EUROPEAN MASTER IN ANIMAL BREEDING AND GENETICS

Estimating the load, allele frequency, and linkage disequilibrium of functional and possibly deleterious variants in different cattle breeds

Stella Aivazidou

January 2024

Main supervisor: Martin Johnsson (SLU)

Co-supervisor(s): Aniek Bouwman (WUR)

Co-funded by the Erasmus+ Programme of the European Union

Estimating the load, allele frequency, and linkage disequilibrium of functional and possibly deleterious variants in different cattle breeds

Stella Aivazidou

Independent project in Animal Science, EX0870 • 30 ECTS
Swedish University of Agricultural Sciences, SLU
Faculty of Veterinary Medicine and Animal Science/ Department of Animal Breeding and Genetics
European Master in Animal Breeding and Genetics
Uppsala, 2024

Estimating the load, allele frequency, and linkage disequilibrium of functional and possibly deleterious variants in different cattle breeds.

Stella Aivazidou

Supervisor:	**Martin Johnsson, Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics**
Assistant supervisor:	Aniek Bouwman, Wageningen University and Research, Department of Animal Breeding and Genomics
Examiner:	Gabriella Lindgren, Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics

Credits:	30 credits
Level:	A2E
Course title:	Independent Project in Animal Science
Course code:	EX0870
Programme/education:	European Master in Animal Breeding and Genetics
Course coordinating dept:	Department of Animal Breeding and Genetics
Place of publication:	Uppsala, Sweden
Year of publication:	2024

Keywords:	deleterious variants, selection, allele frequency, linkage disequilibrium, mutation load

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Abstract

The domestication process and artificial selection can lead to an increased proportion and frequency of deleterious genetic variants affecting health and productivity in cattle. The aim of this project was to estimate the load and allele frequency spectrum of different kinds of functional and possibly deleterious variants (missense, potentially loss-of-function, potentially gene-regulatory) in 3 different cattle breeds (Brahman, Hereford, and Holstein), predict with bioinformatic tools the effect of these variants on proteins, and investigate the pairwise linkage disequilibrium between deleterious variants, and between deleterious variants with their surrounding area.

For the analysis the 1000 Bull Genome dataset was used. Deleterious missense and loss-of-function variants were identified using MutPred2 and MutPredLOF, respectively, and Transcription Start Site and enhancer variants were identified using Cap Analysis Gene Expression sequence data. PLINK was utilized for the allele frequency and linkage disequilibrium computations. The performance of MutPred2 was evaluated by performing MutPred2 analysis on deleterious variants present in the Online Mendelian Inheritance in Animals (OMIA) database.

The results showed low deleterious mutation load and an enrichment of deleterious variants at low frequencies, indicating the effectiveness of purifying selection at purging them from the population. Moreover, balancing selection may be a potential mechanism for the higher frequencies observed for a small amount of deleterious variants. The observed breed-specific differences in load and allele frequencies may be attributed to differences in effective population size, selection pressure, and breeding strategies. Despite lacking proper annotation, the enrichment of regulatory variants at lower frequencies suggests that they are under the influence of selection. The observed low linkage disequilibrium between and around highly deleterious variants may be attributed to their low frequencies and their presence on different haplotypes due to recombination. Evaluation of MutPred2 revealed that the inclusion of Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) in the analysis is necessary for the reliable identification of deleterious variants.

This study, if implemented accurately on a large scale, has the potential to facilitate the development of a genomic database that can contribute to reducing the frequencies of deleterious variant through genomic selection and improving predictions on causative variants linked to genetic defects, ultimately enhancing the effective management of genetic diseases.

Keywords: deleterious variants, selection, allele frequency, linkage disequilibrium, mutation load

Table of contents

List of tables

List of figures

Abbreviations

LD	Linkage Disequilibrium
ChIP-seq	Chromatin Immunoprecipitation followed by sequencing
ATAC-seq	Transposase Accessible Chromatin sequencing
CAGE	Cap Analysis Gene Expression
TSS	Transcription Start Sites
PTM	Posttranslational Modification
MutPred2	Name of software
CADD	Combined Annotation Dependent Depletion
FATHMM	Functional Analysis through Hidden Markov Models
GERP	Genomic Evolutionary Rate Profiling
PolyPhen-2	Polymorphism Phenotyping v2
SIFT	Sorting Intolerant from Tolerant
SNPs&GO	Single Nucleotide Polymorphisms & Gene Ontology
AUC	Area Under the Curve
NutVar	Null and truncating Variant analysis
MutPredLOF	Name of software
LOF	Loss-of-function
Indels	Insertions and deletions
HGMD	Human Gene Mutation Database
VCF	Variant Call Format
VEP	Variants Effect Predictor
SNP	Single Nucleotide Polymorphism
FPR	False Positive Rate

MAF	Minor Allele Frequency
OMIA	Online Mendelian Inheritance in Animals
QTL	Quantitative Trait Locus
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool

Introduction

Each genome contains mutations that may have an impact on fitness and health (Bosse et al., 2019). All genetic variants that lower the fitness of an organism are referred to as deleterious. When these variants are highly deleterious or lethal, they are quickly eliminated from the population through the process of natural selection (Kimura, 1983), whereas the mildly deleterious variants tend to remain at low frequency within a population, primarily in the heterozygous state (Mukai et al., 1972; Zhang et al., 2016)

The domestication of wild animals creates a population bottleneck because only a small subset of the original wild population is chosen to form the initial breeding stock (Lu et al., 2006). Additionally, artificial selection for specific traits further reduces the effective population size during the process of breed formation (Frantz et al., 2020a) and increases the inbreeding (Lush, 1946). Inbreeding, the inheritance of identical copies of genetic material from closely related parents, is more prevalent in small populations, and can negatively impact health and reproduction (Lynch et al., 1995). The reduced fitness observed in inbred offspring, compared to outbred offspring, is called "inbreeding depression" (Keller & Waller, 2002) and it is associated with the accumulation of recessive harmful mutations in the genome and the higher probability of these mutations to become homozygous (Agrawal & Whitlock, 2012; Charlesworth & Willis, 2009a).

Consequently, higher levels of homozygosity, lead to a decrease in the effective recombination rate (Moyers et al., 2018), and to the accumulation of deleterious variants that are in linkage disequilibrium (LD) with loci that undergo strong, positive, artificial selection (Hartfield & Otto, 2011a). Linkage disequilibrium (LD) refers to the nonrandom association of alleles at two or more loci. Genetic

hitchhiking is the process, in which genetic variants increase in frequency due to LD with variants under positive selection, as long as their impact on fitness is lower than the strength of selection acting on the targeted variants (Hartfield & Otto, 2011b). This makes selection less effective at removing mildly deleterious variants, and it is more likely that new slightly beneficial mutations will be lost due to genetic drift (Lande, 1994; Lynch & Gabriel, 1990; Whitlock et al., 2003). As a result, the effectiveness of selection can be limited and, consequently, the genetic gain achieved in breeding programs may be reduced (Moyers et al., 2018).

The domestication process can lead to an increase in the number, frequency, or proportion of deleterious genetic variants that are either fixed or segregating in the genomes of domesticated species (Moyers et al., 2018) Various methods are used to assess the proportion of potentially deleterious variants present in domesticated species' populations, including counting the absolute number of variants at derived sites, calculating the ratio of deleterious to synonymous variants, and observing an increase in the frequency of potentially deleterious variants within a population (Lohmueller, 2014; Moyers et al., 2018). In cattle, breeds are categorized under the subspecies *Bos taurus taurus* and/or *Bos taurus indicus* (Pitt et al., 2019). As the founder population sizes differ among breeds, their deleterious mutation load can also vary (Elsik et al., 2009). In addition, the variation between breeds in the intensity and practices of artificial selection and in the rate of inbreeding can also contribute to the variation in deleterious mutation load among cattle breeds (Frantz et al., 2020a).

Efficient detection and handling of genetic defects could be accomplished with widespread access to genome sequence data from a substantial number of cattle that have been phenotyped for specific traits (Daetwyler et al., 2014). The 1000 Bull Genomes is a collection of complete cattle genome sequences that can account for a significant proportion of global cattle diversity (Hayes & Daetwyler, 2019). This dataset can facilitate the detection of harmful mutations that can affect heath, welfare and productivity of animals and the identification of relationships between specific variants and traits in cattle populations. (Daetwyler et al., 2014).

While significant advancements have been made in annotating protein-coding genes in livestock species, the majority of these genomes consist of noncoding regions that are not well annotated (Halstead et al., 2020). Precise identification and annotation of the gene regulatory elements are crucial for gaining insights into the possible mechanisms that control gene expression (Salavati et al., 2023) and it will enable the identification of causal variants for disease (Halstead et al., 2020) and production traits (Alexandre et al., 2021). Epigenomic methods, such as Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) and the Assay of Transposase Accessible Chromatin sequencing (ATAC-seq), have been utilized for the characterization of functional elements in model organisms and livestock species (Alexandre et al., 2021; Halstead et al., 2020; Shen et al., 2012) . Cap Analysis Gene Expression (CAGE) (Takahashi et al., 2012) has been used successfully for the annotation of Transcription Start Sites (TSS) and enhancers in cattle (Salavati et al., 2023).

1.1 Prediction of possibly deleterious variants

The identification of deleterious variants and their associated functional alterations is challenging. Variants can have a wide range of functional impact, causing a variety of molecular changes even within a single protein. However, most of the existing methods do not offer sufficient information about the potential mechanisms impacted by mutations and they cannot model the type of alteration in protein structure and function (Pejaver et al., 2020). MutPred2 is a machine-learning based software that utilizes both genetic and molecular data to assess the deleteriousness of amino acid substitutions (Pejaver et al., 2020) and it is developed based on a data set that includes Mendelian disease variants present in human. It provides a general pathogenicity score ranging from 0 to 1, and a ranked list of molecular alterations potentially affecting the phenotype. Currently, MutPred2 considers over 50 structural and functional properties, including structure, signal peptide and transmembrane topology, catalytic activity, metal, and macromolecular binding,

PTMs, and allostery. In human, comparative analysis between MutPred2 and other methods such as CADD (Kircher et al., 2014), FATHMM (Shihab et al., 2013), GERP++ (Davydov et al., 2010), MutationTaster2 (Schwarz et al., 2014), MutPred, PhyloP (Pollard et al., 2010), PolyPhen-2 (Adzhubei et al., 2010), SIFT (Ng & Henikoff, 2001), and SNPs&GO (Calabrese et al., 2009) revealed that MutPred2 was the best performing method in terms of AUC and its high sensitivities at lower false positive rates (Pejaver et al., 2020).

Loss-of-function (LOF) variants are variants that disrupt the normal function of a protein. Frameshifting and stop variants are examples of such variants. Frameshifting variants involve insertions and deletions of nucleotides (indels) that are not divisible by three, leading to a shift in the mRNA coding frame. Stop variants involve the gain or loss of stop codons in mRNA. Unlike missense (Cline & Karchin, 2011), LOF variants have not been as extensively studied. Previous methods, such as SIFT Indel (Hu & Ng, 2012) and NutVar (Rausell et al., 2014) encounter challenges in effectively differentiating between various types of loss-of-function variants within the same protein. Moreover, these methods limit their training data to proteins with high quality annotations, limiting their applicability to genes that have not been extensively studied. Therefore, similarly to MutPred2, MutPredLOF was created (Pagel et al., 2017), based on data on pathogenic (disease-causing) stop gain and frameshifting variants from the Human Gene Mutation Database (HGMD), for the identification of potentially LOF variants proposing specific molecular changes. Each mutation input into MutPredLOF yields a score ranging from zero to one, where higher scores indicate a greater likelihood of pathogenicity. Additionally, MutPredLOF provides information about up to five structural and functional mechanisms affected in the implicated region of the protein, supported by significant prior corrected P-values (below 0.05).

1.2 Project Aim

The aim of this project was to estimate the load and allele frequency spectrum of functional and possibly deleterious variants of different kinds (missense, potentially loss-of-function, potentially gene-regulatory) in different cattle breeds, predict with bioinformatic methods the effect of these variants on proteins, and investigate the pairwise LD between deleterious variants, and between deleterious variants with their surrounding area.

Materials and Methods

The dataset used in this study was the publicly available version of the 1000 Bull Genome dataset (Run 9). The data preparation steps, and the software used are explained in detail below.

2.1 Data

Variant Call Format (VCF) files for 1,039 animals in total were obtained from the 1000 Bull Genome dataset (Hayes & Daetwyler, 2019). These animals belonged to 72 breeds in total, including 92 animals of unknown breed. The number of animals in each breed ranged from 1 to 146. There were breeds that belonged to *Bos taurus taurus* and breeds that belonged to *Bos taurus indicus*. Two hundred six animals were females, 349 were males and 484 were of unknown sex, due to insufficient information in the metadata available in the European Nucleotide Archive. The reference genome used in the analysis was ARS-UCD1.2. Genetic variants on sex chromosomes were discarded. To analyze and compare the allele frequency spectrum and LD in different cattle breeds, I decided to incorporate breeds that represented different selection pressures and strategies. Therefore, both *Bos taurus taurus* and *Bos taurus indicus* cattle were included, while both dairy and beef breeds were considered for the Bos taurus taurus cattle. Consequently, three breeds were selected: Hereford (beef breed), Holstein (dairy breed), and Brahman (*Bos taurus indicus*), representing the breeds with the highest number of individuals in each category in order to obtain the best estimates. In total the dataset included 23 Hereford, 146 Holstein and 37 Brahman cattle.

