HAPLOTYPE-BASED SELECTION SIGNATURE ANALYSIS USING UNIVERSITY OF MINNESOTA AND US CONTEMPORARY HOLSTEIN CATTLE

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

JING YANG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Advisor: YANG DA

November 2015

Acknowledgements

I thank my adviser Dr. Yang Da for his guidance, advice and financial support. I also want to thank my committee members, Dr. F. Abel Ponce de León and Dr. Ran Blekhman for their support and commitment. Special thanks to my husband, Li Ma, for his support and advice. Finally, I thank all of those who gave me the possibility to finish this thesis.

This research is supported in part by project MN-16-043 of the Agricultural Experiment Station at the University of Minnesota, by funding from the National Research Initiative Competitive Grants Program/United States Department of Agriculture Grant no. 2008-35205-18846, by United States Department of Agriculture (USDA) National Institute of Food and Agriculture Grant no. 2011-67015-30333, and by a financial contribution from the Holstein Association USA and the Minnesota Supercomputer Institute.

Abstract

Artificial selection in dairy cattle since 1964 has achieved steady increase in milk production that was accompanied by unintended declines in fertility. We conducted selection signature analysis to identify genome changes due to the forty years of selection using direct comparison of 45,878 SNPs between Holstein cattle unselected since 1964 and contemporary Holsteins. The Holstein genome had a landscape change from the unselected to the elite contemporary Holsteins. About 31% of the genome was affected by the forty years of selection, and 230 regions had highly significant changes in long-range allele frequencies and genotypic heterozygosity. From these 230 regions, 197 genes with documented fertility functions mostly in mice and humans were identified, leading to the hypothesis that the unintended declines in fertility since 1964 was due to hitchhiking of selection by negative effects of fertility genes. The female-male ratio of the 197 fertility genes is approximately 5:4, indicating that the fertility problems in the contemporary Holstein population likely was due to decreased fertility in both females and males. The elite Holsteins were more heterozygous than their contemporaries in all thirty regions where the elite cows and their contemporaries had significant heterozygosity differences, including seven regions in or near large clusters of olfactory receptors, zinc fingers, cationic amino acid transporters, sialic acid-binding Ig-like genes, vomeronasal receptors, keratin genes, *EMR2* receptors, and transfer RNA's.

Table of Contents

List of Tables

List of Figures

Symbols and Abbreviations

AFD: allele frequency difference

BTA: *Bos Taurus* chromosome

Chr: chromosome

DPR: daughter pregnancy rate

GWAS: genome-wide association study

IBS: identity by state

Kb: 1000 bases pairs

LD: linkage desequilibrium

Mb: mega base pairs = 1000 kb = 1 million base pairs

MDS: multidimensional scaling

PTA: predicted transmitting ability

PC: principal component

PCA: principal component analysis

QTL: quantitative trait locus

SNP: single nucleotide polymorphism

YD: yield deviation

Chapter 1

Introduction

For the past 50 years, milk production in Holsteins continued to increase but fertility experienced severe declines, as shown by the opposite trends in yield and daughter pregnancy rate (DPR) in Fig. 1 [1-4]. Successful conception in Holstein cows now requires ~30 days longer than those of 50 years ago (−1% DPR = +4 Days Open) [5]. Although reproduction traits generally exhibit low heritability [6-8], the genetic antagonism of milk production and fertility, as well as the unfavorable decline in fertility that occurred during genetic selection, revealed the importance of genetic contribution to fertility traits [4, 7-12]. In addition, variation in fertility between and within breeds indicates the possibility of improving the genetic component of fertility without severely slowing genetic gain for milk production [3, 12-14], which has been supported by the stabilization or even reversal of the declining trend in DPR after 2003 (Fig. 1B), when fertility started to be included in the selection program [5].

Dairy fertility is a complex phenotype affected by many factors that include nutrition, management, environment, and genetics [3, 15-18]. The complex nature of fertility calls for the use of all available resources and tools to understand and improve the genetic component. We hypothesize that genetic factors with additive and non-additive effects contribute to fertility, and sequence-based large sample analysis will lead to the most

productive genomic discovery and application to improve fertility traits. The overall goal is to discover causal/tightly linked genetic variants and apply the genomic discoveries to improve dairy fertility. The specific aims are to: 1) Identify genomic regions under selection by using SNP and haplotype-based selection signature analyses; and 2) Link genomic regions under selection to production and reproduction phenotypes of dairy cattle.

Large-scale sequence-based studies are on-going in humans [19-21] and will be the future for animals [22-24]. Genome sequence covers all genes and genomic regions that affect fertility, hence offering the best power to detect fertility-related variants/regions and to predict fertility traits. The genetic components of fertility range from simple monogenic effects like recessive lethals [25, 26] to complex mechanisms that involve multiple genes with additive and non-additive effects [27-31]. The unprecedented big data in this proposal will provide the highest power to dissect the simple and complex genetic effects for fertility traits out of random noises that mask the underlying genetic mechanism.

Sequence-based genomic selection can potentially improve accuracy by increasing the linkage disequilibrium (LD) between SNP markers and causal mutations, particularly for fertility traits where causal mutations are of low frequency because of selection [23, 32-36]. We have obtained preliminary results using a multi-allelic haplotype approach to integrate functional genomic information into genomic selection that uses ‘causal regions’ than ‘causal variants’ [37, 38], making the research discovery more readily

applicable to genomic selection. While current SNP chips do not have enough SNPs in many genes, sequence data provide the best coverage of large/small functional regions for haplotype analysis. Fertility-related genomic variants and regions discovered by this research can be readily utilized in genomic selection via our new multi-allelic haplotype approach through the implementation and dissection from the USDA Animal Genomics and Improvement Laboratory (AGIL).

1.1 National dairy genomic and phenotypic database at AGIL

The AGIL (Animal Genomics and Improvement Laboratory) maintains a large national dairy genomic and phenotypic database. Research on genomic selection has revolutionized the AI industry: predictions of genetic merit enhanced by the genome-wide SNP data are being calculated and actively used by dairy breeders to make selection decisions. The development of SNP genotyping and imputation have opened up many opportunities [39, 40]. As part of the program AGIL developed with the industry, monthly genomic evaluations are provided for the Holstein, Jersey and Brown Swiss breeds. Breeding organizations (AI studs) and breed associations arrange for the genotyping and distribute the resulting genomic evaluations.

As of March 2015, the USDA national dairy database has 738,829 Holstein cattle with directly measured/imputed 60K SNP data. A total of 2,433 animals have been genotyped with high-density (HD) assays with 770K SNPs. The number of genotyped animals increased exponentially, with 253,149 genotypes being received during the past 12

months. Additionally, the team has acquired access to ~842 cattle whole-genome sequences that can be used to impute sequence variants to all the genotyped animals. Millions of phenotypic records have been collected for over 33 phenotypes that are routinely evaluated, including 13 reproduction related traits: daughter pregnancy rate, heifer and cow conception rates, sire conception rate, sire and dam stillbirth, sire and dam calving ease, sire and dam gestation lengths, age at first calving, services to conception, and calving to first insemination.

1.2 Genome-wide selection signature analysis identifies genes related to fertility

Selection signature analysis detects genome changes due to selection that results in long-range linkage disequilibrium (LD) and haplotype homozygosity at the chromosome regions subjected to selection [41, 42]. This approach has the advantage of being unaffected by phenotypic variations with errors, and has been increasingly used for finding genetic factors associated with phenotypes under selection [41-56]. The combination of GWAS and selection signature results increase the power for finding genetic factors associated with phenotypes [57]. The decline in fertility during the past 50 years of artificial selection for milk yield likely had their selection signatures.

We have obtained results of selection signatures using the bovine 60K SNP chip [58-60]. Several regions showing selection signatures overlapped with fertility-related genes, including *FGF1*, *FBN1*, *DUT*, *PGF*, *ESRRB*, *LHCGR*, *FSHR*, *KITLG*, *FGF6*, *FGF26*, *TIGAR*, *CCNG*, *GHR*, *PRLR*, and a Bovine MHC region on Chr23 (Figure 1.2). With

whole-genome sequences of a large sample with many phenotypes, more selection signatures will be identified with high degree of accuracy and can be used for the discovery of causal/linked genes and variants of fertility. Many of these genes are known to affect fertility. *FGF1* is involved in broad mitogenic and cell survival activities including embryonic development [56], *PGF* plays a key role in embryogenesis [61], *ESRRB* plays an essential role in placenta development [56], *FSHR* is necessary for follicular development and is expressed on the granulosa cells that are closely associated with the developing female gamete in the ovary of mammals [19], and *LHCGR* is necessary for follicular maturation and ovulation [62]. Mouse knockout models showed that *FSHR, KITLG, CCNG2* and *PRLR* were involved in female fertility proteins [63]. These known gene functions related to reproduction and the fact that these genes were in or near chromosome regions with strong selection signals indicate that these genes could be involved in the vast difference in fertility between contemporary Holsteins and the 1964 Holsteins.

1.3 Genome-wide association analysis of fertility traits

Previously, USDA NRI funded genome-wide association study (GWAS) using a 50K SNP chip in over 1,600 contemporary Holstein cows for 31 dairy phenotypes, which identified a number of candidate genes for daughter pregnancy rate (DPR) on Chr 1, 7, 18 and X, for calving ease on Chr18, and for stillbirth on Chr15 and Chr23 [64]. Combining results of selection signature analysis and GWAS led to improved accuracy and increased

power in gene mapping and variants discovery [57]. Therefore, we compared the Holstein genomes of a group of cattle unselected since 1964 to contemporary and elite cows and identified a number of chromosome regions that have been subjected to selection, including a Chr1 region with a large number of SNP effects for DPR. We also discovered evidence of potential involvement of microRNA genes in dairy phenotypes on Chr21 [58].

GWAS analysis using 50K SNP assay in >5,000 Holstein bulls from the AGIL dairy genomics database identified a cluster of SNP markers on Chr18 that have the largest effects on calving ease, conformation traits, longevity, and total merit [65]. Additional sequence data from 11 US Holstein bulls have been used to identify putative causal mutations associated with calving and conformation traits. One duplication CNV and two different tandem duplication events were detected in the Chr18 candidate locus. Predicted tandem duplications present in the carrier animals suggest that the portions of two exons and a connecting intron within the Ig-like protein domains of the *SIGLEC-6* gene may have been duplicated. Some heterozygotes with desirable sire calving ease also have deletions near the N-terminal end of the protein. Additionally, the research team at AGIL has successfully identified causative mutations in several haplotypes associated with spontaneous abortion, embryonic death, and other fertility disorders in dairy cattle using sequences of candidate genes [25, 66, 67].

GWAS results provided useful insights into the genetic mechanism of complex diseases and traits, but most of these associated SNPs are genetic markers linked to

causative mutations. Compared to genotype data, sequence data have the highest power and accuracy by directly measuring every single genetic variant and testing for its effect on a trait of interest, and provides the best coverage for haplotype analysis of functional regions. With the rapid development of techniques and decreasing cost of high-throughput sequencing, sequenced-based GWAS have recently been applied to humans to identify causative genes and mutations of complex traits and diseases [68-70]. Our preliminary results of haplotype analysis using human data indicated that sequence data would provide the best coverage for many tiny functional regions that were not covered by the 500K SNPs [37, 38].

1.4 Integration of functional genomic information with genomic selection

Genomic selection based on evaluation of an individual's SNP markers has been shown to achieve increased accuracy over traditional evaluation [71, 72]. An example for illustrating the usefulness of gene-based selection is the discovery of some low-frequency recessive lethal haplotypes in U.S. dairy breeds [25, 66, 67]. For more complex traits that have no major QTLs, genomic selection using whole-genome SNP markers provides higher accuracy [72]. While the actual effectiveness of using sequence data in genomic selection is under debate [33-35], for fertility traits, sequence-based genomic selection could potentially improve accuracy by better capturing the causal mutations that are generally of low frequencies due to selection [23, 32-36].

We have developed a multi-allelic haplotype approach to integrate functional genomic information with genomic selection [37, 38]. Using the Framingham Heart Study (FHS) data, we have shown that the new approach is superior in prediction accuracy. We compared three methods of genomic prediction using combinations of additive and dominance effects of single SNPs and haplotype blocks. Method I was single SNP analysis of 423,131 SNPs covering all human autosomes with over 6000 individuals. The next two methods add haplotype analysis with functional information: Method II adds haplotype analysis of 595 'cholesterol related genes' with 8,674 SNPs (2% of autosomes); and Method III adds haplotype analysis of 9821 genes with 184,686 SNPs (36% of autosomes) after removing small genes without at least two SNPs. The results from 4-8 validation samples showed that adding haplotype analysis to single SNP analysis improved prediction accuracy in most cases. Method II with cholesterol related genes had the best prediction accuracy for total cholesterol with 4.78% increase in accuracy over single SNP analysis, and Method III using all autosomal genes had the best accuracy for triglyceride with 17.75% increase in accuracy over single SNP. Results were also obtained from one validation sample for adding three other haplotype analyses to single SNP analysis: ChIPseq sites with 375,924 SNPs and average block size of 115.8Kb; non-hotspot blocks with each block between two crossover hotspots with 422,695 SNPs and average block size of 65Kb, and evenly divided blocks with block size of 100Kb of 422,814 SNPs. All three methods improved the prediction accuracy for most phenotypes but ChIPseq blocks mostly had better prediction accuracy than the other two

methods, indicating that ChIPseq sites likely contained useful functional information for genomic prediction not present in anonymous blocks. These preliminary results concluded that a multi-allelic haplotype approach together with functional regions provides more accurate genomic prediction than single-SNP methods.

1.5 Rationale and Sigificance

While dairy production has increased due to intense artificial selection, dairy fertility has experienced severe declines over the past 50 years. Dairy fertility has low additive heritability; however, the negative genetic correlation between milk yield and fertility and the declines in fertility occurred during genetic selection suggest the existence of genetic contribution to fertility. Differences in fertility between breeds and differences between Holstein cattle unselected for 50 years and contemporary Holsteins, as well as the existence of elite cows in both production and fertility, also suggest the existence of genetic contribution to fertility and indicate the possibility of improving fertility without much sacrifice in milk production. Recent efforts in genetic improvement of dairy fertility have already achieved some stability in fertility performance, but a complete reversal of the declining trend remains a difficult task. The USA has the largest quantity of dairy genomic and fertility data and has a unique group of Holstein cattle unselected for 50 years. The rapid advances in sequencing technologies and analytical and computational expertise from our group provide unprecedented powerful tools for genomic discovery and application to improve dairy fertility.

The rationale can be summarized as: 1) Fertility of the Holstein breed has experienced serious declines over the past 50 years and is now faced with the challenge of improvement to stay competitive, 2) Genomic tools have developed to the point that a translational genomics solution to the fertility problem becomes practical, 3) DNA sequencing technology has developed to the point that finding causal polymorphisms affecting phenotypes has become affordable and achievable, 4) US national dairy genomic database has accumulated a huge amount of genotypic and phenotypic data for almost a half million dairy bulls and cows, and the number is rapidly increasing, 5) Access to a considerable amount of dairy sequence data generated from our group and through collaborations, and the ability to impute many cows to whole-genome sequence, and 6) Expertise and experience accumulated so far promise an effective discovery and delivery process of the proposed genomic discovery and prediction system to the dairy industry. The genomic solution will include the use of the state-of-the-art genome sequencing technology for identifying SNP effects and causative/tightly linked polymorphisms underlying fertility and related traits, and improved genomic selection models for reproduction to the dairy industry by adding causative/linked polymorphisms to the existing evaluation markers. The genomic solution to the decreasing fertility is expected to be the most comprehensive and thorough among those thus far available, providing a powerful genome-guided breeding for improving dairy fertility.

