The replicate frequency spectrum can be used to measure heterogeneity across populations (Barghi et al., 2019). Our modified version of the replicate frequency spectrum compares how the 100 markers that changed most in one family also changed similarly in other families. If markers changed differently across families, it suggests redundancy or divergent selection/adaptation. The absolute difference between allele frequency in G0 and G10 was used to identify the 100 SNP that have changed the most in each family.

The proportion of markers that have changed direction within each family (initially increased but later decreased, or initially decreased but later increased) was calculated. Changes were measured by comparing the regression coefficient when regressing allele frequency over G0 to G5, and the regression coefficient when regressing allele frequency over G5 to G10. Instead of using zero as the cut-off point for directional change, we identified those SNP that changed in one direction (increased or decreased) at a rate of at least 0.02 per

Table 3. The number of times a prominent young bull appears as a sire of animals in generation 10 (or generation 9) of each family

Name	Family 1	Family 2	Family 3	Family 4	Family 5
Planet	658 (321)	0 (42)	0 (27)	0 (13)	0 (104)
Goldwyn	0 (99)	449 (399)	0 (92)	0 (43)	0 (158)
Shottle	0 (209)	0 (171)	584 (492)	0 (44)	0 (222)
Domain	22 (0)	42 (0)	118 (0)	276 (7)	43 (1)
BW Marshall	0 (7)	0 (19)	0 (25)	34 (65)	2 (25)
Oman	0 (77)	0 (49)	0 (49)	1 (20)	95 (223)

generation in the first phase, and in the opposite direction at a similar rate in the second phase. Less strict cut-offs of 0.01 or 0.005 per generation were also used to detect more subtle directional changes.

RESULTS AND DISCUSSION

Studies have shown that cross-validation accuracy in determining the usefulness of genomic prediction was lowest when using K-means as clustering (Saatchi et al., 2012; Boddhireddy et al., 2014; Baller et al., 2019). Because the accuracy of genomic predictions depends on the relationships between the training and target populations (Habier et al., 2010; Clark et al., 2012; Pszczola et al., 2012), this suggests that K-means clustering is successful at separating groups that are more related to each other but less related to other clusters. K-means clustering identified 5 clusters as giving shape to the genetic diversity of the young animals in the Holstein breed. A principal component analysis was performed to visualize potential differences between clusters. The principal component analysis plot in Figure 1 is labeled according to the cluster assigned by K-means clustering and reveals different but overlapping subgroups within the population. The first 3 principal components (PC) explain 2.25, 1.86, and 1.63% of the variance of the genomic relationships. The plots of the first 2 PC suggest that cluster 1 is more distinct from the rest. This seems reasonable, as cluster 1 contains mostly direct descendants (i.e., one generation removed from their sire Planet). Whereas clusters 2, 3, and 4 contain animals with more distant and overlapping relationships (i.e., the grand offspring of Goldwyn, Shottle, and a variety of other bulls), cluster 5 appears to be distinct as animals within this cluster are related based upon relationships that are one more generation removed. Comparing the first and third PC, clearer separation is observed between clusters with the exception of cluster 1 and 4. Comparing the second and third PC also shows separation between clusters 1, 2, and 3 but less between the others. Cluster 4 appears to mostly overlap with all other clusters.

Families

Stratification of the Holstein population into different families will continuously change with each successive generation. High numbers of chromosomal segments from high-profile bulls provides a strong selection signal for a limited time period. Table 3 shows the number of times that several prominent bulls appear in different families as a sire of a G10 or G9 animal. Selection signals from earlier generations become pooled together, as most of our dairy animals have multiple shared ancestors. As shown in Figure 4, the proportion of shared ancestors across families increases with each generation back in time. Whereas 82% of the parents of G10 animals are unique to that family, this drops to 46% by G7 and to 35% by G6. With each new generation, ancestral chromosomes are broken and recombined many times and in many ways. The connections of several highly influential bulls appear to be equally distributed across the genotyped ancestors of each family (Table 4). For this reason, as we will discuss later, starting AF across families differ very little.

Differences in the Most Recent Generation

Extensive use of artificial insemination allows the genetics of a prominent sire to be transmitted to the population via different family members. Similar to the 4 paths of selection described by Rendel and Robertson (1950), a prominent bull may transmit his genes to the next generation in 4 ways: sire of sires of production cows (BB), sires of production cows (BC), sires of breeding dams that produce sires of production cows (CB), and sires of breeding dams of production cows (CC). These different paths of selection, coming from animals born in different years, leads to several waves of descendants being born at different time periods. The most direct path is sire to a daughter or a son. A strong selection signal for certain bulls is anticipated, as the number of daughters can easily be in the tens of thousands, and the additive genetic relationship between parent and offspring (0.5) is high. The selection signal of other very prominent bulls, but slightly older,

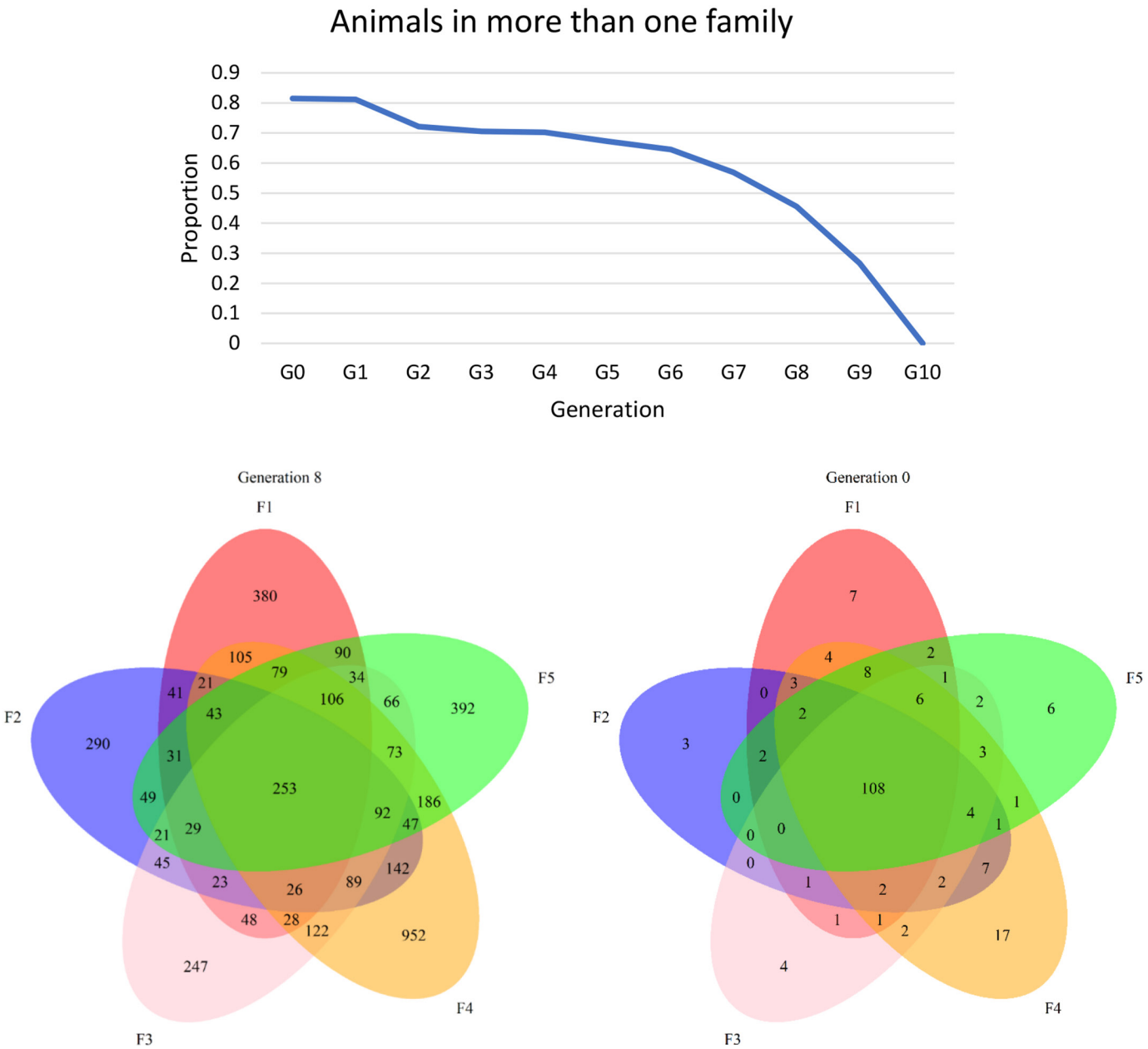


Figure 4. The proportion of animals that appear in more than one family (F) in each generation, and Venn diagrams illustrating the overlapping nature of families in generation 8 (G8) and the founder population (G0).

