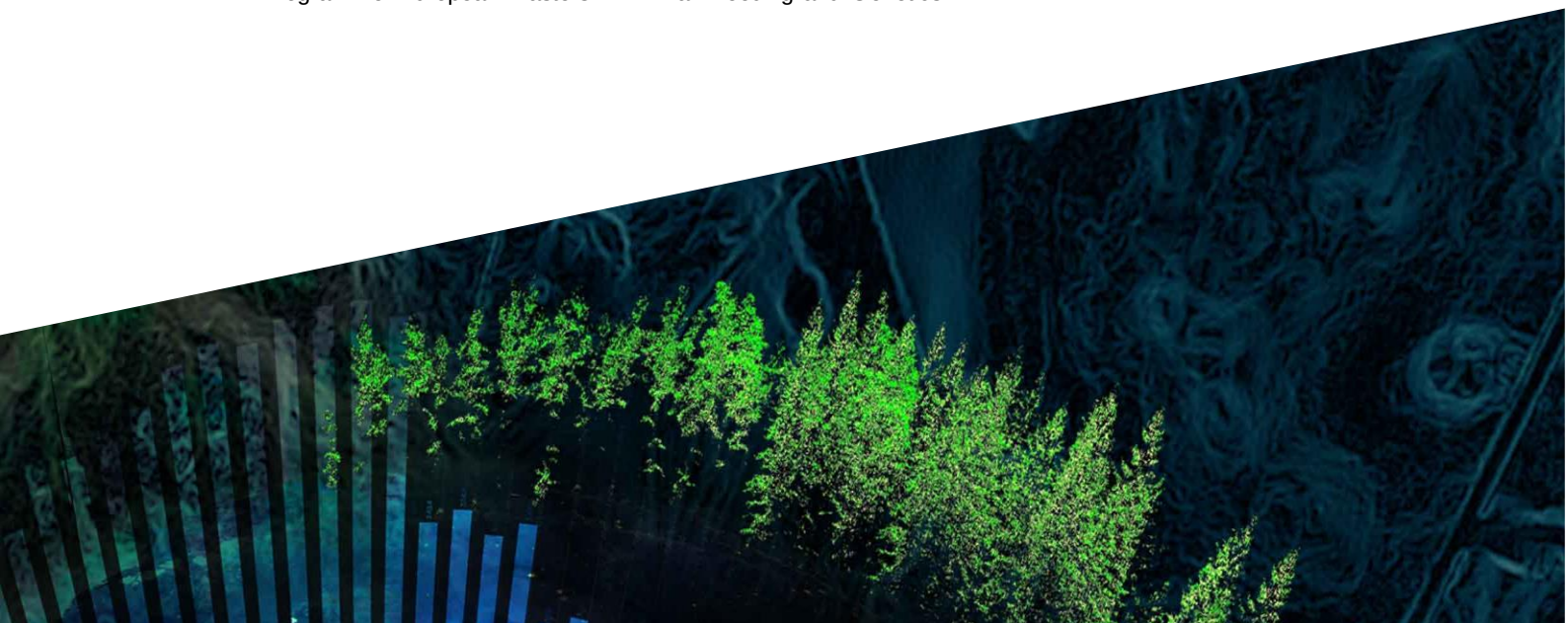
Estimating Variant Intolerance for Genes of Cattle

Simon Lanigan

Supervisor: Martin Johnsson

Co-Supervisor: Martijn Derks

Swedish University of Agricultural Sciences, SLU
Department of Animal Breeding and Genetics
Programme: European Masters in Animal Breeding and Genetics



EUROPEAN MASTER IN ANIMAL BREEDING AND GENETICS

Estimating Variant Intolerance for Genes of Cattle

Simon Lanigan

January 2024

Main supervisor: Martin Johnsson (SLU)

Co-supervisor: Martijn Derks (WUR)

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

Estimating Variant Intolerance for Genes in Cattle

Author's name:	Simon Lanigan
Supervisor:	Martin Johnsson, SLU, Uppsala. Department of Animal Biosciences
Assistant supervisor:	Martijn Derks, Wageningen University and Research, The Netherlands.
Examiner:	Anna Maria Johansson, SLU, Sweden.

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

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

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 you all need to agree on a decision. Read about SLU's publishing agreement here: https://www.slu.se/en/subweb/library/publish-and-analyse/register-and-publish/agreement-for-publishing/.

☒ 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.

Abstract

Cattle play a crucial role in providing nutrition and economic stability worldwide. Threats to their health, particularly infectious diseases, and genetic mutations, can have profound implications for both animal and human livelihoods. With advancements in high-throughput sequencing technologies, the wealth of genomic information available has enabled researchers to explore the relationships between genes, diseases, and phenotypic traits in cattle. Genes vary in their tolerance to mutation, influenced by factors such as functional importance, evolutionary conservation, redundancy, genomic location, presence of mutational hotspots, evolutionary history, regulatory elements, and their essentiality for fundamental biological processes. Using data from the 1000 Bull Genomes Project, encompassing over 1000 individual cattle, a comprehensive analysis of genic functional intolerance within the bovine population was conducted. The primary objective was to assess the tolerance levels of protein-coding bovine genes to functional variation, ranking them based on their level of tolerance using the RVIS method and by analysing the variants present within the dataset. Computational tools such as VEP, PLINK, PANTHER and GALLO were used to prioritize variants and determine their level of tolerance to mutation. The scoring and ranking of genes allowed for the identification of subsets of top and bottom genes, which were subject to further analysis. This contributed to the broader understanding of genomic variants and the impact in which they could potentially have in bovine animals, alongside the implications for selective breeding, disease resistance, and overall herd management, ultimately contributing to the sustainable and resilient future of cattle populations worldwide.

Contents

List of Figures and Tables

Introduction

Domesticated cattle worldwide provide a significant source of both nutrition and economic viability to humankind worldwide (Bradford, 1999). Therefore, any infectious diseases or mutations within genes that could affect phenotypic production traits could potentially threaten the livelihoods of people and other animals (Raszek et al., 2016).Increasing evidence indicates that genes containing disease causal variation have distinct functional and genomic properties (Collins, 2015). Modern technology has allowed high throughput sequencing which allows scientists to have access to a large quantity of genomic information. As a result, an increase in data has abled research to differentiate the relationship between genes and underlying diseases and pathogens which effect the phenotypes of cattle (Raszek et al., 2016). Aiding the understanding of genomic and functional properties of within and between these genes.

The availability of Whole Genome Sequence data has opened new avenues into researching the genetic diversity that exists within and between populations and chromosomal segments of livestock species (Zhang et al., 2015). Analysing genetic variation within sequence data is the basis of selective breeding in livestock and crop species. Genetic variants can result in altered protein structures and can result in altered gene expressions or the protein structure, which is believed to be the main cause for variation that exists within complex traits (Ros-Freixedes et al., 2022). Through sequencing DNA, different mutations across the bovine genome can be identified, which allows for comparisons of gene expression between and across species. This can provide a greater insight into different biological processes and traits can be determined (Raszek et al., 2016).Sequenced data identifies a vast amount of variants of ambiguous significance, however computational tools are needed to accelerate their level of significance (Carter et al., 2013). Therefore, determining which mutations most likely influence disease and their detrimental effect is a major challenge in interpreting genomes (Petrovski et al., 2013).

Until recent years, the clarification of the prioritization process for genes harbouring variants within the genomes of the *Bos taurus* and *Bos indicus* species has been relatively underexplored (Petrovski et al., 2013). As a result, in 2012, the 1000 bull genome project was founded to aid in the global understanding of bovine genetics and to incorporate international collaboration on bovine research (Illumina Inc, 2023). The 1000 Bull Genomes Project is a collection of whole-genome sequences from, what is now and has increased to, 2,703 Bovine individuals from all over the world which capture a significant proportion of the world's cattle diversity. So far, 84 million single-nucleotide polymorphisms (SNPs) and 2.5 million small insertion deletions have been identified in the collection of WGS data alongside a very high level of genetic diversity. The project has greatly accelerated the identification of deleterious mutations for a range of genetic diseases, as well as for embryonic lethals (Hayes & Daetwyler, 2019).

For this analysis, the publicly available data from the 1000 bull genome project will be used to rank all protein-coding bovine genes available in the dataset in terms of their tolerance levels to withstanding functional variation. This dataset includes just over 1000 bovine animals, from

different countries worldwide. Therefore, using this data, an assessment of the overall genetic variation within these genes and how well they tolerate changes without the potential of mutation affecting their function will be carried out. This method will evaluate genes that are needed for biological processes, often in which some genes undergo stronger evolutionary pressure compared to other genes that remain unchanged due to their role within the organism. Therefore, it is hypothesised that genes harbouring a greater diversity of functional variations crucial for survival are more likely to withstand mutations with minimal impact on the organism's health and fitness. This is often reflected in metrics such as the Residual Variation Intolerance Score (RVIS), which evaluates genes based on the observed versus expected number of variants. Lower RVIS scores suggest that a gene is more intolerant to variation than expected, potentially indicating its essentiality for basic cellular functions and highlighting its importance within the genome.

The process of developing a score to rank genes and their variants in terms of their influence on disease susceptibility involves comparing the frequency of variants observed in the genome to the presence of causal variants. Subsets of top and bottom gene names were subsequently be analysed and enriched to analyse the biological effect in which they may have. The purpose of Scoring and Ranking genes based on their tolerance to functional variation will hopefully enhance the understanding of the bovine genome (Gussow et al., 2016), and allow us to compare to other species, such as *Homo sapiens* to give us a contextualised background into evolution and adaption of genes through understanding the necessary genes for phenotypic function within their environment.

Background Literature

Genome sequencing yields extensive knowledge of genetic variation. However, identifying the phenotypically causal variants among the many variants present within the bovine genome remains to be a challenge (Cooper & Shendure, 2011). DNA sequencing has provided a comprehensive genetic map of variation in the genome which includes several million single nucleotide variants (SNVs), thousands of insertion or deletion evens alongside structural variants present in a genome (Bentley et al., 2008). Most of which are common events but often genomes may also contain rare and disease-causing variants when analysing at an individual or at a population level (Durbin et al., 2010). The identification of these variants remains a challenge (Petrovski et al., 2013). Using genetic approaches such as genome wide association studies can identify candidate variants but are not enough to identify the causal variant (Lander, 2011). This limitation arises partly due to linkage disequilibrium (LD), a phenomenon where certain genetic variants tend to be inherited together. LD makes it relatively easy to identify associated genomic regions but challenging to pinpoint the specific causal variant responsible for the observed association. Therefore, the genetic markers that are statistically associated with a disease, do not tell us which specific variant is directly responsible for the association. In order to review and conduct the analysis, the variants present within the data from the 1000 bull genome data were annotated. This was done as gene and

single nucleotide polymorphisms (SNP) annotations is among the first and most important steps in analysing a bovine genome (Florea et al., 2011). Identifying SNPs within a genomic sequence can reveal many features and functions encoded within a genome (Gregory et al., 2006).