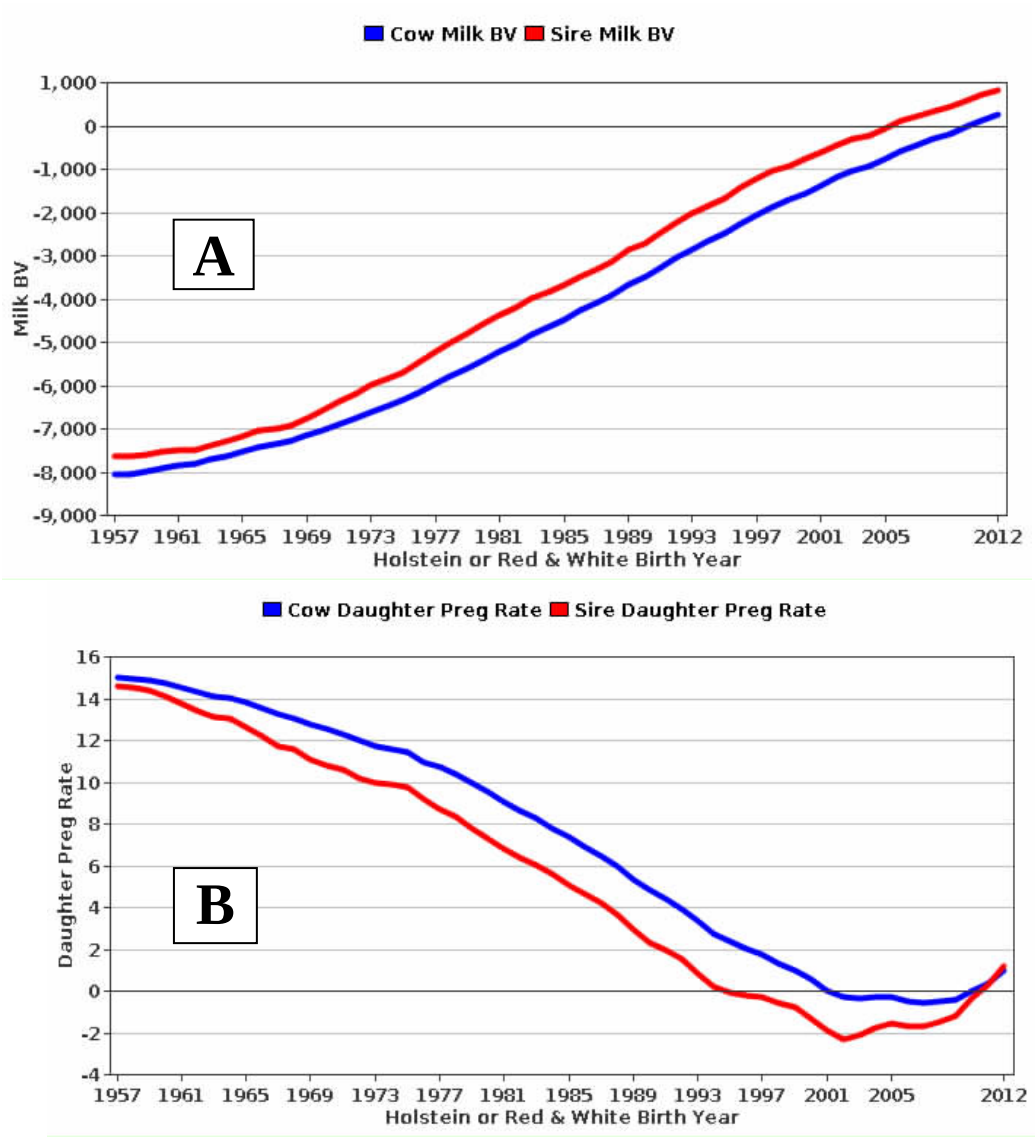


Figure 1.1. Opposite trends in milk yield (A) and daughter pregnancy rate (B) for Holstein or Red & White (Animal Genomics and Improvement Laboratory, USDA/ARS, 2014/12).

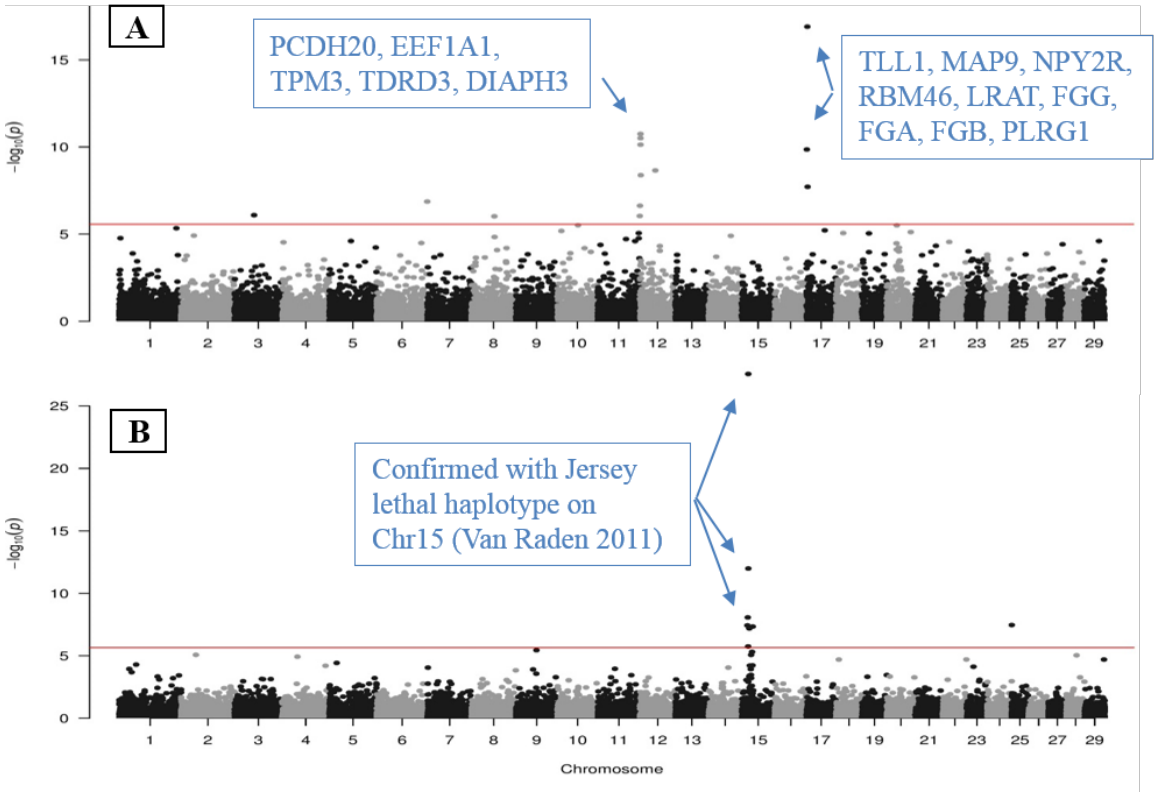


Figure 1.2. Preliminary genome scan of transmission ratio distortion using <60K SNPs in Holstein (A) and Jersey (B). On the y-axis are the $-\log_{10}$(*P*-value) of the test of TRD and on the x-axis are the chromosome locations of the autosomal SNPs.

Chapter 2

SNP-based selection signature analysis in three groups of Holstein cattle that underwent different selection pressures

2.1 Introduction

Genetic selection of U.S. Holstein cattle since 1964 has resulted in tremendous phenotypic changes. By 2004, average annual yields of milk, protein and fat per cow increased 3740 kg, 167 kg, and 193 kg, respectively, but dairy fertility measured by daughter pregnancy rate (DPR) experienced serious declines, with contemporary Holsteins requiring about thirty days longer than the required by the 1964 cows for a successful conception, as shown by the genetic trends [73, 74] shown in Figure 2.1. During the same time period, selection for body appearance traits also resulted in many changes [75]. The genetic trends shown in Figure 2.2 between a group of U.S. Holstein cattle unselected since 1964 and a group of selected contemporary Holstein cows maintained at the same University of Minnesota dairy facility under the same management and environment conditions [76] imply that the phenotypic changes of the Holstein cows during the past forty years were primarily genetic changes due to artificial selection. Genetic selection leaves its signature on the genome. Allele frequency change is the most fundamental change due to selection, and a genomic region subjected to

selection typically has a long-range pattern due to linkage disequilibrium (LD) between the selection target and neighboring variants [42, 47].

We investigate selection signature of the fifty years of artificial selection by comparing direct comparison of the genomes of three groups of Holstein cattle representing no selection, twenty years of selection, and forty years of selection since 1964. We show that the forty years of artificial selection resulted in landscape changes of the Holstein genome and left signatures indicative of genome regions affected by selection.

2.2 Material and Methods

2.2.1 Holstein populations and genotyping data

Three groups of Holstein cattle representing three periods of artificial selection were analyzed. Group I represented Holstein genomes of the 1950's and was used as a sample of the unselected Holstein genomes since 1964. This group included a Holstein unselected line since 1964 with 301 cattle (229 cows and 71 bulls) derived from 20 founder bulls born in 1951-1959, and included 16 bulls born in 1954-1959. Group II consisted of 215 bulls born between 1975 and 1985, and Group III consisted of 1655 contemporary cows representing Holstein cattle approximately subjected to forty years of artificial selection for increased milk yield. Signature of forty years of selection was detected by comparing Groups I and III, while signature of selection during the first and

second 20 years of selection since 1964 was detected by comparing Group II and Group I and Group III, respectively. The SNP genotypes were generated using the Illumina BovineSNP50™ BeadChip (REF). A total of 45,442 SNPs on the 29 autosomes and the X chromosome had a minimal allele frequency difference of 0.02 between the contemporary and unselected groups (Groups I and III) were used in the selection signature analysis.

2.2.2 Selection signature analysis

Selection signature analysis used long-range heterozygosity (H) measures [19], and long-range allele frequency differences (AFD) and standardized AFD in sliding windows of SNP markers for genome-wide scan of selection signatures. Let n = number of SNP markers in a given sliding window of SNP markers, S_j = H or AFD of marker j in this sliding window, j = 1, …, n; l_j = marker index at the lower bound of window for marker j; u_j = marker index at the upper bound of window for marker j; and X_j = long-range H or AFD of marker j calculated as the average of all H or AFD values of SNP markers in the sliding window. Then, X_j is calculated as

$$X_j = \frac{\sum_{i=l_j}^{u_j} S_i}{u_j - l_j + 1}$$

Standardized H and AFD were calculated using the formula of standardized normal variable, i.e., Z = [X – (mean of X)]/(standard deviation of X). To account for different chromosomal averages and variations, within-chromosome mean and standard deviation

was used. For the standardization of long-range AFD or Fst values in sliding windows, the mean and standard deviation of the AFD or Fst values of all SNP markers of the chromosome were used. This type of within-chromosome standardization removes the chromosome mean and variation in AFD and Fst values so that the AFD and Fst values of different chromosomes can be compared taking into account different chromosome means and variations. This within-chromosome standardization can be formulated as:

$$Z_j = \frac{X_j - \overline{X}}{\sigma(X)} \quad , \quad \text{with } \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{and} \quad \sigma(X) = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$$

Linkage disequilibrium (LD) and haplotype analyses [23] as well as extended haplotype homozygosity (EHH) analysis [24] were conducted following the initial scan of selection signatures using long-range heterozygosity and AFD measures. LD and haplotype analyses used Haploview [23] and extended haplotype homozygosity (EHH) analysis was carried out using Sweep-1.1 [24]. Phased genotypic data as input files for Sweep 1.1 were produced by FASTPHASE [25].

The Sweep computer program was used for selection sweep analysis in the selected contemporary population and the unselected population since 1964. The Sweep program implements a long range haplotype homozygosity test that was shown to be more powerful than traditional methods for detecting selection signatures. For each core region (or haplotype block) identified by Sweep, EHH-distance and bifurcation figures were produced for the haplotype that had a selection advantage in the selected population as

measured by higher haplotype frequency in selected population than in the unselected population. This haplotype was considered to have been subjected to positive selection if the EHH-distance and bifurcation figures showed long range EHH patterns. The comparison of EHH patterns between selected and unselected populations excluded cores that could have been considered candidates of selection signature due to similar EHH patterns in both selected population and unselected populations. For cores not matched or partially matched in the unselected population, haplotype frequencies for SNPs in the cores of the selected population were calculated in both selected and unselected populations to study the haplotype diversity between the two populations. Localization of SNPs in selection signatures relevant to known bovine genes was based on the bovine gene database of NCBI and ENSEMBL.

2.3 Results and Discussion

The Holstein genome landscape changes due to forty years of artificial selection comprised of a time shift of genome variation patterns, genome-wide changes in allele frequencies, and localized genome changes that are signatures of genetic selection. In or near selection signatures, about 150 genes with documented fertility functions affecting male and female fertility mostly in mouse and human were identified. We obtained these results by direct comparison of 45,875SNPs in three groups of Holstein cattle at different selection stages: Group I remained unselected since 1964 (no selection for the past

twenty years, n = 301), Group II subjected to twenty years of selection between 1964 and 1985 (n = 215, bulls born 1975-1985), and Group III of contemporary cows subjected to twenty years of selection (n = 1654). Group III has an elite group consisting of 160 half-sibs[75] that are defined as Group IIIb, while the rest of Group III are defined as Group IIIa.

2.3.1 Time trend of the changing Holstein genome

The Holstein genome had a clear time trend of landscape change from the unselected to the contemporary Holstein genomes. In the multidimensional scaling (MDS) plot of the first dimension versus the second dimension of the SNP identity by state (IBS) distances[77], the Holstein genomes unselected since 1964 (Group I) were in the lower left, the Holstein genomes subjected to twenty years of selection (Group II) were in the middle of the figure, and the contemporary Holstein genomes subjected to forty years of selection (Groups IIIa and IIIb) were at the upper right with the elite group (Group IIIb) at the extreme upper-right (Fig. 2.2a). In the plot of the first MDS dimension versus birth year in the 54-year period of 1953-2007, the unselected Holstein genome (Group I) was relatively unchanged across birth years for Group I without selection, shifted to the middle in the upper-right direction for Groups II with twenty years of selection, and reached to the upper-right for the Groups IIIa and IIIb representing forty years of selection with Group IIIb at the extreme upper-right (Fig. 2.2b). We further estimated the genome changes at the SNP level and at regional level showing signatures of selection.

We estimated that approximately 40% of the Holstein genome was affected by the forty years of selection using the non-random AFD method. We first estimate the frequency of random AFD between a random sample of 250 cows and the remaining 1404 cows of Group III. As shown in Table 2.1, the largest random AFD was 0.103 and all AFD above 0.10 except one value of 0.103 were due to selection. Using AFD>0.10 as non-random AFD due to selection and requiring that AFD changes in the first and second twenty years were in the same direction, 18,229 SNPs (39.7%) satisfied these two conditions, or approximately 40% of the Holstein genome were affected by the forty years of artificial selection. The AFD results in Table 2.1 indicate that the first twenty years of selection resulted in more genome changes than the second twenty years of selection. This is consistent with the fact that milk yields were the primary selection targets during the first twenty years, whereas more phenotypes were added for genetic selection during the second twenty years of selection since 1985 [78]. Long-range standardized AFD (Z_{AFD}) also showed that the long-range patterns of genome changes of the first twenty years of selection better aligned with those of forty years of selection than the second twenty years, consistent with the result that the first twenty years of selection had more allele frequency changes than in the second twenty years of selection.