Table 4. The number of times historically prominent ancestors appear in each family among only genotyped animals

Name	Family 1	Family 2	Family 3	Family 4	Family 5
Ivanhoe Star	7	7	6	7	7
Chairman	8	8	6	9	7
Chief	10	12	12	14	11
Elevation	17	20	20	26	18
Bell	29	25	24	33	26
Starbuck	24	32	23	34	26
Blackstar	49	47	44	53	41

can be even stronger as many copies of their genotype will be spread throughout the population by a large number of sons. For example, among animals born from 2010 to 2012, 58,288 daughters of Planet were included in the genetic evaluation for milk production. Goldwyn, Shottle, and O Man were born 3, 4, and 5 yr earlier, respectively. These 3 bulls had 407,709, 532,438, and 416,313 granddaughters from their respective sons. Although the transmission of genetics is one more generation removed (grandsire instead of sire), the high number of granddaughters for these 3 bulls results in a higher number of offspring equivalents and a signal of selection as strong or stronger than Planet.

The K-means clustering placed the majority of descendants of a prominent sire into the same cluster. Family members are grouped together based upon the strength of their relationship to one another. Daughters and sons of the youngest bulls are in the cluster with the highest relationships, granddaughters of slightly older bulls in subsequent groups, and more distant relatives of the older prominent bulls in other groups. That is, daughters and sons of Planet are clustered into one group, granddaughters of Goldwyn and Shottle are largely separated into 2 different clusters, and so on. Cluster groups do not exclusively contain ancestors of only these bulls. Ancestral generations overlap. But the representation of prominent bulls allows for cross-family comparisons to be made. In our study, the female animals in generation 10 of each family were born from 2008 and 2012. The bulls with the highest number of close descendants in generation 10 were Planet, Goldwyn, Shottle, and O Man, respectively, listed by their age. Their descendants are 0, 1, 2, or 3 generations removed from the time when these bulls were first selected. Table 3 shows the number of times a that a prominent bull is a sire or a grandsire of animals in G10. Family 1 contains the most highly related animals in generation 10 and the youngest of these prominent sires. Slightly older bulls, Goldwyn and Shottle, are the primary sires of the G10 and G9 animals for families 2 and 3, respectively. Many of these animals are granddaughters of these 2 bulls. O Man, who is another year older, has granddaughters and great-granddaughters in G10 of family 5. Family 4 does not contain a prominent sire, with less related animals from a variety of sires represented.

Differences in inbreeding when mating generation 10 of each family with each other suggested genetic differences between these families. High within-family inbreeding and low across-family inbreeding indicate that animals have indeed been clustered with animals that are genetically more similar to each other and more different than animals in other clusters. A noticeably smaller than expected inbreeding occurs for all mating

scenarios with animals in generation 10 of F4, whether within- or across-family. This shows that although the most recent generation in F4 is more different than those of other families, it still contains enough variation to allow low inbreeding (Table 1). This suggests that it might be appropriate to select more clusters with the K-means method. Even when 10 clusters were used, one cluster still had lower inbreeding (not shown). More than 5 clusters would have reduced the number of animals per generation for each family, which is a disadvantage for calculating AF and could lead to more sampling bias.

An additional measure of genetic differences, the F_{st} , indicates genetic changes not due to drift alone. The average F_{st} value for markers across generation 10 of the 5 families was 0.03, which is lower than the 0.07 found in a French study that compared 3 different dairy breeds—Holstein, Montbéliarde, and Normande (Flori et al., 2009)—but higher than the expected value (0.00) if populations were uniform. The selected candidates do not represent one large panmictic population. They come from a complex mixture of family subgroups with differences in AF. As shown in the skewed distribution in Figure 5, most SNP have low F_{st} values. No SNP reached statistical significance for F_{st} , although some came close. This suggests that most of the differences in generation 10 across families is not due to strong, divergent selection.

Differences in allele substitution effects across populations can be due to differences in AF, epistasis, dominance, and differences in linkage disequilibrium (Legarra et al., 2021). The SNP effects for stature were estimated when using female animals in generation 10 of each cluster separately or combined. Differences in IGP based on these SNP effects show that allele substitution effects are indeed different across our families. Table 5 presents the adjusted correlations between IGP of animals based on different SNP effects, thus the expression of genotypes of one family within the genetic background of other families. All are compared with the IGP within-family (e.g., IGP of males of generation 10 in family X when SNP effects were based on females of generation 10 of the same family). In general, family 2 is more separated from the other families, with an average correlation of 0.78 with other families. High correlations are apparent between generation 10 of some families, such as 0.99 between family 5 and 3, 0.92–0.99 between Family 1 and Family 3, and 0.94–0.99 between Family 1 and Family 5. Overall, males of family 3 have higher genetic correlations with the females in G10 of all families. Shottle is the prominent male in family 1 and has a larger number of granddaughters across all families (Table 3). Nonadditive gene actions, such as dominance and epistasis, can contribute to these differ-

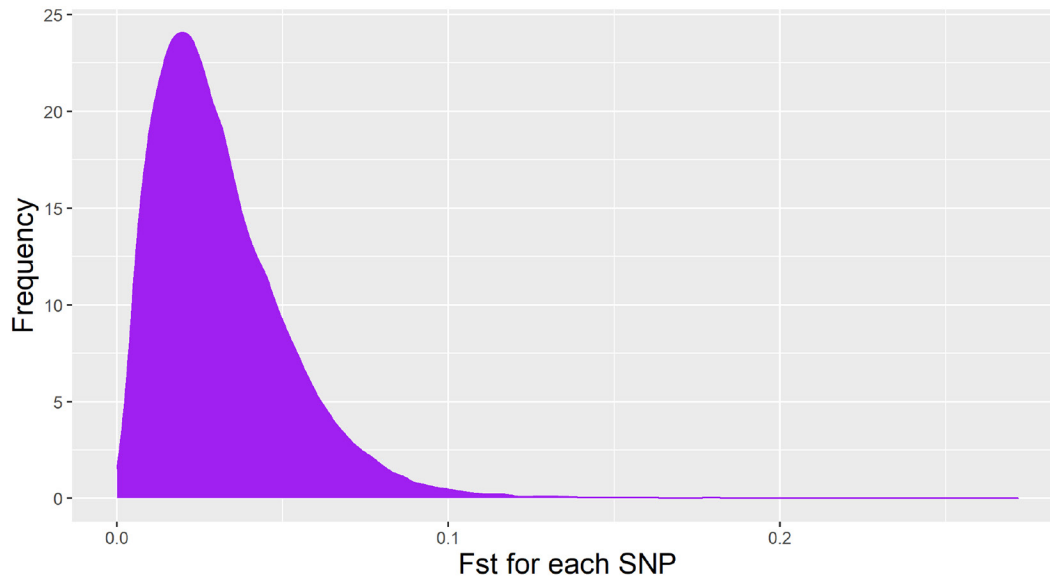


Figure 5. The distribution of fixation index (F_{st}) values of SNP markers across generation 10 of all families.

ences in genetic correlations. A correlation lower than 0.80 is not due to dominance alone. A simulation by Duenk et al. (2020) with realistic epistatic scenarios decreased the correlation between populations to as low as 0.45 without adjustments. Before the adjustment in our study based on Calo et al. (1973), the highest correlation between clusters is 0.45, and the lowest after adjustment is 0.41. Thus, results suggest that epistasis may be present.

Reranking of Indirect Genomic Predictions

An important consideration when calculating the breeding value of an animal is the specification of the population(s) where the animal will be mated. Wade

and Goodnight (1998) reviewed the assumptions used in Fisher's infinitesimal model and Wright's shifting balance theory. Under Fisher's model, the US Holstein would be described as a single, large, panmictic population with a singular set of allele substitution effects. The average effect of an allele would be estimated from all animals combined, and the targeted mates would be from a uniform population. Under Wright's model, the Holstein breed consists of multiple demes (or families) with shifting allelic values depending upon the genetic background of these subpopulations. Genetic redundancy, across families, leads to the additive effect of an allele varying between families and resulting in a unique set of breeding values for each family. Our results are more in line with Wright's shifting balance theory, where reranking is expected within different subpopulations.

Table 6 shows the ranking of 4 specific bulls (2 from generation 10 of family 2 and family 5) based on their IGP when using SNP effects based on generation 10 of all families combined (ALL) or each separate family. Reranking of bulls depending upon the target population is evident. Even though Airlift ranks the highest of all the G10 males when using ALL SNP effects, he only ranks within the top 10 again when using SNP effects of F2. Alleles have different substitution effects in different populations, and therefore, alleles that are more favorable in one population may not be favorable in another. Further investigation is required to determine whether results from this study can be used to improve the accuracy of genetic evaluations for individual subgroups within the population.

Table 5. The additive genetic correlation between generation 10 of each family (F) when the genotype of the validation population (males) is expressed in the genetic background of each family (SNP effects based on females from each cluster)¹

SNP effects used	Validation population				
	F1	F2	F3	F4	F5
F1	1.00	0.89	0.99*	0.89	0.99*
F2	0.99*	1.00	0.86	0.75	0.67
F3	0.92	0.56	1.00	0.89	0.99*
F4	0.81	0.87	0.96	1.00	0.90
F5	0.94	0.59	0.99*	0.77	1.00

¹Each pairwise comparison has 2 correlations: (1) when the indirect genomic prediction of males from family X is based on the SNP effects obtained from females in family Y, and (2) when the indirect genomic prediction of males in family Y is based on the SNP effects obtained from females in family X.