When looking at trait association with genes, there are many variants, each of small effect, which contribute to the overall variation present. As a result, a very large sample size is needed to find significant associations that explain most of the observed genetic variation (Xiang et al., 2019). Most research has been conducted in humans due to large amount of data available and advances in research in the area. However, on the other hand, not much research has been conducted in cattle. There are over 1.46 billion cattle worldwide (FAO, 2017), and an increasing amount are being genotyped as well as phenotyped. Cattle have been domesticated from 2 subspecies (*Bos taurus)* and (*Bos indicus*),which diverged 0.5 million years ago from extinct wild aurochs(*Bos primigenius*) (MacHugh et al., 1997). Due to the increasing amount of information available, cattle are becoming comparable species to that of human, in terms of the overall amount of genetic resources available (Xiang et al., 2019). Furthermore, as a result of the already existing knowledge of phenotypic information due to research for rearing cattle for the agricultural industry worldwide, associations between genes and the phenotype can be made. Nonetheless, identification of the causal variants for a complex trait is still difficult and can also be caused by the small effect size of most causal variants and the Linkage Disequilibrium between variants. Consequently, there are usually many variants in high LD, any one of which could be the cause of the variation in phenotype (Xiang et al., 2019). Prioritization of these variants can be aided by functional information on genomic regions. For example, mutations that change an amino acid are more likely to affect the phenotype than synonymous mutations, and as mentioned before, a genome wide association study in-itself is not sufficient enough to characterize the causal variants.

Scoring genes based on the level of mutation observed using the Residual Variant Intolerance Score (RVIS) method can provide valuable insights into the functional role of genes (Petrovski et al., 2013). Using the RVIS method, one can identify genes that are more or less tolerant to mutation, suggesting their potential functional importance or prioritizing mutating variants. The RVIS method is based on identifying genes that are more or less tolerant to mutation, which suggests their potential functional importance by prioritizing mutating variants. Genes with a lower tolerance score to are more likely to harbour functional mutation than their counterparts with higher scores (Petrovski et al., 2013). Other research methods have been conducted looking at probability of loss-of-function intolerant (pLI) scoring of a gene set, which is designed to assess the intolerance of a gene to loss-of-function mutations. Thus, changes in the DNA sequence that result in a non-functional or impaired protein. This differs from the RVIS method which accounts for the difference between the observed and expected number of all variants in a gene, whereas pLI just focuses on loss-of-function variants. Similar to the RVIS method used by (Petrovski et al., 2013), in the pLI method, identifying genes whose loss is likely to decrease fitness is used to assess whether disrupting mutations are found at lower frequencies than expected in comparison to a reference model (Fuller et al., 2019).

The genetic dissection of quantitative phenotypes into Mendelian-like components, or quantitative trait loci (QTL) analysis, has provided significant insight into how complex traits are regulated and controlled (Rocha et al., 2002). As a result, one can understand interactions between genes and the environment and between genes and other genes. QTL mapping in livestock can also allow the identification of genes that determine the genetic variation affecting traits of economic interest (Miyata et al., 2007). Therefore, the identification of QTLs has significantly increased the potential of the improvement of bovine genetics through implementation of marker assisted selection (MacNeil & Grosz, 2002), and a lot of research is now focused to find markers associated with genes that can be associated traits so that they can be included in breeding programs (Khatkar et al., 2004). However, a QTL analysis will only identify a genomic region associated with a trait. Combining a QTL approach into RVIS scored gene may give a further insight into which traits are most effected by tolerant or intolerant genes.

Furthermore, comparing RVIS results of *bovine* species to that of their human orthologs can provide insights into the evolutionary constraints and functional importance across both species. Comparison of gene scores between species can highlight genes that are highly conserved due to their critical functions. It can also help identify genes highly conserved in one species, but not in another. Signifying the level of gene specific tolerance levels to mutation in and between species. This could indicate that these genes have undergone species-specific adaptation of a gene. If a gene is more constrained depending on the either the environment in which it is in or the level of selection pressure on that particular gene, then depending on the tolerance levels of the gene, it could be more or less susceptible to mutation. Alongside this, genes associated with diseases in humans often have counterparts in other species that may exhibit similar disease phenotypes. Comparing RVIS scores can help identify genes that are crucial for health and disease across different species. All in which allow for the understanding and research of genes across and within species. Comparison of genes between two species can provide valuable insights into the evolutionary conservation and divergence of genes under purifying selection across species. Therefore, investigating the process by which deleterious variants are removed from a population over time. Thus, natural selection preserving functional and well-adapted genetic traits while eliminating those that are detrimental to an organism's fitness. Investigating orthologous genes between bovine and humans allows us to assess the extent of evolutionary conservation in functional elements. Conserved RVIS scores may indicate shared constraints on genetic variation and underscore the importance of these genes in fundamental biological processes. Comparative analysis can reveal if genes with specific RVIS scores are enriched in certain biological pathways or functions that are conserved across species.

Material and Methods

Ranking and scoring variants present in genes was based on the Residual Variation Intolerance Score (RVIS) method, which is a framework that ranks protein-coding genes based on their tolerance level to withstanding functional variation, by comparing the overall number of observed variants in a gene to the observed common functional variants (Gussow et al., 2016). The data that is publicly available data from the 1000 bull genome project was used as this data is a capture of the proportion of worldwide cattle diversity for both the Bos Taurus and Bos Indicus breeds. Within the dataset used for this analyses, there were 497 individuals from Bos Taurus species, 438 Bos Indicus and 106 of unknown origin. The percentage of distribution within the data set used in terms of breed type is show in the figure below.

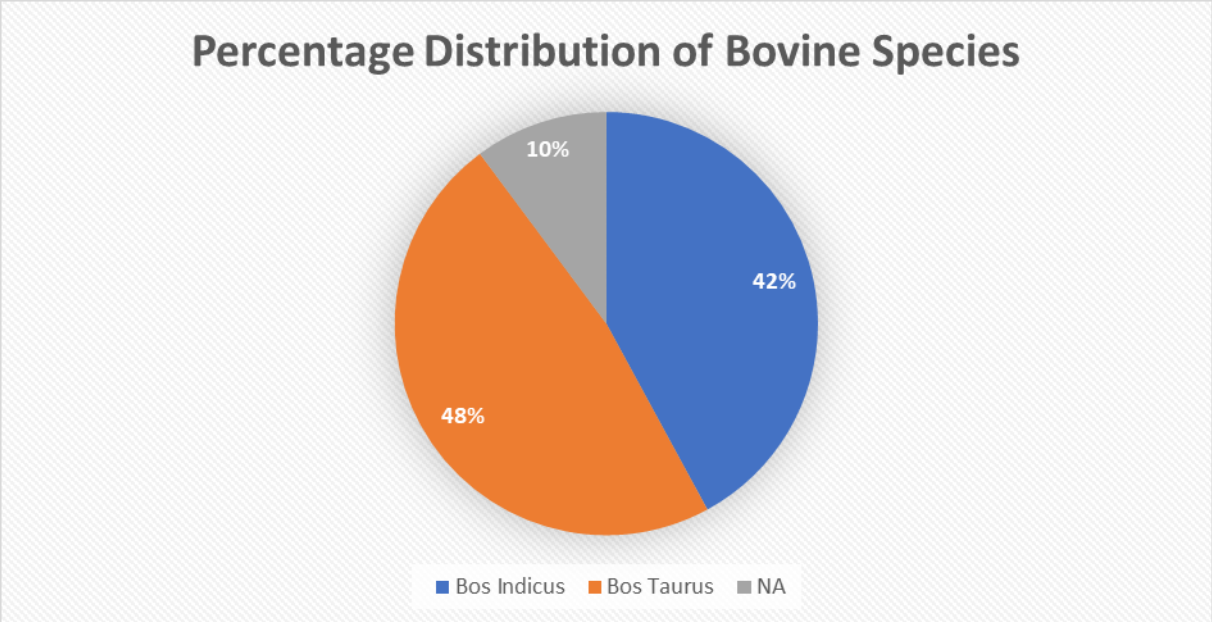


Figure One: Breed type representing worldwide cattle population and the percentage in which the breed makes up within the 1000 bull dataset used for the analysis.

The Variant Effect Predictor calling (VEP)(McLaren et al., 2016) was used to annotate the data to allow for the analyses and manipulation of the genotyped markers from each of the chromosomes of the 1038 cattle in the dataset. PLINK Software (S. G. Gregory et al., 2006) was then used to calculate the Minor Allele Frequency of the variants identified from PLINK. To ensure data quality, genotype data underwent filtering to exclude SNPs with a minor allele frequency (MAF) of less than <0.1, similar to that carried out by *Rajavel et al., 2022 and Petrovski & Wang, 2013*. This threshold of 0.1 was chosen to distinguish between common and rare variants, with a focus on variants deemed biologically significant. SNPs below this threshold were considered rare and thus provided less robust data for further analysis.

Deriving the Residual Variant Intolerance Score

Scoring genes based on the level of mutation observed using the Residual Variant Intolerance Score (RVIS) method can provide valuable insights into the functional role of genes (Petrovski et al., 2013). Using the RVIS method, one can identify genes that are more or less tolerant to mutation, suggesting their potential functional importance or prioritizing mutating variants. Genes with a lower tolerance score to are more likely to harbour functional mutation than their counterparts with higher scores (Petrovski et al., 2013). Therefore, it can be said that analysing the mutational intolerance of genes can provide insights into conservation of genes throughout evolution and tell us more about gene function and importance in the organism. Therefore, the 1000 bull dataset was used to assess the degree to which genes have more or less common functional variation than expected for the genome as a whole given the amount of presumably neutral variation in which they carry. Using a similar method to that of (Petrovski et al., 2013), a regression model was used to score the genes. Using Y as variants that have a greater allele frequency >0.1 and synonymous variants removed. X was defined as the total number of protein coding variants within the gene, including synonymous variants and disregarding allele frequency. Therefore, they are less likely to have a direct impact on the function of the protein produced, and overall purpose of this analysis. Overall, the aim was to prioritize the variants that are more likely to have functional consequences which can be identified in the phenotype and that can be subsequently analysed and discussed further. A count was then derived calculating the total number of X and Y. To derive the RVIS score of the gene, Y was then regressed on X using the MASS package with the *studres* function in R alongside its built-in statistical analysis packages for *lm* regression modelling (R Core Team, 2020).