We identified 217 genome regions with signature of selection using Z_{AFD} and standardized long-range heterozygosity differences (Z_{HD}). A genome region $Z_{AFD} \geq 3$ or $|Z_{HD}| \geq 3$ in sliding windows of 0.5-3Mb is considered a selection signature, and the size of the selection signature is defined as the region with $Z_{AFD} \geq 2$ on both sides of the peak

Z_{AFD} value or $|Z_{HD}| \geq 2$ on both size of the peak $|Z_{HD}|$ value. These 217 genome regions with selection signature contained 2710 genes, average size of 753.7Kb per region in the range of 44.3-7098.6Kb per region. These selection signatures provided evidence to understand the relationships between genome changes due to genetic selection and the unintended consequence of the declined fertility that accompanied genetic selection, to identify genomic regions associated with phenotypes in the contemporary Holstein cattle, and to understand special gnomic features of an elite group of contemporary cows sired by a single bull.

Five examples of such selection signatures on four chromosomes are shown in Fig. 2.3. Three of these five regions (Chr 7,12 and 23) had significant heterozygosity decreases (or homozygosity increases) because the unselected Holsteins had intermediate allele frequencies (near 0.5) and the forty years of selection resulted in more extreme frequencies in these regions, and in contrast, the two Chr16 regions all had heterozygosity increases because the unselected Holsteins had relatively extreme allele frequencies and the forty years of selection resulted in intermediate allele frequencies in these two regions. Most of the AFD was due to the second twenty years of selection for the Chr7 and Chr12 selection signatures. The two Chr16 selection signatures were mainly due to the first 20 years of selection. All forty years of selection affected the Chr23 selection signature that is located in the bovine major histocompatibility (MHC) region that plays a central role in cattle's immunity. The four selection signatures (Fig. 2.3a, Fig. 2.3c-2.3d) were due to selection in all contemporary Holsteins (Group III).). The Chr12

selection signature at 35.9-37.2Mb (Fig. 2.3b) was most pronounced in the elite Holsteins (Group IIIb) although this selection signature was observed in all contemporary Holsteins (Group III). The Chr7, Chr12 and Chr23 selection signatures contained genes with documented fertility functions or effects (genes shaded in yellow in Fig. 2.3), while the Chr16 selection signature at 43.6-44.5Mb was about 1Mb downstream of five fertility genes.

2.3.2 Selection and fertility

The selection signature analysis provided, for the first time, evidence that fertility genes were also under selective pressure during the forty years of genetic selection which was mostly focused on milk production. Out of the 217 genome regions with signature of selection, 117 regions either contained or were close to potential fertility genes, with 84 regions containing fertility genes, and 33 regions were upstream or downstream of fertility genes mostly within 1Mb distance. In these 117 regions, we identified 196 genes with documented fertility functions or effects, including 77 genes for male fertility, 97 genes for female fertility, and 22 genes for fertility in both males and females (Table 2.2). The fact that these fertility genes are in or near targets of direct genetic selection indicates that the unintended decline in dairy fertility since 1964 was due to the fertility genes that accompanied genetic selection, with negative fertility effects hitchhiked by genetic selection for milk production. The fact that male and female fertility genes had a relatively even ratio (77:98, or approximately 4:5 male:female ratio) indicates that

today's dairy fertility problems likely were due to declined fertility in both males and females.

The 196 fertility genes identified by selection signatures were distributed on all autosomes and the X chromosome. Chr3 had the largest number (thirteen genes in eight regions) and Chr6 had the smallest number (1 gene). Several selection signatures each identified multiple fertility genes, including the bovine MHC region of Chr23 with five genes, the *SPATA16-MECOM* region of Chr1 with four genes and the Holstein lethal haplotype 1, the *ELMO1-DPY19L2* region of Chr4 with four genes, the *ACRBP-CCND2* region of Chr5 with four genes, the *NPC2-ESRRB* region of Chr10 with six genes, the *TARDBP-MTHFR* region of Chr16 with five genes, and the *YBX2-SHBG* region of Chr19 with five genes. We recently showed that the middle of a cattle chromosome generally had lower recombination rate. The five selection signatures in Fig. 2.3 except that of Chr10 were in regions with low recombination rate. Other than Chr6 and Chr25, every chromosome had fertility genes in the middle region of the chromosome with lower recombination rate. The middle region of Chr23 corresponding to the bovine MHC region had the lowest recombination rate among all autosomes. The hitchhiking effects of fertility genes in the middle regions of the chromosomes could have been enhanced by the decreased recombination rates in those regions. Many of the fertility genes had multiple journal articles documenting their fertility functions, e.g., *AR*, *BMP15*, *CD9*, *ESR1*, *FSHR*, *HSD17B2*, *LHCGR*, *MSH5*, *NANOS2*, *POF1B*, *POU5F1*, *PRLR*, *SPEF2*, *SRY*, *STAR* and *TSG-6*. The 196 fertility genes have a wide-range of documented fertility

functions, ranging from completion of meiosis, to testis development, spermatogenesis and spermiogenesis, semen mobility and morphology, and male infertility on the male side; to follicle, oocyte and embryo development, embryo implantation and survival, placenta development, uterine receptivity and environment, miscarriage and premature ovarian failure, pregnancy rate, and female infertility on the female side; and to the joint male-female fertility function of fertilization and sperm-egg fusion. These results support the hypothesis that today's Holstein problems were due to negative effects of many genetic factors associated with many aspects of fertility accumulated through hitchhiking of the forty years of genetic selection.

2.3.3 Selection and immunity

The results of selection signature analysis showed that immunological genes also had changes due to hitchhiking of genetic selection, similar to fertility of contemporary Holsteins. The bovine MHC region was subjected to selection in the entire forty-year period (Fig. 2.3). The selection signature at 21.5-22Mb is 0.4Mb upstream a large cluster of T-cell receptors and immunoglobulin kappa and lambda chains. Several other well documented immunity genes were also in or near selection signatures, including interleukin genes (*IL1R1* [79], *IL1R2* [79], *IL1RL2* [79], *IL4R* [80], *IL6R [81]*, *IL6ST [82]*, *IL21R* [83], *IL22RA1 [84]*, *IL33* [85]), cluster of differentiation (CD) genes (*CD9* [86], *CD19* [87], *CD27* [88], *CD28* [89], *CD48* [90], *CD80* [91], *CD84* [92], *CD244* [93]), *BCL11B* [94], *BCL2* [95], *IKBKB* [96], *NFKBIL1* [97], and *TSG-6* [98], noting that

IL6R, *IL6ST*, *IL33*, *CD9*, *CD48*, *BCL2* and *TSG-6* are also among the fertility genes.

2.3.4 Cluster of genes within the same gene family or type

Clusters of several gene families, including olfactory receptors at 58.07-60.54Mb of Chr5 and 79-82Mb of Chr15, T-cell receptors at 21.5-22Mb of Chr10, microRNA genes at 67.3-67.8Mb Chr21, transfer RNA's and olfactory receptors at 30-31Mb of Chr23, and zinc fingers, cationic amino acid transporters, sialic acid-binding Ig-like genes, and vomeronasal receptors at 58.7-61.4Mb of Chr18.

2.3.5 Single genes, gene-sparse regions, CNV

The number of genes in each selection was in the range of 0-166, for a total of approximately 2700 genes in the 217 selection signatures, not counting aforementioned clusters of special types of genes. A selection signature with a single gene identifies the single gene as an apparent target of direct genetic selection. Twenty nine selection signatures each had single genes, including *SLC9A9*, *RND3*, *PARD3B*, *MGC139448*, *DPYD*, *ARPC2*, *HGF*, *ATP2B1*, *TMEM117*, *TLE4*, *CYLC2*, *SPACA1*, *SORL1*, *BRINP3*, *RFWD2*, *USP38*, *FAT4*, *WWOX*, *LOC781392*, *PPAP2A*, *SPEF2*, *SUCLG2*, *MGMT*, *TUSC3*, *CCSER2*, *LOC524642*, *MGC157332*, *AR*, and *DMD*. Twenty three selection signatures were in blank chromosome regions without any coding gene. Two largest gene-sparse regions with signature of selection were the 13-17.4Mb region of Chr1 with *NCAM2*, and the 13.2-18.3Mb region of Chr24 with *PIK3C3*. Forty eight regions

contained copy number variations (CNV), with six regions containing CNV only, and the region with *CYLC2* on Chr8 overlapped with a CNV region. Some of these selection signatures turned out be mostly due to selection in the elite Holsteins (Group IIIb).

2.3.6 The signature of an elite contemporary family

The elite group of 160 contemporary cows (Group IIIb) was from a half-sib family sired by a single bull. Of these 160 cows, 153 were classified into the upper-right cluster in Fig. 2.2a. This group of cows had high milk production, low somatic cells in the milk, high daughter pregnancy rate (DPR) relative to the remaining contemporary cows and least calving problems[75]. Although the elite line has better DPR than the rest of the contemporary Holsteins, the elite line still has much lower DPR than the unselected cattle. Approximately, the elite cows required 5.89 days shorter than the remaining contemporary cows (DPR=1.18 in Group IIIb and DPR=-0.29 in Group IIIa) but 26.36 days longer than the unselected cows (DPR=7.77 in Group I) for successful conception [YD: recalculate using latest PTA].

Long-range standardized AFD and HD identified fifty regions with significant differences between the elite group (Groups IIIb) and the rest of the contemporary cows (Group IIIa). A striking difference between these two groups was the increased heterozygosity in the elite group. The elite group was more heterozygous than the remaining contemporary cows for all the thirty regions with significant difference in heterozygosity ($|Z_{HD}| \geq 3$) between these two groups. These highly heterozygous regions

in the elite group include those in or near the large clusters of olfactory receptors at 58.07-60.54Mb of Chr5; olfactory receptors and *EMR2* receptors at 9.2-12.2Mb of Chr7; zinc fingers, cationic amino acid transporters, sialic acid-binding Ig-like genes at 57.58-62.53Mb of Chr18; and transfer RNAs and olfactory receptors at 28.8-31.3Mb of Chr23. Fifteen of the twenty six regions with fertility genes were more heterozygous in the elite group than in the other contemporary cows. The *LPPR1-CYLC2- SMC2* region of Chr8 had one of the strongest heterozygosity increases, noting that SMC2 has been shown to be the causal gene of Holstein lethal haplotype 3[67].

Among these regions, those with fertility genes and significant AFD between the elite and unselected cows but not between the remaining contemporary cows and the unselected cow could be regions associated with decreased fertility in the elite cows but not in the remaining contemporary cows, based on the assumption that fertility genes with similar frequencies in the selected and unselected cattle should have similar fertility performance. Such regions include fertility genes of MECOM, STATA3, CD48-VANG12-ATP1A4, CYLC2-SMC2, SPAG6, MTDT and BMP15. Similarly, the elite cows' better fertility than the remaining contemporary cows could be contributing to fertility genes where the elite cows had insignificant AFD but the remaining contemporary cows had significant AFD in comparison to the unselected cattle. Such fertility genes include GPR37, SHB, SPAG1, and IL33-ERMP1. For the SOX2-MFN1-PIK3CA region, Groups IIIa and IIIb had significant AFD when compared to each other and to the unselected cattle in opposite directions.

Table 2.1. Distribution of allele frequency differences (AFD) of 45,878 SNPs.

|AFD|	Number of SNPs			
	Random AFD	I vs. II (1^{st} 20 years)	II vs. III (2^{nd} 20 years)	I vs. III (40 years)
<0.05	44,746	17,482	28,217	15,214
[0.05,0.10)	1131	12,520	12,871	11,937
[0.10,0.15)	1	7912	3917	8330
[0.15,0.20)	0	4437	802	4974
[0.20,0.25)	0	2187	64	3906
[0.25,0.30)	0	908	6	1504
[0.30,0.35)	0	297	1	611
≥0.35	0	135	0	366
>0.10	1	15876 (35.0%)	4790 (10.6%)	18,229 (43.4%)

Table 2.2. List of genes with documented fertility functions or effects in or near genome regions with signature of selection.

Male fertility (76 genes)	*AATK, ACRBP, AGO4, AGTPBP1, ATP1A4, AZIN2, BIRC6, BSP5, CATSPER2, CATSPERB, CTNNB1, DHCR24, DPY19L2, EHD1, ELMO1, ELSPBP1, EPAS1, ESRRB, FGF1, FGF4, FGF9, FKBP6, GFRA1, GPR37, GPX5, HMGB2, IGF1R, IMMP1L, JMJD1C, LUZP2, MAATS1, MLH3, MTDH, MYBL1, NANOS2, NDRG2, NKAPL, PGAP1, PRKAR1A, PRSS37, PTCHD3, RAE1, RBM5, RNF17, RPGR, SEMA3F, SEPT7, SFPQ, SH3GLB1, SHBG, SHCBP1L, SKIV2L, SKIV2L2, SMC6, SOX30, SPACA1, SPAG16, SPAG6, SPATA16, SPATA3, SPATA33, SPATC1, SPEF2, SPEM1, SRM, SRY, SUMO1, TEKT2, TMEM95, TRPC2, TSPY, TSSK4, TTLL5, VRK1, VRK2, YBX2*
Female fertility (97 genes)	*ADAM19, AKT1, ALKBH3, ANG2, AR, ARFGEF2, ARPC2, ASH2L, BCL2, BMP2, BMP7, BMP15, CCL28, CCND2, CD9, CD48, CDC20B, CSE1L, DACH2, DIAPH2, DICER1, DYNLT3, EIF2B2, ERMP1, ESR1, FGF16, FGF23, FLRT3, FMN2, GHR, GREM1, HMGCR, HSF1, HSPA1A, IGFBP1, IGFBP3, IGFBP7, IL33, IL6ST, INSL5, ISM1, KAT8, KIF16B, KITLG, LAMC1, LATS1, LATS2, LLGL1, LMO4, MAPK3, MECOM, MFN1, MMADHC, MST1, MTOR, MYO18B, NELL2, NEURL4, NTRK2, NTRK3, NUPR1, PAPPA2, PCDH11X, PCDH12, PDE4B, PELO, PGF, PIK3C3, PIK3CA, POF1B, POLR3G, POU5F1, POU6F2, RAB10, RAPGEF2, RBX1, REV3L, RSBN1L, SFMBT2, SHB, SMAD4, SMC2, SMG7, SMN2, SOX4, SPAG1, STIM1, SULF1, SULF2, TARDBP, TMED2, TMEM60, TSG6, UBIAD1, UCHL3, VANGL2, ZP4*
Fertility in both males and females (22 genes)	*ATHFR, ATP6V0A2, FANCC, FSHR, GHSR, HMGB2, HSD17B12, IL6R, IMMP2L, LHCGR, LOXL4, MSH5, NPC2, PRLR, PUM1, SLC39A10, SOX2, SPO11, STAR, TYRO3, UBB*