2.2 Functional annotation and identification of deleterious variants

Ensembl Variant Effect Predictor (VEP, (McLaren et al., 2016)) was used to obtain the functional consequences of the variants. Variants within a gene with a missense annotation were retained in order to predict their deleterious effect using MutPred2 (Pejaver et al., 2020). The necessary input files for MutPred2 were prepared, formatted in the standard FASTA format with the substitutions specified in each sequence's header along with the sequence ID, followed by the protein sequence. Every protein sequence was of length >30 and <30,000 residues, in order to be analyzed by MutPred2, therefore proteins of length <30 residues were removed. In addition, to accommodate reading limitations of MutPred2, the amino acid selenocysteine (U) was replaced by cysteine (C) in all protein sequences. Missense indels were excluded from the analysis, focusing solely on missense single nucleotide polymorphisms (SNPs). To reduce computation time only missense variants that belonged to Ensembl canonical transcripts were retained for the proteins with multiple transcripts. Canonical transcripts are the ones that are the most conserved, most highly expressed, have the longest coding sequence and are represented in other resources, such as NCBI and UniProt. In total 436,944 SNPs were assessed by MutPred2 (Table 1). MutPred2 provides information about possible structural changes due to amino-acid alterations, along with the posterior probabilities of the loss or gain of certain structural and functional properties due to the substitution (*Pr*) and empirical p-values (P) calculated as the fraction of benign substitutions in MutPred2's training set with *Pr* values >= to the *Pr* value for the given substitution. It also provides a pathogenicity score between zero and one that indicates the probability that the amino acid substitution is pathogenic. A score threshold of 0.50 would suggest pathogenicity. However, a threshold of 0.68 yields a false positive rate (FPR) of 10% and that of 0.80 yields an FPR of 5%. Therefore, in this study, substitutions with a score higher than 0.80 were considered highly deleterious and substitutions with a score higher than 0.50 and lower than 0.80 were considered as moderate.

Similarly to missense, variants within a gene that were labelled by VEP as frameshift and stop gain were retained in order to predict their deleterious effect using MutPredLOF (Pagel et al., 2017). The input files were prepared in a modified FASTA format with each variant represented by a pair of two ordered sequences: the unmodified wildtype protein sequence and then the mutant protein sequence. To generate the mutant protein sequence, the Biostrings R package (Pagès et al., 2019) was employed to create the mutated coding sequence and subsequently translate it into the mutant protein. The header did not need to conform to any particular format in MutPredLOF. Every protein sequence that had a length <30 and >30,000 residues was removed. To reduce computation time only frameshift and stop gain variants on Ensembl canonical transcripts were retained for the proteins that had multiple transcripts. In total 43,001 number of variants were assessed by MutPredLOF. MutPred-LOF assigns a score ranging from zero to one for each mutation input, with higher scores indicating a higher likelihood of pathogenicity. Furthermore, the model reports up to five structural and functional mechanisms affected in the region of the protein influenced by the mutation. To facilitate classification, three score thresholds are offered to help distinguish between pathogenic and neutral variants, considering different levels of FPR: 0.40 (10% FPR), 0.50 (5% FPR, recommended), 0.70 (1% FPR). In this study, mutations with a score >= 0.50 were classified as highly deleterious, while mutations with a score between 0.40 and 0.50 were classified as mildly deleterious. This approach aimed to maintain consistency between the classification of missense and loss-of-function variants.

2.3 Identification of variants in gene regulatory regions

To identify genetic variants within promoter and enhancer regions, we utilized predictions of transcription start site (TSS) and TSS-enhancer sets derived from on CAGE sequence data that were available in the study of (Salavati et al., 2023). In their study, CAGE sequencing was employed to identify TSS across a set of 24

tissues from 3 different cattle populations: one dairy (Belgian Holstein Friesian), one beef– dairy cross (German Charolais × Holstein F2), and the Canadian Kinsella composite cattle (beef). They considered a putative TSS or TSS-Enhancer region, as valid only when it was present across at least two-thirds of the tissues (Salavati et al., 2020) resulting in the detection of 51,295 TSS and 2,328 TSS-Enhancers. All the coordinates from the TSS and TSS-Enhancers were extracted, and subsequently, all genetic variants located within these specified genomic regions were identified by cross-referencing with the original VCF files in our dataset. This process involved determining which positions of the variants in our VCF files coincided with the coordinates of TSS and enhancers. The total number of TSS and TSS-enhancer variants are shown in Table 1.

2.4 Allele frequency spectrum for different classes of variants

For the allele frequency analysis, minor allele frequencies (MAF) were computed separately for each breed (Brahman, Holstein, and Hereford) using PLINK (Purcell et al., 2007). Initially, the analysis focused on determining the distribution of allele frequencies for different types of variants, including synonymous, missense, frameshift, and stop-gain. For each breed, a MAF histogram was generated, and a different color was assigned to each variant type. In addition, to investigate whether predicted deleterious mutations showed generally lower allele frequencies or if there were deleterious variants enriched for high frequencies in specific breeds, the distribution of allele frequencies of variants with different MutPred2 and MutPredLOF scores were determined. For missense variants, MAF histograms were generated for each breed, with different colors representing each MutPred2 score category (low <0.50, moderate 0.50-0.80, high >0.80). Similarly, for loss of function variants, MAF histogram were created for each breed, and each MutPredLOF score category was distinguished by a different color (low <0.40, moderate 0.40-0.50, high >0.50). Lastly, the distribution of allele frequencies for

the variants within the TSS and TSS-enhancer regions was determined. MAF histograms were generated for each breed, with TSS and enhancer variant types represented by a different color. To identify any potential enrichment of rare variants in the TSS and TSS-enhancer regions, all SNPs annotated as intergenic by VEP were used as control for comparison. For the MAF histograms in the X axis I used 10 bins with a bin size of 0.05, and the Y axis represented density, indicating the number of variants in each bin divided by the total number of variants in that specific category. To ensure the comparison of allele frequencies of variants that are still segregating within the breeds, excluding those already fixed, all variants with a MAF of 0 were removed.

2.5 Mutation Load

To determine the deleterious mutation load, the count of high MutPred2 score and high MutPredLOF score variants was assessed for individuals within the Holstein, Hereford, and Brahman breeds. The process involved creating a list of individual sample IDS that belonged to the three breeds of interest and extracting their genotypes from the VCF files at each position using BCFtools (Li, 2011). Subsequently, the positions of all missense variants predicted to be highly deleterious by MutPred2 and all loss of function variants predicted to be highly deleterious by MutPredLOF were identified and the highly deleterious alleles were counted based on the genotypes. For instance, a genotype of 0/0 indicated a deleterious mutation load of 0 for that position, a genotype of 0/1 or 1/0 indicated a deleterious mutation load of 1 for that position, and a genotype of 1/1 resulted in a deleterious mutation load of 2 for that position. The counts for all positions were then summed to calculate the total deleterious mutation load for each individual.

2.6 Linkage Disequilibrium (LD)

To examine LD patterns and to assess the presence of highly deleterious variants in LD, pairwise LD analysis was conducted by calculating r^2 values between all missense variants identified as highly deleterious by MutPred2 and among all loss of function variants that were classified as highly deleterious by MutPredLOF. The LD computations were executed using PLINK with specific parameters: 1) r^2 LD window (--ld-window-r2) set to 0 to make sure that even the lowest r^2 values will be reported, 2) LD window in kilobases (kb) (--ld-window-kb) set to 1000000, and 3) an LD window (--ld-window) set to 1000000. Regarding the last 2 parameters, by default in Plink, every pair of variants with at least (10-1) variants between them, or more than 1000 kilobases apart is ignored, therefore the window was increased to include all the pairwise comparisons. Initially, pairwise LD was computed for all individuals in the dataset and subsequently, pairwise LD was computed separately for the individuals belonging to the three breeds of interest (Holstein, Hereford, and Brahman). The results were visualized through LD decay plots generated separately for the variants of each chromosome. Furthermore, an integrated LD decay plot was constructed by combining the data from all chromosomes.

In addition, to assess LD between highly deleterious variants and surrounding variants, r2 was calculated between missense variants with high MutPred2 scores and all variants in the surrounding area. Similarly, r2 was computed between loss of function variants with high MutPredLOF score and their surrounding variants. For this purpose, an LD window of 100 kb was used. Individual LD decay plots were created for each variant, depicting the LD patterns in the specific areas. Moreover, a collective LD decay plot was constructed using the average r2 values. The 100 kb distance was divided into 20 bins, each with a size of 5 kb. Within each bin, the average LD between a variant with a high MutPred score and its surrounding variants was computed and presented in the plot. For instance, for a variant with a high MutPred2 score, all variants within a distance <5 kilobases with a known r2 were identified and included in the first bin. The average LD was then computed for these variants and illustrated in the plot. This process was repeated

for subsequent bins, each representing a specific distance range (e.g. 5-10 kb, 10-15kb etc.). The same approach was applied to the surrounding variants of loss-of-function variants with a high MutPredLOF score.

2.7 Evaluation of MutPred2 using predicted deleterious variants from the OMIA database

Numerous variants predicted as deleterious have been associated with specific genetic defects. Online Mendelian Inheritance in Animals (OMIA) is a comprehensive, annotated catalogue of inherited disorders and other traits in animals, including associated genes and variants in 498 animal species (Nicholas et al., 1995). OMIA provides an extensive resource of phenotypic information on heritable animal traits and genes, establishing strong comparative connections between traits and genes.

Within OMIA, a table provides details on 272 likely causal variants in cattle, including information such as gene, breed, type of variant and position. To assess MutPred2's performance on variants previously predicted as deleterious and associated with specific genetic diseases, a similar analysis to the one performed with missense variants from the 1000 Bull Genome dataset was conducted, this time using variants from the OMIA database. Specifically, only missense variants were retained to maintain consistency with the prior MutPred2 analysis. Non-deleterious variants were excluded, along with variants on sex chromosomes and those based on a different reference genome than the one used in our analysis. A total of 80 OMIA missense variants were included in this analysis. Initially, the MutPred2 output was examined to determine whether prediction scores were assigned to the OMIA missense variants, by cross-referencing the gene and position of these variants in the MutPred2 output file. For the remaining variants that were absent from the MutPred2 output, indicating they likely were not present in the 1000 Bull Genome dataset and thus lacked a prediction score from MutPred2, an input file was generated. Subsequently, MutPred2 analysis was conducted for these variants.

Finally, the prediction scores provided by MutPred2 for all the OMIA deleterious missense variants were assessed in order to determine the effectiveness of MutPred2 in identifying deleterious variants and to evaluate the reliability of the generated scores.

Table 1. Variant Counts by Type and Score Category

Variant Type	**Number of variants**	**MutPred2 score**			**MutPredLOF score**		
		Low	**Moderate**	**High**	**Low**	**Moderate**	**High**
Missense	436,944	436,224	374	346	-	-	-
Loss of function	43,001	-	-	-	37,058	5,652	291
TSS	1,106,377	-	-		-	-	-
Enhancer	36,899	-	-		-	-	-

Results

3.1 Identification of deleterious variants

Following the analysis of MutPred2 on missense variants, among the total 436,944 missense variants, 436,224 were assigned scores below 0.50, indicating non-deleterious variants, 374 obtained scores between 0.50-0.80, suggesting mild deleteriousness, and 346 received scores above 0.8, representing highly deleterious variants (Table 1). In the case of MutPredLOF applied to potentially loss-of-function variants, out of the total 43,001 potential loss-of-function variants, 37,058 received scores below 0.40, indicating non-deleterious variants, 5,652 obtained scores between 0.40-0.50, indicating mild deleteriousness, and 291 obtained scores above 0.50, representing highly deleterious variants (Table 1). In the case of missense variants, mildly and highly deleterious variants constitute 0.08% of the total variants each. On the other hand, in loss-of-function variants, mildly and highly deleterious variants represent 13% and 0.67% respectively. This implies that there is a higher proportion of highly deleterious variants in the loss-of-function category.

3.2 Minor Allele Frequency (MAF) distributions

3.2.1 MAF distributions of missense and potential loss-of-function variants compared to synonymous

The analysis of MAF for missense and potential loss-of-function variants showed that, compared to synonymous variants, they were enriched for low frequencies. **Figure 1** shows the distribution of Minor Allele Frequency (MAF) for four variant types: synonymous, missense, frameshift, and stop-gain variants. There was a common trend across all variant types, with a higher proportion of variants observed at lower frequencies, particularly with MAF close to 0. These patterns were consistently observed across all breeds. Notably, synonymous variants showed a slightly different distribution pattern with a more even distribution across the entire allele frequency spectrum. Moreover, in Brahman, there was a unique pattern observed for missense, stop-gained, and frameshift variants, with the highest proportion of deleterious variants observed at allele frequencies close to 0.05. This was in contrast to the other two breeds, where the highest proportion was observed at MAF close to 0.