The 'MASS' package then takes the residual of each gene, and it is divided by an estimate of the overall standard deviation of all the residuals of the genes based on their variant counts, considering the leverage of each of the data points. This accounts for the differences in variability that come with differing mutational burdens. Therefore, standardising and normalizing the residuals of each calculated genes linear regression residual score, considering for differences in gene expression that is influenced by genetic influence. This procedure of normalizing the data will allow for meaningful comparisons in the results. This subsequent RVIS score produced in the output is then a measure of the overall genome wide average number of common functional mutations found in genes with a similar amount of mutational burden. To avoid overlapping of NCBI genes and Ensembl, which could subsequently end up being the same gene, the dataset was further filtered to only include Ensembl genes. 22,706 Ensembl genes were subsequently identified within the dataset. For the purpose of keeping the genes of interest with the most amount of data available, the dataset was further filtered to only include Ensembl canonical transcripts genes. That is, keeping the genes with their relevant transcripts that are considered most representative or biologically relevant to reference phenotypes to maintain consistency in the overall analyses. From this, 22,620 unique genes were kept when filtered with Ensembl canonical genes, from the Ensembl online database. For the purposes of analytics, the top and bottom 10 percent of RVIS scores with subsequent gene

names were extracted from the dataset for further investigation. This was conducted to give us an idea of the most tolerant and intolerant genes within the dataset. The top and bottom genes were then analysed further using PANTHER for gene ontology and Gallo for gene set enrichment.

When analysing the RVIS scores produced, genes with a score of less than 0, means that the gene has less common functional variation than predicted in comparison to scores greater than 0, which means that there is more functional variation than predicted. If there is more functional variation than predicted, it suggests that the genes are more tolerant to variation than expected based on the model provided. PANTHER was used to gain insight into the biological relevance of genomic data derived from the 1000 bull dataset, with a comprehensive look at the top 10% and bottom 10% of genes. To achieve this, a PANTHER GO Biological Process Overrepresentation Test (PANTHER version 18.0 Released 2023-08-01) using the Fisher Exact test. This test was aimed to look at the significance of over-representation and under-representation of gene function within the sub-set data. The reference list for this analysis consisted of Bos Taurus genes, derived from the PANTHER database, which ensured a relevant comparison. The choice of the Fisher exact test was motivated by the smaller sub-set data of the top and bottom proportion of genes and its suitability for categorical data.

The QTL enrichment analysis within this analysis was reviewed to see the traits associated with the derived RVIS scored genes from. Based on this information, it was then investigated the associated traits and areas of overall bovine performance that have higher or lower RVIS scores, and the impact in which this may have. To do this, the R-package 'GALLO' was used (Fonseca et al., 2020). GALLO was used for the integration of multiple data sources in livestock for positional candidate loci and was used for QTL annotation and enrichment of the RVIS scored genes.

For comparison to human orthologs, the derived bovine RVIS scores from the previous steps were aligned to the human RVIS scores (Wang et al., 2023). The BioMart database was used to get Human Genes with corresponding cattle names and ensemble IDs which allowed the production of a translation table of human and bovine gene identifiers which were merged with the same Ensembl genes present from RVIS scored genes, and subsequently analysed. Outlier genes that scored high in one species and low in the other species were focused on and functional relevance analysed.

Results

The RVIS values were calculated to assess the tolerance of these genes to genetic variation, shedding light on their potential functional significance. 22,620 unique genes were scored, which were then further subset to look at the top and bottom genes of interest. As previously mentioned, genes with a score of less than 0, means that the gene has less common functional variation than predicted in comparison to scores greater than 0, which means that there is more functional variation than predicted. If there is more functional variation than predicted, it

suggests that the genes are more tolerant to variation than expected based on the model provided. Figure 2 is a representation of the number of variants present in the dataset. Where X is the total number of variants, including synonymous variants and disregarding allele frequency. Y and the filtered number of variants that have a greater allele frequency >0.1 and synonymous variants removed.

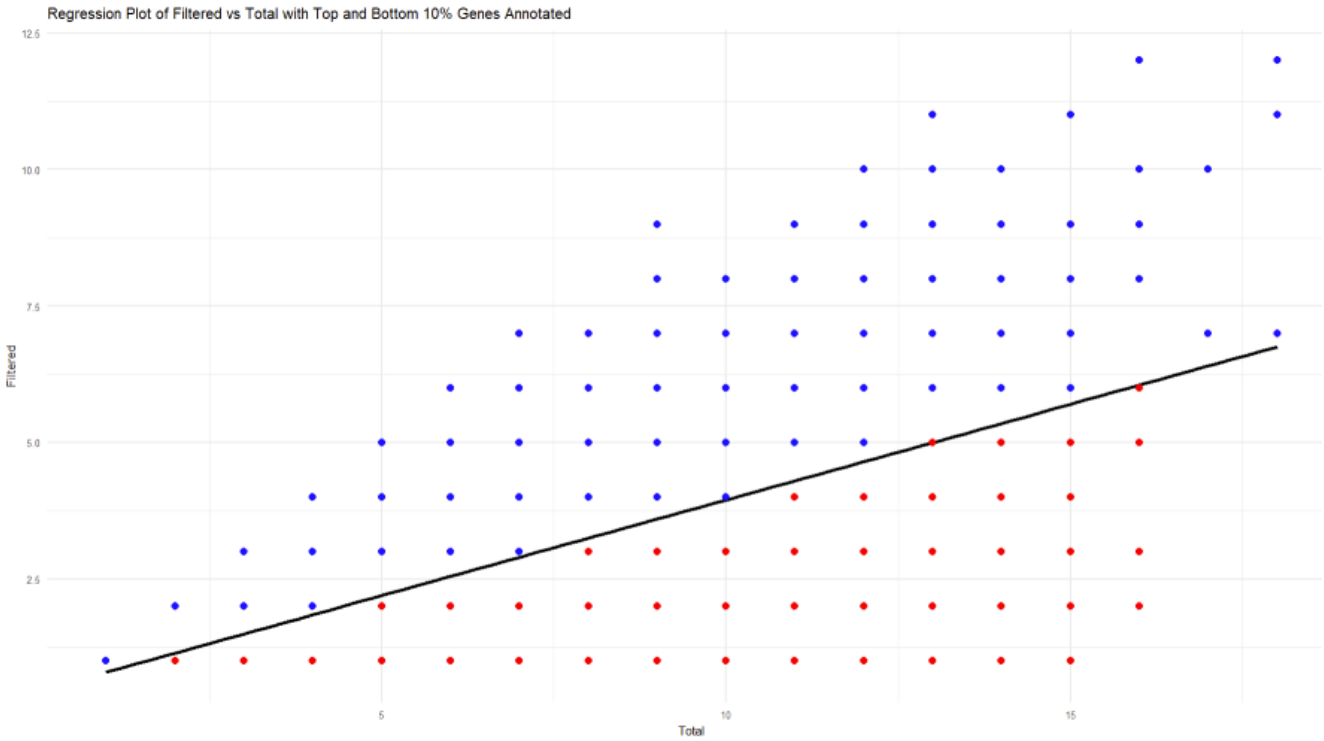


Figure 2: A regression plot illustrating the total number of variants present in the gene vs the number of filtered variants present, which include the top and bottom 10% most extreme genes. Regression line represented in black, which summarizes the relationship between the independent variables and dependant variables, in this case the total number of variants versus the filtered number. (Y and X). Red=10% of genes most intolerant, Blue=10% of genes most tolerant.

The calculated RVIS scores ranged from -4.288 to +3.158. Table 1 and Table 2 below show the highest ranking 10 genes per bottom and top RVIS scored genes as a result of the studentized variant intolerance scoring process. In each of the tables, the gene name is presented in Ensembl format, alongside the RVIS score represented by studentized residuals. The total variants column is the total amount of variants present per gene, and the filtered variants column is the total number of variants present per gene after variants that have a greater allele frequency >0.1 and synonymous variants were removed.

Table 1: Bottom RVIS Scored Genes showing gene name in Ensembl format, alongside the calculated RVIS score in the studentized residual column. The total amount of variants present per gene are represented in the total variant's column, with the filtered variant column showing the amount of variants when synonymous variants were removed, had allele frequency above >0.1 and had a High, Medium, or Low impact. Finally, the gene association column is the associated description for that protein.

GENE	*Studentized Residuals*	*Total Variants*	*Filtered Variants*	*Gene Association*
ENSBTAG00000013334	-4.286706	15	1	*CSF3R*
ENSBTAG00000040318	-3.950731	14	1	*Novel Gene*
ENSBTAG00000014643	-3.950731	14	1	*EEF1D*
ENSBTAG00000011642	-3.950731	14	1	*RBM27*
ENSBTAG00000011491	-3.950731	14	1	*WDR81*
ENSBTAG00000046684	-3.945329	16	2	*FOXN3*
ENSBTAG00000055316	-3.614764	13	1	*SNRPD1*
ENSBTAG00000050334	-3.614764	13	1	*CD83*
ENSBTAG00000035587	-3.614764	13	1	*ARRB2*
ENSBTAG00000020860	-3.614764	13	1	*CHD1*

Table 1 represents genes which are RVIS scored to be intolerant to functional variation. Included within the table is the *CSF3R* gene which is the receptor for colony stimulating factor 3, a cytokine that controls the production, differentiation, and function of granulocytes. Alongside this, the *EEF1D* gene encodes a subunit of the elongation factor-1 complex, which is responsible for the enzymatic delivery of aminoacyl tRNAs to the ribosome, and the *RBM27* gene enables RNA binding activity (Sayers et al., 2022). Interestingly, these genes are mainly involved in crucial cellular processes such as cytokine signalling (*CSF3R*), protein synthesis (*EEF1D*), and RNA binding (*RBM27*), which are all processes involved in normal cellular function. The intolerance to variation in these genes indicates that they are sensitive to genetic changes if mutation occurs. This sensitivity implies that changes in these genes may have significant consequences for their functionality.