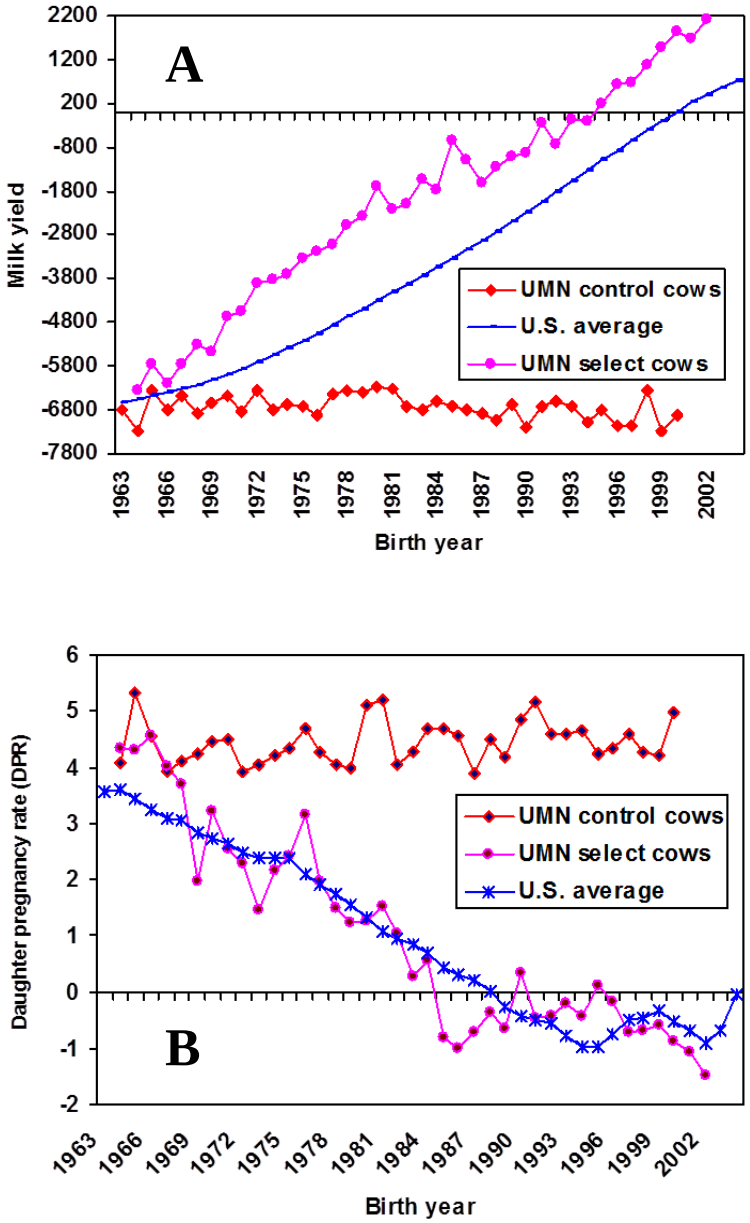


Figure 2.1. Phenotypic changes due to selection since 1964 for milk yield (A) and daughter pregnancy rate (B). The genetic merit of milk yield increased but daughter pregnancy rate decreased steadily for the U.S. Holstein cows and the University of Minnesota (UMN) selected cows. The UMN cows unselected since 1964 remained relatively unchanged for milk yield and daughter pregnancy rate.

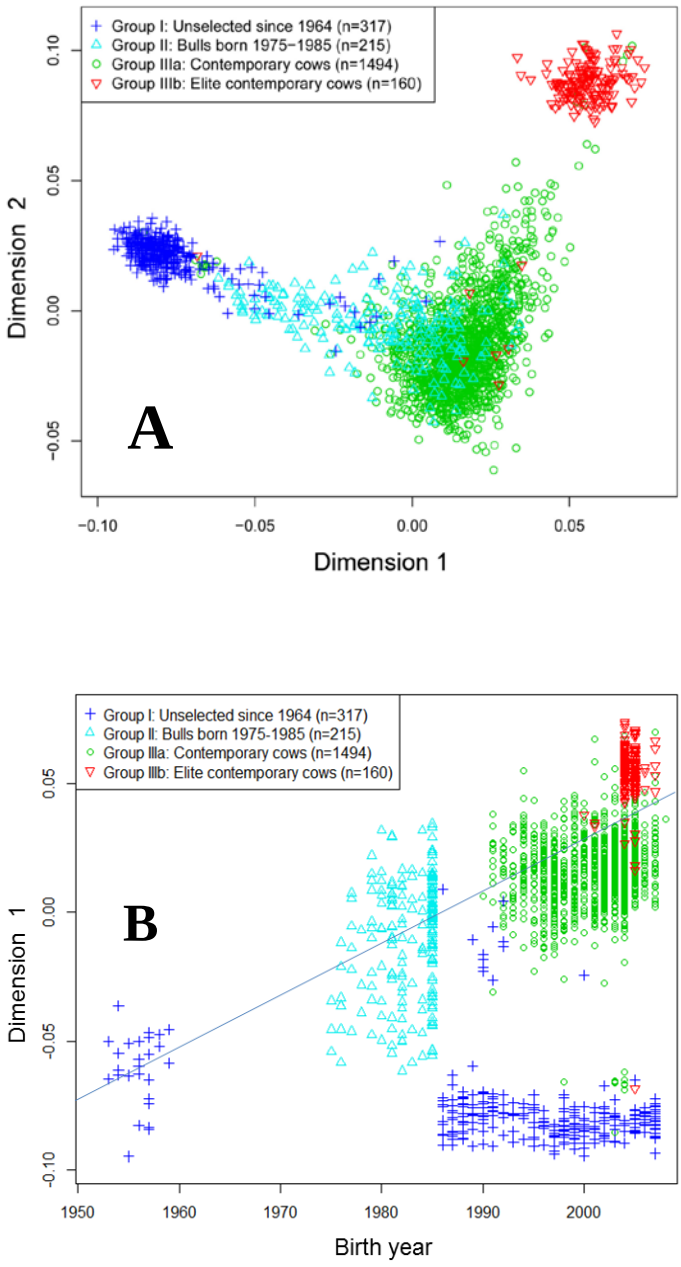


Figure 2.2. Multidimensional scaling plot of the first dimension versus the second dimension of the SNP identity by state distances: Genome landscape shift from the unselected Holsteins at the lower left to the elite Holsteins at the upper right (A) and Time trend of the changing Holstein genome (B).

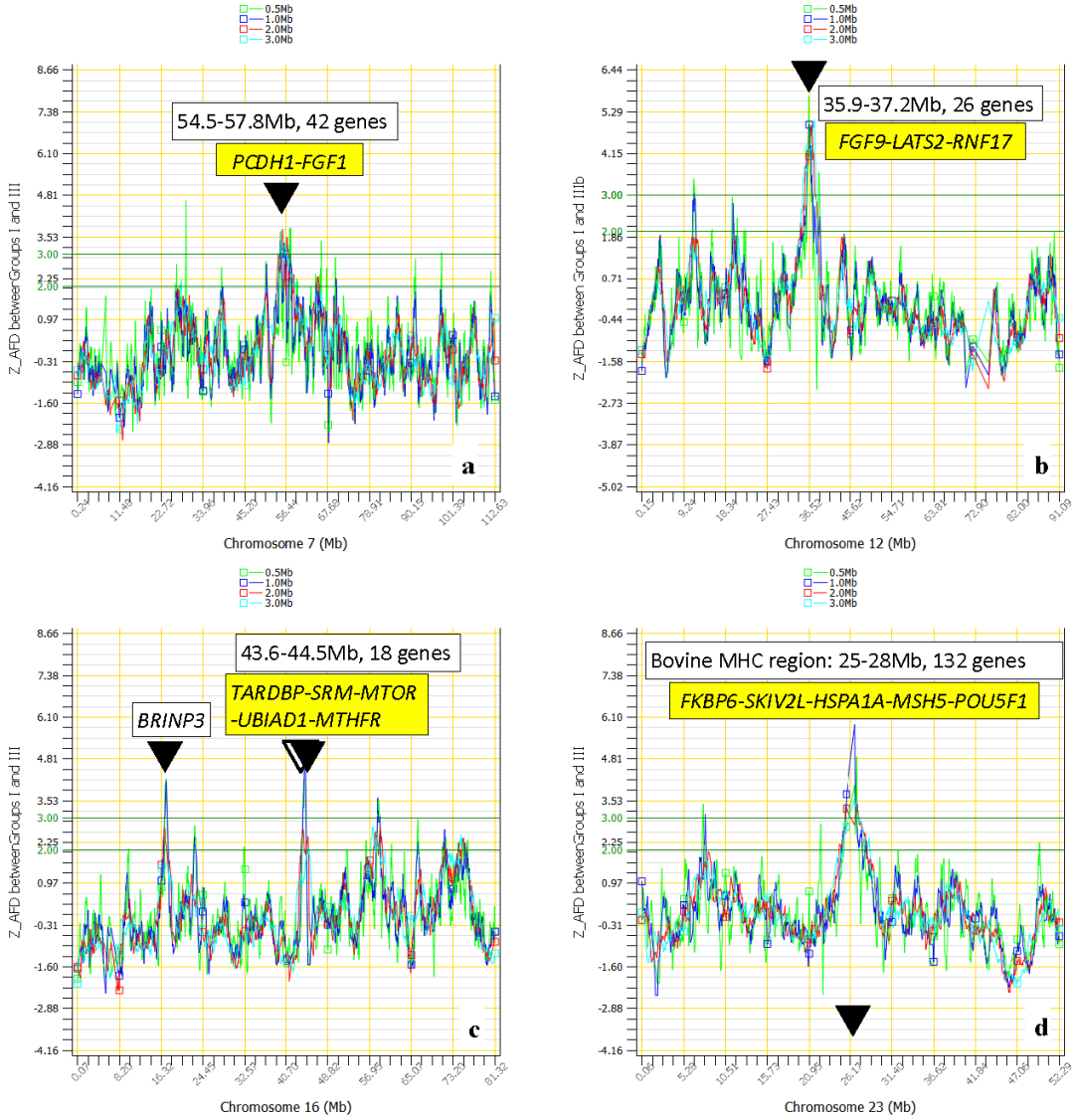


Figure 2.3. Examples of selection signatures identified by all four measures of selection signatures, long-range standardized allele frequency differences and heterozygosity differences between Groups III and I, and between Groups IIIb and I. a, The 54.5-57.8Mb region of Chr07. **b**, The 35.9-37.2Mb region of Chr12. **c**, The 16.1-17.9Mb and 43.6-44.5Mb regions of Chr16. **d**, The 25-28Mb region of Chr23 that is approximately the bovine major histocompatibility (MHC) region.

Chapter 3

Haplotype-based selection signature analysis in three Holstein cattle groups

3.1 Introduction

Haplotype data have been widely used in GWAS and selection signature analysis, as well as in many other studies including imputation. Haplotype-based analysis is more powerful than SNP-based analysis because a haplotype may carry or is linked to more QTLs than a single SNP. In dairy cattle population where the linkage desequilibirum is high due to the small effective population size (N_e) and strong artificial selection, haplotype-based selection signature analysis has even more advantages than single-SNP analysis.

In this chapter, we generated the haplotype data for three groups of Holstein cattle, and compared the haplotype-based statistics between those groups. As a result, we validated a number of selection signatues from SNP-based analysis, and found new selection signaures from haplotype-based analysis.

3.2 Material and Methods

3.2.1 Study population

A total of 2,366 Holstein cattle were genotyped using the Bovine 50K SNP array, as described in Chapter 2. Based on animal herd information and birth year, we split them into four groups of interest, control group (C; N = 301), middle group (G; N = 216), UMN selected group (S; N = 151), and contemporary group (M; 1990). The control group includes mostly cattle from the UMN control herd and 10 old bulls born before 1958 that were no longer under artificial selection since then. The middle group includes bulls born between the years from 1975 to 1985. The UMN selected group includes cows maintained and selected in UMN at the same time as the UMN control herd. The contemporary group includes cows born after 1990. Using this design of four groups, we anticipated to show genome differences between C and S or M with G in the middle if they are associated with the complex traits under artificial selection including mostly milk production and possibly cattle reproduction.

3.2.2 Haplotype phasing, frequency and diversity

After quantity control procedures, the 2,366 genotype data were phased into haplotypes using the Findhap program at the Animal Genomics and Improvement Laboratory (AGIL), USDA. Findhap was reported to be fast and have high accuracy in

cattle studies. Using Findhap, the 50K SNPs were divided into 621 genome segments with about 100 SNPs in each segment. For each segment, two haplotypes were identified and assigned to an animal based on the SNP genotype.

Because many forms of haplotypes (alleles) at each segment can be identified, we focused on the five most frequent ones in the C and M groups in our analysis. Within each of the four groups (C, G, S, and M), we calculated the frequency for the five most frequent haplotypes in C and M, respectively (Figures 3.1 and 3.2). We also calculated and summarized the site frequency spectrum of the haplotypes within each segment. In particular, we counted the number of haplotypes with frequency greater than 0.01 within each of the 621 genome segments.

3.2.3 Pairwise comparison of haplotype frequencies

To identify genomic regions that differ due to selection, we compared the frequency of haplotypes between the four groups (pairwise). Since we are mostly interested in the difference between control and selected groups, the five most frequent haplotypes in C and five in M were included in the analysis. Under the model of neutrality (by chance), we expect to identify haplotypes with large frequency difference between C and M: 1) high frequency in C and low frequency in M from the 5 most frequent haplotypes in C, and 2) low frequency in C and high frequency in M from the 5 most frequent haplotypes

in M. Other haplotypes not included here will be less different between C and M than those included.

3.3 Results and Discussion

3.3.1 Asymmetric distribution of haplotypes with different frequencies in control and selected groups

Our analysis is unbiasedly designed to detect two types of frequency differences, high frequency in C and low frequency in M, and low frequency in C and high frequency in M. However, large frequency differences are much more enriched in the latter category, low frequency in C and high frequency in M (Figures 3.2 and 3.4). Using 0.2 as a cutoff value, we identified a total of 54 haplotypes that are more frequent in M than in C (Tables 3.1 and 3.2), none of which are the other way around (more frequent in C than in M).

Such an asymmetric distribution of haplotypes with different frequencies in the control and selected groups could be related to the reduced effective population size and strong positive selection on a few “optimum” haplotypes that carry favorable mutations associated with best milk production. Because the increasing intensity of selection in cattle breeding, the chance is much higher to see a haplotype (likely from an influential bull with a large number of offspring) to have a high frequency in the contemporary group than in the control group.

To further verify the hypothesis of decreased effective population size with time, we checked the site frequency spectrum of the haplotypes. First, we compared the frequencies of the most frequent haplotypes in each of the four groups. As shown in Figure 3.5, the most frequent haplotypes in the control group tend to have larger frequency than those in the contemporary group, indicating a larger effective population size for the control group. Consistently, additional analyses using other haplotypes (the 2nd most frequent to the fifth most frequent) showed the same trend (Figure 3.5). Second, we calculated the number of unique haplotypes in each segment within each of the four groups. To reduce possible genotyping and phasing errors, we filtered all haplotypes with frequency not greater than 0.01. Except for the UMN selected group that has the smallest sample size and possibly largest variation due to random noise, the other three groups (C, G, and M) showed a consistent pattern with previous results that the control group (C) has the largest number of unique haplotypes, the contemporary group (M) has the smallest number of haplotypes, and the middle group (G) is in between (Figure 3.6), suggesting a continuous reduction of effective population size in the past.

3.3.2 Genomic loci mostly different between the control and selected groups

Using Manhattan plots of the haplotype frequency, frequency difference, and number of haplotypes (Figures 3.1-4 and 3.7), we identified 10 genomic loci that exhibit large difference between the control and selected groups on Chromosomes 1, 2, 5, 7, 8, 10, 11,

13, 16, and 20, one locus on each chromosome (Tables 3.1 and 3.2). For most of these 10 loci, results from haplotype frequency difference and number of unique haplotype are consistent, suggesting stronger evidence supporting the findings from multiple analyses (Tables 3.1 and 3.2). We compared this result with the SNP-based results in Chapter 2, and six loci out of 10 (Chromosomes 2, 5, 7, 8, 10, and 20) are confirmed to show some reproduction-related evidence.