*Results exceeded 1.00.

Table 6. The ranking of 2 bulls from each cluster (C) when indirect genomic predictions for stature are based on different SNP effects¹

Bull	Group SNP effects were based on					
	All	C1	C2	C3	C4	C5
Bulls in G10 of cluster 2						
Airlift	1	101	3	64	44	276
G.W. Atwood	22	81	8	1,150	212	446
Bulls in G10 of cluster 5						
Monreal	2	46	97	140	30	2
Broch	4	52	146	124	19	57

¹The SNP effects were based on female animals in generation 10 (G10) of each family separately or all combined.

Genome Changes

Evolve and resequence studies observe changes in AF over generations. An example is a study by Barghi et al. (2019), where they observed a natural outcross population as it adapted to a higher temperature. This led to different subpopulations with their own genetic solutions (redundancy) converging to the same phenotypic goal. Genetic redundancy in dairy cows is expected as breeders define an overall breeding objective or fitness measure and select breeding animals from the available candidates representing several different subpopulations.

To fairly compare the genomic changes across families from a similar starting point, the earliest generation needs to contain similar animals, whereas the more recent generations are different. The large proportion of shared ancestors in the earlier generations allows for greater similarity. This can be seen in Tables 3 and 4. Prominent bulls, such as Planet, Goldwyn, Shottle, BW Marshall, and Oman, are not exclusive to a specific family as they appear in the different families and generations, albeit in different proportions. However, these younger, prominent sires drive the genetic diversity within the most recent generation of each family (G10). Overall, differences in their proportional influence upon a family as a whole allow them to shift the frequency of different alleles over time. Heterogeneous changes in AF across families can come from similar ancestors contributing a varying percentage of descendants. Tracing back to some of the historically most prominent ancestors, such as Elevation, Chief, and Blackstar, it can be seen that they are more evenly represented in the genotyped animals of each family. This provides a homogeneous early genetic base with similar initial AF across all families.

Our analysis concentrated on temporal changes in AF over several generations among 5 different families. Heterogeneity in allele frequency changes across families indicates that different sets of SNPs are changing

over time. Having several distinct families with diverse genomes helps maintain genetic variation over time. Comparing allele frequency changes in our study allowed us to identify family-specific signatures of selection. Fixation of alleles was infrequent across the whole population (3 alleles); however, it was greater and varied within families. The number of SNP markers that became fixed within each family were 38, 22, 22, 59, and 40 for F1, F2, F3, F4, and F5, respectively. None could be described as a selective sweep as all had initial frequencies near fixation.

A subset of results is presented here; Figures 6 and 7 show selected genes, Figures 8 and 9 show markers with the most change based on the regression coefficient, Figure 10 shows the marker with the greatest variance, Figure 11 the marker with the greatest range, and Figure 12 the marker with the greatest F_{st} . Results for other chosen markers are presented in the supplemental material (https://figshare.com/projects/Non-parallel_genetic_change/145899; Steyn, 2022).

The frequency of alleles can change due to factors including selection, migration, random genetic drift, linkage, epistasis, and pleiotropy (Falconer and Mackay, 1996; Barghi et al., 2020). All factors can also contribute to genetic redundancy, especially when traits are polygenic and complex. Changes can be rapid, such as selective sweeps when genes have a large effect, or slow and unpredictable, such as subtle shifts when traits are highly polygenic. Markers surrounding a SNP undergoing rapid change can also change similarly simply due to its proximity to the marker of interest, a phenomenon known as linkage drag or hitchhiking (Santiago and Caballero, 1998). This hitchhiking can reduce genetic diversity because it can lead to higher genomic inbreeding (Pedersen et al., 2010). The direction of frequency change of a hitchhiker depends on the phase. If the favorable allele started at a low frequency and was in opposite phase with the hitchhiker, their changes will be in opposite directions. Possible hitchhiking was observed in markers surrounding the SNP we chose for

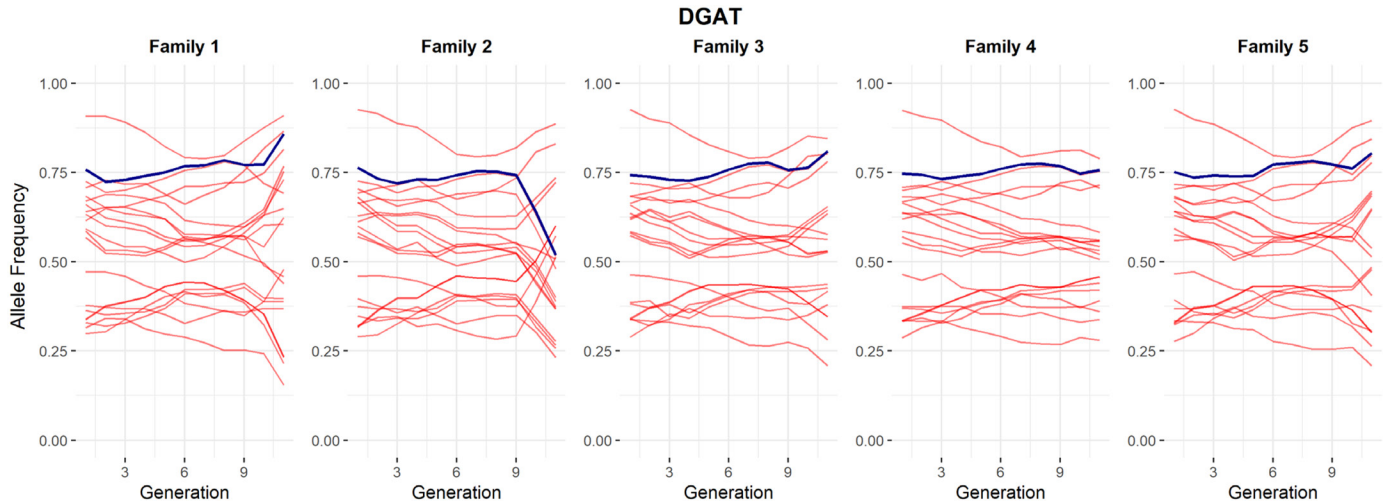


Figure 6. The allele frequency of the *DGAT* gene (blue) and surrounding 20 SNP markers (red) per generation within each family.

observation based on our criteria, including Figures 6, 8, 10, and 11. Pleiotropy is when one gene contributes to more than one trait, such as the *DGAT1*, which will be discussed in more detail below. Antagonistic pleiotropic effects can prevent the fixation of an allele when the desired phenotype is a compromise between the different effects. The expression of a gene can be affected by the expression of another (epistasis). Both pleiotropy and environmental conditions can affect epistasis (Arjan et al., 2011). Estimating gene interactions is statistically and computationally challenging due to the large number of possible pairwise interactions and is generally ignored in genetic evaluations. A recent study on US Holstein by Prakapenka et al. (2021) showed that the degree of epistasis varies depending on trait, with pro-

duction traits having less than 40% interchromosome epistatic effects, but over 80% for daughter pregnancy rate. A change in direction or magnitude of SNP effects can be as a result of epistatic interactions (Mackay, 2014). The frequency may increase, or decrease initially but change if an antagonistic relationship with another gene is present. These changes could also be due to changes in selection pressure, whether due to artificial or natural selection or hitchhiking. Table 7 shows the number of SNP markers that have changed direction when comparing the regression coefficient of allele frequency over G0 to G5, with the regression coefficient of allele frequency over G5 to G10. Whereas the number of SNP that reversed direction over time numbered in the thousands, the proportion of SNP markers that initially

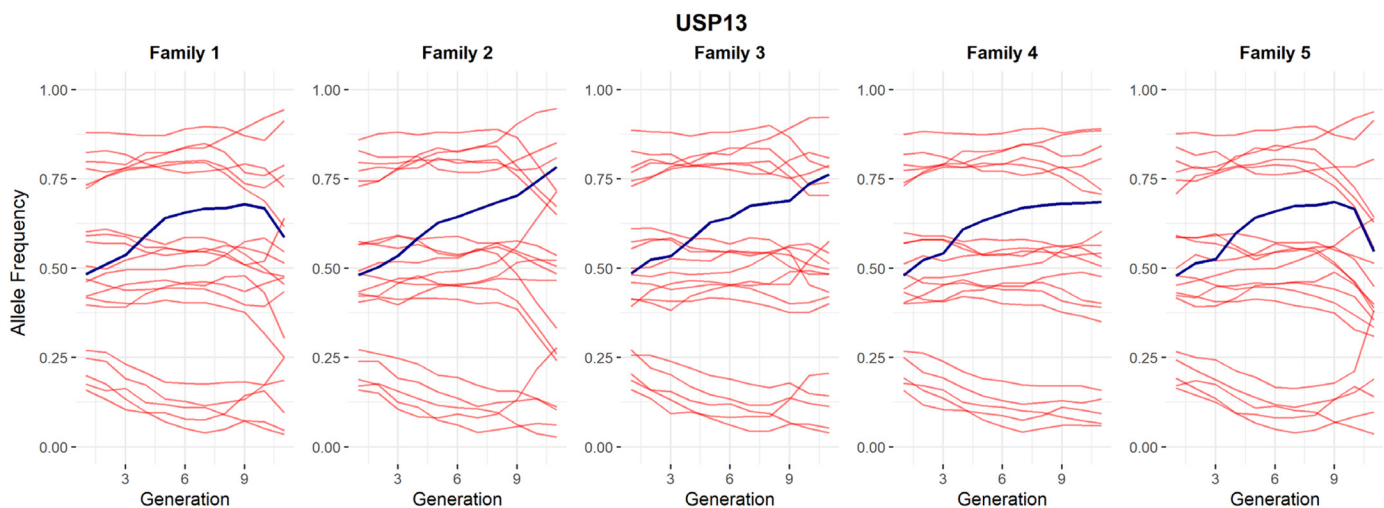


Figure 7. The allele frequency of the *USP13* gene (blue) and surrounding 20 SNP markers (red) per generation within each family.