Table 2: Top RVIS Scored Genes showing gene name in Ensembl format, alongside the calculated RVIS score in the studentized residual column. The total amount of variants present per gene are represented in the total variant's column, with the filtered variant column showing the number of variants once genes with synonymous variants were removed, had allele frequency above >0.1 and had a High, Medium and Low impact. Finally, the gene association is shown in the last column.

GENE	Studentized Residuals	Total Variants	Filtered Variants	Gene Association
ENSBTAG00000014365	3.158	13	11	*Novel Gene*
ENSBTAG00000052577	3.147	9	9	*Pseudogene*
ENSBTAG00000053748	3.147	9	9	*Pseudogene*
ENSBTAG00000001476	2.827	16	12	*Novel Gene*
ENSBTAG00000007075	2.82	16	12	*Novel Gene*
ENSBTAG00000020116	2.827	16	12	*JSP.1*
ENSBTAG00000019884	2.816	12	10	*PKD1L3*
ENSBTAG00000023309	2.486	15	11	*Novel Gene*
ENSBTAG00000038797	2.486	15	11	*Novel Gene*
ENSBTAG00000006748	2.480	13	10	*DMXL1*

The presence of pseudogenes, which are non-functional copies of genes due to accumulated mutations, having high RVIS scores doesn't necessarily mean these genes are still being actively maintained or tolerating mutations. Instead, it reflects the evolutionary history of the original functional gene. Pseudogenes form when mutations make a gene non-functional, allowing more mutations to accumulate over time without being filtered out by natural selection. So, while pseudogenes might show high RVIS scores, it doesn't mean they're still functionally important or adapting to mutations; it just shows the legacy of the original gene's evolutionary journey.

The identification of novel genes such as ENSBTAG00000001476, ENSBTAG00000007075, and ENSBTAG00000038797 were found to include genes that code for proteins that are involved in the presentation of antigens to the immune system, otherwise known as Major Histocompatibility Complex (MHC) genes. This underscores the essential role in which they have in vertebrate immune systems. Their high variant intolerance levels suggest a heightened adaptability to various pathogens and environmental challenges, indicative of the critical importance of their functionality in the bovine immune system. Moreover, their increased tolerance to mutation further highlights their evolutionary significance in maintaining immune system integrity and response efficiency. However, understanding the functional significance of the genes is essential to the complex biological and cellular processes involved. The PANTHER database was a vital resource database used and is designed to facilitate gene enrichment analysis. For this, the top 10% and bottom 10% of genes were analysed against the

reference cattle genes in the database. This analytical method plays a crucial role in determining whether specific genes in the dataset were biologically relevant and to see if they are differentially expressed or not. PANTHER, as a comprehensive online gene analysis platform, integrates the ability to analyse gene function, ontology, pathways, and statistical analyses tools that enable analyses of large-scale genome wide data (Mi et al., 2013). The outcome of the analysis can be seen in table 1 and table 2.

Table 3: Bottom 10% of Genes from PANTHER Gene Ontology showing the functional relevance of the genes present in the Dataset in comparison to reference, alongside their if they are over or underrepresented within the data.

PANTHER GO-Slim Biological Process	Ref	Data Set	expected	Fold Enrichment	+/-	P value
Unclassified	10096	716	825.03	.87	-	0.00E00
regulation of macromolecule metabolic process	2794	307	228.32	1.34	+	6.82E-04
regulation of nitrogen compound metabolic process	2573	285	210.26	1.36	+	1.04E-03
regulation of RNA metabolic process	2028	233	165.72	1.41	+	1.23E-03
regulation of nucleobase-containing compound metabolic process	2103	240	171.85	1.40	+	1.48E-03
biological process	13742	1232	1122.97	1.10	+	2.54E-03
regulation of primary metabolic process	2612	285	213.45	1.34	+	3.75E-03
regulation of gene expression	2349	259	191.96	1.35	+	5.77E-03
regulation of RNA biosynthetic process	1879	213	153.55	1.39	+	1.01E-02
regulation of DNA-templated transcription	1877	212	153.39	1.38	+	1.24E-02
regulation of metabolic process	2973	314	242.95	1.29	+	1.28E-02
regulation of cellular biosynthetic process	2010	224	164.25	1.36	+	1.57E-02
regulation of macromolecule biosynthetic process	2005	223	163.85	1.36	+	1.87E-02
regulation of biosynthetic process	2037	224	166.46	1.35	+	3.77E-02
regulation of transcription by RNA polymerase II	1592	181	130.10	1.39	+	4.86E-02

Table 1 shows the bottom 10% of genes from the Panther Go-Slim Biological Process results. Go-Slim Biological test is a subset of the gene ontology that consists of a reduced set of terms in order to provide a broader overview of the functional categories. The P-value signifies the level of statistical significance of the enrichment. The smaller the level of significance, the more unlikely the observed enrichment has occurred by random chance. Within the bottom 10% of scored genes, we can see that the regulation of different metabolic processes, regulation of gene expression and biological processes are most significant. Meaning that, the genes within this subset of data have an important role within these categories. The +/- indicates the direction of enrichment. The "+" denotes that the genes are over-represented in the data set while a "-" indicates underrepresentation.

Table 4: Top 10% of genes from PANTHER gene ontology alongside their functional relevance as a result of the PANTHER GO-Slim Biological Process.

PANTHER GO-Slim Biological Process	Ref	Data Set	Expected	Fold Enrichment	+/-	P value
Unclassified	10096	605	695.01	.87	-	0.00E00
cellular process	9090	741	625.75	1.18	+	4.13E-05
sensory perception of chemical stimulus	478	8	32.91	.24	-	1.51E-03
regulation of transcription by RNA polymerase II	1592	62	109.59	.57	-	1.86E-03
biological_process	13742	1036	945.99	1.10	+	2.26E-02
sensory perception	535	13	36.83	.35	-	2.56E-02
establishment of localization	1957	187	134.72	1.39	+	2.75E-02
cellular component organization or biogenesis	2592	237	178.43	1.33	+	2.97E-02
transmembrane transport	606	73	41.72	1.75	+	3.02E-02

The results of the statistical over-representation analysis in the biological enrichment examination provide information on the functional distinctions between genes identified as top and bottom scorers. This analysis discloses the biological processes associated with these gene subsets, indicating either heightened (over-represented) or diminished (under-represented) number of genes compared to random expectation. In table 4, the top 10% of RVIS scored genes are shown in relation to their biological relevance. As these are the top scored genes, they are most likely more tolerant to mutation than those shown in table 3 and are thus, less conserved. Activities related to cellular processes are most conserved and overrepresented in the dataset, outlining their functional relevance. This is followed by two classes of sensory perception which although significantly relevant are both under-represented in the dataset in comparison to the reference genome dataset.

Following this, The QTL analysis results were reviewed from GALLO to see the traits associated with the derived RVIS scored genes. The associated traits were investigated and areas of overall bovine performance that have higher or lower RVIS scores, and the impact in which this may have been studied. Combining a QTL approach into RVIS scored gene gave a further insight into which traits are most effected by tolerant or intolerant genes.

Table 5: QTL enrichment results from Bottom 10% of RVIS Scored Genes

QTL	N_QTLs	N_QTLs_db	Total_annotated_QTLs	Total_QTLs_db	pvalue	adj.pval	QTL_type
Non-return rate	2119	2312	80558	161221	0.00E+00	0.00E+00	Reproduction
Milk myristic acid content	708	902	80558	161221	3.58E-70	8.25E-68	Milk
Interval from first to last insemination	379	445	80558	161221	5.68E-55	8.73E-53	Reproduction
Milk fat yield	4773	8221	80558	161221	1.08E-51	1.24E-49	Milk
Milk C14 index	2657	4437	80558	161221	2.24E-41	2.07E-39	Milk
Milk myristoleic acid content	1867	3047	80558	161221	6.72E-37	5.17E-35	Milk
Milk capric acid content	640	912	80558	161221	1.69E-35	1.11E-33	Milk
Interval to first estrus after calving	716	1053	80558	161221	1.24E-32	7.17E-31	Reproduction
Milk zinc content	297	384	80558	161221	2.27E-28	1.17E-26	Milk
Longissimus muscle area	856	1328	80558	161221	8.60E-27	3.96E-25	Meat and Carcass
Gestation length	430	629	80558	161221	6.34E-21	2.66E-19	Reproduction
Milk lauric acid content	268	366	80558	161221	9.05E-20	3.48E-18	Milk
Milk yield	3440	6189	80558	161221	1.08E-19	3.82E-18	Milk
Average daily gain	1817	3140	80558	161221	1.95E-19	6.44E-18	Production
Milk protein yield	1749	3093	80558	161221	7.82E-14	2.40E-12	Milk

Table 5 shows the QTL enrichment results of the bottom 10% of RVIS scored genes from GALLO. The dataset includes the number of identified QTLs (N_QTLs), both within this study (N_QTLs) and in external databases (N_QTLs_db), which provides a comprehensive overview of the genomic loci associated with the traits under investigation. Focusing on annotated QTLs (Total_annotated_QTLs) to enhance the precision of our findings due to additional information being present in the database, and the total QTLs present in external databases (Total_QTLs_db) serve as a reference for the broader genetic landscape. The statistical significance of these associations can be seen through p-values, and the adjusted p-values (adj.pval) account for multiple comparisons. The classification of QTLs into distinct types (QTL_type) allows for an interpretation of the genetic architecture underlying various traits. This comprehensive QTL analysis provides a foundation for understanding the genetic contributions to the traits studied, which can aid future research and breeding strategies. The bottom scored RVIS traits in table 3 above, which include reproductive efficiency and milk quality traits collectively highlight the genetic vulnerability and sensitivity to mutations in crucial agricultural and livestock characteristics. The traits related to milk and reproduction exhibit consistently significant p-values, emphasizing the substantial genetic influence on these traits, which therefore, could be subject to more variation with an increase in selection pressure is on them. Notable to the fact that many of these traits have economic consequences in terms of production.