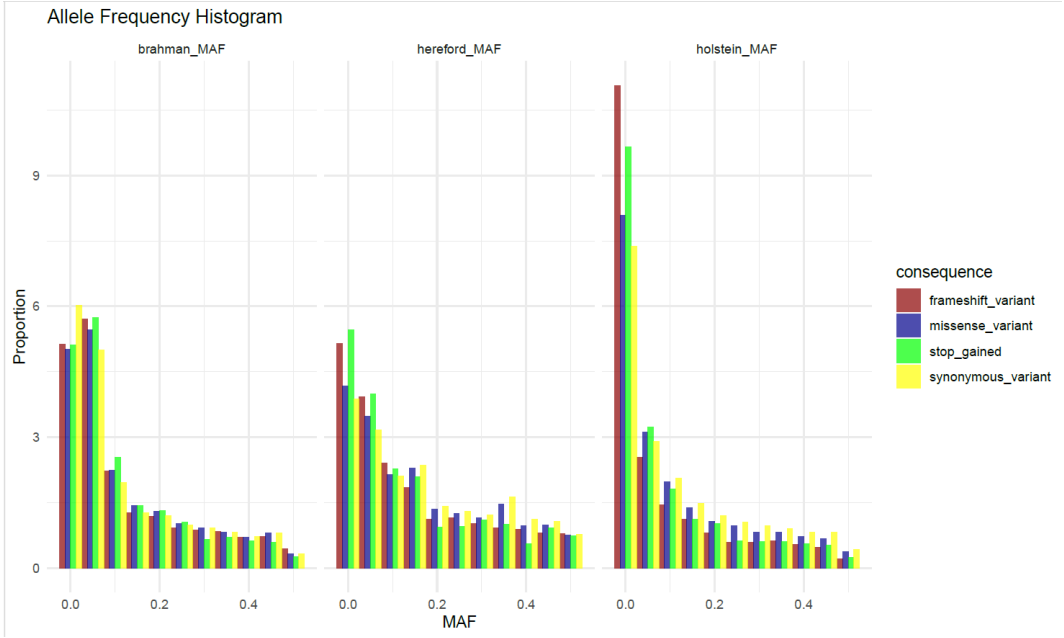


Figure 1. MAF distribution for different functional classes of variants (missense, synonymous, stop-gained and frameshift) across the 3 breeds (Brahman, Hereford, and Holstein).

3.2.2 MAF distribution of missense variants based on their MutPred2 score

The analysis of MAF for missense variants of different MutPred2 scores showed an enrichment of mildly and highly deleterious variants at lower frequencies. **Figure 2** shows the allele frequency distribution of missense variants according to their assigned MutPred2 scores. In Hereford and Brahman, there was an absence of highly deleterious variants with frequencies above 0.4. In Holstein, no mildly deleterious variants were observed at frequencies exceeding 0.3. While the allele frequency distribution followed a similar pattern across all three breeds, different variants were identified at high frequencies in different breeds. Table 2 shows the allele frequencies of the highly deleterious missense variants, with the highest allele frequencies across the 3 breeds. It was observed that certain variants were present at high frequencies only in Brahman (L94R and V26A) and at frequency of 0 in other breeds, some variants had high frequencies in Holstein and Hereford and frequency close to 0 in Brahman (R46G), and other variants had high frequencies across all breeds (Table 2).

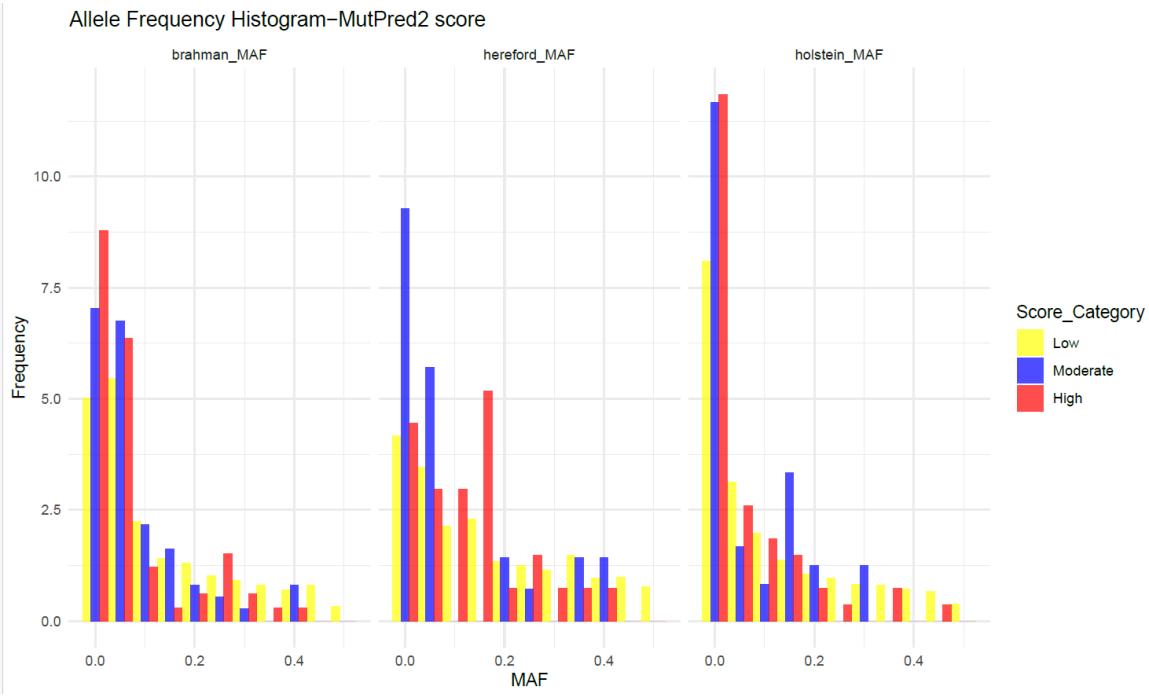


Figure 2 MAF distribution of missense variants across different classes of MutPred2 scores for the 3 breeds (Brahman, Hereford, and Holstein).

Table 2 Highest allele frequencies of highly deleterious missense variants across the 3 breeds.

Ensembl Transcript ID	Gene Name	Amino-acid Substitution	MutPred2 score	Hereford MAF	Holstein MAF	Brahman MAF
ENSBTAT00000051986	*RPS15A*	L94R	0.977	0	0	0.3143
ENSBTAT00000045443	*DYNLL1*	A11D	0.953	0.1739	0.4692	0.2568
ENSBTAT00000045443	*DYNLL1*	M13I	0.946	0.3913	0.3733	0.2297
ENSBTAT00000087168	*RPL39*	R46G	0.936	0.3696	0.3475	0.02941
ENSBTAT00000056532	*H2AC25*	S2P	0.904	0.2391	0.2568	0.2297
ENSBTAT00000056472	*CBX1*	V26A	0.868	0	0	0.2407
ENSBTAT00000031096	*CHIC2*	N82I	0.842	0.225	0.1884	0.3889

3.2.3 MAF distribution of potentially loss-of-function variants based on their MutPredLOF score

The analysis of MAF for potentially loss-of-function variants of different MutPredLOF scores showed an enrichment of mildly and highly deleterious variants at lower frequencies. **Figure 3** shows the allele frequency distribution of loss-of-function variants according to their assigned MutPredLOF scores. In Hereford and Holstein, there was an absence of highly deleterious variants at higher frequencies, with the highest observed frequency being approximately 0.2 in Holstein and 0.15 in Hereford. In contrast, in Brahman highly deleterious variants were observed even at frequencies near 0.3 and 0.35. While the allele frequency distribution of loss-of-function variants followed a similar pattern across all three breeds, different variants were identified at high frequencies in different breeds, similarly to missense variants. Table 3 shows the allele frequencies of the highly deleterious loss-of-function variants, with the highest allele frequencies across the 3 breeds. It was observed that high frequency variants in one breed had frequencies of 0 or close to 0 in the other breeds (Table 3).

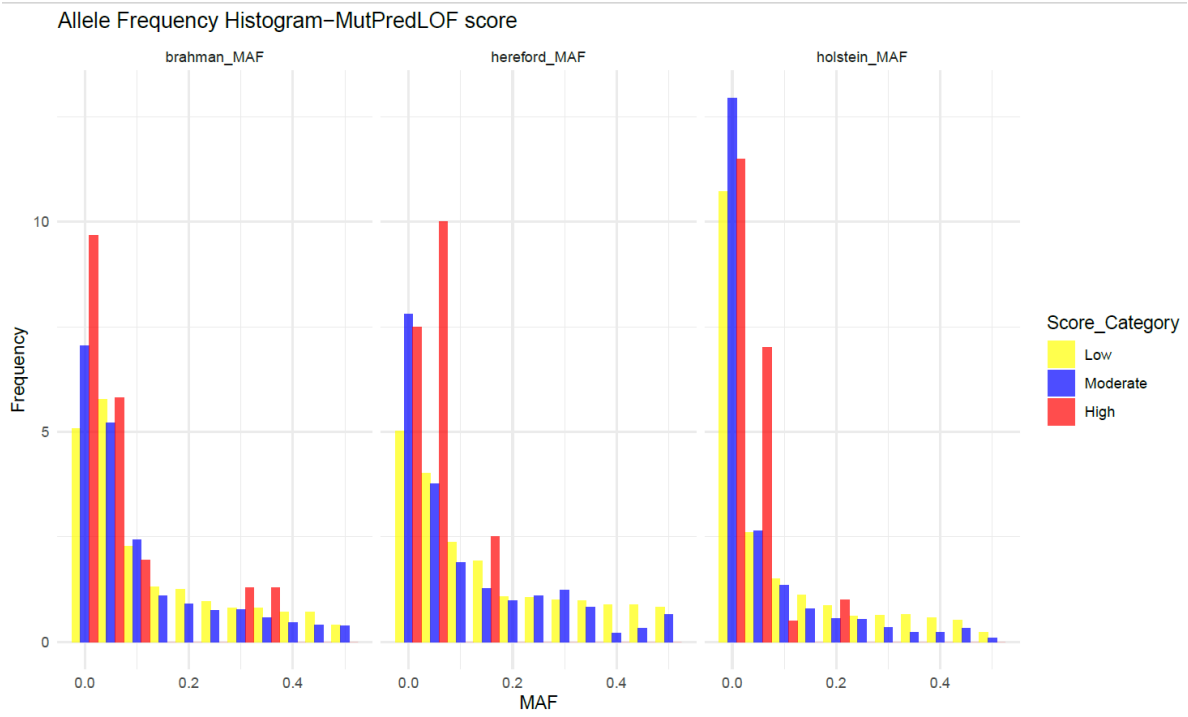


Figure 3 MAF distribution of highly deleterious loss-of-function variants across different classes of MutPredLOF scores for the 3 breeds (Brahman, Hereford, and Holstein).

Table 3 Highest allele frequencies of highly deleterious loss-of-function variants across the 3 breeds. The dashes (-) represent deletions.

Ensembl Transcript ID	**Gene Name**	**allele**	**CSD position**	**MutPredLOF score**	**Hereford MAF**	**Holstein MAF**	**Brahman MAF**
ENSBTAT00000086078	*NEB*	-	8372-8375	0.50191	0	0	0.3571
ENSBTAT00000086078	*NEB*	CAAA	8379-8380	0.50187	0	0	0.3676
ENSBTAT00000077684	*FBN1*	TC	1152-1153	0.53885	0.1739	0.02672	0.08824
ENSBTAT00000003173	*MKI67*	AA	7773-7774	0.50771	0	0	0.2778
ENSBTAT00000003173	*MKI67*	T	7771-7773	0.50815	0	0	0.2778
ENSBTAT00000083797	*MUC2*	-	6587-6644	0.51017	0.02174	0.1761	0.05172
ENSBTAT00000083797	*MUC2*	-	6647-6651	0.50784	0.04348	0.1889	0.05357

3.2.4 MAF distributions of Transcription Start Site (TSS) and enhancer variants

The analysis of MAF showed an enrichment in rare TSS and enhancer variants. **Figure 4** shows the distribution of TSS, enhancer and intergenic variants, with intergenic variants serving as the reference. The highest proportion of TSS and enhancer variants was concentrated around frequencies near 0 and 0.05 across all three breeds, decreasing as we moved to higher frequencies.