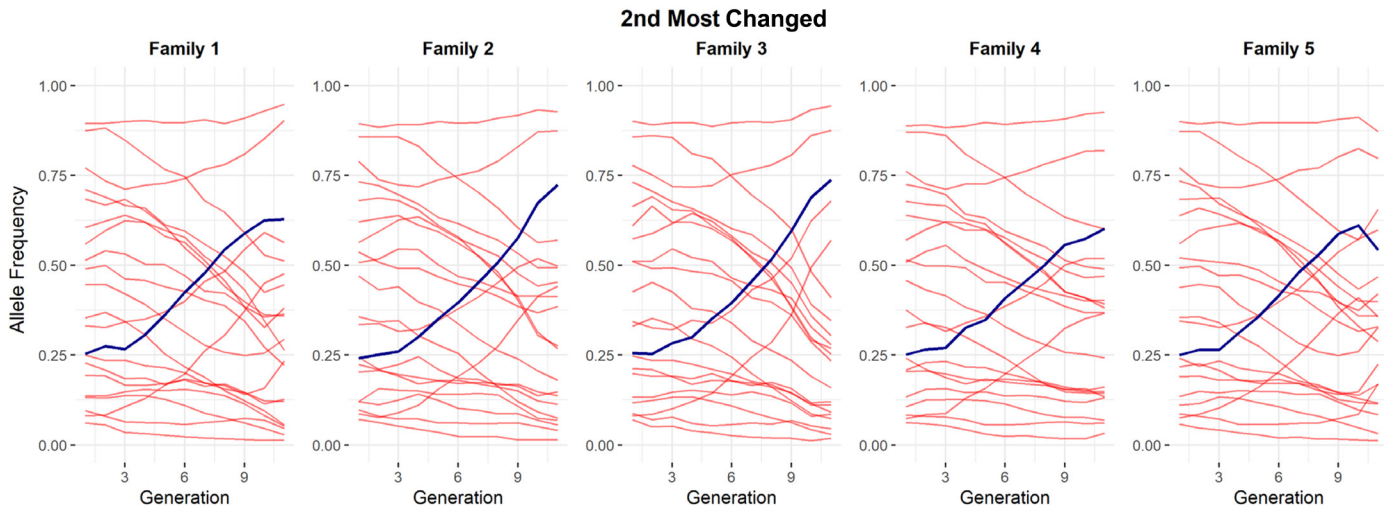


Figure 8. The allele frequency of the second selected SNP among the 20 SNP markers that have changed the most over time based on the regression coefficient (blue) and the surrounding 20 SNP markers (red) per generation within each family.

increased (or decreased) by 0.005 per generation until G5, and decreased (or increased) by 0.005 until G10, ranged from 0.14 in family 4 to 0.24 in family 1. Due to the large number of possible gene-by-gene interactions, estimating their presence is statistically challenging, especially when the number of genotyped animals is low. A possible measure of nonadditive effects over time is the correlation of SNP effects estimated per generation within a family/line (Legarra et al., 2021, personal communication with Jorge Hidalgo). This method cannot be applied within our families as the number of animals per generation in our older generations is too small to estimate SNP effects with confidence.

The genome trajectory taken by each family involves a combination of the above-mentioned factors. These can be observed in plots of specific alleles selected based on these different criteria. All figures showing changes over time are presented in Supplemental Data S1 (https://figshare.com/projects/Non-parallel_genetic_change/145899; Steyn, 2022), while selected examples are shown for discussion purposes in this manuscript.

Genes. The allele frequency changes for *DGAT1* for all families are presented in Figure 6. It starts at a high frequency and remains relatively unchanged or shows small increases for 4 families. This gene is known to significantly affect fat yield (Spelman et al.,

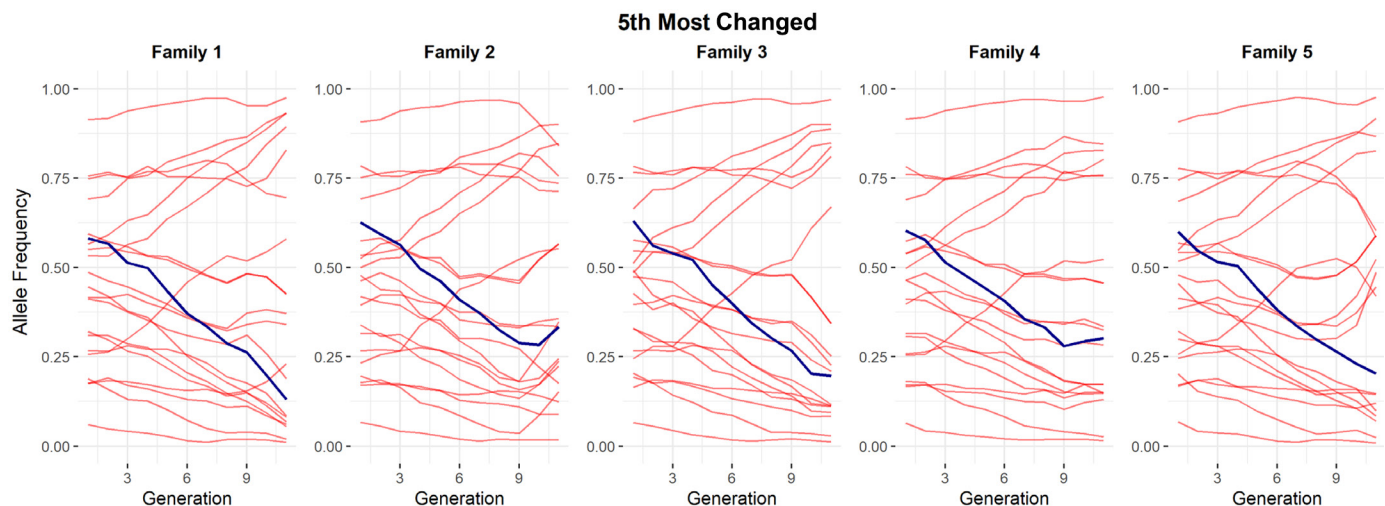


Figure 9. The allele frequency of the fifth selected SNP among the 20 SNP markers that have changed the most over time based on the regression coefficient (blue) and the surrounding 20 SNP markers (red) per generation within each family.

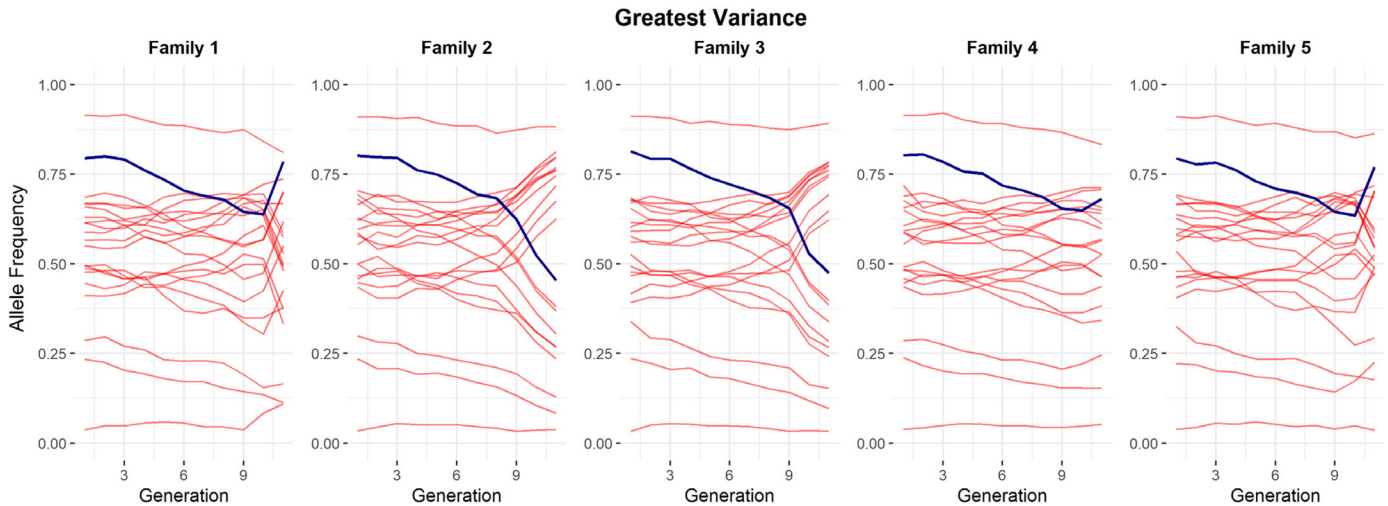


Figure 10. The allele frequency of the first selected SNP among the 20 SNP markers that have shown different changes across families (blue) and the surrounding 20 SNP markers (red) per generation within each family.