Table 6: QTL enrichment results from Top 10% of RVIS Scored Genes

QTL	N_QTLs	N_QTLs_db	Total_annotated_QTLs	Total_QTLs_db	pvalue	adj.pval	QTL_type
Milk C14 index	3569	4437	84784	161221	0.000000e+00	0.000000e+00	Milk
Milk myristoleic acid content	2505	3047	84784	161221	3,92E-256	8,92E-254	Milk
Milk fat yield	5820	8221	84784	161221	2,06E-255	3,12E-253	Milk
Non-return rate	1984	2312	84784	161221	1,52E-249	1,73E-247	Reproduction
Milk yield	4371	6189	84784	161221	3,22E-185	2,93E-183	Milk
Tenderness score	2566	3483	84784	161221	1,00E-140	7,58E-139	Meat and Carcass
Milk C16 index	1601	2002	84784	161221	2,90E-139	1,89E-137	Milk
Milk palmitoleic acid content	1837	2422	84784	161221	8,15E-119	4,64E-117	Milk
Milk myristic acid content	799	902	84784	161221	7,30E-114	3,69E-112	Milk
Milk capric acid content	793	912	84784	161221	4,35E-103	1,98E-101	Milk
Gestation length	563	629	84784	161221	4,77E-83	1,97E-81	Reproduction
Connective tissue amount	2184	3142	84784	161221	9,76E-79	3,70E-77	Meat and Carcass
Milk protein yield	2085	3093	84784	161221	4,25E-58	1,49E-56	Milk
Milk protein percentage	5382	8796	84784	161221	8,48E-57	2,76E-55	Milk
Milk caprylic acid content	653	811	84784	161221	4,23E-56	1,28E-54	Milk

Table 6 shows the output from GALLO QTL enrichment output and provides information about significant traits based on QTL enrichment results. QTL trait enrichment analyses in bovine cattle can identify genomic regions associated with economically important production traits such as milk yield, meat quality, and reproductive performance. This information is valuable for selective breeding programs aimed at enhancing livestock productivity. The information in the table provides us with insights into traits in which may be more tolerant to mutation due to their higher RVIS score. Although more milk quality traits seem to be present in table 6 than in table 5 it is noticeable that some traits overlap between the two. QTL traits that exhibit both high and low RVIS scores suggests a more complex genetic landscape with potential implications for trait variation and adaptability. RVIS scores are used to assess the intolerance of genes to functional genetic variation, and they are indicative of how much genetic diversity a gene can tolerate without affecting its function. However, when traits are high and low scored, it could suggest adaptability and flexibility of the trait through evolutionary pressures evolutionary pressures, the genomic region in which they are in within the organism, or their functional relevance in the population.

Comparison to Human Orthologous Genes

Investigating orthologous genes between bovine and humans allows us to assess the extent of evolutionary conservation in functional elements. Comparative analysis can reveal if genes with specific RVIS scores are enriched in certain biological pathways or functions that are conserved across species. The BioMart database was used to get Human Genes with corresponding cattle names and ensemble IDs which allowed the production of a translation table of human and bovine gene identifiers which were merged with the same ENSEMBL genes present from RVIS scored genes, and subsequently analysed.

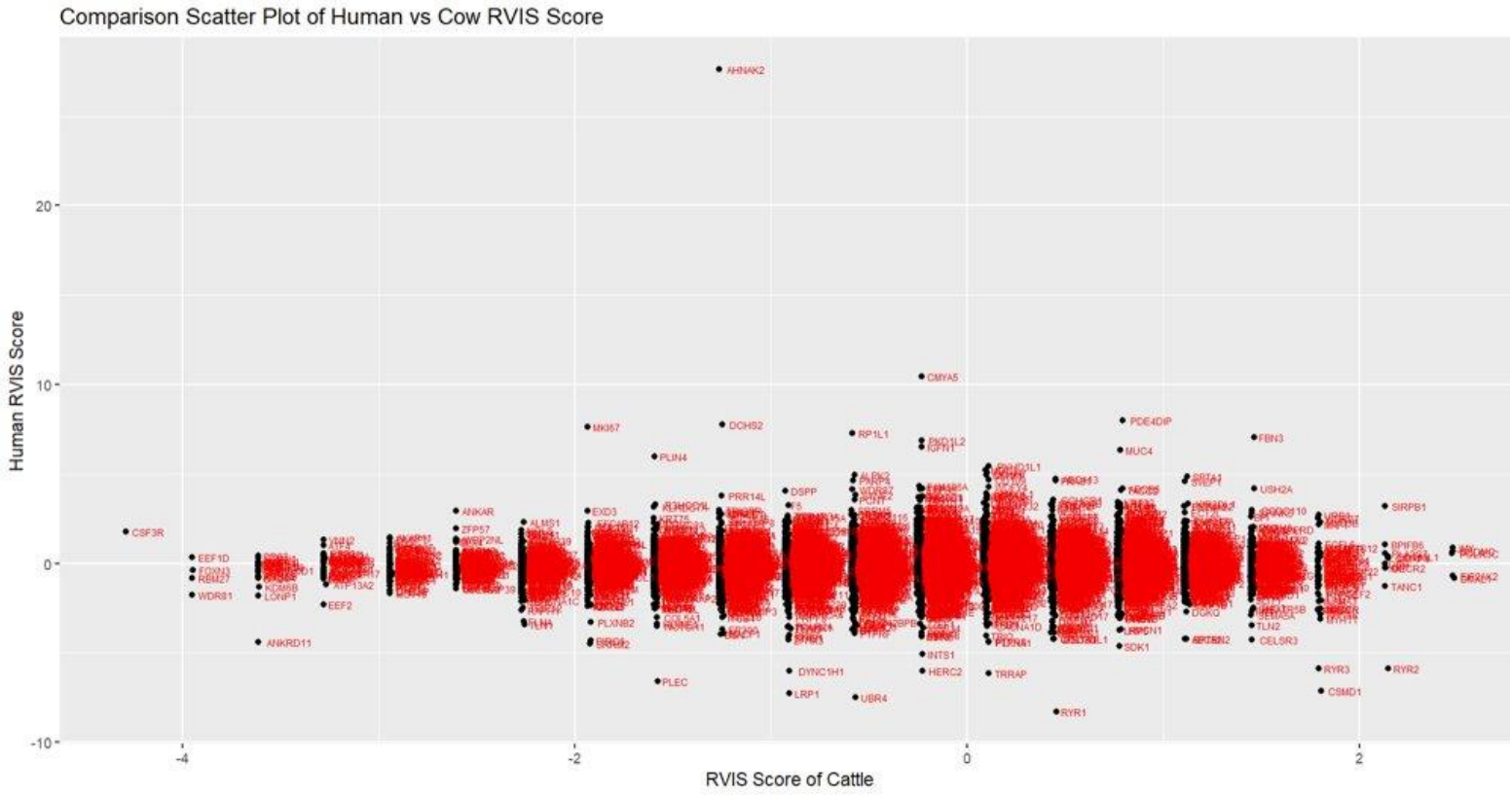


Figure 3: A Scatter Plot which represents the RVIS score of Cattle (x-axis) in comparison to Human (y-axis).

Figure 3 illustrates a scatter plot of RVIS scored human versus bovine genes. Although many clusters of genes are present with similar scores for each species, some outliers prevail. These include the *AHNAK2* gene in human which is a novel prognostic marker and correlates with immune infiltration in papillary thyroid cancer, the RYR2 gene which if mutations occurs are associated with stress-induced polymorphic ventricular tachycardia and arrhythmogenic right ventricular dysplasia, and the CSDM1 gene which is involved in several processes, including learning or memory, mammary gland development involved in pregnancy and reproductive

structure development (Sayers et al., 2022).The calculated correlated coefficient score between the human and bovine RVIS scored genes was 0.143.

Discussion

The analysis of Residual Variation Intolerance Scores (RVIS) across a diverse set of genes associated with the 1000 Bull Genome Project aimed to identify key genes and their associated functions based on their tolerance to withstanding genetic variation within a diverse population of bovine cattle. A total of 22,620 unique genes were scored, and further investigation focused on the top and bottom genes of interest. In the context of RVIS, a score less than 0 indicates that a gene is more intolerant to withstanding functional variation and in contrast to this , a score greater than 0 suggests that a gene is considered more tolerant to genetic variation. Genes with higher RVIS scores are often associated with essential biological functions and are more conserved throughout evolution (Gussow et al., 2016). This is essential for the interpretation of results discussed below. From the analysis, genes associated with cellular processing showed low RVIS scores, indicating intolerance to functional variation. These genes play crucial roles in normal cellular function and, when mutated, may have significant consequences. On the other hand, top-scored genes included pseudogenes and immunity genes, suggested a higher tolerance to mutations. Highlighting the adaptability of immunity genes to different pathogens and their essential role in the immune system. PANTHER gene ontology results revealed that highly scored genes were associated with macromolecule and biosynthetic processes, emphasizing their capacity to endure and potentially adapt to mutations without compromising essential functions. In contrast, lower scored genes were linked to cellular regulatory processes, indicating potential vulnerability and reduced tolerance levels. QTL enrichment analysis demonstrated associations between highly scored genes and milk trait, and lower scored genes with fertility traits, showcasing potential constraints, and intensified selective pressures. Comparing RVIS scores between cattle and human orthologs revealed a weak positive correlation, suggesting some shared evolutionary constraints but would need to be reviewed at a per-gene level, as gene function could be species specific.

Calculated RVIS Scores

The calculated RVIS scores ranged from -4.288 to +3.158. As can be seen in table 1, a variety of cellular processing genes are present, in which their RVIS scores suggest them to be intolerant to functional variation. Interestingly, these genes are mainly involved in crucial cellular processes such as cytokine signalling (*CSF3R*), protein synthesis (*EEF1D*), and RNA binding (*RBM27*), which are all processes involved in normal cellular function. This is further backed up by the PANTHER gene ontology in table 3 where cellular processes are found to be most enriched. The intolerance to variation in these genes indicates that they are sensitive to genetic changes if mutation occurs. This sensitivity implies that changes in these genes may have significant consequences for their functionality. All these genes have significant

consequences within animal breeding practices and veterinary medicine. For example, *RBM27* which is involved in Cellular processing, including RNA binding, which plays a vital roles in reproduction. *RBM27* was found to be candidate for weight and male fertility traits in Nellore cattle, particular scrotal circumference in males (Utsunomiya et al., 2014). Furthermore, *RBM27* was also found to have differentially expressed mRNA isoforms in 12 milk somatic cells (Asselstine et al., 2022). The negative RVIS score indicating intolerance suggests that *RBM27* is less tolerant to genetic variations that could potentially have an impact on its function. The presence of various mRNA isoforms differentially expressed in milk somatic cells underscores the sensitivity of RBM27 to genetic variations, as alterations in its expression profiles may significantly influence its involvement in RNA-related activities. When looking at the QTL enrichment results in table 5, milk quality and reproduction are most dominant QTL trait types present, indicating again, the phenotypic traits in which genes such as the *RBM27* may affect. It is important to consider the animal impact this may have in terms of productivity, health and welfare, and economic sustainability to the farmer. Breeding for genetic resilience in these processes may positively impact productive success of the bovine animal, leading to healthier and more efficient herds.