3.3.3 Conclusions

Collectively, our conclusions include: 1) Selection since 1964 resulted in genome landscape changes; 2) Diversity in Holstein has drastically reduced due to selection; 3) Haplotype-based selection signature analysis is a good compliment of the SNP-based analysis because haplotype-based selection signature analysis confirms identified signatures by SNP-based test and haplotype-based selection signature analysis can detect additional selection signals. In the future, those candidate regions under selection can be used in genomic selection to increase prediction accuracy for production, reproduction and health traits in dairy cattle.

Table 3.1. Haplotypes show large frequency difference between control and selected groups based on the 5 most frequent haplotypes from the control group. Listed haplotypes include the top 10 haplotypes between control haplotype 1 and modern group as well as any other haplotypes that show an allele frequency difference larger than 0.2 from the control group. D_cm1 is the frequency difference between C and M for the most frequent haplotype from C (Hap#1 in C). N_C is the number of the unique haplotype in C with frequency greater than 0.01.

#Seg	Chr	Start	End	D_cm1	D_cs1	D_cg1	D_cm2	D_cm3	D_cm4	D_cm5	N_C	N_G	N_S	N_M
60	2	114262794	117642353	-0.20737	-0.15511	-0.0949	0.024202	0.056	0.047652	0.046	26	21	22	18
61	2	117735021	122019234	-0.20707	-0.16022	-0.11222	0.05	0.047303	0.042	-0.03178	27	21	25	16
62	2	122043704	126341038	-0.21546	-0.15435	-0.09622	0.059561	0.052	0.041303	0.034774	28	18	28	17
63	2	126372006	129768904	-0.17831	-0.14057	-0.07417	0.052	-0.0182	0.046	0.024321	25	20	30	19
65	2	133737741	136305920	-0.25686	-0.23495	-0.17101	0.084955	-0.01229	0.02308	0.048606	28	17	20	16
144	5	97166662	101442885	0.058955	0.06	0.052991	0.012007	0.056	0.052955	-0.20192	31	18	28	14
145	5	101493043	105621447	0.105909	0.104622	0.103327	0.080258	0.020641	-0.22325	0.043303	27	18	25	13
190	7	42262165	47717578	0.02301	0.04673	0.033121	-0.21023	0.056	-0.04277	0.051303	25	18	19	11
267	10	52814637	57297166	-0.20596	-0.07662	-0.1571	0.034578	0.045652	0.035031	0.042	25	20	26	13
268	10	57427043	62511957	0.095561	0.084486	0.083981	-0.05001	-0.2582	0.053652	0.040683	26	18	23	11
269	10	62544551	68199588	0.066	0.066	0.061327	-0.06109	-0.21517	0.020718	0.051561	27	19	28	14
270	10	68231955	72769319	0.101714	0.082216	0.059925	-0.22698	0.075895	0.085303	0.053909	29	20	27	17
281	11	6188577	9905475	-0.14482	-0.10505	-0.0989	0.074	0.022913	0.010369	0.042	28	17	25	18
331	13	28095457	33201457	0.104509	0.091054	0.068542	0.073728	0.07	-0.20228	0.002188	24	21	23	16
469	20	27530784	32045791	-0.19808	-0.07635	-0.06729	0.059303	0.044	0.042	0.035909	26	11	27	14
470	20	32074342	35639448	-0.15557	-0.1013	-0.08204	-0.01376	0.060955	0.056	0.049652	20	9	19	11
471	20	35684389	39727100	-0.23326	-0.20259	-0.13136	0.102077	0.076927	0.056955	0.048	24	14	20	11
473	20	44351953	49784650	0.118592	0.117865	0.085944	0.104606	-0.21921	0.056167	0.052	24	14	19	9

Table 3.2. Haplotypes show large frequency difference between control and modern groups based on the 5 most frequent haplotypes from the modern group. Listed haplotypes include all haplotypes that show an allele frequency difference larger than 0.2 between the two groups. D_cm1 is the frequency difference between C and M for the most frequent haplotype from C (Hap#1 in C). N_C is the number of the unique haplotype in C with frequency greater than 0.01.

#Seg	Chr	Start	End	D_cm1	D_cs1	D_cg1	N_C	N_G	N_S	N_M
11	1	46518842	52007154	-0.27118	-0.30081	-0.14187	31	16	23	19
12	1	52030982	56181944	-0.25099	-0.232	-0.12686	28	20	21	19
13	1	56193336	61469035	-0.27805	-0.27365	-0.13551	25	21	24	19
14	1	61495059	65624310	-0.22753	-0.23649	-0.13084	29	24	27	22
15	1	65663517	70159840	-0.23528	-0.23111	-0.14052	27	18	21	17
56	2	95141342	101409902	-0.20192	-0.19532	-0.09179	21	13	20	16
60	2	114262794	117642353	-0.20737	-0.15511	-0.0949	26	21	22	18
61	2	117735021	122019234	-0.20707	-0.16022	-0.11222	27	21	25	16
62	2	122043704	126341038	-0.21546	-0.15435	-0.09622	28	18	28	17
64	2	129809374	133699531	-0.21024	-0.20711	-0.10484	30	18	24	19
65	2	133737741	136305920	-0.25686	-0.23495	-0.17101	28	17	20	16
144	5	97166662	101442885	-0.20192	-0.19662	-0.25374	31	18	28	14
145	5	101493043	105621447	-0.22325	-0.22289	-0.19665	27	18	25	13
190	7	42262165	47717578	-0.21023	-0.1307	-0.09856	25	18	19	11
191	7	47779803	54513838	-0.23668	-0.14665	-0.11482	26	19	24	13
192	7	54567919	59405737	-0.23284	-0.15678	-0.1195	26	15	21	12
193	7	59447780	63542128	-0.21472	-0.13651	-0.11015	27	17	23	13
212	8	30407640	35339110	-0.21786	-0.15003	-0.13819	24	23	28	18
215	8	45113463	49837145	-0.21612	-0.14327	-0.13819	25	21	28	17
216	8	49925026	54083456	-0.20559	-0.18119	-0.14153	27	18	27	16
217	8	54105059	59054244	-0.21709	-0.22511	-0.15789	25	19	22	15
218	8	59085873	63875597	-0.25855	-0.24876	-0.15088	21	24	21	15
219	8	63931756	68588863	-0.21691	-0.18181	-0.13852	25	23	22	19
220	8	68623725	73881694	-0.23798	-0.21284	-0.16822	26	18	24	18
221	8	73907982	78890670	-0.28915	-0.21449	-0.16725	25	15	20	13
261	10	26146027	29967808	-0.22127	-0.10011	-0.10415	23	20	32	21
264	10	40432932	45207315	-0.20174	-0.125	-0.08178	31	22	29	18
265	10	45237046	48926752	-0.21772	-0.16043	-0.07379	25	23	23	17
266	10	49013809	52814326	-0.28302	-0.16354	-0.19893	27	21	30	14
267	10	52814637	57297166	-0.20596	-0.07662	-0.1571	25	20	26	13

268	10	57427043	62511957	-0.2582	-0.13857	-0.1703	26	18	23	11
269	10	62544551	68199588	-0.21517	-0.10416	-0.14994	27	19	28	14
270	10	68231955	72769319	-0.22698	-0.162	-0.12695	29	20	27	17
277	10	97300873	100974235	-0.22265	-0.19932	-0.16121	27	17	26	16
331	13	28095457	33201457	-0.20228	-0.13457	-0.11256	24	21	23	16
395	16	34994367	38944217	-0.20636	-0.17368	-0.05407	27	20	22	16
468	20	23052519	27503083	-0.23114	-0.16657	-0.12353	24	15	24	17
471	20	35684389	39727100	-0.23326	-0.20259	-0.13136	24	14	20	11
473	20	44351953	49784650	-0.21921	-0.11795	-0.05751	24	14	19	9

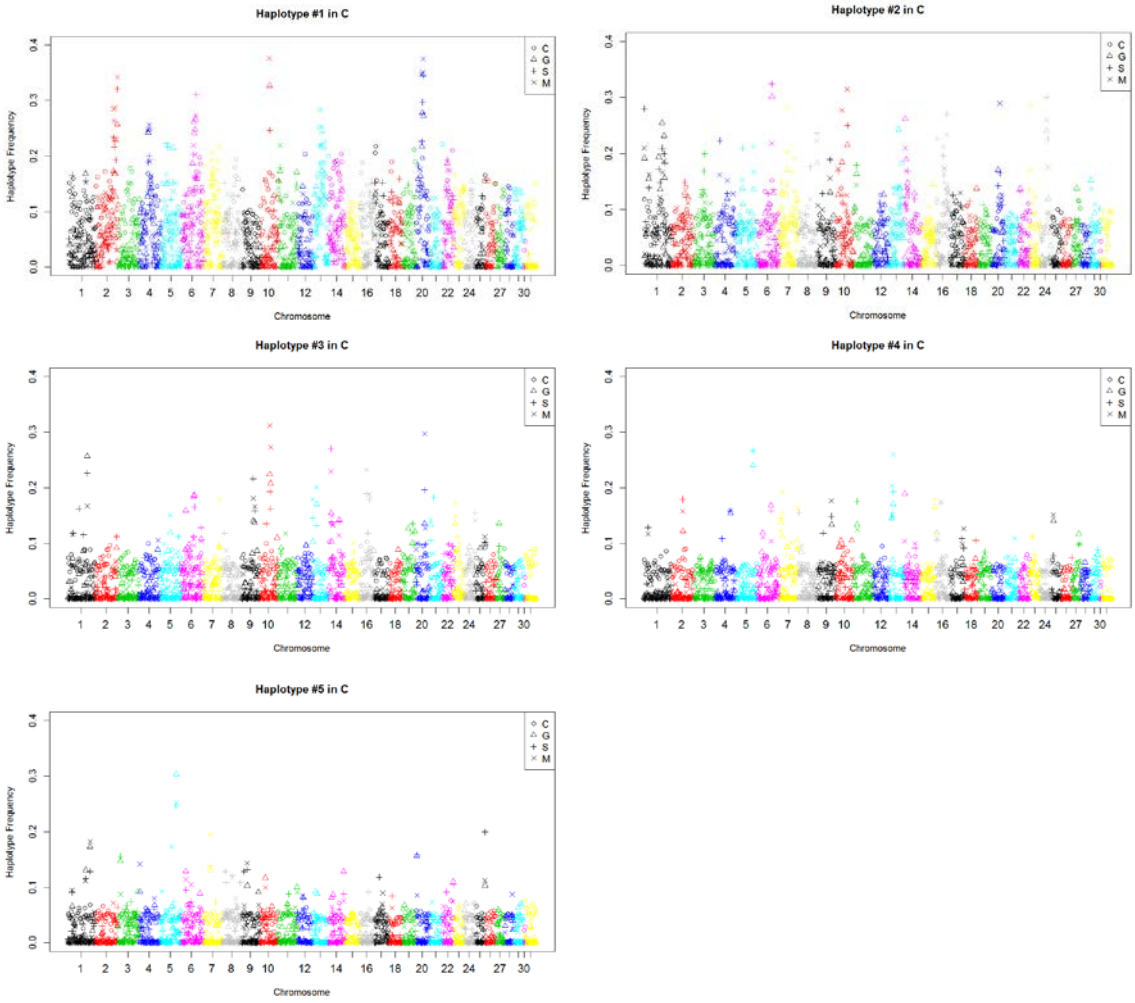
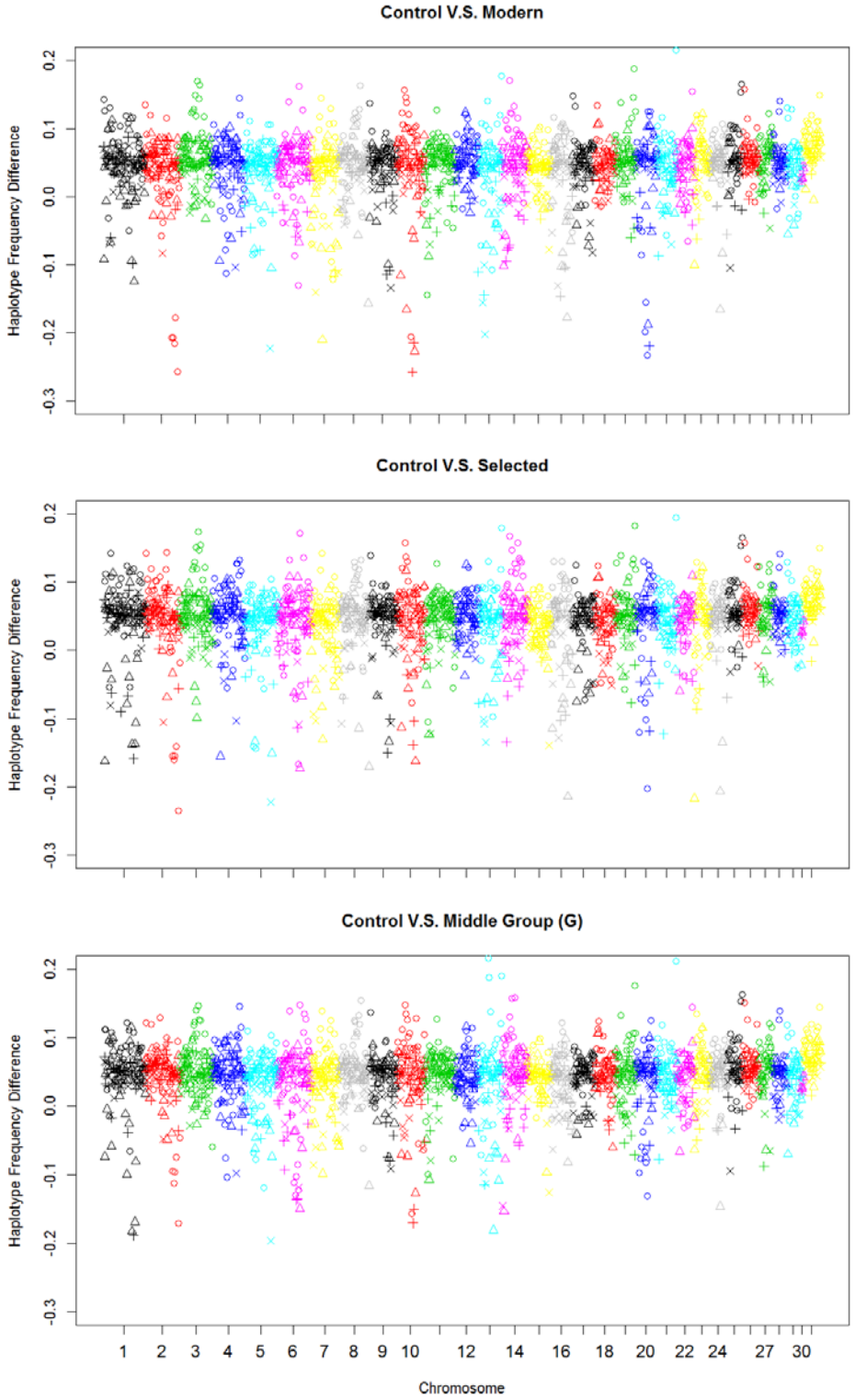


Figure 3.1 Manhattan plots of haplotype frequencies in the four groups. Only the top 5 most frequent haplotypes from the control group are included: A) the most frequent haplotypes in control; B) the second most frequent haplotypes; C) the third most frequent haplotypes; D) the fourth most frequent haplotypes; and E) the fifth most frequent haplotypes.