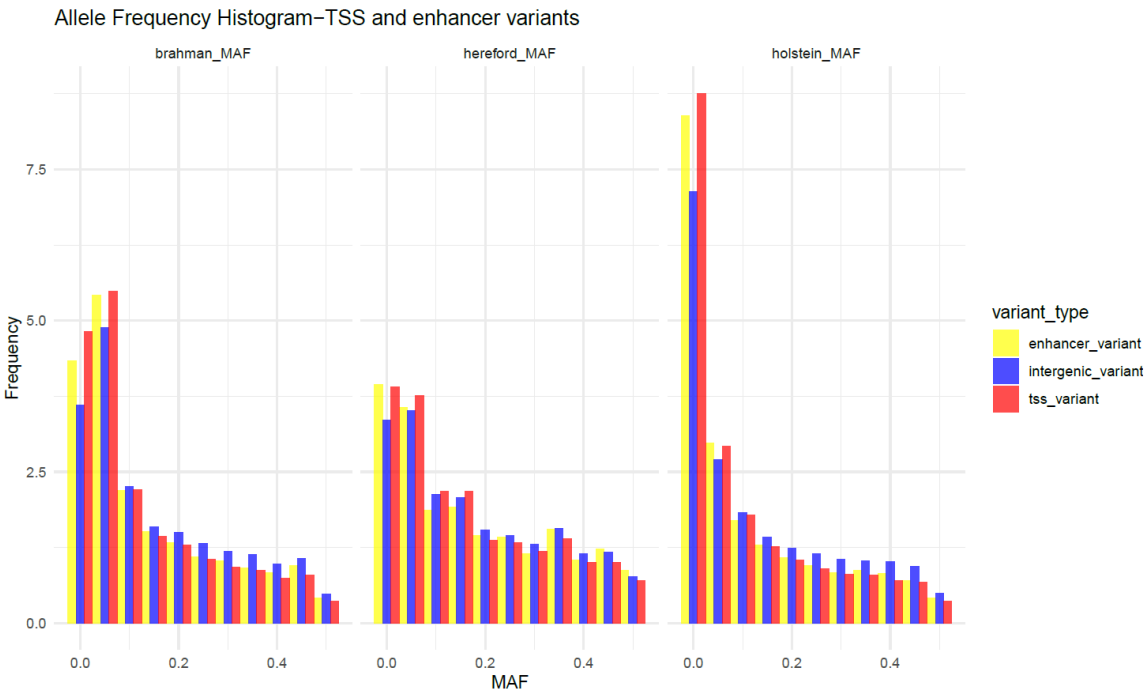


Figure 4 MAF distribution of TSS, enhancer, and intergenic variants across the 3 breeds (Brahman, Hereford, and Holstein).

3.3 Mutation load

To examine the load of highly deleterious mutations, the count of high MutPred2 score (346 in total) and high MutPredLOF score (291 in total) variants was assessed for individuals within the Holstein, Hereford, and Brahman breeds. Analysis of the mutation load at each genomic position revealed the absence of homozygous animals for these deleterious variants; all animals carrying such variants were found to be heterozygous. Subsequently, the total deleterious mutation load was computed for each individual separately for highly deleterious missense and highly deleterious loss of function variants. **Figure 5** shows the distribution of deleterious mutation loads across all animals from the three breeds of interest, indicating the count of animals for each mutation count. The results showed that for Brahman, all animals had 3-17 missense mutations and 0-7 loss-of-function mutations; Hereford animals had 3-11 missense and 0-3 loss-of-function; and for Holstein animals, all had 2-12 missense and 0-5 loss-of-function mutations, except for one individual with 15 loss of function variants. The results suggested a higher per-individual load in missense mutations compared to loss-of-function.

Regarding the missense mutation load, in Brahman 13.5% of the animals had 3-5 deleterious mutations, 73% had 6-10, and 13.5% of the animals had 11-17 deleterious mutations. In Hereford 26% of the animals had 3-5 deleterious mutations, 69.5% had 6-10, and 4.5% of the animals had 11 deleterious mutations. In Holstein, 30% of the animals had 2-5 deleterious mutations, 65% had 6-10, and 5% had 11-12 deleterious mutations.

Regarding the deleterious mutation load for loss-of-function variants in Brahman, 2 animals (approximately 5% of animals) had 0 mutations. This percentage was significantly higher for Hereford and Holstein, with 65% of Hereford and 48% of Holstein animals having 0 mutations. In Brahman, 92% of animals had 1-5 mutations, and one animal had 7 mutations. In Hereford 35% of the animals had 1-3 mutations, with 3 being the highest number of individual mutations recorded. In Holstein 51.3% of all animals had 1-5 mutations and one animal had 15 mutations.

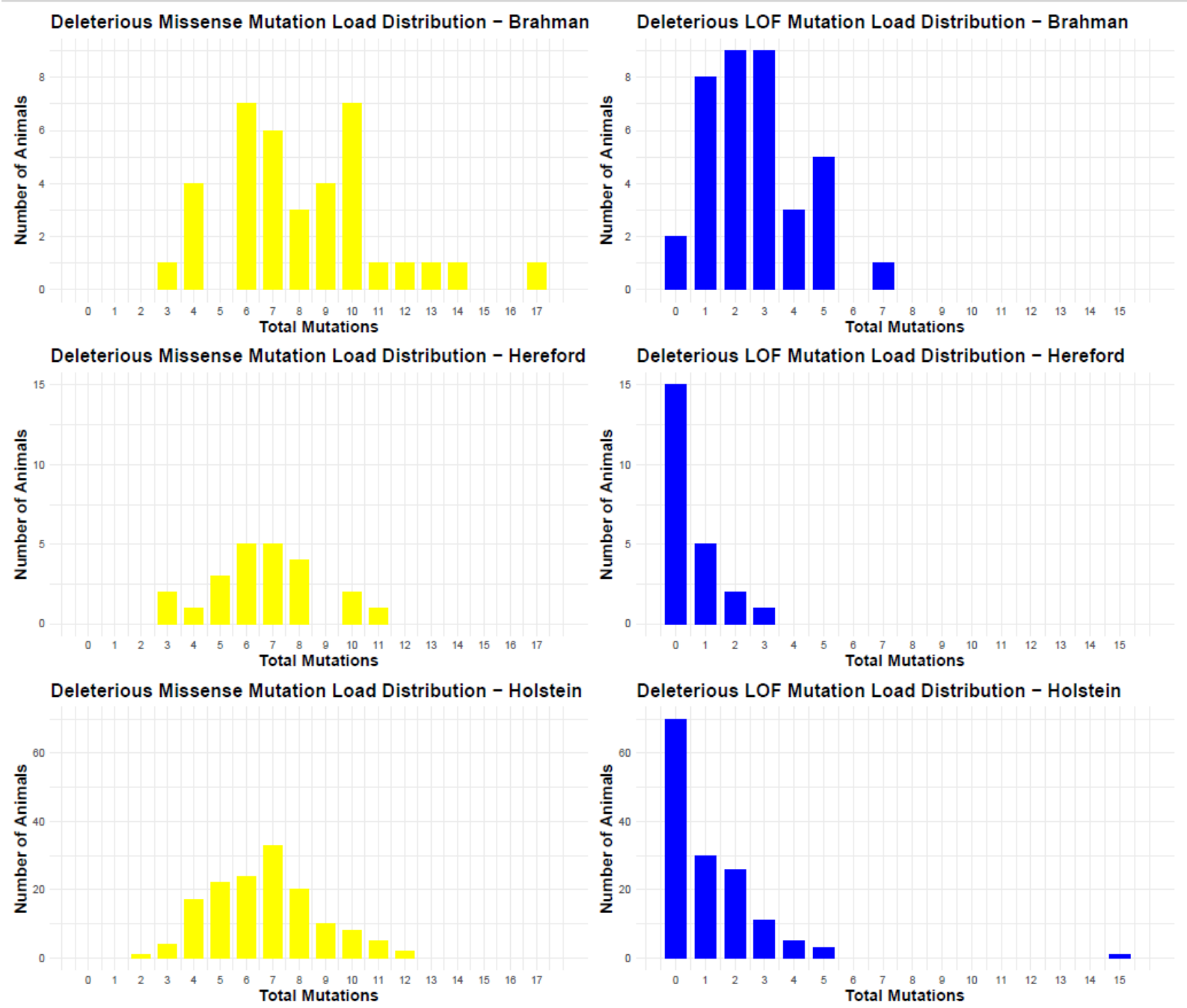


Figure 5 Distribution of highly deleterious mutation loads for missense (left side-yellow) and loss-of-function (right side-blue) for the 3 breeds of interest (Brahman, Hereford, and Holstein).

3.4 Linkage Disequilibrium (LD)

3.4.1 Pairwise LD between highly deleterious variants

To examine pairwise LD between the highly deleterious variants, r^2 was estimated and the LD decay was visualized in a plot. **Figure 6** shows the decay in pairwise LD among all highly deleterious missense variants (top) and among all highly deleterious loss of function variants (bottom). This plot was generated using the r^2 values calculated from the genotypes of the entire dataset (all 1,039 animals in the dataset). The results indicated that as the distance between variants increased, the LD decreased, and variants with high LD were mostly in proximity to each other. The decay in LD appeared to be relatively rapid, especially for the loss of function variants. Notably, some variants maintained high LD even at considerable distances. **Figure 7** and **Figure 8** illustrate the decay in pairwise LD among highly deleterious missense and loss-of-function variants focusing specifically on the three breeds of interest Hereford, Holstein, and Brahman. Similar to the overall dataset trend, LD decay was observed to be rapid for all three breeds.

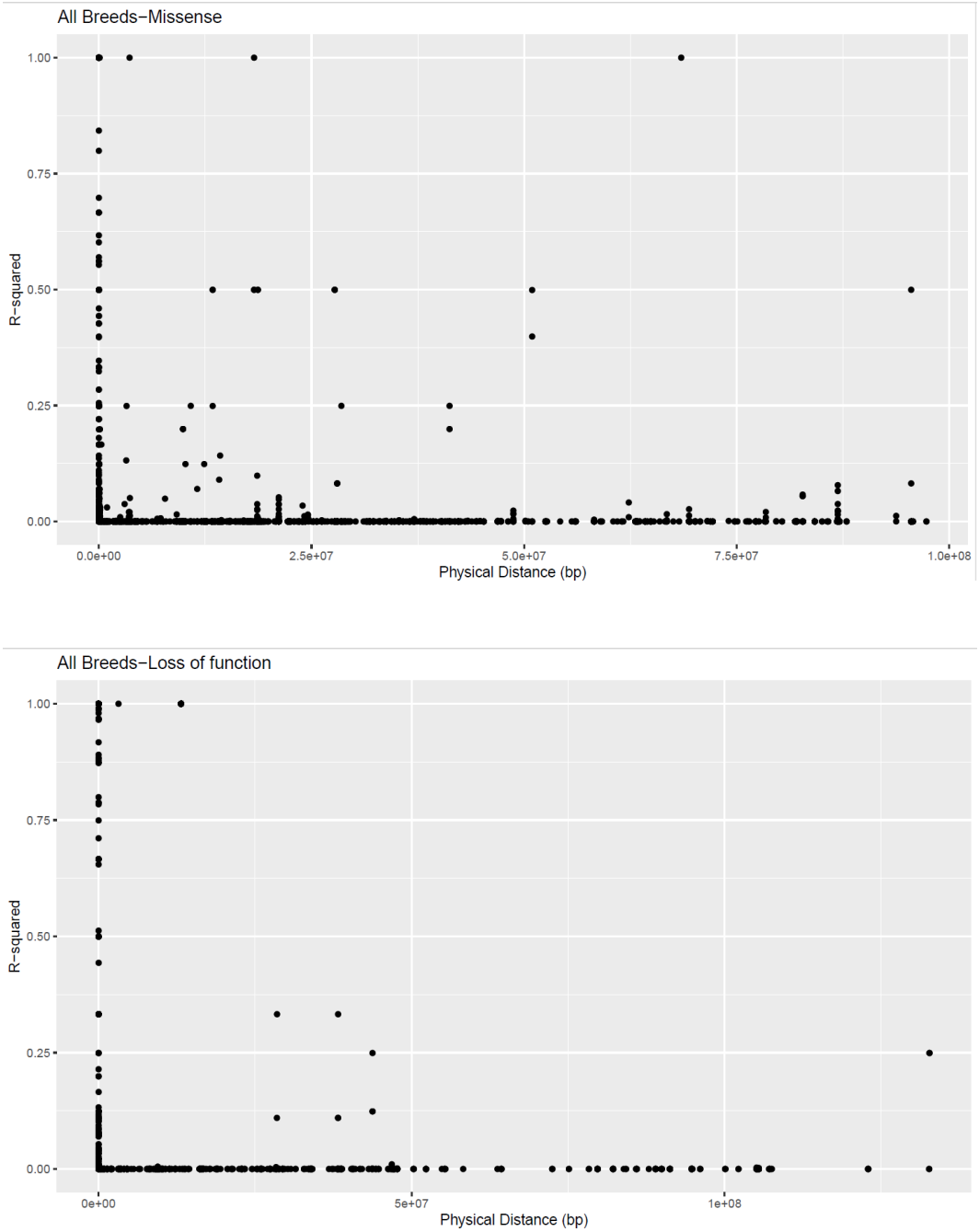


Figure 6 Pairwise LD decay plot for highly deleterious missense (top) and loss-of-function (bottom) variants using the whole dataset.