The presence of highly tolerant immunity genes in table 2, including Major Histocompatibility Complex (MHC) genes, implies that these genes can undergo mutations without severely compromising their function due to their ability to withstand functional variation. MHC genes that are defined as a group of genes that play a crucial role in the immune system of vertebrates, including humans and cattle. The presence of Immunity genes with high variant intolerance levels suggests adaptability of these to different pathogens and landscape, which is crucial for the immune system of the bovine animal. MHC genes are mainly involved in encoding for proteins for cellular and extracellular antigen presentation to circulating T cells, inflammatory and immune-responses, heat shock, complement cascade systems, cytokine signalling, and the regulation of various aspects of cellular development, and cell differentiation (Kulski et al., 2019). Highly tolerant immunity genes, including those within the MHC, contribute to the robustness and versatility of the immune response, allowing organisms to fend off a wide array of pathogens through evolutionary time. This tolerance is crucial for the adaptability of the immune system, allowing it to respond to a constantly changing landscape of pathogens. The evolvement of MHC genes is driven by the need to cope with diverse pathogens alongside evolutionary time. Positive selection acts on MHC genes to favour alleles that enhance the ability of the immune system to recognize and combat a broad range of infectious agents.

The presence of pseudogenes means that the gene has recently lost its function within the cattle lineage. Pseudogenes are DNA sequences that bear significant homology to functional genes, but they lack promoter sequences for their transcription or contain other mutations rendering them incapable of producing functional proteins or carrying out their original biological roles (Wilde, 1986). Pseudogenes can serve as sources of genetic variation. While a particular pseudogene may not be functional, the mutations it accumulates can contribute to the overall

genetic diversity of a population. This diversity can be important for adaptation to changing environments. Pseudogenes are considered once-functional genes where throughout evolution, genetic changes and duplications have occurred, leading to the creation of pseudogenes. Due to the mutations the gene has accumulated over time, the genetic diversity of the population, can provide insights into the historical changes that have shaped the genome (Martínez-Arias et al., 2001). Mutations can accumulate in these pseudogenes without affecting the organism's fitness, as they do not perform any essential functions. Given the high tolerance of pseudogenes to functional variation, as indicated by RVIS scores in table 2, suggests that these pseudogenes do not currently play a critical role in the functioning of the organism. However, the mutations they accumulate can still contribute to genetic diversity within the population providing a reservoir of potential genetic changes that could be beneficial in changing environments over time.

Panther Gene Ontology

In order to get a greater understanding of the impact of these genes, it was important to conduct further analysis, in particular looking at the biological role in which the genes have within the phenotype. If the gene plays a crucial role in key biological processes, this variation that we see in table 1 and table 2 may have functional consequences. When observed at a population level, this observed variation could explain evolutionary factors that may be due to historical environmental and geographic factors that have shaped the genetic makeup of an organism or population over time (Schierenbeck, 2017). Furthermore, genes that under-go strong selective pressure in one population might be more tolerant to functional variation in comparison to another population facing different selective pressure where the need for that gene may not be as crucial within everyday life (T. R. Gregory, 2009).

The exploration of the top-scored gene ontology using the PANTHER database sheds light on the biological processes essential to gene activity. The presence and emphasis on macromolecule and biosynthetic processes offers valuable insights into the significance of these pathways in gene function. The over-representation of these functions not only underscores their biological significance but also suggests their capacity to endure and potentially adapt to mutations without compromising their integral roles and functions. Within the biological processes, genes related to cytoskeleton organisation, biosynthetic processes and metabolic processes comprise of some of the most present genes in the top scored genes. This suggests a focus on maintaining cellular structure, producing essential molecules, managing energy, and supporting cellular functions in the bovine biological system are key genes that are highly tolerant to variation and is due to the organisms' daily needs.

The exploration of the bottom scored gene ontology using the PANTHER database shows us the genes associated with less tolerance to functional variation and biological processes in which they are involved in that are essential to gene activity. The genes identified in the bottom

10% of RVIS-scored genes indicate reduced tolerance to mutation and are more constrained in comparison to the top scored genes. The genes associated with these lower scored genes reveal a spectrum of cell regulatory processes, that could potentially have implications due to biological cell processes due their reduced tolerance, in particular genes that are involved in transcription by RNA polymerase II . The presence of these genes could lead to increased sensitivity and have reduced adaptability to variant tolerance. The most significant result was that based on regulation of macromolecule metabolic processes. When further investigated within the database, the genes are related to any process that controls or influences the chemical reactions involving large, complex molecules in living organisms. Genes such as *ZNF774*, *LBX1*, and *TBX10* all in which control the regulation of transcription by RNA polymerase II were present in these bottom scored genes. These proteins are involved in the processing of other macromolecules, and in the case of the genes mentioned above, RNA specific genes. This process can ultimately affect the frequency, rate, or extent of these reactions and subsequent pathways, alongside gene expression. If the process involving these complex genes is prone to mutating, it can have effects on the genetic information, protein structure, cellular processes, and overall health and adaptability of the organism (Morris et al., 2022). The over-representation of genes associated with metabolic processes and in particular, regulation of cellular activity suggests vulnerability in this area of cellular processing. The genes associated with these activities were lower RVIS scored, representing their reduction in being able to withstand functional variation if mutations arise, which could compromise the balance required for metabolism, impacting the health and function of the cellular environment. Metabolic pathways provide the precursor molecules necessary for gene expression, alongside providing ATP, which is the primary fuel driving gene expression. (Carthew, 2021). Given that metabolic pathways provide precursor molecules for gene expression and are involved in gene expression, the compromised tolerance in these genes implies potential challenges in maintaining the necessary balance for cellular health and functionality.

When comparing the top vs, the bottom genes in the PANTHER output, the higher scored genes reveal the role of sensory perception in animals Although, highly scored these genes are underrepresented within the output. This could be due to the low sample size of genes versus the amount present in the reference genome used in the database and that these higher scored genes are found less frequently than expected by chance in the analysed dataset. Nonetheless, it's interesting to note that there is a deficiency of these genes compared to what would be predicted based on random chance. The included genes are mainly involved in olfactory receptors. These genes might have specialized functions related to sensory perception, and their underrepresentation could be due to the dataset being more broadly representative of various biological processes within that class. For example, the presence of *Beta-crystallin B1* gene under sensory perception has a main role is the progression of the lens over time, from its formation to the mature structure, where the lens is a transparent structure in the eye through which light is focused onto the retina, in order to form an image to the sound. The placement of the bovine's eye allows for a wide field of view, which is advantageous for detecting potential predators in their surroundings in grazing conditions and/or modern agricultural.

Therefore, this gene can be deemed essential for the bovine animal and has been conserved throughout evolution. Therefore, it can be considered highly conserved if it has remained relatively unchanged and maintains a similar sequence or function across different species over evolutionary time. High conservation often indicates that the gene performs crucial functions essential for the survival or development of organisms, which can explain the presence of olfactory receptors in the gene set, that are deemed tolerant to withstanding mutation.

QTL enrichment analysis

To get a more practical relevance of the phenotypic traits in which these genes represent a QTL enrichment analysis was carried out. The QTL analysis results were reviewed to see the traits associated with the derived RVIS scored genes, which was shown in table 5 and table 6. The associated traits were investigated and areas of overall bovine performance that have higher or lower RVIS scores, and the impact in which this may have. Intriguingly, non-return rate and gestation length, indicative of reproductive processes, were observed in both highly and lowly scored genes, indicating that they are not only highly conserved, but are sensitive to adaption, perhaps depending on the environment in which they find themselves in. For future analysis, this could be further tested by splitting up the *Bos taurus* and *Bos Indicus* breeds into separate analysis. The existence of both highly and lowly tolerant genes associated with these traits suggests the characteristics and associations involved could be specific to the gene region linked. On the other hand, other regions of these genes may have greater tolerance to genetic variations, implying potential flexibility or adaptability to diverse environmental conditions. Genes capable of maintaining essential functions while accommodating variation may have advantages in specific environmental conditions, thereby contributing to a greater understanding of gene evolution to its environment. The difference in variability could be explained by breeds and countries that are present within the dataset, or the differences between *Bos taurus* and *Bos indicus* breeds, and as previously mentioned, would be an interesting area to investigate.

Within the 1000 bull genome dataset, there is a varying range of breeds from all over the world. And although some fertility traits seem to be highly tolerant, non-return rate and gestation length seem to be also present within the bottom scored genes. When looking to dairy cattle, fertility has generally declined since around 1980 due to intensive selection on negatively correlated traits to fertility, such as milk production (Pryce et al., 2014). Barriers to genetic improvement of such fertility traits include variables such as the low heritability of fertility, the insufficient selection intensity on fertility, the negative association between level of milk production and the genetic correlation of fertility and milk production and the impact of inbreeding on reproductive performance (Pryce et al., 2014). In recent years, this issue has been identified and an improvement in phenotypic and genetic trends for fertility has been observed as a result of introducing breeding values for fertility with increased emphasis on fertility in breeding objectives alongside management practices such as improved reproductive management through technology, and nutrition (Crowe et al., 2018). These genes are

characterized by a heightened resilience to mutational events and subsequently, play a pivotal role in upholding health and robustness within a cattle population. The elevated tolerance of genes to mutations is integral to mitigating potential deleterious health effects associated with genetic variations. Such genomic resilience has profound implications for breeding programs and farm management, as genetically resilient cattle may yield substantial advantages, including a potential reduction in veterinary costs and enhanced productivity. Noteworthy among the findings is the identification of phenotypic traits such as non-return rate, interval from first to last insemination, interval to first oestrus after calving, and gestation length as intolerant. The genetic resilience observed in cattle with fertility traits underscores the potential for reduced reproductive challenges, contingent upon environmental factors. This insight holds particular relevance for breeders and farm managers in their decision-making processes, as cattle possessing genetically resilient fertility traits may experience fewer complications related to reproduction, thereby contributing to overall herd health and productivity. However, bovine species tend to exhibit considerable variability in their responses to diverse disease challenges, with a substantial portion of this variability attributed to genetic factors, alongside environmental conditions (Morris, 2007). Therefore, when looking to fertility as a trait that is susceptible to change, the use of genetics and increasing the emphasis of level of understanding around the trait and its mechanisms can help for increased tolerance towards mutation and enhanced breeding program structures.