Control V.S. Modern
Haplotype Frequency Difference
0.2
0.1
0.0
-0.1
-0.2
-0.3
Control V.S. Selected
Haplotype Frequency Difference
0.2
0.1
0.0
-0.1
-0.2
-0.3
Control V.S. Middle Group (G)
Haplotype Frequency Difference
0.2
0.1
0.0
-0.1
-0.2
-0.3
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 27 30
Chromosome

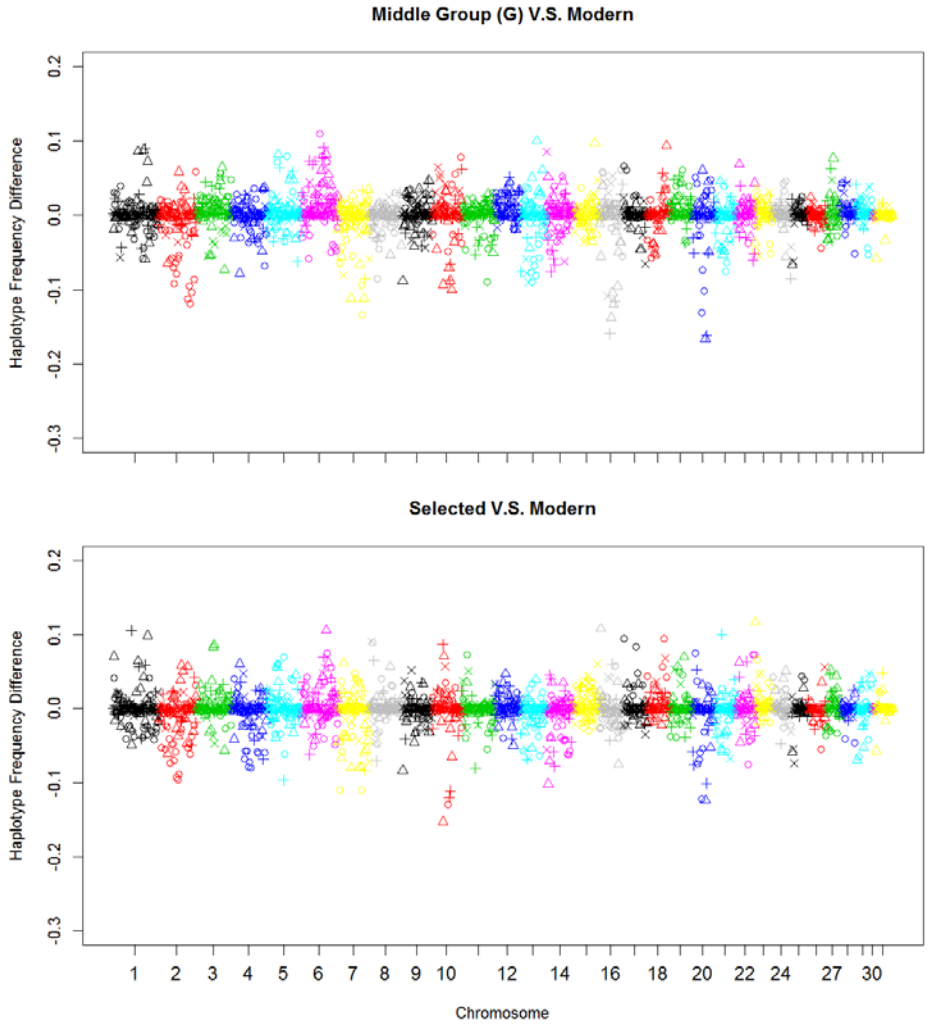


Figure 3.2. Pairwise haplotype-frequency differences between four groups based on five most frequent haplotypes from the control group.

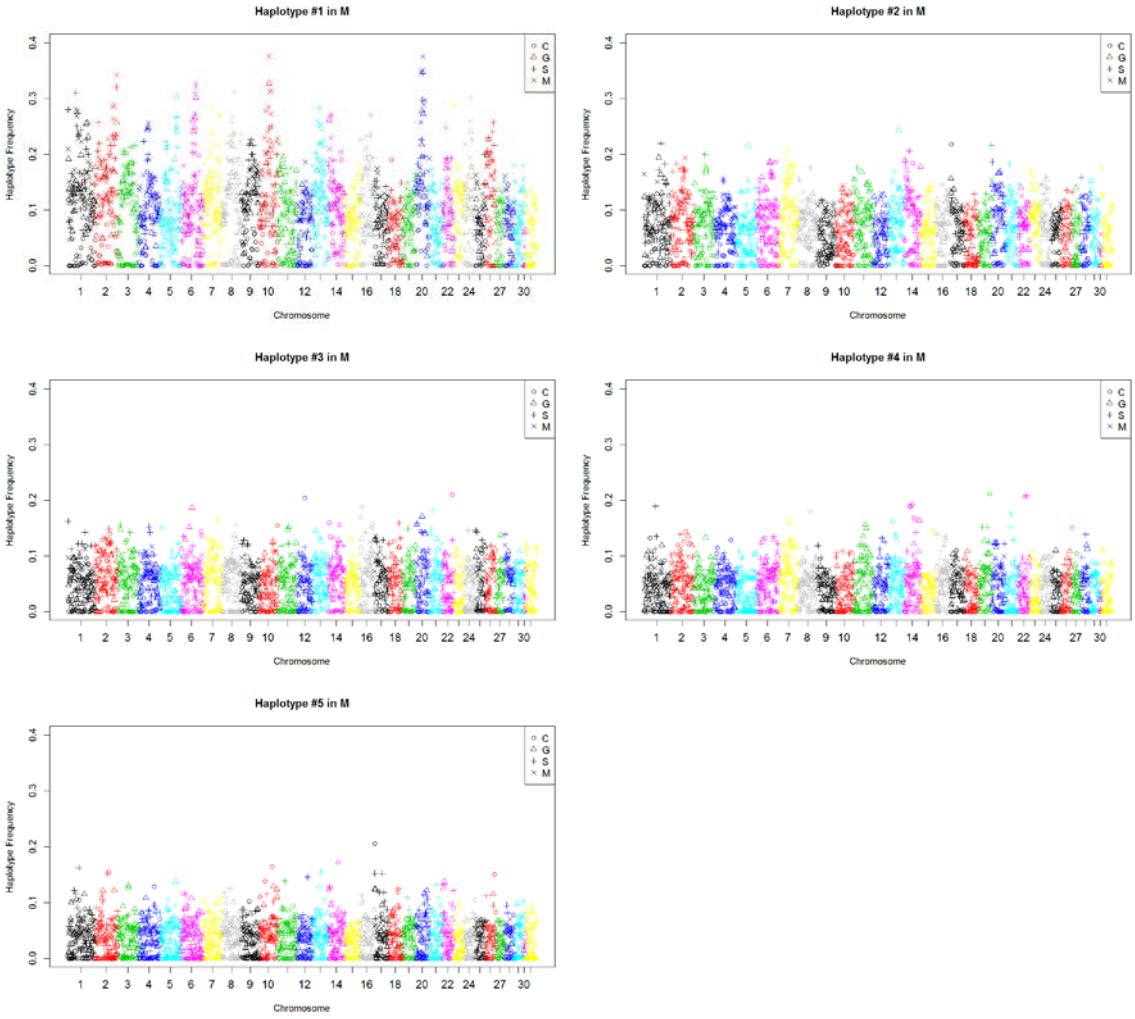
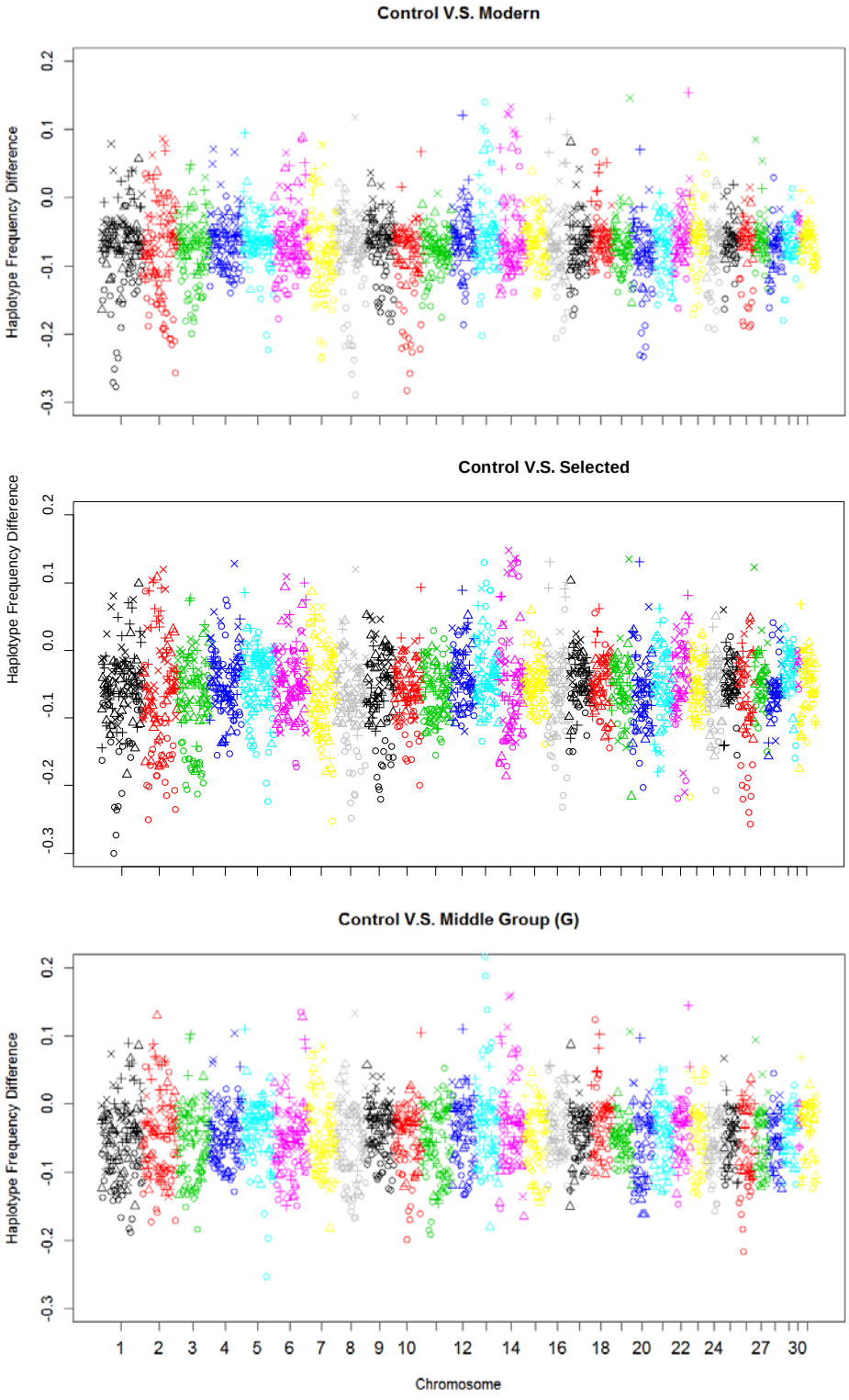


Figure 3.3. Manhattan plots of haplotype frequencies in the four groups. Only the top 5 most frequent haplotypes from the contemporary group are included: A) the most frequent haplotypes in control; B) the second most frequent haplotypes; C) the third most frequent haplotypes; D) the fourth most frequent haplotypes; and E) the fifth most frequent haplotypes.


Control V.S. Modern
Haplotype Frequency Difference
0.2
0.1
0.0
-0.1
-0.2
-0.3
Control V.S. Selected
Haplotype Frequency Difference
0.2
0.1
0.0
-0.1
-0.2
-0.3
Control V.S. Middle Group (G)
Haplotype Frequency Difference
0.2
0.1
0.0
-0.1
-0.2
-0.3
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 27 30
Chromosome

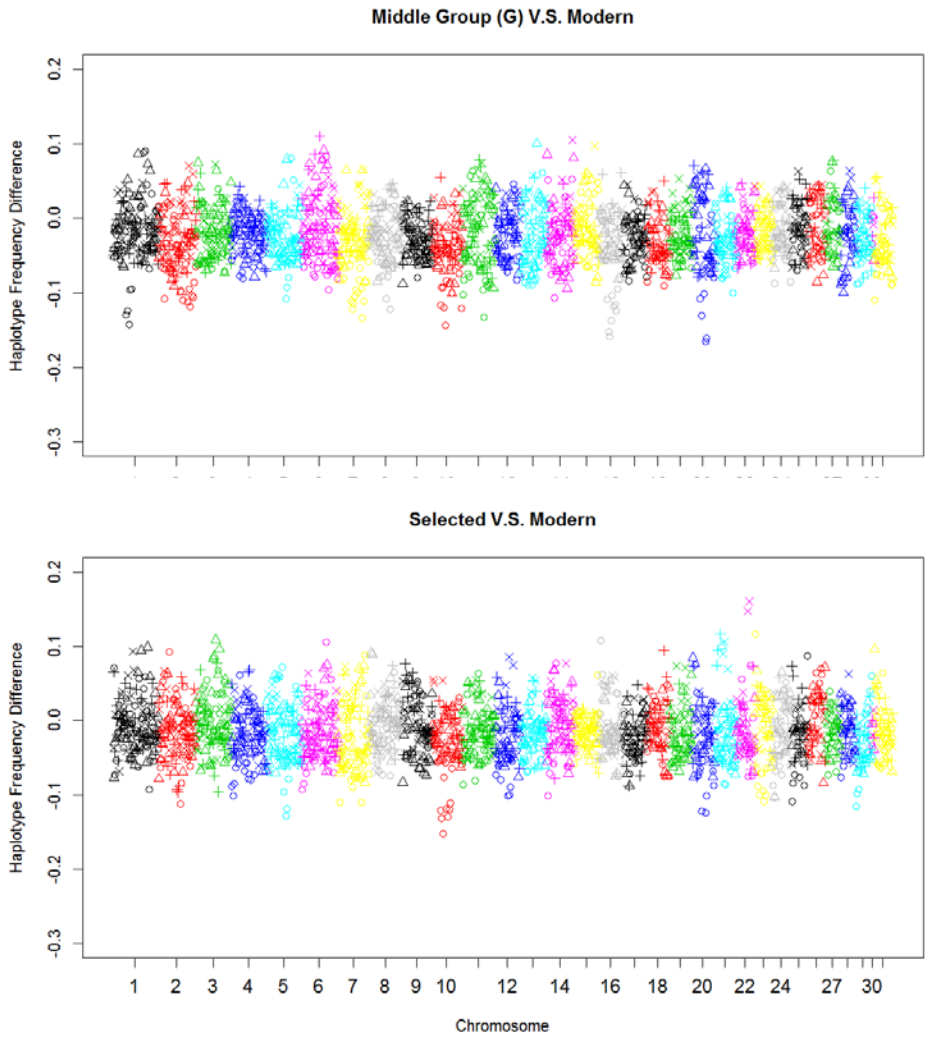


Figure 3.4. Pairwise haplotype-frequency differences between four groups based on five most frequent haplotypes from the contemporary group.