2002; Thaller et al., 2003; Jiang et al., 2019). The high starting point of allele frequency in G0 is not surprising given that milk production traits have undergone selection decades before the birth of our G0. If this gene was acting alone, the expectation would have been that it would be fixed in the population. However, due to its antagonistic effects on milk yield, its frequency has remained similar over time (Jiang et al., 2018). The change for F2 is distinctly different, clearly decreasing. Surrounding SNP markers also show different behavior compared with other families. This may reflect a change in breeding objectives in the dairy industry because consumers' attitude toward fat in human diets has changed over time.

The *AVEN* gene is associated with male fertility (Laurentino et al., 2011) and shows similar increasing trends in allele frequency changes, with surrounding SNP markers showing different behavior in F2 (Supplemental Data S1; https://figshare.com/projects/Non-parallel_genetic_change/145899; Steyn, 2022). The *ERBB4* gene is involved in embryonic lethality (Tidcombe et al., 2003). The F1 and F5 show sharper increases in allele frequency compared with other families, but the overall trend may also be considered similar. The surrounding SNP markers in F1 show hitchhiking signals (Supplemental Data S1; https://figshare.com/projects/Non-parallel_genetic_change/145899; Steyn, 2022). This may only be observed in F1 because linkage

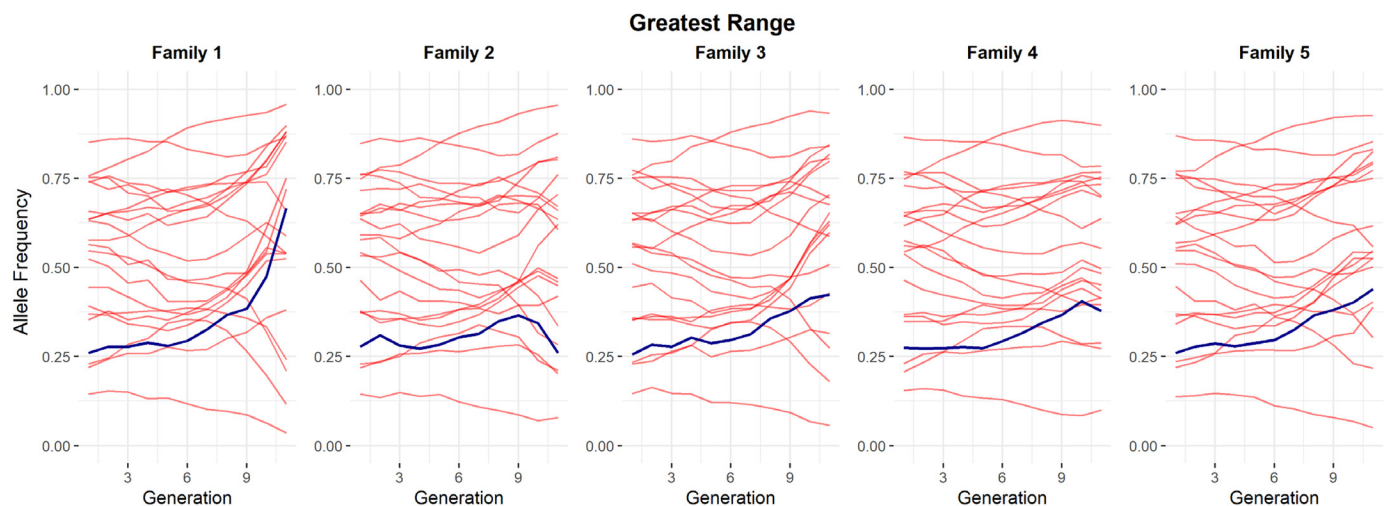


Figure 11. The allele frequency of the first selected SNP among the 20 SNP markers that have shown small changes in at least one family and large changes in another (blue) and the surrounding 20 SNP markers (red) per generation within each family.

Table 7. The number and proportion of SNP markers that have changed direction in each family¹

Family	Rate of allele frequency change					
	>0.02		>0.01		>0.005	
	Number	Proportion	Number	Proportion	Number	Proportion
Family 1	1,086	0.02	6,765	0.11	14,132	0.24
Family 2	780	0.01	5,986	0.10	13,012	0.22
Family 3	833	0.01	6,238	0.11	13,161	0.22
Family 4	56	0.00	2,172	0.04	8,121	0.14
Family 5	839	0.01	6,285	0.11	13,450	0.23

¹These are SNP that changed at a rate of 0.02, 0.01, or 0.005 allele frequencies per generation (G) in one direction (positive or negative) from G0 to G5, and the same (but opposite) rate from G5 to G10.

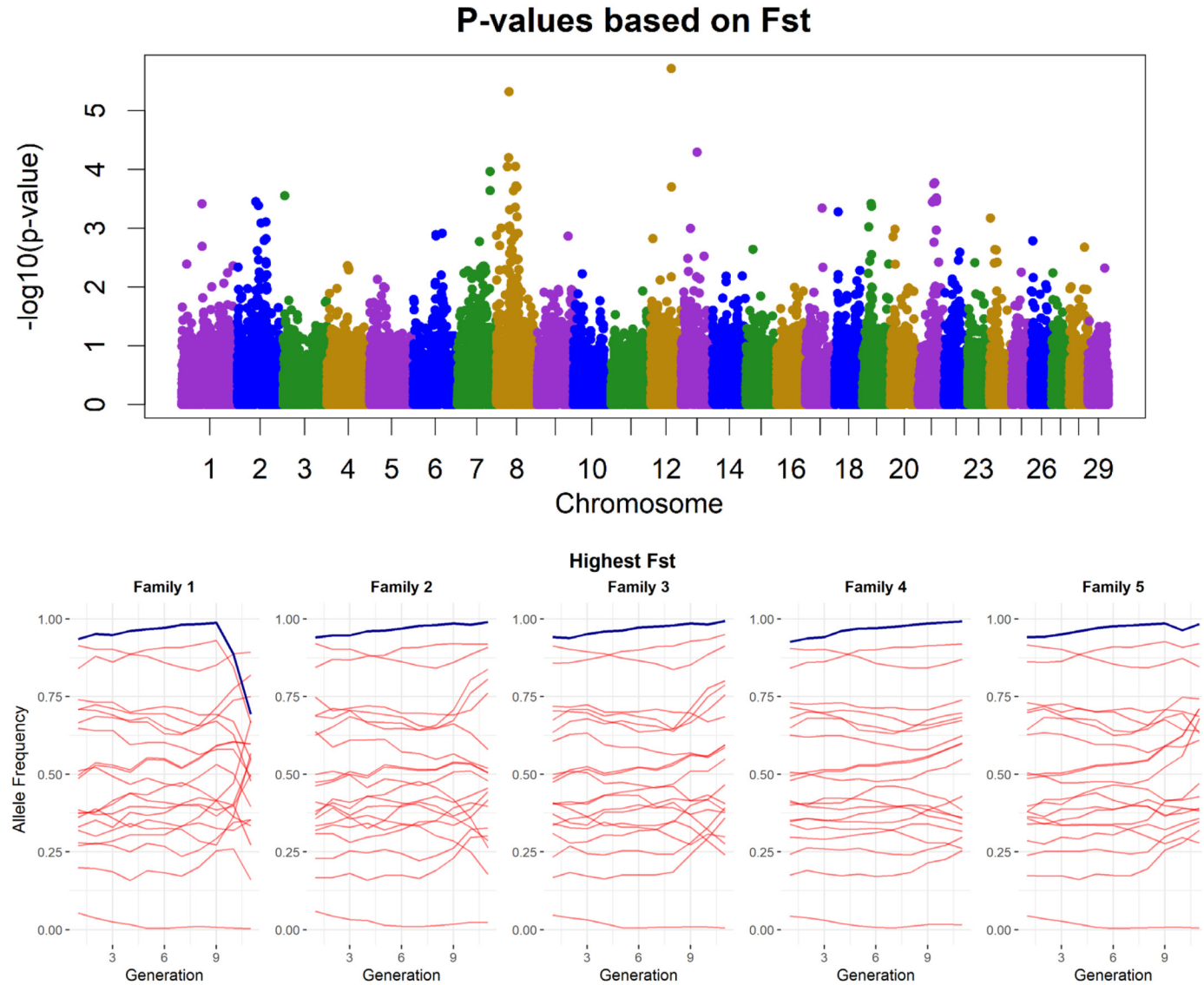


Figure 12. The Manhattan plot with *P*-values based on the Lewontin and Krakauer extension of the fixation index (*F_{st}*) test, and the allele frequency of the SNP with the highest Lewontin and Krakauer value (blue) and the surrounding 20 SNP markers (red) over generations within each family.

disequilibrium was lost due to recombination in other families.