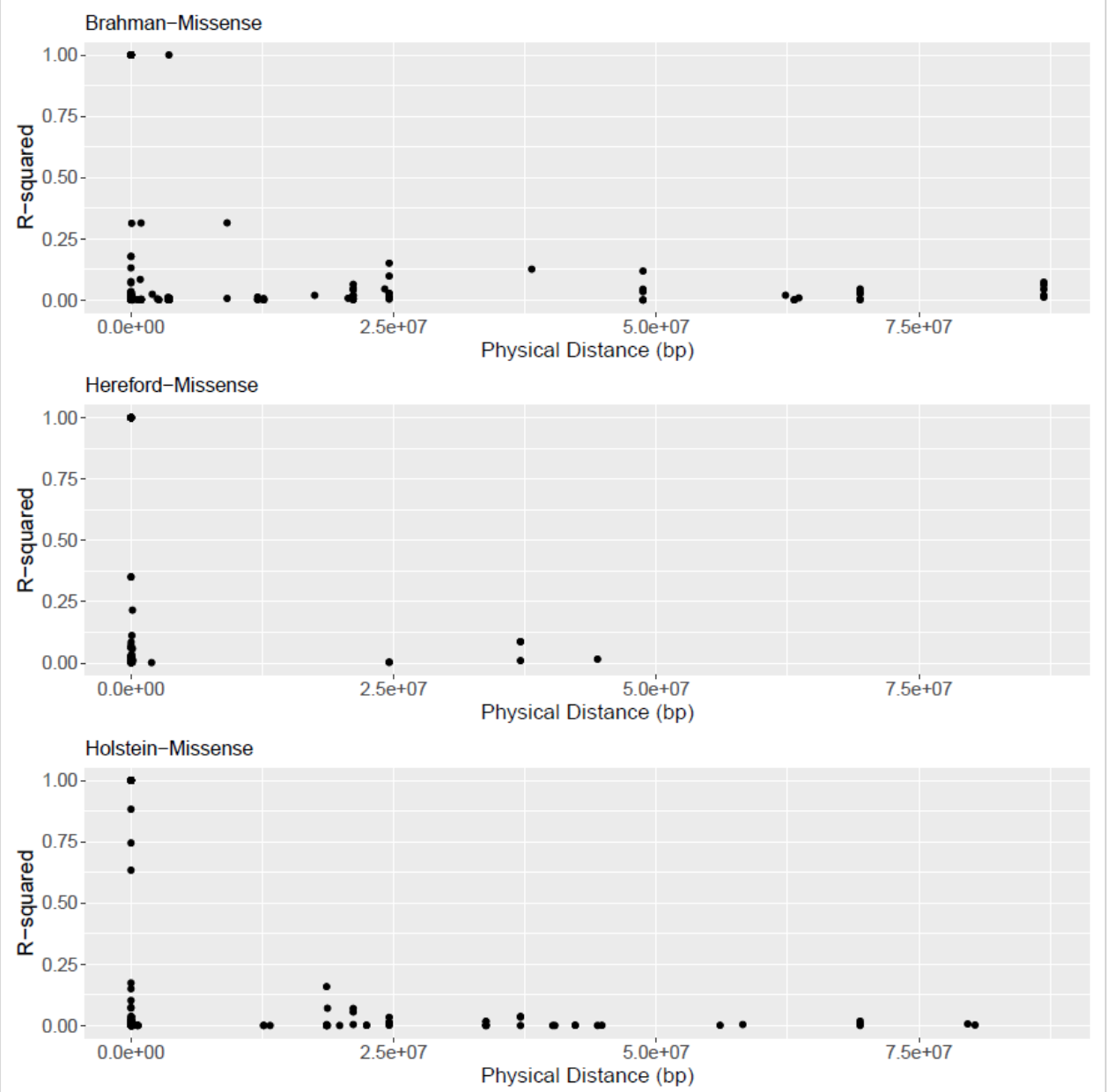


Figure 7 Pairwise LD decay plot for highly missense variants in Brahman, Hereford, and Holstein cattle.

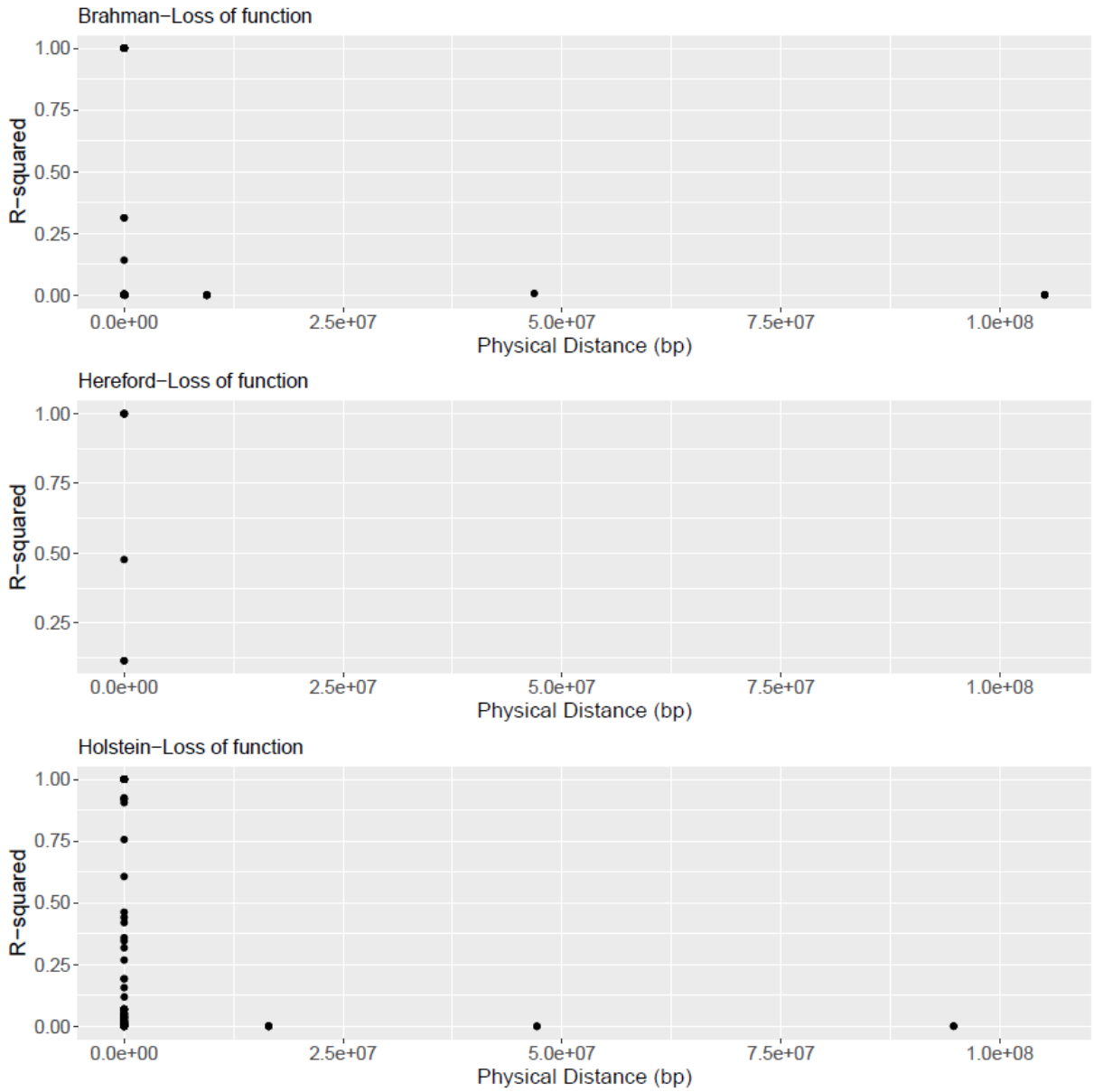


Figure 8 Pairwise LD decay plot for highly Loss-of-function variants in Brahman, Hereford, and Holstein cattle.

3.4.2 LD between highly deleterious variants and variants in their surrounding area

LD analysis between highly deleterious variants and their surrounding area showed that LD decayed rapidly around highly deleterious variants (Appendix 1). **Figure 9** shows the average LD between each highly deleterious variant and the variants in its surrounding area, divided into different bins based on the distance from the highly deleterious variant. Separate plots were generated for missense (top) and loss of function (bottom) variants. Each point in the plot represented the average LD between a highly deleterious variant and the surrounding variants within the same distance bin. The results revealed consistently low average LD, even in small distances, with r^2 below 0.25 for missense and below 0.3 for loss of function variants. To compare LD patterns around deleterious variants and around random variants, the average LD decay of variants in enhancer regions is shown as reference in **Figure 10**. Notably, LD around enhancer variants was higher compared to LD around deleterious missense and loss-of-function variants, with r^2 reaching values even close to 1.

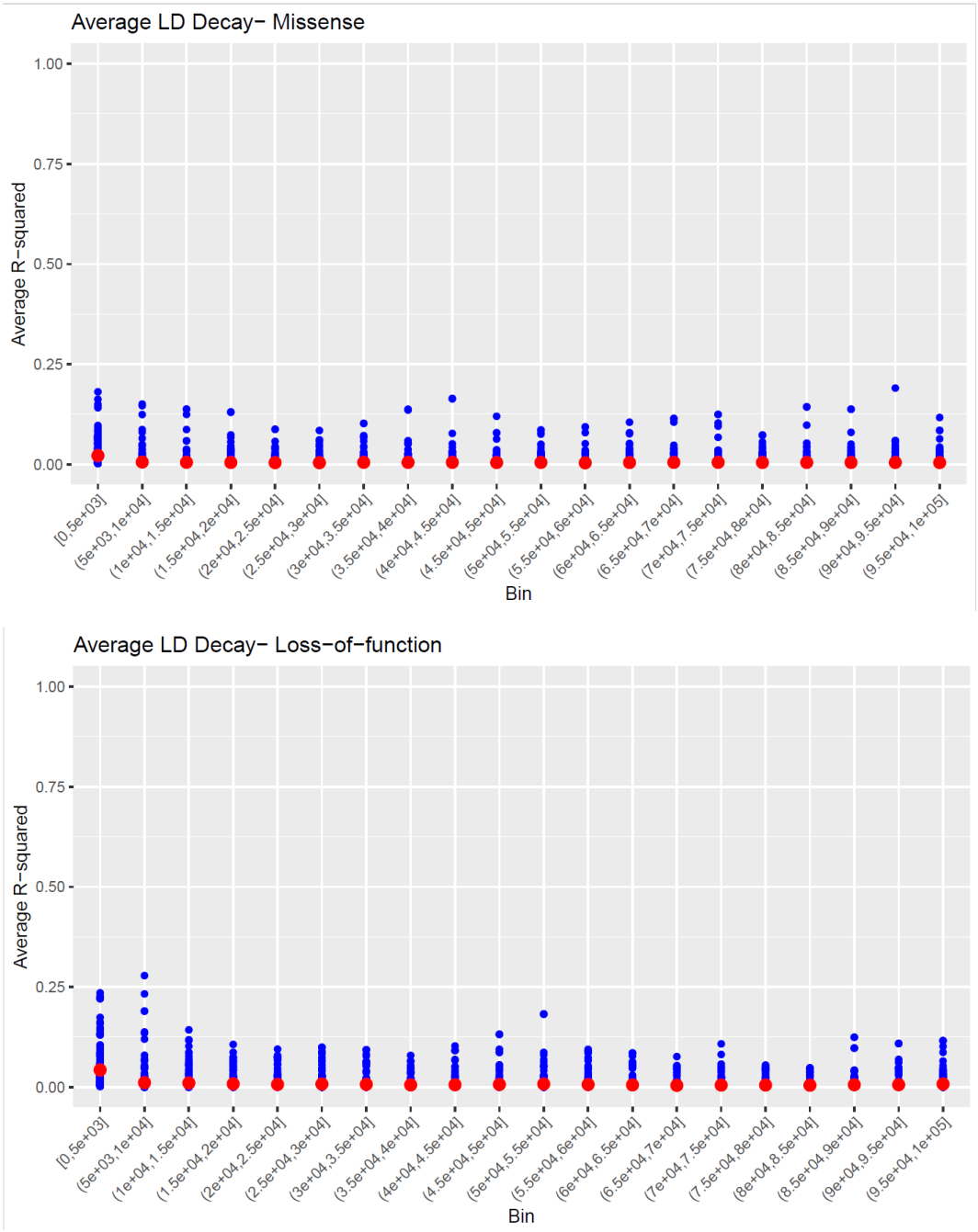


Figure 9 LD decay plot showing the average LD between each highly deleterious variant and the variants in their surrounding area, divided into two different bins based on the distance from the highly deleterious variant: missense (top) and loss-of-function (bottom) variants using the whole dataset. The red point represents the average value for the respective distance bin.

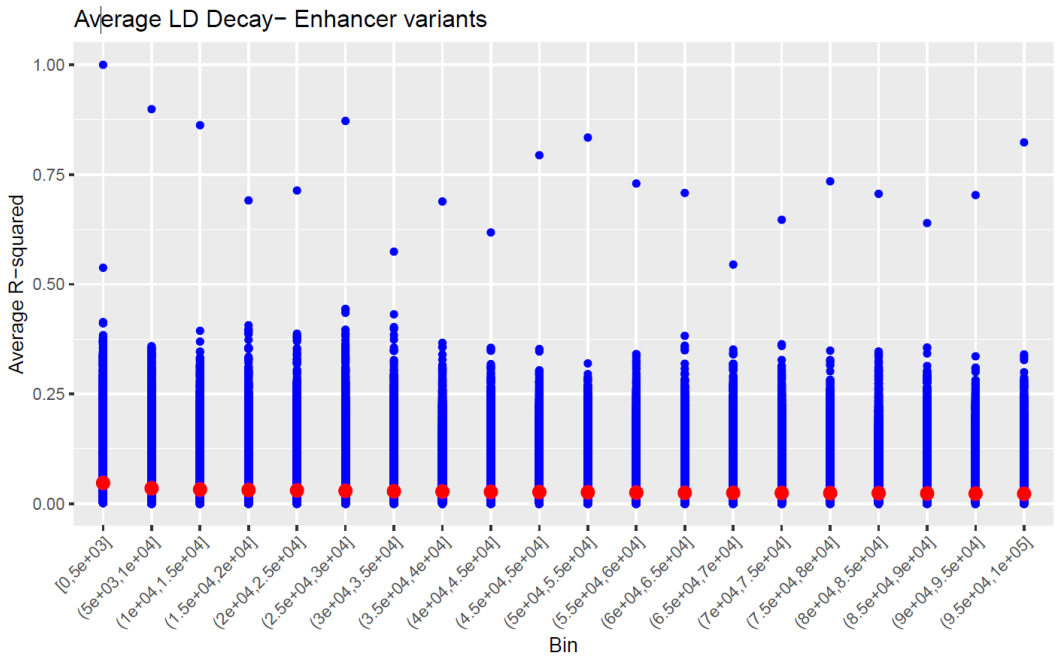


Figure 10 LD decay plot showing the average LD between each variant in enhancer regions and the variants in their surrounding area, divided into different bins based on the distance. The red point represents the average value for the respective distance bin.