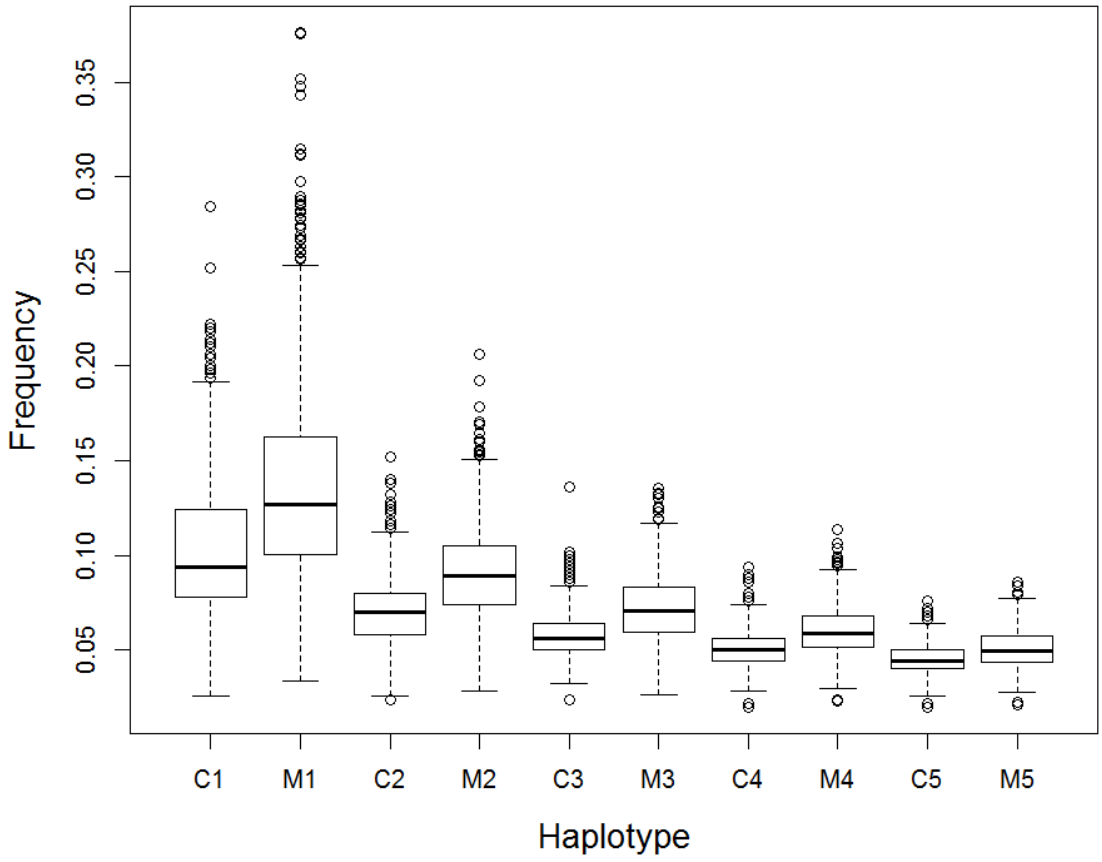


Figure 3.5. Boxplot of frequencies of five most frequent haplotypes in C (N = 5) and M (N = 5).

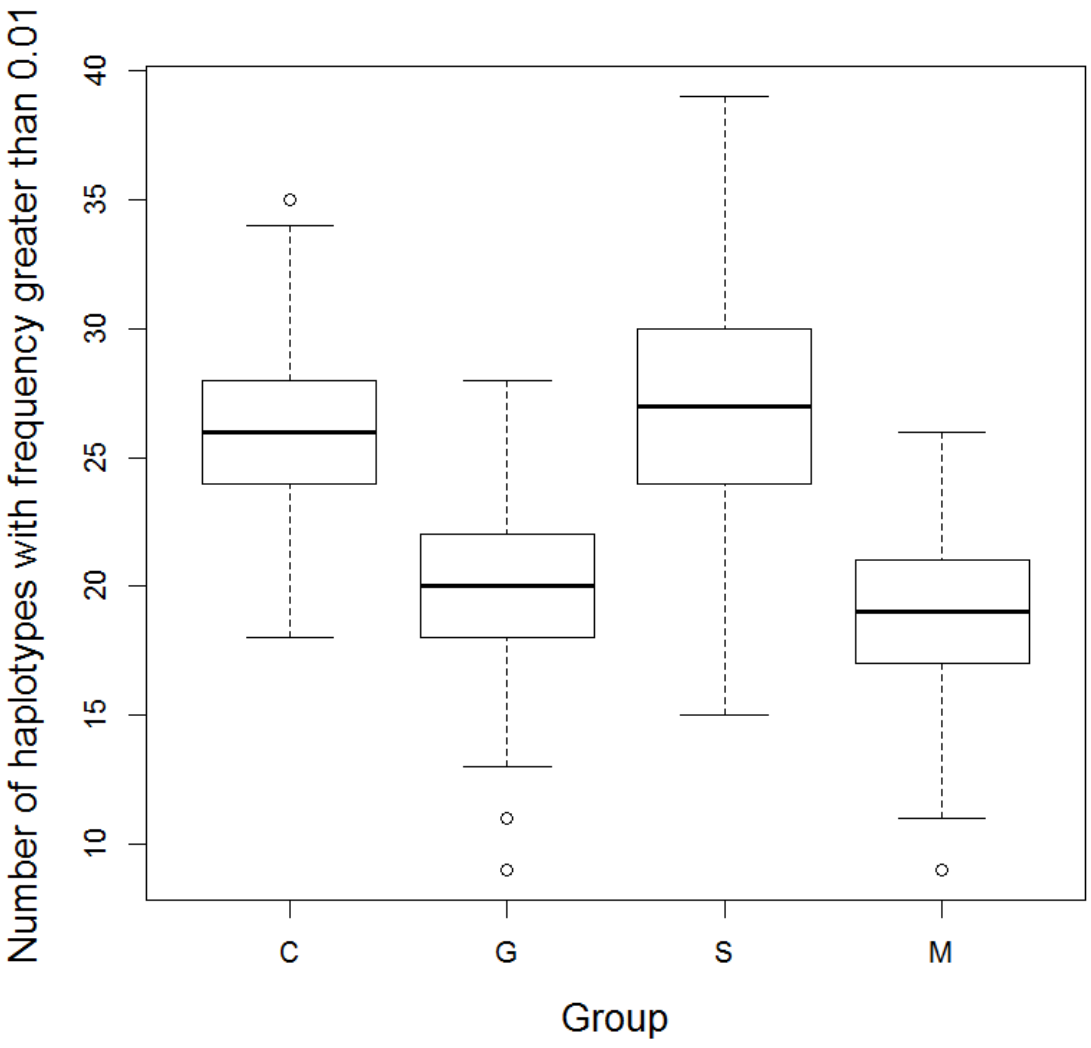
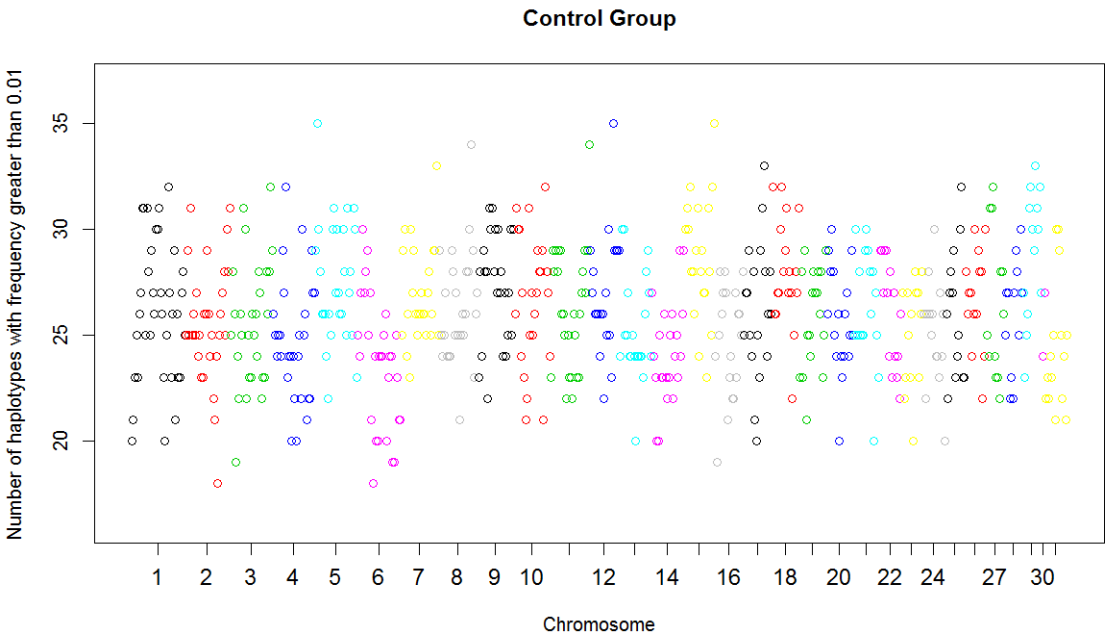
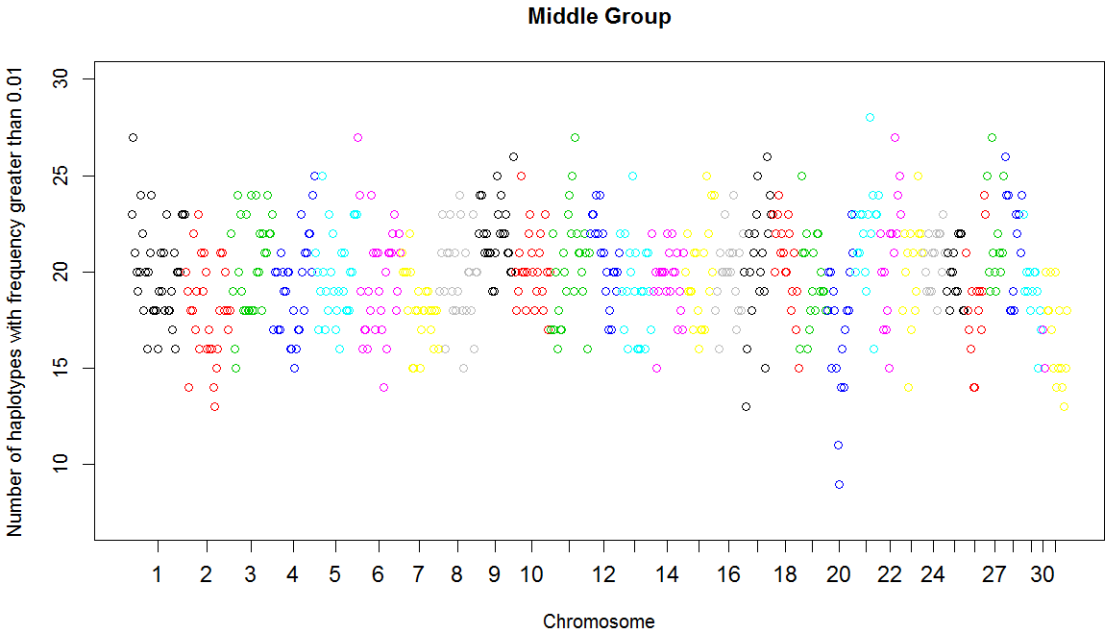


Figure 3.6. Boxplot of number of haplotypes with frequency greater than 0.01 across all genome segments in the four groups.


Control Group
Number of haplotypes with frequency greater than 0.01
20
25
30
35
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
27
30
Chromosome


Middle Group
Number of haplotypes with frequency greater than 0.01
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
27
30
Chromosome

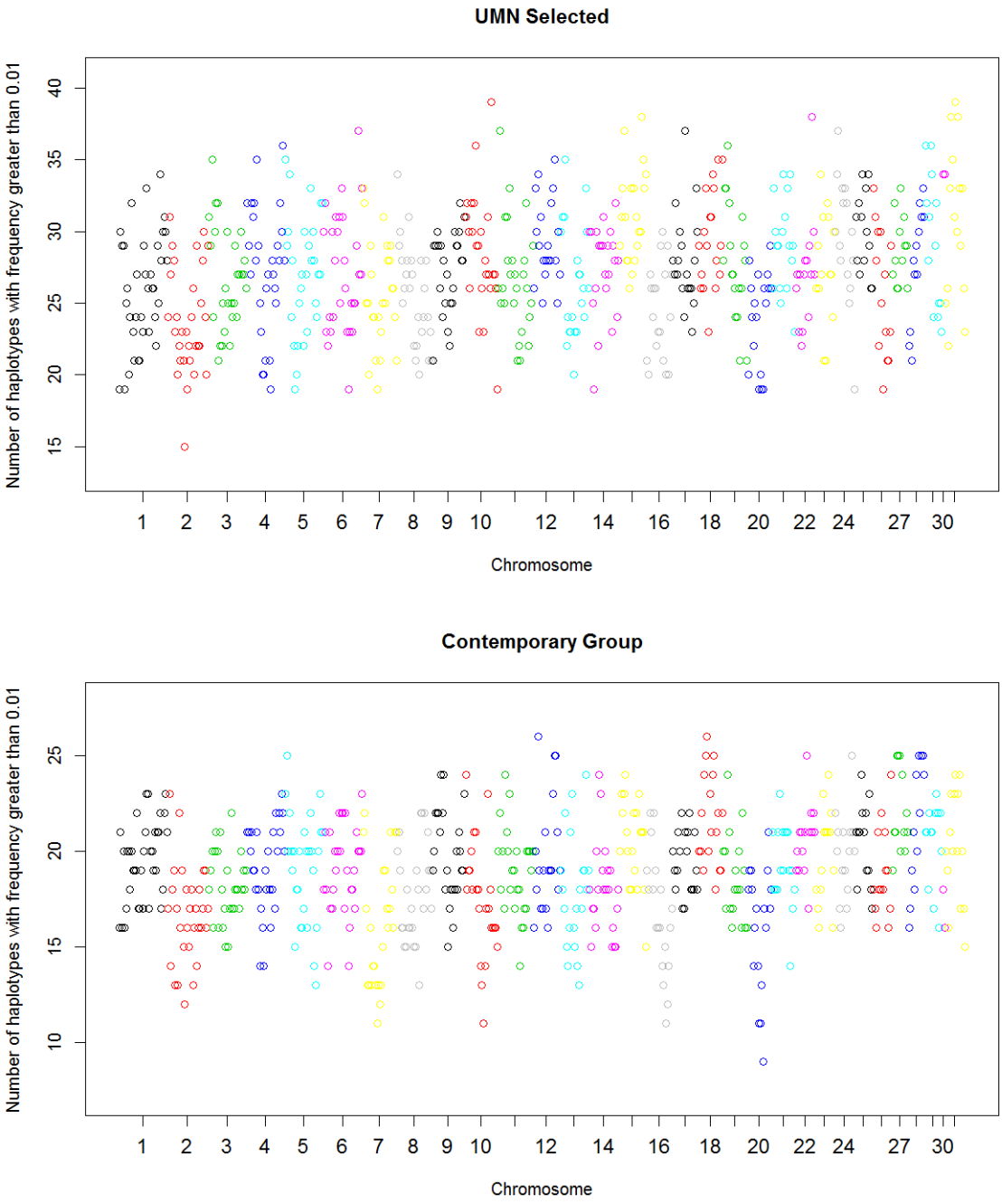


Figure 3.7. Manhattan plots of the number of haplotypes with frequency greater than 0.01 in the four groups, C, G, S, and M.