When analysing the QTL enrichment traits with lower scores, as shown in table 5, a notable trend emerges wherein the traits predominantly related to milk composition. This observation implies a heightened degree of constraint and intensified selective pressures acting upon these traits to uphold their functionality. The genes associated with these bottom-scored traits are presumed to be more susceptible to mutations, potentially having an adverse influence on the overall health and fitness, or in the context of this study, the productivity levels of the individuals in terms of milk production. Specifically, the Milk C14 index, milk myristoleic acid content, and milk fat content emerge as the three most significant traits among the bottom scored, all of which are linked to the fat content of milk. Changes in milk composition, particularly in fat content, can have direct implications for milk production and its commercial value, depending on how a farmer is paid for their milk. Evidence suggests that when looking at milk C14 index, both diet and genetics has a big influence (Schennink et al., 2007). To estimate genetic variation in milk fatty acid unsaturation indices, a study conducted by (Schennink et al., 2008) looked at measuring milk fatty acid composition of 1,933 Dutch Holstein heifers and unsaturation indices were calculated. This represents the concentration of the unsaturated product proportional to the sum of the unsaturated product and the saturated substrate. The authors genotyped the cows for the SCD1 A293V and DGAT1 K232A polymorphisms, which have been shown to alter milk fatty acid composition. The SCD1 V allele was found to be associated with lower C10, C12, and C14 indices, and with higher C16, C18, and CLA indices in comparison to the SCD1 A allele, with no differences in total unsaturation index. In comparison to the DGAT1 K allele, the DGAT1 A allele is associated with lower C10, C12, C14, and C16 indices and with higher C18, CLA, and total indices.

Therefore, they concluded that selective breeding could contribute to higher unsaturation indices in milk and that genetics has a key role to play, with differences among cows and among breeds (Schennink et al., 2008). Due the many different breeds present in the 1000 bull genome project, each from different environments, this could explain the variation of the milk C14 index within the lower scored RVIS genes, as they are either tolerant or intolerant to functional variation depending on the environment in which they are in. When looking at nutrition aspects, there was also evidence that hydrogenated fats versus hydrogenated fats present in the diet of the bovine altered the C14 expression of milk content and are mainly derived from de novo fatty acid synthesis in the body (Grummer, 1991). This study concluded that significant changes in milk fat composition can be achieved on farm via nutritional modifications. With different countries, different breeds and different dietary requirements per breed, Milk C14 content is therefore constantly changing depending on the environment in which the cow is in. Therefore, it is more intolerant to functional variation and is constantly changing.

Human Ortholog Comparison

Investigating shared genes helps identify evolutionarily conserved elements. These are genes that have been maintained across species over long periods of evolutionary time. Discovering conserved elements can highlight fundamental biological processes essential for life. The RVIS scored genes were compared to that of *homo sapiens* calculated from the BioMart database, to identify genes that show similar or divergent levels of tolerance to genetic variation. This was done to investigate the functional importance and evolutionary conservation of specific genes. The correlation score of 0.143 between Human and Bovine genes suggests a positive relationship between the variables, but it's not a very strong correlation. The weak positive correlation in RVIS scores might indicate some similarities in the evolutionary constraints or functional importance of certain genes between humans and bovines. It could also imply that some genes share a common level of intolerance to genetic variation across the two species. However, on the other hand, it also could suggest that there are factors contributing to variations in RVIS scores that are not shared or are influenced differently in the two species. This could be due to functional differences of the same gene within the two species, or it could relate to the differences in the evolutionary history and selective pressures on genes acting in the human and bovine. Previous research carried out by (Petrovski et al., 2013) showed research in humans where developmental disorders were being caused by genes that do not tolerate functional variation and immunological disorders were caused by genes with an excess of common functional variation. In the context of the scatter plot in figure 3 comparing RVIS scores of cattle genes (x-axis) to human genes (y-axis), the outlier gene *RYR2*, associated with stress-induced polymorphic ventricular tachycardia and arrhythmogenic right ventricular dysplasia, exhibits a distinctive pattern. The RVIS score for *RYR2* is notably high in cattle, suggesting a higher tolerance to functional genetic variation in this species, while concurrently, its RVIS score is comparatively low in humans. The contrasting RVIS scores for *RYR2* in cattle and humans imply potential species-specific differences in the genetic constraints on this gene and the disorders which may arise if the gene was to undergo mutation. Given *RYR2's*

associations with cardiac disorders, the observed variation in RVIS scores could be representative of selection pressures within the two species and thus could impact disease susceptibility. While these scores hint at potential differences in the developmental or functional aspects of *RYR2* between cattle and humans, a more comprehensive investigation would be needed. There are numerous genes present in which further investigation could be conducted to outline the differences between the effect of functional variation in specific genes that are present in different species. However, it's important to recognize that limitations must be recognised when comparing the same genes between species as even if a gene has a similar sequence between the two species, it could potentially have a different function. Therefore, different selection pressures would also need to be taken into account.

Conclusion

The analysis of RVIS method of scoring genes from the 1000 Bull Genome Project provides valuable insights into the genetic landscape of bovine cattle. The RVIS scores, ranging from -4.288 to 3.158, serve as indicators of the genes' tolerance to genetic variation within the population. Genes with lower RVIS scores are considered less tolerant and more susceptible to the impact of mutations, potentially affecting their functions. While higher RVIS scores suggest greater tolerance to genetic variation, often associated with genes essential for biological functions and conserved throughout evolution. The top and bottom genes, highlighted in Table 1 and Table 2 based on RVIS scores, offer an insight into the potential functional significance and evolutionary constraints of these genes. The bottom scored genes showed to have an array of cellular processing genes and within the top scored genes, many were found to be either pseudogenes or immunity related genes. This suggests that cellular processing genes may be more intolerant to functional variation if mutation occurs while pseudogenes and immunity genes may exhibit greater variant tolerance if they undergo mutation.

When looking at the Gene Ontology output, the genes associated with macromolecule metabolic processes and regulation of cellular activity exhibit reduced tolerance in the bottom-scored genes, suggesting vulnerability in cellular processing activity. The compromised tolerance in these genes may impact the balance required for metabolism, influencing the health and function of the cellular environment. On the other hand, the higher scored genes reveal a capacity for tolerance in sensory perception, particularly in olfactory receptors. The underrepresentation of these genes in the dataset suggests their specialized functions within the biological role they have. For example, genes related to sensory perception, like Beta-crystallin B1, are highly conserved, reflecting their essential roles in the survival and development of organisms.

The QTL enrichment analysis showed trends with fertility traits exhibiting both highly and lowly tolerant genes. Reproductive processes, such as non-return rate and gestation length, show associations with genes of varied tolerance levels, indicating potential adaptability depending on environmental factors. The observed genetic resilience in cattle with fertility traits has significant implications for breeding programs, impacting overall herd health and productivity. In contrast to this, traits related to milk composition, such as Milk C14 index, milk myristoleic acid content, and milk fat content, show a trend of heightened constraint and

intensified selective pressures. The genes associated with these traits may be more susceptible to mutations, potentially influencing productivity levels, particularly in terms of milk production.

Finally, comparisons with human genes reveal a weak positive correlation in RVIS scores, suggesting some similarities in the evolutionary constraints or functional importance of certain genes between humans and bovines. However, this correlation also implies species-specific differences influenced by perhaps evolution or species-specific gene function could also play a role.

The exploration of variant intolerance and/or tolerance rate in bovine species allows for a greater understanding of gene function, evolutionary conservation, and trait associations provides a comprehensive understanding of the genetic landscape in bovine cattle. These findings contribute to the knowledge base for breeding programs, farm management, and the potential impact of genetic variations on the health, adaptability, and productivity of cattle populations.

References

Asselstine, V., Medrano, J. F., & Cánovas, A. (2022). Identification of novel alternative splicing associated with mastitis disease in Holstein dairy cows using large gap read mapping. *BMC Genomics*, *23*(1), 222. https://doi.org/10.1186/s12864-022-08430-x

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Bradford, G. E. (1999). Contributions of animal agriculture to meeting global human food demand. *Livestock Production Science*, *59*(2), 95–112. https://doi.org/https://doi.org/10.1016/S0301-6226(99)00019-6

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., & Karchin, R. (2013). Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics*, *14*(3), S3. https://doi.org/10.1186/1471-2164-14-S3-S3

Carthew, R. W. (2021). Gene Regulation and Cellular Metabolism: An Essential Partnership. *Trends in Genetics*, *37*(4), 389–400. https://doi.org/10.1016/j.tig.2020.09.018

Collins, A. (2015). The genomic and functional characteristics of disease genes. *Briefings in Bioinformatics*, *16*(1), 16–23. https://doi.org/10.1093/bib/bbt091

Cooper, G. M., & Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, *12*(9), 628–640. https://doi.org/10.1038/nrg3046

Crowe, M. A., Hostens, M., & Opsomer, G. (2018). Reproductive management in dairy cows - the future. *Irish Veterinary Journal*, *71*(1), 1. https://doi.org/10.1186/s13620-017-0112-y

Durbin, R. M., Altshuler, D., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., … Institute, T. T. G. R. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

FAO, F. (2017). *Food and Agriculture Organization of the United Nations–FAOSTAT "https://www. fao. org/faostat/en/# data*. QCL.