The *SKIV2L* gene is also involved in fertility (Ma et al., 2019). Here, more heterogeneity is observed as the allele frequency in F3 start to decrease from G9 whereas others continue to increase. Differences can also be observed in the surrounding SNP markers (Supplemental Data S1, https://figshare.com/projects/Non-parallel_genetic_change/145899; Steyn, 2022). The *SPATA6* gene is associated with sperm quality (Yuan et al., 2015) and also shows heterogeneous changes. The allele frequency increased up to G7 for all families, after which they plateau or slowly decrease in F2, F3 and F4, decrease more sharply in F5, and increase sharply in F1. The surrounding SNP also behave differently in F1. The *USP13* gene is involved in the immune system (Zhang et al., 2013) and shows a change in different directions across the families (Figure 7). This difference in the direction of change may indicate different epistatic effects in the families.

Change in Allele Frequencies Over Generations, Greatest Variance of Change Over Generations Within Families, and Greatest Range of Change Over Generations Within Families. The regression coefficient when regressing allele frequency over generations for the whole population was used to identify SNP markers that have changed the most over time. Thus, this is not family-specific. The change in allele frequency is fairly consistent in direction and magnitude across all 5 families for 2 of the top 5 selected SNP, such as in Figures 8 and 10. This rapid change in AF for all families may reflect a partial selection sweep for a gene that has undergone selection. Some surrounding SNP markers change at a similar rate but in opposite directions and are more pronounced in some cases. This can be a hitchhiking effect by nearby markers. Our selected marker may be, in fact, the hitchhiker instead of the gene under selection.

The variance of change was used to identify scenarios where at least one family shows small changes while at least one other family shows a large change. For all 5 selected SNP, F1 and F5 follow similar trends, whereas F2 and F3 have trends similar to each other, but in a different direction and pattern than F1 and F5 (such as Figure 10). A change in direction can be due to a change in selection, but can also be influenced by an interaction with another gene of interest on the same or even different chromosomes (Mackay, 2014). Surrounding SNP markers show stronger linkage in Figure 11.

The range between the family with the least change and the family with the most change was also used to identify SNP that behave differently. The top 3 selected SNP showed the greatest change in F1 and stronger responses in the surrounding SNP (Figure 11 shows

the SNP with the greatest range). The fourth and fifth selected SNP markers show changes in different directions across families (Supplemental Data S1; https://figshare.com/projects/Non-parallel_genetic_change/145899; Steyn, 2022).

Lewontin Krakauer Test and Fixation Index.

Figure 12 includes the resulting Manhattan plot of the $-\log_{10}(P\text{-value})$ of the F_{st} of each SNP marker. Peaks are observed, and some approach significance, but none of the SNP markers met the criteria for statistically significant differences ($P < 0.05$ with a Bonferroni adjustment for the number of markers, or false discovery rate). The allele frequency changes for 5 SNP nearest to the significance threshold were investigated. All were markers that started with an allele frequency close to fixation. Even though the allele frequency remained stable for 4 families, one family (such as Figure 12 for the SNP marker with the highest F_{st}) showed rapid changes. These may have been the result of strong divergent selection. Figure 5 shows the distribution of F_{st} values of all markers. Although most markers are smaller than 0.05, some markers reach F_{st} values higher than 0.10.

Nonparallel Changes

Table 8 compared changes of the 100 SNP with the greatest absolute change from G0 to G10 within each specific family, instead of the population as a whole. The allele frequency change must be greater than 0.20 or greater than 0.30. Family 4 has the fewest number of SNP that changed more than 0.30. However, differences across families are more similar when change is greater than 0.20 instead of 0.30. Figure 13 presents Venn diagrams showing which markers change similarly, or differently, across families (F5 is not shown but follows patterns similar to F1, F2, and F3). In all families but F4, more than 10 SNP markers changed by more than 0.30 only in the family the top SNP are based on and none that changed in only another family. This is not the case in F4, where only 3 SNP markers changed by more than 0.30 in only F4, 3 only in F3, 4 only in F2, and none only in another family. Based on these results, we conclude that genetic redundancy is indeed present.

Limitations

Replicate populations over time are useful to observe adaptation (Franssen et al., 2017). Ideally, these replicates should be from the same environment, share founding populations, and evolve independently to the same environmental stressor (Barghi et al., 2020). In our study, the high overlap of animals in our families in older generations allowed us to have reasonable found-

Table 8. The number of times the 100 SNP markers with the greatest allele frequency change (AFC) in 1 family change by more than 0.20 or 0.30 from oldest to youngest generation in each family

AFC	Family 1	Family 2	Family 3	Family 4	Family 5
Top 100 SNP in family 1					
>0.20	100	89	100	92	95
>0.30	100	69	82	31	73
Top 100 SNP in family 2					
>0.20	90	100	100	95	84
>0.30	61	100	78	28	57
Top 100 SNP in family 3					
>0.20	82	92	100	93	85
>0.30	64	71	100	31	52
Top 100 SNP in family 4					
>0.20	86	93	83	100	89
>0.30	69	72	82	50	67
Top 100 SNP in family 5					
>0.20	87	90	99	98	100
>0.30	65	78	82	32	98

ing populations that are similar across families. Over time, differences from this starting point reflect both divergent selection and genetic redundancy.

Genotypes of older Holstein animals in our data do not reflect a true baseline of animals that were part of the gene pool at the time because they are generally animals that were predominantly bulls and considered to be best. However, these bulls were widely used in artificial insemination (AI) programs. Therefore, their genetic material is expected to be present in large proportions of the population. In 2015, it was shown that all AI bulls could be traced back to only 2 bulls born in 1880. Two highly influential bulls, Pawnee Farm Arlinda Chief and Round Oak Rag Apple, shared Y-chromosomes with 48.78 and 51.06% of the Holstein bull population in the 2010s (Yue et al., 2015). Both these bulls are included in our study. Our imbalance of sexes among older genotypes is not unlike the study by Barghi et al. (2019), where only females were genotyped in the founder population. During more recent decades, genotyping costs have decreased enough for breeders to genotype most of their animals, regardless of their genetic merit or sex. This will enable future studies to use AF that are more reflective of the whole population over time.

Other limitations of our study include the small number of replicates and generations. More generations and replicates are required to detect the genetic change in different subpopulations and determine the reason for changes. Additionally, our families are not closed families. Not all parents were genotyped, and therefore, not all genetic changes over time are captured by our animals. This is also similar to the study by Barghi et al. (2020), where not all replicate members were genotyped.

Sampling bias is possible in our initial clusters. Different sampling methods could select different animals.

Although different changes would be observed, and thus different selection signatures might be identified, genetic redundancy is still expected to be observed. Faster changes during the second half of our time period are not surprising as fewer shared ancestors are found, genotyping females became more popular, and more breeders could afford genotyping.

Our work may serve as a preliminary study to encourage the investigation of adaptation and genetic redundancy in the future once these obstacles are overcome. We have explored changes in AF over generations, but did not identify the cause of change or the extent to which selection, drift, epistasis, gene-by-environment interactions, or redundancy contributed to change. Considerably more genotypes have been collected since 2014. Their inclusion in future studies will greatly increase the ability to investigate the differences among subpopulations more thoroughly. Including all genotypes to date would allow greater numbers to calculate AF per generation, observe change over a longer period of time, and capture changes that have occurred.

Implications

Changes over time, whether in the overall population or multiple subpopulations, could aid in identifying selection signatures. Markers that change only in specific subpopulations could be associated with different breeding objectives in different regions, epistatic interactions, or adaptation to different environments. Those that change similarly in all subpopulations could be associated with genes that have undergone universal selection or are vital in biological mechanisms.