References

1. Hayes, B., et al., *Invited review: Genomic selection in dairy cattle: Progress and challenges.* Journal of dairy science, 2009. **92**(2): p. 433-443.
2. De Vries, M. and R. Veerkamp, *Energy balance of dairy cattle in relation to milk production variables and fertility.* Journal of dairy science, 2000. **83**(1): p. 62-69.
3. Royal, M., et al., *Declining fertility in dairy cattle: changes in traditional and endocrine parameters of fertility.* Animal science, 2000. **70**(3): p. 487-501.
4. Hansen, L., A. Freeman, and P. Berger, *Yield and fertility relationships in dairy cattle.* Journal of Dairy science, 1983. **66**(2): p. 293-305.
5. VanRaden, P., et al., *Development of a national genetic evaluation for cow fertility.* Journal of dairy science, 2004. **87**(7): p. 2285-2292.
6. Seykora, A. and B. McDaniel, *Heritabilities and correlations of lactation yields and fertility for Holsteins.* Journal of dairy science, 1983. **66**(7): p. 1486-1493.
7. Pryce, J., et al., *Genetic aspects of common health disorders and measures of fertility in Holstein Friesian dairy cattle.* Animal Science, 1997. **65**(3): p. 353-360.
8. Kadarmideen, H.N., et al., *Genetic parameters and evaluations from single-and multiple-trait analysis of dairy cow fertility and milk production.* Livestock Production Science, 2003. **81**(2): p. 183-195.
9. Walsh, S., E. Williams, and A. Evans, *A review of the causes of poor fertility in high milk producing dairy cows.* Animal reproduction science, 2011. **123**(3): p. 127-138.
10. Peñagaricano, F. and H. Khatib, *Association of milk protein genes with fertilization rate and early embryonic development in Holstein dairy cattle.* Journal of Dairy Research, 2012. **79**(01): p. 47-52.
11. Veerkamp, R. and B. Beerda, *Genetics and genomics to improve fertility in high producing dairy cows.* Theriogenology, 2007. **68**: p. S266-S273.
12. Lucy, M., *Reproductive loss in high-producing dairy cattle: where will it end?* Journal of dairy science, 2001. **84**(6): p. 1277-1293.
13. Philipsson, J., *Genetic aspects of female fertility in dairy cattle.* Livestock production science, 1981. **8**(4): p. 307-319.
14. Cochran, S.D., et al., *Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle.* BMC genetics, 2013. **14**(1): p. 49.
15. Wathes, D., et al., *Influence of negative energy balance on cyclicity and fertility in the high producing dairy cow.* Theriogenology, 2007. **68**: p. S232-S241.
16. Roche, J.F., *The effect of nutritional management of the dairy cow on reproductive efficiency.* Animal reproduction science, 2006. **96**(3): p. 282-296.
17. Berry, D., E. Wall, and J. Pryce, *Genetics and genomics of reproductive performance in dairy and beef cattle.* animal, 2014. **8**(s1): p. 105-121.

18. Butler, W., *Nutritional interactions with reproductive performance in dairy cattle.* Animal Reproduction Science, 2000. **60**: p. 449-457.
19. Consortium, G.P., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
20. Morrison, A.C., et al., *Whole Genome Sequence-Based Analysis of a Model Complex Trait, High Density Lipoprotein Cholesterol.* Nature genetics, 2013. **45**(8): p. 899.
21. Georgi, B., et al., *Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate.* PLoS genetics, 2014. **10**(3): p. e1004229.
22. Daetwyler, H.D., et al., *Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle.* Nature genetics, 2014. **46**(8): p. 858-865.
23. Georges, M., *Towards sequence-based genomic selection of cattle.* Nature genetics, 2014. **46**(8): p. 807-809.
24. Pausch, H., et al. *Exploiting Whole Genome Sequence Data for the Identification of Causal Trait Variants in Cattle*. in *10th World Congress on Genetics Applied to Livestock Production*. 2014. Asas.
25. VanRaden, P., et al., *Harmful recessive effects on fertility detected by absence of homozygous haplotypes.* Journal of dairy science, 2011. **94**(12): p. 6153-6161.
26. Kadri, N.K., et al., *A 660-kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in nordic red cattle: additional evidence for the common occurrence of balancing selection in livestock.* PLoS genetics, 2014. **10**(1): p. e1004049.
27. Wakasugi, N., *A genetically determined incompatibility system between spermatozoa and eggs leading to embryonic death in mice.* Journal of reproduction and fertility, 1974. **41**(1): p. 85-96.
28. Raadsma, H.W., *Quantitative trait loci mapping in dairy cattle: review and meta-analysis.* Genet. Sel. Evol, 2004. **36**: p. 163-190.
29. Sahana, G., et al., *Genome-wide association mapping for female fertility traits in Danish and Swedish Holstein cattle.* Animal Genetics, 2010. **41**(6): p. 579-588.
30. Huang, W., et al., *A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle.* Animal genetics, 2010. **41**(6): p. 570-578.
31. Peñagaricano, F., K. Weigel, and H. Khatib, *Genome-wide association study identifies candidate markers for bull fertility in Holstein dairy cattle.* Animal genetics, 2012. **43**(s1): p. 65-71.
32. Meuwissen, T. and M. Goddard, *Accurate prediction of genetic values for complex traits by whole-genome resequencing.* Genetics, 2010. **185**(2): p. 623-631.
33. Hickey, J., *Sequencing millions of animals for genomic selection 2.0.* Journal of Animal Breeding and Genetics, 2013. **130**(5): p. 331-332.

34. MacLeod, I.M., B.J. Hayes, and M.E. Goddard, *The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data.* Genetics, 2014. **198**(4): p. 1671-1684.
35. Clark, S.A., J.M. Hickey, and J.H. Van der Werf, *Different models of genetic variation and their effect on genomic evaluation.* Genet Sel Evol, 2011. **43**: p. 18.
36. Druet, T., I. Macleod, and B. Hayes, *Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions.* Heredity, 2014. **112**(1): p. 39-47.
37. Maddox, J.F., et al., *An enhanced linkage map of the sheep genome comprising more than 1000 loci.* Genome research, 2001. **11**(7): p. 1275-1289.
38. Tan, C., et al., *Integration of haplotype analysis of functional genomic information with single SNP analysis improved accuracy of genomic prediction. ADSA/ASAS 2015, Orlando, July 12-16 2015. Abstract accepted.*
39. VanRaden, P.M., et al., *Genomic evaluations with many more genotypes.* Genet. Sel. Evol, 2011. **43**(10).
40. VanRaden, P.M. and C. Sun, *Fast Imputation Using Medium- or Low-Coverage Sequence Data.* Proceedings, 10th World Congress of Genetics Applied to Livestock Production, 2014.
41. Sabeti, P.C., et al., *Detecting recent positive selection in the human genome from haplotype structure.* Nature, 2002. **419**(6909): p. 832-837.
42. Rubin, C.-J., et al., *Whole-genome resequencing reveals loci under selection during chicken domestication.* Nature, 2010. **464**(7288): p. 587-591.
43. Consortium, B.H., *Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds.* Science, 2009. **324**(5926): p. 528-532.
44. Cutter, A.D. and B.A. Payseur, *Genomic signatures of selection at linked sites: unifying the disparity among species.* Nature Reviews Genetics, 2013. **14**(4): p. 262-274.
45. Larkin, D.M., et al., *Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle.* Proceedings of the National Academy of Sciences, 2012. **109**(20): p. 7693-7698.
46. Worley, K., et al., *Detecting the signature of selection on immune genes in highly structured populations of wild sheep (Ovis dalli).* Molecular Ecology, 2006. **15**(3): p. 623-637.
47. Sabeti, P.C., et al., *Genome-wide detection and characterization of positive selection in human populations.* Nature, 2007. **449**(7164): p. 913-918.
48. Hayes, B., et al., *The origin of selection signatures on bovine chromosome 6.* Animal genetics, 2008. **39**(2): p. 105-111.
49. MacEachern, S., et al., *An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle.* BMC genomics, 2009. **10**(1): p. 181.

50. DAYO, G.K., et al., *Detection of selection signatures within candidate regions underlying trypanotolerance in outbred cattle populations.* Molecular ecology, 2009. **18**(8): p. 1801-1813.
51. Prasad, A., et al., *Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle.* Animal genetics, 2008. **39**(6): p. 597-605.
52. Fontanesi, L., et al., *Genetic heterogeneity and selection signature at the KIT gene in pigs showing different coat colours and patterns.* Animal genetics, 2010. **41**(5): p. 478-492.
53. Qanbari, S., et al., *A genome-wide scan for signatures of recent selection in Holstein cattle.* Animal genetics, 2010. **41**(4): p. 377-389.
54. Stella, A., et al., *Identification of selection signatures in cattle breeds selected for dairy production.* Genetics, 2010. **185**(4): p. 1451-1461.
55. Qanbari, S., et al., *Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle.* BMC genomics, 2011. **12**(1): p. 318.
56. Zhang, H., et al., *Selection signature analysis implicates the PC1/PCSK1 region for chicken abdominal fat content.* PloS one, 2012. **7**(7): p. e40736.
57. Schwarzenbacher, H., et al., *Combining evidence of selection with association analysis increases power to detect regions influencing complex traits in dairy cattle.* BMC genomics, 2012. **13**(1): p. 48.
58. Sonstegard, T., et al., *Forty years of artificial selection in US Holstein cattle had genome-wide signatures.* Poster presentation at 9th World Congr Genet Appl Livest Prod, Leipzig, Germany, 2010.
59. Kim, E.-S., et al., *Effect of Artificial Selection on Runs of Homozygosity in US Holstein Cattle.* PloS one, 2013. **8**(11): p. e80813.
60. Ma, L., et al., *Selection Signature Analysis in Holstein Cattle Identified Genes Known to Affect Reproduction.* ADSA-ASAS Joint Annual Meeting, Orlando, Florida, 2015.
61. Randhawa, I.A., et al., *Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep.* BMC genetics, 2014. **15**(1): p. 34.
62. Grossman, S.R., et al., *A composite of multiple signals distinguishes causal variants in regions of positive selection.* Science, 2010. **327**(5967): p. 883-886.
63. Matzuk, M.M. and D.J. Lamb, *Genetic dissection of mammalian fertility pathways.* Translocations, 2002. **45**: p. 46XY.
64. Cole, J.B., et al., *Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows.* BMC genomics, 2011. **12**(1): p. 408.
65. Cole, J., et al., *Distribution and location of genetic effects for dairy traits.* Journal of dairy science, 2009. **92**(6): p. 2931-2946.

66. Sonstegard, T.S., et al., *Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle.* PloS one, 2013. **8**(1): p. e54872.
67. McClure, M.C., et al., *Bovine Exome Sequence Analysis and Targeted SNP Genotyping of Recessive Fertility Defects BH1, HH2, and HH3 Reveal a Putative Causative Mutation in SMC2 for HH3.* PloS one, 2014. **9**(3): p. e92769.
68. Peloso, G.M., et al., *Association of Low-Frequency and Rare Coding-Sequence Variants with Blood Lipids and Coronary Heart Disease in 56,000 Whites and Blacks.* The American Journal of Human Genetics, 2014. **94**(2): p. 223-232.
69. Teslovich, T.M., et al., *Re-sequencing Expands Our Understanding of the Phenotypic Impact of Variants at GWAS Loci.* PLoS Genetics, 2014. **10**(1): p. e1004147.
70. Holmen, O.L., et al., *Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk.* Nature genetics, 2014. **46**(4): p. 345-351.
71. Hayes, B., et al., *< i> Invited review</i>: Genomic selection in dairy cattle: Progress and challenges.* Journal of dairy science, 2009. **92**(2): p. 433-443.
72. Meuwissen, T., B. Hayes, and M. Goddard, *Prediction of total genetic value using genome-wide dense marker maps.* Genetics, 2001. **157**(4): p. 1819-1829.
73. Wiggans, G., I. Misztal, and L. Van Vleck, *Implementation of an animal model for genetic evaluation of dairy cattle in the United States.* Journal of Dairy Science, 1988. **71**: p. 54-69.
74. VanRaden, P., *Invited review: Selection on net merit to improve lifetime profit.* Journal of dairy science, 2004. **87**(10): p. 3125-3131.
75. Su, G., et al., *Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers.* PloS One, 2012. **7**(9): p. e45293.
76. Weber, W., et al., *Effects of genetic selection for milk yield on somatotropin, insulin-like growth factor-I, and placental lactogen in Holstein cows.* Journal of dairy science, 2007. **90**(7): p. 3314-3325.
77. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* The American Journal of Human Genetics, 2007. **81**(3): p. 559-575.
78. *History of USDA dairy evaluations.* *http://www.aipl.arsusda.gov/aipl/history/hist_eval.htm.*
79. Sims, J.E. and D.E. Smith, *The IL-1 family: regulators of immunity.* Nature Reviews Immunology, 2010. **10**(2): p. 89-102.
80. Gabrilovich, D.I. and S. Nagaraj, *Myeloid-derived suppressor cells as regulators of the immune system.* Nature Reviews Immunology, 2009. **9**(3): p. 162-174.
81. Jones, S.A., *Directing transition from innate to acquired immunity: defining a role for IL-6.* The Journal of Immunology, 2005. **175**(6): p. 3463-3468.

82. Betz, U.A., et al., *Postnatally induced inactivation of gp130 in mice results in neurological, cardiac, hematopoietic, immunological, hepatic, and pulmonary defects.* The Journal of experimental medicine, 1998. **188**(10): p. 1955-1965.
83. John, S.Y., M. Du, and A.J. Zajac, *A vital role for interleukin-21 in the control of a chronic viral infection.* Science, 2009. **324**(5934): p. 1572-1576.
84. Endam, L.M., et al., *Polymorphisms in the interleukin-22 receptor alpha-1 gene are associated with severe chronic rhinosinusitis.* Otolaryngology-Head and Neck Surgery, 2009. **140**(5): p. 741-747.
85. Oboki, K., et al., *IL-33 is a crucial amplifier of innate rather than acquired immunity.* Proceedings of the National Academy of Sciences, 2010. **107**(43): p. 18581-18586.
86. Rubinstein, E., et al., *CD9, CD63, CD81, and CD82 are components of a surface tetraspan network connected to HLA-DR and VLA integrins.* European journal of immunology, 1996. **26**(11): p. 2657-2665.
87. Tedder, T.F., M. Inaoki, and S. Sato, *The CD19–CD21 complex regulates signal transduction thresholds governing humoral immunity and autoimmunity.* Immunity, 1997. **6**(2): p. 107-118.
88. Hendriks, J., et al., *CD27 is required for generation and long-term maintenance of T cell immunity.* Nature immunology, 2000. **1**(5): p. 433-440.
89. Lenschow, D.J., T.L. Walunas, and J.A. Bluestone, *CD28/B7 system of T cell costimulation.* Annual review of immunology, 1996. **14**(1): p. 233-258.
90. Kato, K., et al., *CD48 is a counter-receptor for mouse CD2 and is involved in T cell activation.* The Journal of experimental medicine, 1992. **176**(5): p. 1241-1249.
91. Lanier, L.L., et al., *CD80 (B7) and CD86 (B70) provide similar costimulatory signals for T cell proliferation, cytokine production, and generation of CTL.* The Journal of Immunology, 1995. **154**(1): p. 97-105.
92. Cannons, J.L., et al., *Optimal germinal center responses require a multistage T cell: B cell adhesion process involving integrins, SLAM-associated protein, and CD84.* Immunity, 2010. **32**(2): p. 253-265.
93. Kim, J., et al., *Altered expression of signalling lymphocyte activation molecule (SLAM) family receptors CS1 (CD319) and 2B4 (CD244) in patients with systemic lupus erythematosus.* Clinical & Experimental Immunology, 2010. **160**(3): p. 348-358.
94. Wakabayashi, Y., et al., *Bcl11b is required for differentiation and survival of αβ T lymphocytes.* Nature immunology, 2003. **4**(6): p. 533-539.
95. Cory, S. and J.M. Adams, *The Bcl2 family: regulators of the cellular life-or-death switch.* Nature Reviews Cancer, 2002. **2**(9): p. 647-656.
96. Nenci, A., et al., *Epithelial NEMO links innate immunity to chronic intestinal inflammation.* Nature, 2007. **446**(7135): p. 557-561.

97. Chiba, T., et al., *NFKBIL1 confers resistance to experimental autoimmune arthritis through the regulation of dendritic cell functions.* Scandinavian journal of immunology, 2011. **73**(5): p. 478-485.
98. Milner, C.M. and A.J. Day, *TSG-6: a multifunctional protein associated with inflammation.* Journal of cell science, 2003. **116**(10): p. 1863-1873.