3.5 Evaluation of MutPred2 based on scores assigned to deleterious variants in the OMIA database

Initially, the MutPred2 output was examined to determine whether prediction scores were assigned to the OMIA missense variants. Among the 80 OMIA deleterious missense variants, 30 had already been predicted and assigned a score by MutPred2. For the remaining 50 variants that were not present in the MutPred2 output and lacked prediction scores, MutPred2 analysis was conducted to evaluate the prediction scores for these variants. 7 variants could not be predicted due to an error in MutPred2, indicating the presence of a different wild-type amino acid in the mutation. The results for the variants that received a score revealed remarkably low prediction scores, all falling below 0.1.

Discussion

In this study, I utilized MutPred2 and MutPredLOF to predict highly deleterious missense and loss of function variants, respectively. I estimated the individual load of highly deleterious variants (missense and loss-of-function), the allele frequency spectrum of functional and possibly deleterious variants of different kinds (missense, synonymous, potentially loss of function, potentially gene-regulatory) in different cattle breeds, and investigated the pairwise LD between highly deleterious variants, and LD between highly deleterious variants and their surrounding area. Lastly, I evaluated the performance of MutPred2 by conducting analysis on missense variants predicted as deleterious and present in the OMIA database. The impact of purifying selection and the possible effects of balancing selection on the load and frequency of deleterious and gene-regulatory variants, the differences between breeds, the low LD patterns between and around deleterious variants in relation to recombination and selection, and the parameters affecting MutPred2's performance are discussed in detail below.

4.1 The effect of purifying selection on the load and frequency of deleterious variants

The unique allele frequency spectrum of the deleterious mutations compared to non-deleterious and synonymous, as well as the low deleterious mutation load indicate that deleterious variants are subject to purifying selection. When comparing the allele frequency spectrum of synonymous, potentially loss-of-function (frameshift and stop-gain), and missense variants, the results showed an excess of rare missense and potentially loss of function variants compared to

synonymous. Moreover, the allele frequency distribution of missense variants indicated an enrichment of mildly and highly deleterious variants at lower frequencies. In contrast, the non-deleterious variants followed a similar distribution of allele frequencies as synonymous variants. In the context of population genetic theory, the amount of deleterious alleles originating from mutation is expected to be equal to the amount of deleterious alleles eliminated by selection, leading to a mutation - selection balance (Charlesworth & Willis, 2009b). This balance ensures that harmful alleles are kept at a low frequency in the population, due to the elimination through purifying selection (Old, 1993), suggesting that the large majority of these missense variants are evolutionary tolerated (Derks et al., 2018). Similar patterns of allele frequency distributions have been reported in studies on chicken (Derks et al., 2018) and maize (Mezmouk & Ross-Ibarra, 2014).

The same trends were observed in loss of function variants with mildly and highly deleterious variants being enriched at lower frequencies, indicating that they are also subject to purifying selection. Frameshift and stop gain variants can be highly disruptive, as they can lead to the formation of a substantially different protein either by altering the coding frame or introducing a premature stop codon. However, some of these variants were classified as non-deleterious and may actually be tolerated. For instance, frameshift variants located at the N-terminal of the protein may still allow a functional protein to be created through an alternate start codon, effectively rescuing a substantial portion of the protein (Ng et al., 2008). Similarly, frameshift or stop-gained variants at the C-terminal end may be tolerated, given that they often result in the production of an almost complete protein. Therefore, the position of these variants in the amino-acid sequence is significant, as their impact on the protein is likely influenced by this position (Derks et al., 2018). Lastly, the mutation load analysis showed a lower load of deleterious loss of function variants compared to deleterious missense variants, despite the loss-of-function category having a higher proportion of highly deleterious variants compared to missense. This suggests that missense variants might be more tolerable, whereas loss-of-function variants could potentially be more harmful and, consequently, more susceptible to elimination through selection.

4.2 The effect of purifying selection on the frequency of TSS and enhancer variants

The distribution of allele frequencies in TSS and enhancer variants reveals that the non-coding genome also undergoes purifying selection. It was observed that variants in TSS and enhancers were enriched for low frequencies compared to intergenic variants. Genetic variants that do not directly modify protein sequences exhibit diverse functions, taking part in various gene regulation processes from transcription to post-translation (Makrythanasis & Antonarakis, 2013). However, the characterization of regulatory variants as deleterious is challenging due to the fact that most of the algorithms widely used to assess the impact of mutations are designed to estimate changes in protein structure or sequence conservation (Kaplun et al., 2016). This limitation may result in overlooking a substantial number of potentially deleterious or functionally relevant variants (Derks et al., 2018). Ongoing methods such as whole-genome association tests, linkage analysis, and quantitative trait locus mapping aim to detect causal regulatory variants (M. J. Li et al., 2014). Moreover, comparative genomics could be a significant approach by quantifying the evolutionary conservation of regulatory elements across different species. If divergent species show high conservation of their regulatory elements, it suggests that these elements are more likely to be functional, affecting the transcription of specific genes. Andrews et al., 2023 reported that variants in conserved regions among mammals tend to explain a larger proportion of heritability of human traits and that disease and trait associated variants are most enriched in highly conserved cis-regulatory elements. However, given the limitations of these methods, there is a need for alternative approaches to assess the potential impact of the non-coding genetic variations.

4.3 Balancing selection as a potential mechanism for the higher frequencies of a limited number of deleterious variants

Despite the generally low mutation load and allele frequencies of most deleterious variants, some of these variants were present at relatively high frequencies. This can be attributed to genetic drift. However, another possible explanation involves antagonistic pleiotropy, in which harmful alleles simultaneously impact multiple traits. These alleles may have beneficial effects on one trait and detrimental effects on another trait (Hedrick, 1999) and therefore they are under balancing selection. Heterozygous advantage or overdominance is a mechanism of antagonistic pleiotropy that leads to a balance between purifying selection against mutant homozygotes, and positive artificial selection on heterozygotes (Hedrick, 2015). These pleiotropic variants can remain at moderate to high frequency offering advantages for favorable traits in heterozygotes, while proving harmful in homozygotes, leading to lethality and reduced fitness. The mutation load analysis in this study indicated that all animals carrying deleterious variants were heterozygous, providing a potential explanation for this phenomenon. Besides heterozygote advantage, variants may undergo balancing selection when they impact multiple traits (allelic pleiotropy) that are (negatively) correlated. The linkage of deleterious variants to genes subject to balancing selection may also result in the excessive presence of deleterious variants in specific regions. One example of such linkage is the antagonistic relationship between fertility and milk production. While a quantitative trait locus (QTL) for fertility with effects in milk production is under balancing selection in nature, the transition to directional selection for milk production has resulted in lower fertility in cows (Kadri et al., 2014)

4.4 Differences in mutation load and allele frequencies among breeds

While the allele frequency distribution followed a similar pattern across all three breeds, different variants within different genes were identified at high frequencies in different breeds. However, the number of genes that deviated between breeds was very small, thus there were no obvious functional connections among genes within the same breed. If there had been more genes that were systematically different between breeds, Gene Set Enrichment Analysis could have been applied in order to determine whether particular functions of genes were more common in one breed compared to the others. These variations observed among breeds, particularly between taurine and indicine breeds, may be explained by the differences in the effective population sizes of their ancestral population before domestication (Gibbs et al., 2009), and in the magnitude of the bottleneck occurred during the domestication and breed formation (Frantz et al., 2020b), as well as differences in selection pressure and selection strategies associated with their production purposes.

Differences in mutation load were evident among breeds, with Brahman exhibiting an overall higher mutation. Specifically for missense variants in Brahman, 13.5% of individuals had more than 10 mutations, a higher percentage compared to 4.5% and 5% in Hereford and Holstein, respectively. Regarding loss of function variants in Brahman, 5% of the individuals had 0 mutations, whereas the percentage was significantly higher for Hereford and Holstein (65% and 48% respectively). Brahman is a mixture of different breeds, including Guzerat, Nellore, Gir, Indu-Brazil, and other Zebu breeds (Sanders, 1980), indicating the substantial genetic diversity present in this breed. Moreover, it has larger effective population size, compared to Holstein and Hereford. These factors may have contributed to Brahman's tolerance to deleterious mutations, resulting in a higher mutation load. In contrast, a similar study on deleterious single nucleotide variants (SNVs) utilizing groups of breeds showed that, on average, taurine cattle breeds had a higher mutation load than indicine breeds (Subramanian, 2021). However, this

study used numerous breeds within each category (taurine and indicine) and the diversity ratio (ω) between nonsynonymous and synonymous SNVs as a measure of deleterious mutational load, rather than the count of deleterious mutations per breed. In our study, we focused on one breed for indicine and two breeds for taurine cattle, and the sizes of each breed varied, potentially not fully capturing the entire variation. Further research involving a more extensive range of breeds within each category will provide more insights into the mutation load in cattle.

4.5 Low LD between deleterious variants and between deleterious variants and their surrounding area

The analysis of pairwise LD between deleterious missense and between deleterious loss-of-function variants revealed low LD that decays rapidly. The power of LD depends on the allele frequencies, therefore the low frequencies of deleterious variants led to a low LD between them. Garcia & Lohmueller, 2021 reported in a human study that LD of nonsynonymous variants was lower compared to synonymous suggesting that nonsynonymous variants tend to appear on different haplotypes. Linkage disequilibrium indicates how selection at one locus impacts other loci. Moreover, when two variants are on the same haplotype they interfere with each other (Hill & Robertson, 1966) in order to become fixed, a mechanism known as Hill-Robertson interference. As a result, selection at one locus may reduce or increase the chance of fixation at a second locus. Genetic recombination plays a significant role in reducing interference by enabling selected sites to segregate independently, creating new haplotypes, and thereby slowing down the accumulation of rare harmful variants (Keightley & Otto, 2006). As a result, newly emerging deleterious mutations in regions with higher recombination rates will be more efficiently purged by natural selection (Hussin et al., 2015).

Another way that purifying selection can affect LD patterns between deleterious variants is negative synergistic epistasis. Negative synergistic epistasis occurs when

each additional deleterious mutation reduces fitness by a greater amount than the reduction in fitness caused by each mutation independently (Lewontin, 1964). This results in negative selection eliminating haplotypes containing multiple deleterious alleles. The remaining deleterious alleles are more likely to segregate on different haplotypes compared to neutral mutations, resulting in negative LD (Sohail et al., 2017). Sohail et al. (2017) investigated the patterns of signed LD among rare loss of function mutations in humans and fruit flies. The study revealed that LOF variants had significantly lower LD compared to synonymous sites, suggesting that these mutations not only undergo purifying selection, but also non-independently affect fitness, indicating negative synergistic epistasis.

The analysis of LD between deleterious variants and surrounding variants revealed a rapid LD decay around deleterious variants. LD is limited by the allele frequencies of the participating variants. This pattern is evident in the average LD decay distributions around rare deleterious variants, in contrast to enhancer variants found at various frequencies, indicating that the LD around deleterious variants is significantly lower than that around enhancer variants. The rapid LD decay around deleterious variants suggests that the effectiveness of tagging deleterious variants with SNP chips may be limited, as fewer nearby SNPs may effectively capture information from these variants. Consequently, estimating the effect of these deleterious variants on health-related traits and selecting against these variants using genomic predictions may be challenging, as their presence cannot be easily predicted based on the genotyping of nearby SNPs. Gaining insights into how recombination, selection and mutation interact to distribute deleterious mutations across the genome, as well as the differences in the history and size of different populations will enhance our understanding of mutations that contribute to disease (Hussin et al., 2015).

4.6 Performance of MutPred2

For the identification of deleterious missense variants, the standalone version of MutPred2 was utilized and accessed through the server. The allele frequency analysis and the enrichment of rare deleterious variants verified the effectiveness of MutPred2 in identifying strongly deleterious variants. However, when assessing variants known to be deleterious in the OMIA database, the prediction scores were consistently very low, all below 0.1. This indicates that MutPred2 may fail to detect certain deleterious variants.