Florea, L., Souvorov, A., Kalbfleisch, T., & Salzberg, S. (2011). Genome Assembly Has a Major Impact on Gene Content: A Comparison of Annotation in Two Bos Taurus Assemblies. *PloS One*, *6*, e21400. https://doi.org/10.1371/journal.pone.0021400

Fonseca, P. A. S., Suárez-Vega, A., Marras, G., & Cánovas, Á. (2020). GALLO: An R package for genomic annotation and integration of multiple data sources in livestock for positional candidate loci. *GigaScience*, *9*(12), giaa149. https://doi.org/10.1093/gigascience/giaa149

Fuller, Z. L., Berg, J. J., Mostafavi, H., Sella, G., & Przeworski, M. (2019). Measuring intolerance to mutation in human genetics. *Nature Genetics*, *51*(5), 772–776. https://doi.org/10.1038/s41588-019-0383-1

Gregory, S. G., Barlow, K. F., McLay, K. E., Kaul, R., Swarbreck, D., Dunham, A., Scott, C. E., Howe, K. L., Woodfine, K., Spencer, C. C. A., Jones, M. C., Gillson, C., Searle, S., Zhou, Y., Kokocinski, F., McDonald, L., Evans, R., Phillips, K., Atkinson, A., … Bentley, D. R. (2006). The DNA sequence and biological annotation of human chromosome 1. *Nature*, *441*(7091), 315–321. https://doi.org/10.1038/nature04727

Gregory, T. R. (2009). Understanding Natural Selection: Essential Concepts and Common Misconceptions. *Evolution: Education and Outreach*, *2*(2), 156–175. https://doi.org/10.1007/s12052-009-0128-1

Grummer, R. R. (1991). Effect of Feed on the Composition of Milk Fat. *Journal of Dairy Science*, *74*(9), 3244–3257. https://doi.org/https://doi.org/10.3168/jds.S0022-0302(91)78510-X

Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S., & Goldstein, D. B. (2016). The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biology*, *17*(1), 9. https://doi.org/10.1186/s13059-016-0869-4

Hayes, B. J., & Daetwyler, H. D. (2019). 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annual Review of Animal Biosciences*, *7*(1), 89–102. https://doi.org/10.1146/annurev-animal-020518-115024

Illumina Inc. (2023). *Large-Scale Bull Genome Sequencing Enables Rapid Livestock Improvement*. Https://Emea.Illumina.Com/Science/Customer-Stories/Icommunity-

Customer-Interviews-Case-Studies/Daetwyler-Latrobe-Interview-Hiseq-1000bulls.Html#:~:Text=In%202012%2C%20Ben%20Hayes%2C%20PhD,Genetics%20and%20foster%20international%20collaboration.

Khatkar, M., Thomson, P., Tammen, I., & Raadsma, H. (2004). Quantitative trait loci mapping in dairy cattle: Review and meta-analysis. *Genetics Selection Evolution*, *36*, 163–190. https://doi.org/10.1186/1297-9686-36-2-163

Kulski, Shiina, & Dijkstra. (2019). Genomic Diversity of the Major Histocompatibility Complex in Health and Disease. *Cells*, *8*(10), 1270. https://doi.org/10.3390/cells8101270

Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature*, *470*(7333), 187–197. https://doi.org/10.1038/nature09792

MacHugh, D. E., Shriver, M. D., Loftus, R. T., Cunningham, P., & Bradley, D. G. (1997). Microsatellite DNA Variation and the Evolution, Domestication and Phylogeography of Taurine and Zebu Cattle (Bos taurus and Bos indicus). *Genetics*, *146*(3), 1071–1086. https://doi.org/10.1093/genetics/146.3.1071

MacNeil, M. D., & Grosz, M. D. (2002). Genome-wide scans for QTL affecting carcass traits in Hereford × composite double backcross populations1. *Journal of Animal Science*, *80*(9), 2316–2324. https://doi.org/10.1093/ansci/80.9.2316

Martínez-Arias, R., Calafell, F., Mateu, E., Comas, D., Andrés, A., & Bertranpetit, J. (2001). Sequence variability of a human pseudogene. *Genome Research*, *11*(6), 1071–1085. https://doi.org/10.1101/gr.167701

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), 122. https://doi.org/10.1186/s13059-016-0974-4

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, *8*(8), 1551–1566. https://doi.org/10.1038/nprot.2013.092

Miyata, M., Gasparin, G., Coutinho, L. L., Martinez, M. L., Machado, M. A., Silva, M. V. G. B. da, Campos, A. L., Sonstegard, T. S., Rosário, M. F. do, & Regitano, L. C. de A. (2007). Quantitative trait loci (QTL) mapping for growth traits on bovine chromosome 14. *Genetics and Molecular Biology*, *30*.

Morris, R., Black, K. A., & Stollar, E. J. (2022). Uncovering protein function: from classification to complexes. *Essays in Biochemistry*, *66*(3), 255–285. https://doi.org/10.1042/EBC20200108

Petrovski et al. (2013). Correction: Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLOS Genetics*, *9*(8), null. https://doi.org/10.1371/annotation/32c8d343-9e1d-46c6-bfd4-b0cd3fb7a97e

Pryce, J. E., Woolaston, R., Berry, D. P., Wall, E., Winters, M., Butler, R., & Shaffer, M. (2014). World trends in dairy cow fertility. *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production*, *10*, 6.

R Core Team. (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* . Https://Www.R-Project.Org/.

Rajavel, A., Klees, S., Hui, Y., Schmitt, A. O., & Gültas, M. (2022). Deciphering the Molecular Mechanism Underlying African Animal Trypanosomiasis by Means of the 1000 Bull Genomes Project Genomic Dataset. *Biology*, *11*(5), 742. https://doi.org/10.3390/biology11050742

Raszek, M. M., Guan, L. L., & Plastow, G. S. (2016). Use of Genomic Tools to Improve Cattle Health in the Context of Infectious Diseases. *Frontiers in Genetics*, *7*. https://doi.org/10.3389/fgene.2016.00030

Rocha, J. L., Pomp, D., & Van Vleck, L. D. (2002). QTL Analysis in Livestock. In N. J. Camp & A. Cox (Eds.), *Quantitative Trait Loci: Methods and Protocols* (pp. 311–346). Humana Press. https://doi.org/10.1385/1-59259-176-0:311

Ros-Freixedes, R., Valente, B. D., Chen, C.-Y., Herring, W. O., Gorjanc, G., Hickey, J. M., & Johnsson, M. (2022). Rare and population-specific functional variation across pig lines. *Genetics Selection Evolution*, *54*(1), 39. https://doi.org/10.1186/s12711-022-00732-8

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Schennink, A., Heck, J. M. L., Bovenhuis, H., Visker, M. H. P. W., van Valenberg, H. J. F., & van Arendonk, J. A. M. (2008). Milk Fatty Acid Unsaturation: Genetic Parameters and Effects of Stearoyl-CoA Desaturase (SCD1) and Acyl CoA: Diacylglycerol Acyltransferase 1 (DGAT1). *Journal of Dairy Science*, *91*(5), 2135–2143. https://doi.org/10.3168/jds.2007-0825

Schennink, A., Stoop, W. M., Visker, M. H. P. W., Heck, J. M. L., Bovenhuis, H., Van Der Poel, J. J., Van Valenberg, H. J. F., & Van Arendonk, J. A. M. (2007). *DGAT1* underlies large genetic variation in milk-fat composition of dairy cows. *Animal Genetics*, *38*(5), 467–473. https://doi.org/10.1111/j.1365-2052.2007.01635.x

Schierenbeck, K. A. (2017). Population-level genetic variation and climate change in a biodiversity hotspot. *Annals of Botany*, *119*(2), 215–228. https://doi.org/10.1093/aob/mcw214

Utsunomiya, Y. T., Carmo, A. S., Neves, H. H. R., Carvalheiro, R., Matos, M. C., Zavarez, L. B., Ito, P. K. R. K., Pérez O'Brien, A. M., Sölkner, J., Porto-Neto, L. R., Schenkel, F. S., McEwan, J., Cole, J. B., da Silva, M. V. G. B., Van Tassell, C. P., Sonstegard, T. S., & Garcia, J. F. (2014). Genome-Wide Mapping of Loci Explaining Variance in Scrotal Circumference in Nellore Cattle. *PLoS ONE*, *9*(2), e88561. https://doi.org/10.1371/journal.pone.0088561

Wang, D., Perera, D., He, J., Cao, C., Kossinna, P., Li, Q., Zhang, W., Guo, X., Platt, A., Wu, J., & Zhang, Q. (2023). cLD: Rare-variant linkage disequilibrium between genomic

regions identifies novel genomic interactions. *PLOS Genetics*, *19*(12), e1011074-. https://doi.org/10.1371/journal.pgen.1011074

Wilde, C. D. (1986). Pseudogenes. *CRC Critical Reviews in Biochemistry*, *19*(4), 323–352.

Xiang, R., van den Berg, I., Macleod, I., Hayes, B., Prowse-Wilkins, C., Wang, M., Bolormaa, S., Liu, Z., Rochfort, S., Reich, C., Mason, B., Vander Jagt, C., Daetwyler, H. D., Lund, M., Chamberlain, A., & Goddard, M. (2019). Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proceedings of the National Academy of Sciences*, *116*, 201904159. https://doi.org/10.1073/pnas.1904159116

Zhang, Q., Guldbrandtsen, B., Bosse, M., Lund, M. S., & Sahana, G. (2015). Runs of homozygosity and distribution of functional variants in the cattle genome. *BMC Genomics*, *16*(1), 542. https://doi.org/10.1186/s12864-015-1715-x

Popular Science Summary

The global cattle population plays a critical role in providing nutrition for the growing human population and ensuring economic stability through the agricultural industry. However, threats such as diseases and genetic mutations can have far-reaching consequences for both human and cattle health. With advances in genetic technology, researchers can now explore how genes affect cattle at the molecular level and subsequently influence traits expression. The 1000 Bull Genomes project was initiated to leverage publicly available whole-genome sequence data of cattle for research purposes. This project facilitated the investigation of variant intolerance rates, which assess the sensitivity of genes to changes within their DNA sequences. By utilizing this data, genes were ranked based on their residual variant intolerance scores, considering the susceptibility of variants within the genes to mutation and their impact on phenotype, relative to the total number of variants in protein-coding genes. This analysis aimed to understand the effect of genetic variations on crucial biological processes and evolutionary adaptability.

The results revealed a spectrum of genes ranging from highly sensitive to changes in DNA sequence to those with greater tolerance. Genes involved in cellular processing exhibited lower scores, indicating intolerance to mutations, while genes related to sensory perception, immunity, and non-functional genes exhibited higher tolerance. This suggests the critical role of sensory and immune genes in survival, while non-functional genes can tolerate changes as they have no impact on the organism. Traits associated with fertility and milk production were found in both tolerant and intolerant genes, highlighting the complexity of these genes, particularly in response to environmental factors. Comparison with human genes revealed shared genes with similar tolerance scores, underscoring the importance of validating gene function across species.

In conclusion, studying variant intolerance rates in bovine species provides valuable insights into gene tolerance to mutation, evolutionary processes, and their impact on production traits in cattle. These findings can inform breeding programs by guiding selective breeding strategies to avoid adverse effects and aid in identifying causal variants in genetic disorders.

Acknowledgements

I would like to express my gratitude to my supervisors, examiner, and everyone who contributed to the work and reviews carried out in relation to this thesis project. I am sincerely grateful to have been afforded the opportunity to study within the Swedish University of Agricultural Sciences in Uppsala as part of the European Masters in Animal Breeding and Genetics programme. Finally, I would like to acknowledge my parents in Ireland who have allowed me the opportunity to pursue my passion for Animal Genetics.