Nonparallel changes across families reveal genetic redundancy; however, because families may differ in genetic merit, this may be confounded with divergent selection. Due to selection or redundancy, nonparallel

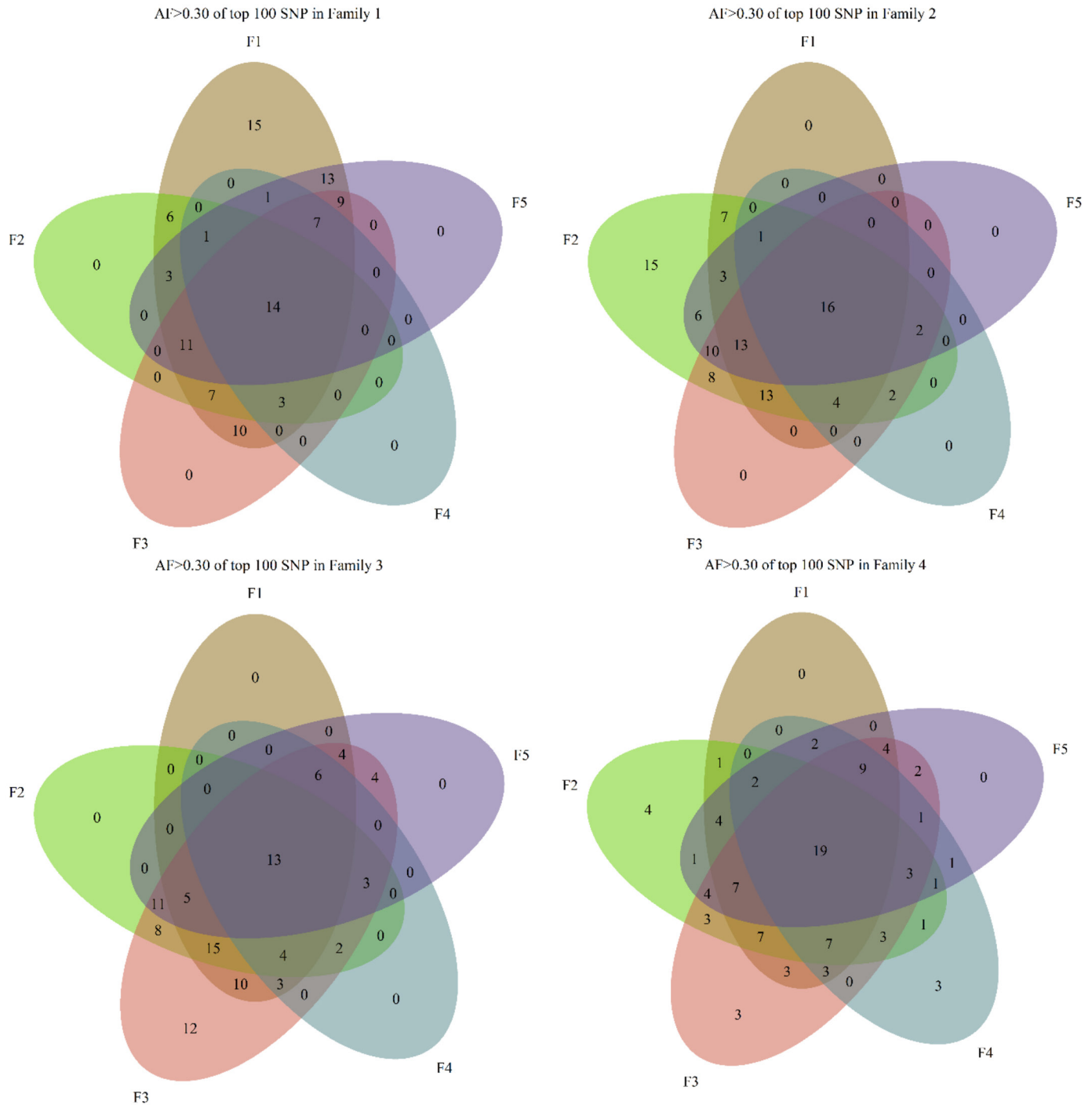


Figure 13. The number of times the 100 SNP that changed the most in a particular family (here family 1 to family 4) changed in allele frequency (AF) by more than 0.30 from generation 0 to generation 10 in the 5 different families (F1 to F5).

changes show underlying genetic diversity within the US Holstein breed. These results have important implications for the projection of long-term genetic response in the breed and other panmictic populations. Genetic selection in Holstein cattle has achieved a continued increase in milk production with no sign of

reaching a selection limit. A question arises whether selection should continue treating the breed as one large population with AF that can converge to the best overall average allele frequency for the population or as separate lines to increase the genetic distance between families. The latter can potentially prevent the loss of

alleles that may be beneficial in the future and allow outcrossing that can take advantage of heterosis.

Existing tools are available to breeders to actively reduce the rate of inbreeding while still achieving genetic improvement. Optimal contribution selection for livestock production was proposed by Meuwissen (1997) for this purpose. The recent availability of genomic information allows the use of a genomic relationship matrix for optimal contribution selection instead of a traditional pedigree matrix (Clark, 2013). A genomic matrix based on linkage analyses (IBD-based) shows promise in managing homozygosity and drift-based inbreeding while increasing genetic gains but can be computationally expensive (Meuwissen et al., 2020). The combined use of subpopulations within the US Holstein and optimal contribution selection can manage genetic diversity within separate lines. Management tools can also allow breeders to set an inbreeding limit that they are comfortable with. Additional considerations in the industry include limiting the number of progenies a sire can have or the time period in which their semen is available for purchase.

CONCLUSIONS

The Holstein breed is a complex mixture of family subgroups with different allele frequencies and gene combinations. The different families offer redundant solutions to the goals of modern-day breeders. Genetic redundancy allows for the value of individual alleles to shift over time in unique ways within a specific family. The substitution value of different alleles and, consequently, the breeding value will differ for different target populations such as a specific family versus the overall combined population. Stratification of selection candidates into unique subpopulations promotes genetic redundancy, can maintain diversity, and lowers the risk of the fixation or loss of alleles.

ACKNOWLEDGMENTS

The 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>.
- Arjan, J., G. M. de Visser, T. F. Cooper, and S. F. Elena. 2011. The causes of epistasis. *Proc. Biol. Sci.* 278:3617–3624. <https://doi.org/10.1098/rspb.2011.1537>.
- 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>.
- Barbosa da Silva, M. V. G., T. S. Sonstegard, R. M. Thallman, E. E. Connor, R. D. Schnabel, and C. P. Van Tassell. 2010. Characterization of DGAT1 allelic effects in a sample of North American Holstein cattle. *Anim. Biotechnol.* 21:88–99. <https://doi.org/10.1080/10495390903504625>.
- Barghi, N., J. Hermisson, and C. Schlötterer. 2020. Polygenic adaptation: A unifying framework to understand positive selection. *Nat. Rev. Genet.* 21:769–781. <https://doi.org/10.1038/s41576-020-0250-z>.
- Barghi, N., R. Tobler, V. Nolte, A. M. Jakšić, F. Mallard, K. A. Otte, M. Dolezal, T. Taus, R. Kofler, and C. Schlötterer. 2019. Genetic redundancy fuels polygenic adaptation in *Drosophila*. *PLoS Biol.* 17:e3000128. <https://doi.org/10.1371/journal.pbio.3000128>.
- 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>.
- Boyle, E. A., Y. I. Li, and J. K. Pritchard. 2017. An expanded view of complex traits: From polygenic to omnigenic. *Cell* 169:1177–1186. <https://doi.org/10.1016/j.cell.2017.05.038>.
- Buffalo, V., and G. Coop. 2020. Estimating the genome-wide selection to temporal allele frequency change. *Proc. Natl. Acad. Sci. USA* 117:20672–20680. <https://doi.org/10.1073/pnas.191903911>.
- Calo, L. L., R. E. McDowell, L. D. VanVleck, and P. D. Miller. 1973. Genetic aspects of beef production among Holstein-Friesians pedigree selected for milk production. *J. Anim. Sci.* 37:676–682. <https://doi.org/10.2527/jas1973.373676x>.
- Clark, S. A. 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>.
- 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>.
- Csilléry, K., A. Rodríguez-Verdugo, C. Rellstab, and F. Guillaume. 2018. Detecting the genomic signal of polygenic adaptation and the role of epistasis in evolution. *Mol. Ecol.* 27:606–612. <https://doi.org/10.1111/mec.14499>.
- Duenk, P., P. Bijma, M. P. L. Calus, Y. C. J. Wientjes, and J. H. J. van der Werf. 2020. The impact of non-additive effects on the genetic correlation between populations. *G3 (Bethesda)* 10:783–795. <https://doi.org/10.1534/g3.119.400663>.
- Falconer, D., and T. Mackay. 1996. *Introduction to Quantitative Genetics*. 4th ed. Pearson.
- Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.* 52:399–433. <https://doi.org/10.1017/S0080456800012163>.
- Flori, L., S. Fritz, F. Jaffrézic, M. Boussaha, I. Gut, S. Heath, J.-L. Foulley, and M. Gautier. 2009. The genome response to artificial selection: A case study in dairy cattle. *PLoS One* 4:e6595. <https://doi.org/10.1371/journal.pone.0006595>.
- Franssen, S. U., R. Kofler, and C. Schlötterer. 2017. Uncovering the genetic signature of quantitative trait evolution with replicated time series data. *Heredity* 118:42–51. <https://doi.org/10.1038/hdy.2016.98>.
- Goldstein, D. B., and K. E. Holsinger. 1992. Maintenance of polygenic variation in spatially structured populations: Roles for local mating and genetic redundancy. *Evolution* 46:412–429. <https://doi.org/10.1111/j.1558-5646.1992.tb02048.x>.
- Gualdrón Duarte, J. L., R. J. Cantet, R. O. Bates, C. W. Ernst, N. E. Raney, and J. P. Steibel. 2014. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. *BMC Bioinformatics* 15:246. <https://doi.org/10.1186/1471-2105-15-246>.
- Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breed-