MutPred2 utilizes precomputed databases containing multiple sequence alignments and conservation scores to compute conservation-based protein features. These data were precomputed for humans. However, when dealing with input substitutions from novel protein sequences where conservation-based features might not be available, MutPred2 offers an option to predict them from sequence information and PSI-BLAST position-specific scoring matrices. It was observed that models incorporating conservation features outperformed those lacking conservation features by two percentage points (Pejaver et al., 2020). Due to the long computation time and increased memory requirements of MutPred2, I decided to skip PSI-BLAST, which may have an impact on the reliability of the scores. To test this hypothesis, I conducted MutPred2 analysis again for several OMIA deleterious variants that previously received low MutPred2 scores, this time without skipping PSI-BLAST. The results showed that the scores were significantly higher, placing the majority of these variants within the deleterious score category (Appendix 2). For example, the D128G substitution associated with Leukocyte adhesion deficiency, type I in the *ITGB2* gene, initially received a score of 0.056. This score was found to be 0.831 when PSI-BLAST was included in the analysis. This indicates that some cattle proteins not present in humans would benefit from the inclusion of PSI-BLAST.

The development of software similar to MutPred2 specifically for cattle would require a large genomic database containing both pathogenic and putatively neutral

variants in cattle. For reference, more than 50,000 human pathogenic and 200,000 neutral variants were utilized for the development of MutPred2 (Pejaver et al., 2020). Generating a comparable dataset in cattle would take several years, making the development of a similar software challenging. MutPred2, particularly with the inclusion of PSI-BLAST was shown to be effective in predicting deleterious variants. Moreover, given that different variant types besides missense are potentially deleterious, it might be more practical to investigate existing human prediction software with proven effectiveness in cattle.

4.7 Implications

In this study, all deleterious missense, frameshift and stop-gain variants present in the 1000 Bull Genome dataset were identified. This effort aims to contribute to the creation of a comprehensive catalogue including all deleterious variants. Accurate detection of deleterious variants and the application of predictions for functional variants on a large scale offer the potential to integrate this information into genomic predictions. For this purpose, it is necessary to increase the number of sequenced animals, especially those diagnosed with genetic defects within each breed. This would enable us to capture a more significant proportion of deleterious variants and to detect rare or breed-specific variants. Subsequently, the identified deleterious variants could be utilized to develop SNP chips incorporating them. Moreover, it is essential to expand predictions to include a wider range of potentially deleterious variant types, such as inframe insertions and deletions. Similarly to MutPred2 and MutPredLOF, MutPredIndel has been developed for the identification of inframe insertions and deletions (Pagel et al., 2019).

Further research should be focused on the detection of LD between deleterious variants and variants in the commercially used SNP panels. For instance, if we establish that deleterious variants are in high LD with SNPs on a standard SNP chip, selection against them is feasible. However, if LD between deleterious variants and

SNPs is low, using SNPs as markers for the deleterious variants may be less efficient. In this case, customized chips including deleterious variants, as mentioned previously, may prove to be a more effective alternative. This approach would enable the incorporation of rare deleterious variants into genomic predictions, enhancing the efficiency of genomic predictions in reducing the frequency of harmful variants within the population.

In addition, the accurate identification of deleterious variants can be applied to develop screenings for potential genetic defects. Various stages of this study could integrate into pipelines designed for the prediction of causative variants associated with genetic defects. Screening entire populations may be beneficial when a reduction in fitness or lethality (e.g. infertility, embryonic lethality etc.) is observed within the same related population or family, which can be an indication of a genetic disease. In addition, if a specific genomic region or gene is associated with a particular trait or disorder, screening for deleterious variants within this candidate region or gene can assist in identifying potential causative variants.

Moreover, in cases where a known causative variant is linked to a genetic disease and phenotypes are available, the information provided by MutPred2 and MutPredLOF regarding consequences in protein functions can provide insights on the pathogenicity of the disease. These consequences can reveal pathways or biological processes that may be affected, providing information about the potential impact of these variants on cellular functions. By integrating this information with clinical data and the results from laboratory examinations on patient samples, it is possible to confirm the connection between the predicted consequences and observed phenotypes in individuals affected by genetic diseases. Understanding the molecular mechanisms influencing certain genetic defects will assist the development of targeted therapies aimed at reducing the impact of deleterious variants on protein function, enhancing the efficacy of managing and treating the associated disease.

.

Conclusions

This study aimed to estimate the load and allele frequency spectrum of functional and possibly deleterious variants in different cattle breeds, and investigate the LD between deleterious variants, and between deleterious variants and their surrounding areas. The low mutation load of deleterious variants and their enrichment at lower frequencies highlighted the significant role of purifying selection in eliminating deleterious variants from the population. Furthermore, it was shown that the non-coding genome also undergoes selection, yet further research should focus on the functional annotation of the gene regulatory variants. The LD analysis revealed that deleterious variants tend to occur on different haplotypes, suggesting that recombination prevents the accumulation of deleterious variants. Breed-specific differences in load and allele frequencies underscored the influence of effective population size, selection pressure, and breed-specific selection strategies.

This study has the potential to contribute to a comprehensive catalogue of deleterious variants in cattle if it is expanded to include a wider range of variant types and a larger number of sequenced animals. The detection of deleterious variants will not only facilitate their integration into genomic predictions but also it will provide a better understanding of the mechanisms by which these mutations are contributing to diseases. As a result, the effectiveness of genomic selection against harmful mutations and the overall management and treatment of genetic defects will be enhanced.

References

Agrawal, A. F., & Whitlock, M. C. (2012). Mutation load: The fitness of individuals in populations where deleterious alleles are abundant. In *Annual Review of Ecology, Evolution, and Systematics* (Vol. 43, pp. 115–135). https://doi.org/10.1146/annurev-ecolsys-110411-160257

Alexandre, P. A., Naval-Sánchez, M., Menzies, M., Nguyen, L. T., Porto-Neto, L. R., Fortes, M. R. S., & Reverter, A. (2021). Chromatin accessibility and regulatory vocabulary across indicine cattle tissues. *Genome Biology*, *22*(1). https://doi.org/10.1186/s13059-021-02489-7

Andrews, G., Fan, K., Pratt, H. E., Phalke, N., Zoonomia Consortium §, Karlsson, E. K., ... & Zhang, X. (2023). Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. Science, 380(6643), eabn7930.

Bosse, M., Megens, H. J., Derks, M. F. L., de Cara, Á. M. R., & Groenen, M. A. M. (2019). Deleterious alleles in the context of domestication, inbreeding, and selection. *Evolutionary Applications*, *12*(1), 6–17. https://doi.org/10.1111/eva.12691

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., & Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutation*, *30*(8), 1237–1244. https://doi.org/10.1002/humu.21047

Charlesworth, D., & Willis, J. H. (2009a). The genetics of inbreeding depression. In *Nature Reviews Genetics* (Vol. 10, Issue 11, pp. 783–796). https://doi.org/10.1038/nrg2664

Charlesworth, D., & Willis, J. H. (2009b). The genetics of inbreeding depression. In *Nature Reviews Genetics* (Vol. 10, Issue 11, pp. 783–796). https://doi.org/10.1038/nrg2664

Cline, M. S., & Karchin, R. (2011). Using bioinformatics to predict the functional impact of SNVs. In *Bioinformatics* (Vol. 27, Issue 4, pp. 441–448). https://doi.org/10.1093/bioinformatics/btq695

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brøndum, R. F., Liao, X., Djari, A., Rodriguez, S. C., Grohs, C., Esquerré, D., Bouchez, O., Rossignol, M. N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P. J., Coote, D., … Hayes, B. J. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, *46*(8), 858–865. https://doi.org/10.1038/ng.3034

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, *6*(12). https://doi.org/10.1371/journal.pcbi.1001025

Derks, M. F. L., Megens, H. J., Bosse, M., Visscher, J., Peeters, K., Bink, M. C. A. M., Vereijken, A., Gross, C., De Ridder, D., Reinders, M. J. T., & Groenen, M. A. M. (2018). A survey of functional genomic variation in domesticated chickens. *Genetics Selection Evolution*, *50*(1). https://doi.org/10.1186/s12711-018-0390-1

Elsik, C. G., Tellam, R. L., Worley, K. C., Gibbs, R. A., Muzny, D. M., Weinstock, G. M., Adelson, D. L., Eichler, E. E., Elnitski, L., Guigó, R., Hamernik, D. L., Kappes, S. M., Lewin, H. A., Lynn, D. J., Nicholas, F. W., Reymond, A., Rijnkels, M., Skow, L. C., Zdobnov, E. M., … Zhao, F. Q. (2009). The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science*, *324*(5926), 522–528. https://doi.org/10.1126/science.1169588

Frantz, L. A. F., Bradley, D. G., Larson, G., & Orlando, L. (2020a). Animal domestication in the era of ancient genomics. In *Nature Reviews Genetics* (Vol. 21, Issue 8, pp. 449–460). Nature Research. https://doi.org/10.1038/s41576-020-0225-0

Frantz, L. A. F., Bradley, D. G., Larson, G., & Orlando, L. (2020b). Animal domestication in the era of ancient genomics. In *Nature Reviews Genetics* (Vol. 21, Issue 8, pp. 449–460). Nature Research. https://doi.org/10.1038/s41576-020-0225-0

Garcia, J. A., & Lohmueller, K. E. (2021). Negative linkage disequilibrium between amino acid changing variants reveals interference among deleterious mutations in the human genome. *PLoS Genetics*, *17*(7). https://doi.org/10.1371/journal.pgen.1009676

Halstead, M. M., Kern, C., Saelao, P., Wang, Y., Chanthavixay, G., Medrano, J. F., Van Eenennaam, A. L., Korf, I., Tuggle, C. K., Ernst, C. W., Zhou, H., & Ross, P. J. (2020). A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-020-07078-9

Hartfield, M., & Otto, S. P. (2011a). Recombination and hitchhiking of deleterious alleles. *Evolution*, *65*(9), 2421–2434. https://doi.org/10.1111/j.1558-5646.2011.01311.x

Hartfield, M., & Otto, S. P. (2011b). Recombination and hitchhiking of deleterious alleles. *Evolution*, *65*(9), 2421–2434. https://doi.org/10.1111/j.1558-5646.2011.01311.x

Hayes, B. J., & Daetwyler, H. D. (2019). *1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes*. https://doi.org/10.1146/annurev-animal-020518

Hedrick, P. W. (1999). Antagonistic pleiotropy and genetic polymorphism: a perspective. *Heredity*, *82*, 126–133. https://doi.org/10.1046/j.1365-2540.1999.00440.x

Hedrick, P. W. (2015). Heterozygote advantage: The effect of artificial selection in livestock and pets. *Journal of Heredity*, *106*(2), 141–154. https://doi.org/10.1093/jhered/esu070

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research*, *8*(3), 269–294. https://doi.org/10.1017/S0016672300010156

Hu, J., & Ng, P. C. (2012). Predicting the effects of frameshifting indels. *Genome Biology*, *13*(2). https://doi.org/10.1186/gb-2012-13-2-r9

Hussin, J. G., Hodgkinson, A., Idaghdour, Y., Grenier, J. C., Goulet, J. P., Gbeha, E., Hip-Ki, E., & Awadalla, P. (2015). Recombination affects accumulation of

damaging and disease-associated mutations in human populations. *Nature Genetics*, *47*(4), 400–404. https://doi.org/10.1038/ng.3216

Kadri, N. K., Sahana, G., Charlier, C., Iso-Touru, T., Guldbrandtsen, B., Karim, L., Nielsen, U. S., Panitz, F., Aamand, G. P., Schulman, N., Georges, M., Vilkki, J., Lund, M. S., & Druet, T. (2014). 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*, *10*(1). https://doi.org/10.1371/journal.pgen.1004049

Kaplun, A., Krull, M., Lakshman, K., Matys, V., Lewicki, B., & Hogan, J. D. (2016). Establishing and validating regulatory regions for variant annotation and expression analysis. *BMC Genomics*, *17*. https://doi.org/10.1186/s12864-016-2724-0

Keightley, P. D., & Otto, S. P. (2006). Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, *443*(7107), 89–92. https://doi.org/10.1038/nature05049

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Lande, R. (1994). RISK OF POPULATION EXTINCTION FROM FIXATION OF NEW DELETERIOUS MUTATIONS. In *Evolution* (Vol. 48, Issue 5). https://academic.oup.com/evolut/article/48/5/1460/6870203

Lewontin, R. C. (1964). *THE INTERACTION OF SELECTION AND LINKAGE. I. GENERAL CONSIDERATIONS; HETEROTIC MODELS'*.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, M. J., Yan, B., Sham, P. C., & Wang, J. (2014). Exploring the function of genetic variants in the non-coding genomic regions: Approaches for identifying

human regulatory variants affecting gene expression. *Briefings in Bioinformatics*, *16*(3), 393–412. https://doi.org/10.1093/bib/bbu018

Lohmueller, K. E. (2014). The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. *PLoS Genetics*, *10*(5). https://doi.org/10.1371/journal.pgen.1004379

Lu, J., Tang, T., Tang, H., Huang, J., Shi, S., & Wu, C. I. (2006). The accumulation of deleterious mutations in rice genomes: A hypothesis on the cost of domestication. *Trends in Genetics*, *22*(3), 126–131. https://doi.org/10.1016/j.tig.2006.01.004

Lynch, M., & Gabriel, W. (1990). Mutation load and the survival of small populations. *Evolution*, *44*(7), 1725–1737. https://doi.org/10.1111/j.1558-5646.1990.tb05244.x

Makrythanasis, P., & Antonarakis, S. (2013). Pathogenic variants in non-protein-coding sequences. In *Clinical Genetics* (Vol. 84, Issue 5, pp. 422–428). https://doi.org/10.1111/cge.12272

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0974-4

Mezmouk, S., & Ross-Ibarra, J. (2014). The pattern and distribution of deleterious mutations in maize. *G3: Genes, Genomes, Genetics*, *4*(1), 163–171. https://doi.org/10.1534/g3.113.008870

Moyers, B. T., Morrell, P. L., & McKay, J. K. (2018). Genetic costs of domestication and improvement. *Journal of Heredity*, *109*(2), 103–116. https://doi.org/10.1093/jhered/esx069

Mukai, T., Chigusa, S. I., Mettler, L. E., & Crow, J. F. (1972). *MUTATION RATE AND DOMINANCE OF GENES AFFECTING VIABILITY IN DROSOPHILA MELANOGASTERI*.
https://academic.oup.com/genetics/article/72/2/335/5990661

Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research*, *11*(5), 863–874. https://doi.org/10.1101/gr.176601

Ng, P. C., Levy, S., Huang, J., Stockwell, T. B., Walenz, B. P., Li, K., Axelrod, N., Busam, D. A., Strausberg, R. L., & Venter, J. C. (2008). Genetic variation in

an individual human exome. *PLoS Genetics*, *4*(8). https://doi.org/10.1371/journal.pgen.1000160

Pagel, K. A., Antaki, D., Lian, A., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M., Mooney, S. D., & Radivojac, P. (2019). Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. *PLoS Computational Biology*, *15*(6). https://doi.org/10.1371/journal.pcbi.1007112

Pagel, K. A., Pejaver, V., Lin, G. N., Nam, H. J., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M., Mooney, S. D., & Radivojac, P. (2017). When loss-of-function is loss of function: Assessing mutational signatures and impact of loss-of-function genetic variants. *Bioinformatics*, *33*(14), i389–i398. https://doi.org/10.1093/bioinformatics/btx272

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H. J., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M., Mooney, S. D., & Radivojac, P. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19669-x

Pitt, D., Sevane, N., Nicolazzi, E. L., MacHugh, D. E., Park, S. D. E., Colli, L., Martinez, R., Bruford, M. W., & Orozco-terWengel, P. (2019). Domestication of cattle: Two or three events? *Evolutionary Applications*, *12*(1), 123–136. https://doi.org/10.1111/eva.12674

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–121. https://doi.org/10.1101/gr.097857.109

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rausell, A., Mohammadi, P., McLaren, P. J., Bartha, I., Xenarios, I., Fellay, J., & Telenti, A. (2014). Analysis of Stop-Gain and Frameshift Variants in Human

Innate Immunity Genes. *PLoS Computational Biology*, *10*(7). https://doi.org/10.1371/journal.pcbi.1003757

Salavati, M., Caulton, A., Clark, R., Gazova, I., Smith, T. P. L., Worley, K. C., Cockett, N. E., Archibald, A. L., Clarke, S. M., Murdoch, B. M., & Clark, E. L. (2020). Global Analysis of Transcription Start Sites in the New Ovine Reference Genome (Oar rambouillet v1.0). *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.580580

Salavati, M., Clark, R., Becker, D., Kühn, C., Plastow, G., Dupont, S., Moreira, G. C. M., Charlier, C., & Clark, E. L. (2023). Improving the annotation of the cattle genome by annotating transcription start sites in a diverse set of tissues and populations using Cap Analysis Gene Expression sequencing. *G3: Genes, Genomes, Genetics*, *13*(8). https://doi.org/10.1093/g3journal/jkad108

Sanders, J. O. (1980). *HISTORY AND DEVELOPMENT OF ZEBU CATTLE IN THE UNITED STATES ~*.

Shen, Y., Yue, F., Mc Cleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Ren, B., & Lobanenkov, V. V. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, *488*(7409), 116–120. https://doi.org/10.1038/nature11243

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M., & Gaunt, T. R. (2013). Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation*, *34*(1), 57–65. https://doi.org/10.1002/humu.22225

Sohail, M., Vakhrusheva, O. A., Sul, J. H., Pulit, S. L., Francioli, L. C., Veldink, J. H., De Bakker, P. I. W., Bazykin, G. A., Kondrashov, A. S., Shamil, ‡, & Sunyaev, R. (2017). *EVOLUTIONARY GENETICS*. https://www.science.org

Subramanian, S. (2021). Deleterious protein-coding variants in diverse cattle breeds of the world. *Genetics Selection Evolution*, *53*(1). https://doi.org/10.1186/s12711-021-00674-7

Takahashi, H., Kato, S., Murata, M., & Carninci, P. (2012). CAGE (Cap analysis of gene expression): A protocol for the detection of promoter and

transcriptional networks. *Methods in Molecular Biology*, *786*, 181–200. https://doi.org/10.1007/978-1-61779-292-2_11

Whitlock, M. C., Griswold, C. K., & Peters, A. D. (2003). Compensating for the meltdown: The critical effective size of a population with deleterious and compensatory mutations. In *Annales Zoologici Fennici* (Vol. 40, Issue 2).

Zhang, M., Zhou, L., Bawa, R., Suren, H., & Holliday, J. A. (2016). Recombination Rate Variation, Hitchhiking, and Demographic History Shape Deleterious Load in Poplar. *Molecular Biology and Evolution*, *33*(11), 2899–2910. https://doi.org/10.1093/molbev/msw169

Popular science summary

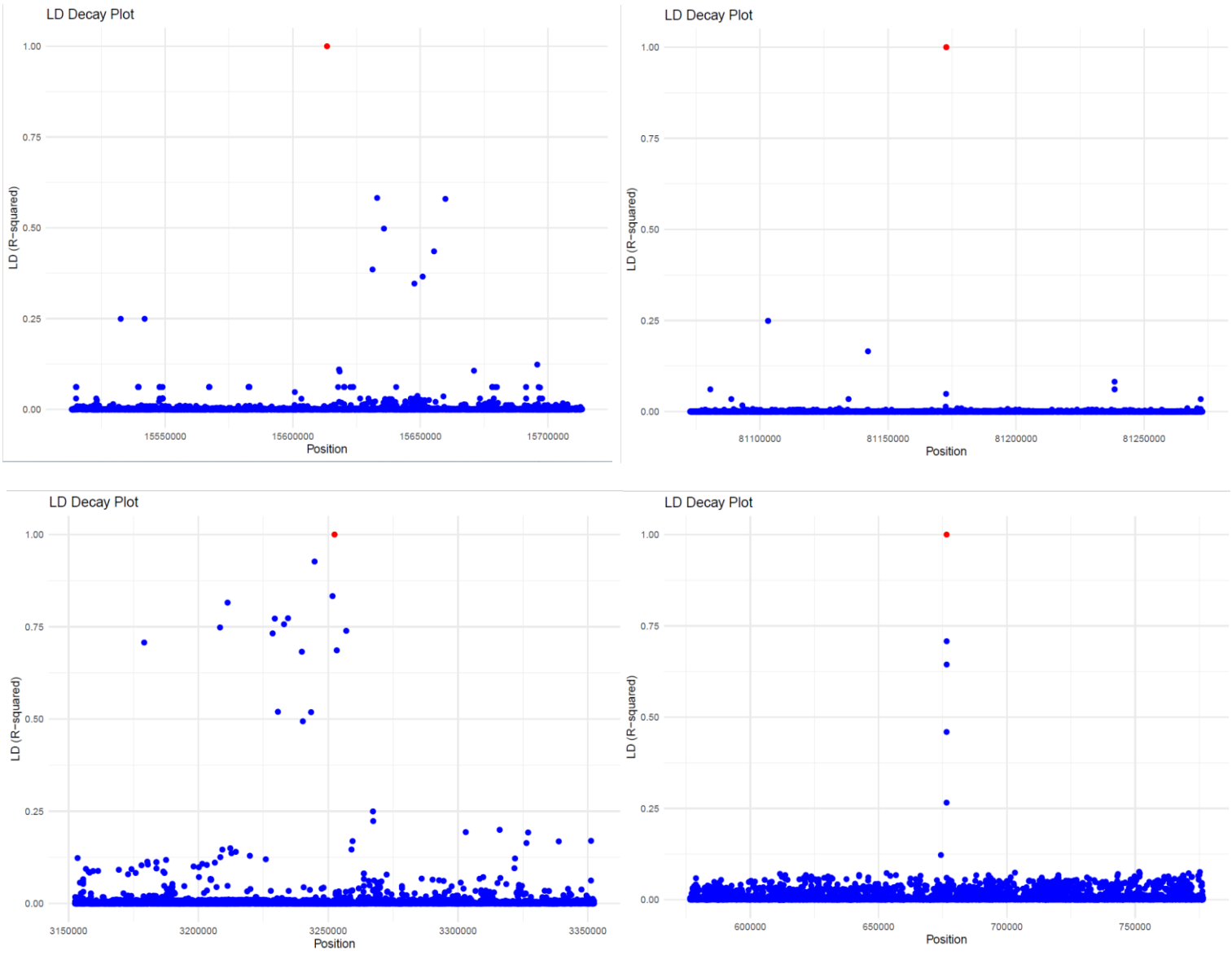
The DNA of all organisms undergoes changes, known as mutations, which can have negative consequences on health and production. Therefore, the aim of this project was to investigate the frequency of these harmful mutations in different cattle breeds, determine the number of harmful mutations that each individual carried, and examine whether these mutations are linked, i.e. whether they are located in adjacent positions on the DNA that are inherited together, as well as if these harmful mutations are linked and inherited together with non-detrimental mutations in the surrounding area of the DNA. Different software programs were used to identify harmful mutations, compute their frequencies, and assess their linkage. The results showed that the frequencies and the number of harmful mutations per individual were relatively low in all breeds, indicating the effectiveness of natural selection in removing these mutations from the population, as animals carrying such mutations have lower chances of survival and reproduction. However, different mutations were present in different breeds, highlighting their different evolution histories and production purposes (meat, milk etc). Due to the low frequencies of the harmful mutations, the linkage both between these mutations and with their surrounding ones were low, suggesting that these harmful mutations are located on separate DNA fragments inherited independently. If applied accurately on a large scale, this study has the potential to assist in DNA-based selection in order to reduce the frequency of harmful mutations from cattle populations and enhance the effective management of genetic diseases by identifying which variants are harmful, where they are located, and the mechanisms through which they contribute to genetic diseases.

Acknowledgements

First, I would like to express my gratitude to my supervisors, Martin Johnsson and Aniek Bouwman for all the help they provided. Martin, thank you for trusting me through my ups and downs during the thesis and for giving me the space to learn and improve. Aniek, thank you for your genuine interest in my work and for always providing me with new perspectives during our monthly meetings.

Then, I would like to thank my family for continuously believing in me and encouraging me despite the distance. I am also grateful to my friends in Greece and the Netherlands for their constant support, always being just a phone call away, and my friends in Sweden, including the EMABG group for making my thesis experience more fun.

Appendix 1



LD decay plots between different deleterious variants and the variants in their surrounding area.

Appendix 2

MutPred2 scores with and without the inclusion of PSI-BLAST for several OMIA deleterious variants.

Ensembl Transcript ID	**Gene Name**	**Substitution**	**MutPred2 score**		**Variant Phenotype**
			Without PSI-BLAST	**With PSI-BLAST**	
ENSBTAT00000008593	*ATP2A1*	G211V	0.045	0.934	Pseudomyotonia, congenital
ENSBTAT00000084344	*COL5A2*	G789V	0.030	0.930	Ehlers-Danlos syndrome, classic type, 2
ENSBTAT00000001290	*IARS*	V79L	0.041	0.855	Perinatal weak calf syndrome
ENSBTAT00000015674	*MSTN*	L64P	0.048	0.777	Muscular hypertrophy (double muscling)
ENSBTAT00000028571	*EDN2*	C50Y	0.057	0.800	Growth and respiratory lethal syndrome
ENSBTAT00000024865	*TBXT*	K66E	0.054	0.913	Vertebral and spinal dysplasia
ENSBTAT00000043649	*MYBPC1*	L295R	0.060	0.833	Arthrogryposis, distal, type 1B
ENSBTAT00000046583	*SLC12A1*	P372L	0.050	0.788	Hydrallantois

Publishing and archiving

Approved students' theses at SLU are published electronically. As a student, you have the copyright to your own work and need to approve the electronic publishing. If you check the box for **YES**, the full text (pdf file) and metadata will be visible and searchable online. If you check the box for **NO**, only the metadata and the abstract will be visible and searchable online. Nevertheless, when the document is uploaded it will still be archived as a digital file. If you are more than one author, the checked box will be applied to all authors. You will find a link to SLU's publishing agreement here:

- https://libanswers.slu.se/en/faq/228318.

☒ YES, I/we hereby give permission to publish the present thesis in accordance with the SLU agreement regarding the transfer of the right to publish a work.

☐ NO, I/we do not give permission to publish the present work. The work will still be archived, and its metadata and abstract will be visible and searchable.