- ing 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>.
- Höllinger, I., P. S. Pennings, and J. Hermisson. 2019. Polygenic adaptation: From sweeps to subtle frequency shifts. *PLoS Genet.* 15:e1008035. <https://doi.org/10.1371/journal.pgen.1008035>.
- Jiang, J., L. Ma, D. Prapapenka, P. M. VanRaden, J. B. Cole, and Y. Da. 2019. A large-scale genome-wide association study in U.S. Holstein cattle. *Front. Genet.* 10:412. <https://doi.org/10.3389/fgene.2019.00412>.
- Jiang, J., D. Prapapenka, L. Ma, J. Cole, P. VanRaden, and Y. Da. 2018. Extreme antagonistic pleiotropy effects of DGAT1 on fat, milk, and protein yields. Page 142 in *Proceedings of the World Congress on Genetics Applied to Livestock Production*.
- Laurentino, S., J. Gonçalves, J. E. Cavaco, P. F. Oliveira, M. G. Alves, M. de Sousa, A. Barros, and S. Socorro. 2011. Apoptosis-inhibitor Aven is downregulated in defective spermatogenesis and a novel estrogen target gene in mammalian testis. *Fertil. Steril.* 96:745–750. <https://doi.org/10.1016/j.fertnstert.2011.06.009>.
- Legarra, A., C. A. Garcia-Baccino, Y. C. J. Wientjes, and Z. G. Vitezica. 2021. The correlation of substitution effects across populations and generations in the presence of non-additive functional gene action. *Genetics* 219:iyab138. <https://doi.org/10.1093/genetics/iyab138>.
- Legarra, A., D. A. L. Lourenco, and Z. G. Vitezica. 2022. Bases for genomic prediction. Accessed Jan. 16, 2023. <http://genoweb.toulouse.inra.fr/~alegarra/GSIP.pdf>.
- Lewontin, R. C., and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74:175–195. <https://doi.org/10.1093/genetics/74.1.175>.
- Liu, X., Y. I. Li, and J. K. Pritchard. 2019. Trans effects on gene expression can drive omnigenic inheritance. *Cell* 177:1022–1034.E6. <https://doi.org/10.1016/j.cell.2019.04.014>.
- Lourenco, D. A. L., I. Aguilar, A. Legarra, S. Miller, S. Tsuruta, and I. Misztal. 2019. Genomic accuracy for indirect predictions based on SNP effects from single-step GBLUP. *Annual meeting of the European Association for Animal Production (EAAP)* 25:717. Wageningen Academic Publishers.
- Ma, L., T. S. Sonstegard, J. B. Cole, C. P. VanTassell, G. R. Wiggans, B. A. Crooker, C. Tan, D. Prapapenka, G. E. Liu, and Y. Da. 2019. Genome changes due to artificial selection in U.S. Holstein cattle. *BMC Genomics* 20:128. <https://doi.org/10.1186/s12864-019-5459-x>.
- Mackay, T. F. C. 2014. Epistasis and quantitative traits: Using model organisms to study gene–gene interactions. *Nat. Rev. Genet.* 15:22–33. <https://doi.org/10.1038/nrg3627>.
- McGaugh, S. E., A. J. Lorenz, and L. E. Flagel. 2021. The utility of genomic prediction models in evolutionary genetics. *Proc. Biol. Sci.* 288:20210693. <https://doi.org/10.1098/rspb.2021.0693>.
- 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. Gebregiorgis, 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. University of Georgia. Accessed Jan. 16, 2023. http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf.
- Nowak, M. A., M. C. Boerlijst, J. Cooke, and J. M. Smith. 1997. Evolution of genetic redundancy. *Nature* 388:167–171. <https://doi.org/10.1038/40618>.
- Pedersen, L. D., A. C. Sorensen, and P. Berg. 2010. Marker-assisted selection reduces expected inbreeding but can result in large effects of hitchhiking. *Genetics* 127:189–198. <https://doi.org/10.1111/j.1439-0388.2009.00834.x>.
- Pickett, F. B., and D. R. Meeks-Wagner. 1995. Seeing double: Appreciating genetic redundancy. *Plant Cell* 7:1347–1356. <https://doi.org/10.1105/tpc.7.9.1347>.
- Prapapenka, D., Z. Liang, J. Jiang, L. Ma, and Y. Da. 2021. A large-scale genome-wide association study of epistasis effects of production traits and daughter pregnancy rate in U.S. Holstein cattle. *Genes (Basel)* 12:1089. <https://doi.org/10.3390/genes12071089>.
- Pryce, J. E., M. Haile-Mariam, M. E. Goddard, and B. J. Hayes. 2014. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. *Genet. Sel. Evol.* 46:71. <https://doi.org/10.1186/s12711-014-0071-7>.
- 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>.
- Rendel, J. M., and A. Robertson. 1950. Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. *J. Genet.* 50:1–8. <https://doi.org/10.1007/BF02986789>.
- Rowan, T. N., H. J. Durbin, C. M. Seabury, R. D. Schnabel, and J. E. Decker. 2021. Powerful detection of polygenic selection and evidence of environmental adaptation in US beef cattle. *PLoS Genet.* 17:e1009652. <https://doi.org/10.1371/journal.pgen.1009652>.
- Saatchi, M., R. D. Schnabel, M. M. Rolf, J. F. Taylor, and D. J. Garrick. 2012. Accuracy of direct genomic breeding values for nationally evaluated traits in US Limousin and Simmental beef cattle. *Genet. Sel. Evol.* 44:38. <https://doi.org/10.1186/1297-9686-44-38>.
- Santiago, E., and A. Caballero. 1998. Effective size and polymorphism of linked neutral loci in populations under directional selection. *Genetics* 149:2105–2117. <https://doi.org/10.1093/genetics/149.4.2105>.
- Spelman, R. J., C. A. Ford, P. McElhinney, G. C. Gregory, and R. G. Snell. 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. *J. Dairy Sci.* 85:3514–3517. [https://doi.org/10.3168/jds.S0022-0302\(02\)74440-8](https://doi.org/10.3168/jds.S0022-0302(02)74440-8).
- Steyn, Y. 2022. Non-parallel genetic change. Figshare. Figures. https://figshare.com/projects/Non-parallel_genetic_change/145899.
- Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt, and R. Fries. 2003. Effects of DGAT1 variants on milk production traits in German cattle breeds. *J. Anim. Sci.* 81:1911–1918. <https://doi.org/10.2527/2003.8181911x>.
- Tidcombe, H., A. Jackson-Fisher, K. Mathers, D. F. Stern, M. Gassmann, and J. P. Golding. 2003. Neural and mammary gland defects in ErbB4 knockout mice genetically rescued from embryonic lethality. *Proc. Natl. Acad. Sci. USA* 100:8281–8286. <https://doi.org/10.1073/pnas.1436402100>.
- Tsuruta, S., T. Lawlor, D. A. L. Lourenco, and I. Misztal. 2021. Bias in genomic predictions by mating practices for linear type traits in a large-scale genomic evaluation. *J. Dairy Sci.* 104:662–677. <https://doi.org/10.3168/jds.2020-18668>.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423. <https://doi.org/10.3168/jds.2007-0980>.
- Wade, M. J., and C. J. Goodnight. 1998. Perspective: The theories of Fisher and Wright in the context of metapopulations: When nature does many small experiments. *Evolution* 52:1537–1553. <https://doi.org/10.1111/j.1558-5646.1998.tb02235.x>.
- Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res.* 94:73–83. <https://doi.org/10.1017/S0016672312000274>.
- Wright, S. 1943. Isolation by distance. *Genetics* 28:114–138. <https://doi.org/10.1093/genetics/28.2.114>.
- Yuan, S., C. J. Stratton, J. Bao, H. Zheng, B. P. Bhetwal, R. Yanagimachi, and W. Yan. 2015. *Spata6* is required for normal assembly of the sperm connecting piece and tight head–tail conjunction. *Proc. Natl. Acad. Sci. USA* 112:E430–E439. <https://doi.org/10.1073/pnas.1424648112>.

- 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>.
- Zhang, J., P. Zhang, Y. Wei, H. Piao, W. Wang, S. Maddika, M. Wang, D. Chen, Y. Sun, M.-C. Hung, J. Chen, and L. Ma. 2013. Deubiquitylation and stabilization of PTEN by USP13. *Nat. Cell Biol.* 15:1486–1494. <https://doi.org/10.1038/ncb2874>.

ORCIDS

- Y. Steyn  <https://orcid.org/0000-0001-5467-9555>
- T. Lawlor  <https://orcid.org/0000-0002-4458-1025>
- Y. Masuda  <https://orcid.org/0000-0002-3428-6284>
- S. Tsuruta  <https://orcid.org/0000-0002-6897-6363>
- A. Legarra  <https://orcid.org/0000-0001-8893-7620>
- D. Lourenco  <https://orcid.org/0000-0003-3140-1002>
- I. Misztal  <https://orcid.org/0000-0002-0382-1897>